第 2 章

統計データ解析および並列化プログラム

本章では，本研究においてプログラムを開発した統計データ解析の手法と，それぞれの手法の並列化方法について述べる．また，開発に用いた環境についても述べる．

2.1 統計データ解析手法の分類

統計データ解析の手法は，大きく分けて，与えられたデータから目的の変量を予測したり，指標を求めたりする処理と，与えられたデータを分類する処理とに分けられる．
予測や指標を求める処理は更に，外的基準 (目的変量) の有無，外的基準が数値データか質的データか，説明変量が数値データか質的データか等により，図 2.1に示すように分類出来る．

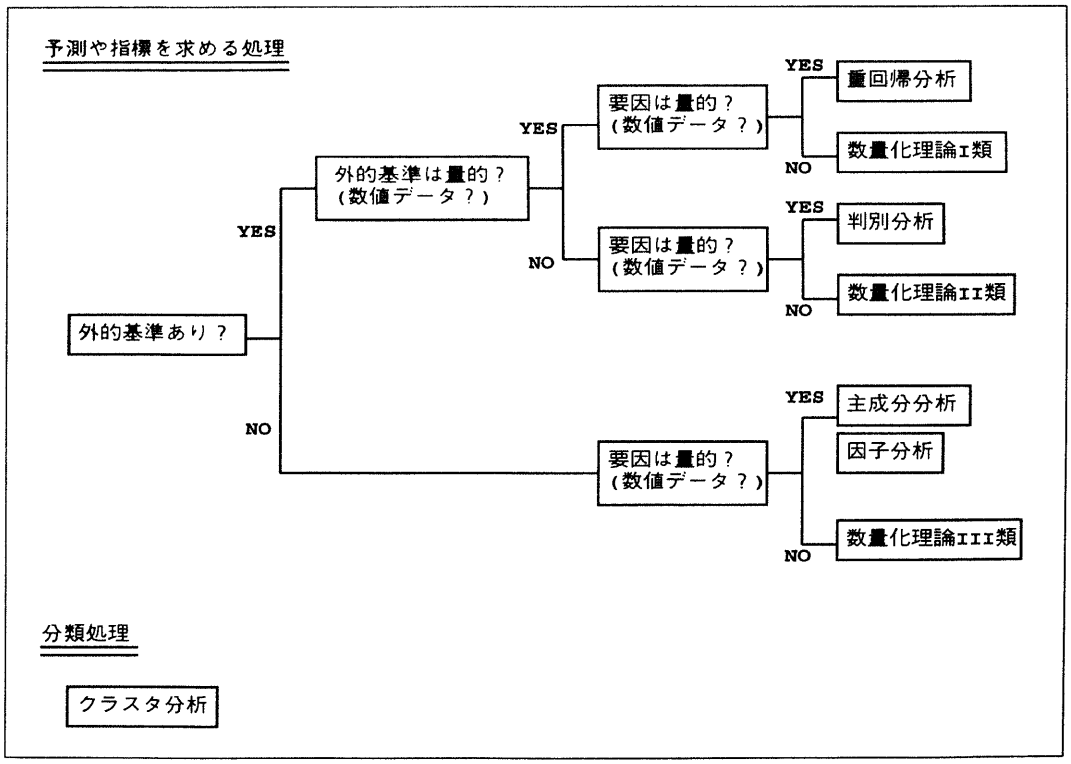

図 2.1: 統計データ解析の手法の分類

2.2 基本統計量

2.2.1 基本統計量とは

統計量のうち、最大値、最小値、平均、分散、標準偏差、変動係数、共分散、相関係数をまとめて基本統計量と呼ぶことにする。それぞれの計算式は式(2.1) ~ (2.6)に示す通りである。[10],[11]

1. 平均

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

2. 分散

$$s_x{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.2}$$

3. 標準偏差

$$s_x = \sqrt{s_x{}^2} \tag{2.3}$$

4. 変動係数
標準偏差は平均値の回りの散らばりの大きさを表す量であるが、この値が例えば同じ1cmであったとしても、平均値が10cmのときと1mのときとでは実質的にはかなりの違いがある。そこで、標準偏差を平均値で割ったものを変動係数(または変異係数)と呼び、平均値に対する相対的な散らばりの大きさを表すために用いる。ただし、この係数はデータが全て正の値から成り立っているときに用いるのが普通である。

$$Cov = \frac{s_x}{\overline{x}} \tag{2.4}$$

5. 共分散
変数xとyとの間の共分散は式(2.5)で表される。

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{2.5}$$

6. 相関係数
変数xとyとの間の相関係数は式(2.6)で表される。

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y} \tag{2.6}$$

2.2.2 基本統計量の並列化

プログラムではまず、PU[0]で各PU毎のデータの割り当て個数を計算し、PU[0]がファイルからデータを読み込み各PUへ割り当て個数分のデータを転送する。そして、全PUへのデータの入力が完了した後一斉に計算をスタートし、その結果を各PU毎の表として出力した。
データの入出力に関する処理はPU[0]で行った。これは、NFSなどの環境がないような場合でも適用出来るような汎用性を考慮したからである。また、データの割り当て方法に関しては、全部でN個のデータをP台のPUにN/P個ずつ均等に割り当てるという方法をとった。(図2.2)基本統計量の処理全体の流れを図2.3に示す。

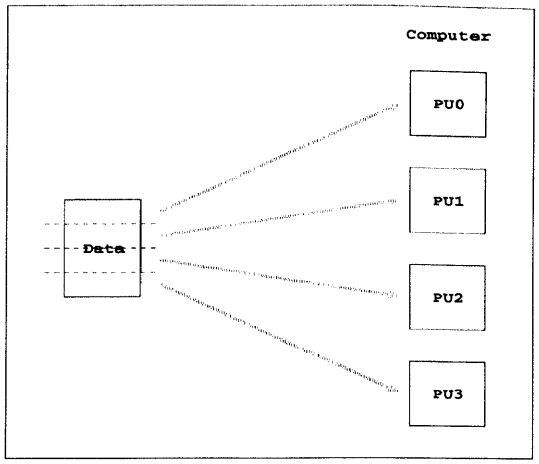

図 2.2: データの各ノードへの割り当て方法 1

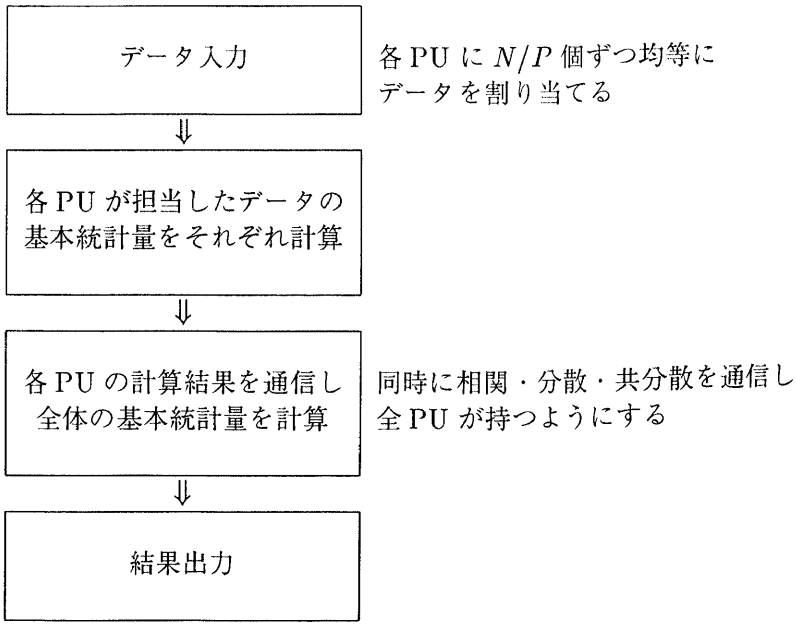

図 2.3: 基本統計量の並列処理の流れ

2.2.3 相関係数を求める処理の拡張

相関係数を求める処理に関しては、より精密な解析を行えるような拡張を行った。全サンプル数に対して適当な区間を決め、その区間をずらしていくことにより、一定区間における相関を解析出来るようにしたのである。計算の割り当て方法は、変量$x[0]$と$x[1] \sim x[m]$との相関をPU[0]が、$x[1]$と$x[2] \sim x[m]$との相関をPU[1]がという要領で割り当てていき、$x[P-1]$まで割り当てるとまたPU[0]から割り当てていくという方法をとった。
この処理により、図2.4に示したような、全体としては相関が低い変量どうしでも、ある区間(図2.4では区間(a))では非常に相関が高いなどという、データ全体の処理では見落としがちな特性の発見を可能にした。

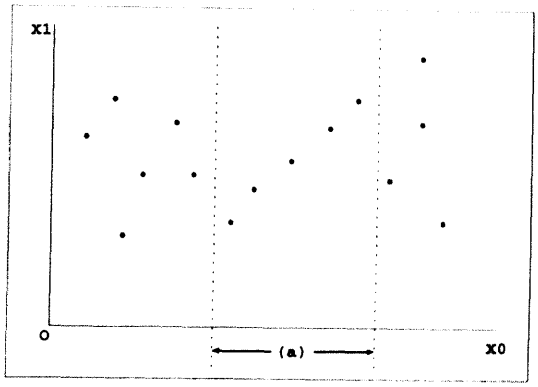

図 2.4: ある区間でのみ相関係数の高い例

2.3 重回帰分析

2.3.1 重回帰分析とは

ある変量y(目的変量と呼ぶ)と、それに影響すると考えられる変量$x_1, \cdots, x_m$(説明変量と呼ぶ)の間の関係式を求め、それに基づいて$x_1, \cdots, x_m$の値からyの値を予測したり、その際の各xの影響の大きさを評価したりする分析を回帰分析と呼ぶ。特に、説明変数が1つの場合を単回帰分析、2つ以上の場合を重回帰分析という。[12], [13]
いま、目的変量yと説明変量$x_1, \cdots, x_m$に関して、表2.1のようなn個の観測値が得られている

表 2.1: 重回帰分析のデータ (m変量nサンプルの観測値)

	y	x_1	$\cdots\cdots$	x_m
1	y_1	x_{11}	$\cdots\cdots$	x_{m1}
2	y_2	x_{12}	$\cdots\cdots$	x_{m2}
$\cdot$	$\cdot$	$\cdot$	$\cdots\cdots$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdots\cdots$	$\cdot$
n	y_n	x_{1n}	$\cdots\cdots$	x_{mn}

とする。目的変量yの値を説明変量$x_1, \cdots, x_m$から予測するため、ある関数fを用いてyと$x_1, \cdots, x_m$

との間に,

$$y_i = f(x_{1i}, \cdots, x_{mi}) + \epsilon_i \quad (i = 1, 2, \cdots, n) \tag{2.7}$$

のような関係を想定する. ここで, ϵ_i は説明変量の値 $x_{1i}, \cdots, x_{mi}$ によって説明出来ない誤差項を表す.

重回帰分析では $f(x_{1i}, \cdots, x_{mi})$ として, ふつう $x_{1i}, \cdots, x_{mi}$ の線形式を考え, 次のような線形重回帰分析モデル

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi} + \epsilon_i \quad (i = 1, 2, \cdots, n) \tag{2.8}$$

が成り立つものと仮定する.
ここで, $\beta_1, \cdots, \beta_m$ は偏回帰係数, β_0 は定数項と呼ばれる.
この仮定した式と実際のデータとの間で, 最小二乗法により誤差が最も小さくなるように係数を推定し, y を $x_1, \cdots, x_m$ により予測する式 (線形重回帰式) を求めるのが重回帰分析の手法である.

2.3.2 重回帰分析の適合性の指標

ここでは, 重回帰分析により求められたモデルの適合性を示す指標として, 決定係数 R^2, 重相関係数 R, 自由度調整済み重相関係数 $\overline{R^2}$, 赤池情報量規準 (AIC) について説明する.

決定係数 R^2 と重相関係数 R

係数の推定値が求まると, 観測された n 組の $\{(x_{1i}, \cdots, x_{mi}), i = 1, \cdots, n\}$ に対して, 重回帰式に基づく予測値 $\{F_i, i = 1, \cdots, n\}$ を計算することが出来る.
この時, 目的変量 y の偏差平方和に関して,

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(F_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - F_i)^2 \tag{2.9}$$

のような分解が成り立つ. 左辺は, もとの観測値 $\{y_i\}$ の平方和 (これを S_T とする) を表し, 右辺の第 1 項はそのうち回帰式で説明される部分の大きさ (これを S_R とする), 第 2 項は回帰式で説明されない部分の大きさ (これを S_e とする) を表すものとする.

総平方和 S_T の中で, 回帰式で説明される部分 S_R の割合,

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T} \tag{2.10}$$

を決定係数といい, R^2 が大きければ回帰モデルがよく当てはまっており, 小さければあまり当てはまっていないと判断する. R^2 の正の平方根 R は観測値 $\{y_i\}$ と予測値 $\{F_i\}$ との間の相関係数に等しく, 重相関係数という. [12]

自由度調整済み重相関係数 $\overline{R^2}$

前記のように, 決定係数はモデルの総合的な説明力を表す指標であり, また, データとの適合性の指標である. しかし, 重回帰分析において説明変量の数を増加させると決定係数はいくらでも 1 に近くなるという性質がある. このため, 説明変量の数が異なるモデルの説明力を比較する場合には, モデルに使用した変量の数でモデルの説明力を調整する必要がある. このように, 調整した決定係数を自由度調整済み決定係数といい, 次式で定義される. [12]

$$\overline{R^2} = 1 - \frac{S_e/(n - m - 1)}{S_T/(n - 1)} \tag{2.11}$$

赤池情報量規準 AIC

目的変量の将来の観測値 y の分布に対する推定量として、$f(y|b_0,\cdots,b_m)$ を考え、y の真の分布 $g(y)$ と予測分布 $f(y|b_0,\cdots,b_m)$ の間の Kullback-Leibler の情報量[1]で測った距離を出来るだけ小さくするという観点から導入された、モデル選択の一般的基準として、赤池情報量規準 (AIC: Akaike Information Criterion) がある.

m 個の説明変量をもつ重回帰モデルの場合、

$$\begin{aligned} AIC(m) &= -2\times(\text{最大対数尤度})+2\times(\text{モデルの自由パラメータ数}) \\ &= n(\log_e 2\pi+1)+n\log_e\frac{S_e(m)}{n}+2(m+2) \end{aligned} \tag{2.12}$$

と表される. AIC はその値自体に意味があるのではなく、相対的に小さいほど望ましいモデルであるということを表している.
上式(2.12)の第1項はモデルに関係しない定数であるが、第2項はモデルの当てはまりの良さを、第3項は変量の増加に対するペナルティを表すと解釈することが出来る. [12], [14], [15]

2.3.3 重回帰分析の並列化

重回帰分析では、まず2.2.2節で述べた手順で基本統計量を求め、その後 Beaton 法により解を求めた. 処理全体の流れを図2.5に示す.

逆行列を求める処理の並列処理の方法

逆行列を求めるルーチンの並列化方法を示す.
まず、行列が 2×2 以下の場合は、1PU の中で逆行列を求めた. これは、データの通信時間などを考慮に入れると、1つの PU の中で処理した方が効率が良いと判断したからである.
行列が 2×2 以上の時は、行列の各行を各 PU に割り当てた. つまり、行列の第0行を PU[0] に、第1行を PU[1] にというように順に割り当てていき、第 $P-1$ 行を PU[$P-1$] に割り当てるまで行う. そして、第 P 行からは再び PU[0] から順に割り当てていくという方法である. 各 PU は、自分に割り当てられた行がピボット行になった場合は、その行を全 PU にブロードキャスト (複数 PU への同時通信) する. ピボット行以外の行を割り当てられていた PU は、受けとったピボット行のデータを用いて一斉に計算を行う.

2.4 主成分分析

2.4.1 主成分分析とは

主成分分析とは、多数の変量の間の相関関係に着目して、それらの変量の変動に共通する成分を抽出するための統計的手法である. また、複数の成分が含まれている場合には、互いに無相関になるように成分を抽出する. このように抽出された成分を主成分という. 言い換えると、主成分分析とは1個または小数個 (m 個) の主成分によって、多数の変量の変動を要約する総合的特性を求めるための方法であると言える. p 変量 (p 次元) の観測値を m 個 (m 次元) の主成分に縮約するという意味で、次元を縮小する方法と言うことも出来る. [1], [12], [13], [16]

[1] [$Kullback-Leibler$情報量]

$$I(g;f)=\int_{-\infty}^{\infty} g(y)\log\frac{g(y)}{f(y)}dy$$

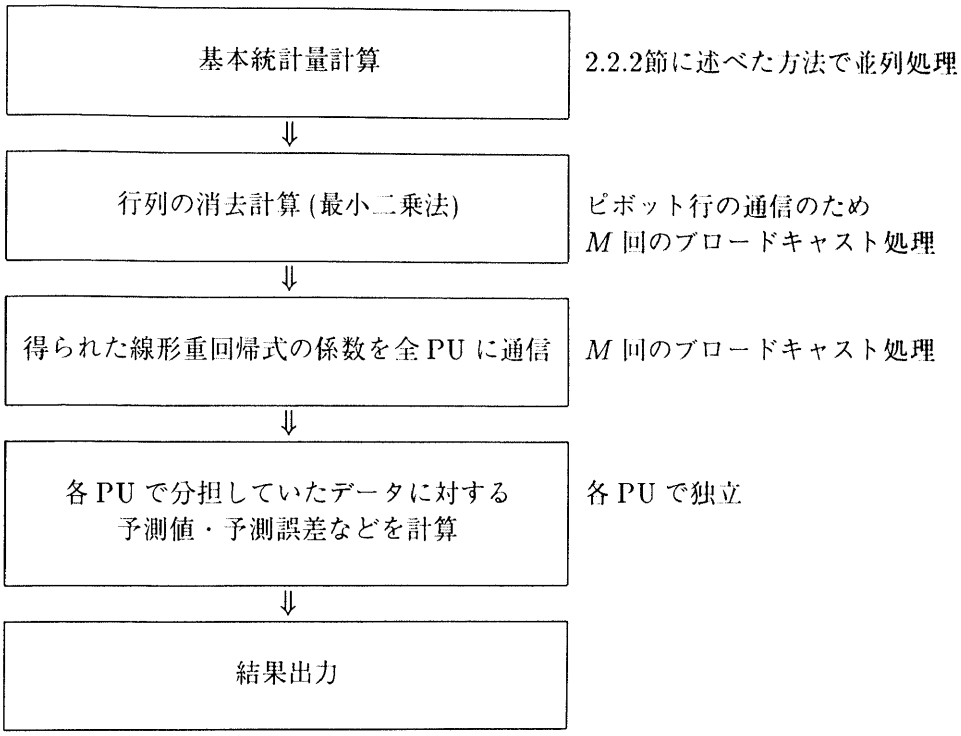

図 2.5: 重回帰分析の並列処理の流れ

今, 表 2.2のようなデータが与えられているとする.

表 2.2: p 変量 n サンプルの観測値

	x_1	x_2	$\cdots$	x_p
1	x_{11}	x_{21}	$\cdots$	x_{p1}
2	x_{12}	x_{22}	$\cdots$	x_{p2}
$\cdot$			$\cdots$	
n	x_{1n}	x_{2n}	$\cdots$	x_{pn}

例として, 次のような場合が考えられる.

- n 人の男性についての垂直とび, ボール投げなどの p 種類の体力テストの成績
- n 国についての人口, GDP などの p 種類の調査データ

そして, これらのデータに基づいて,「総合的な体力」,「総合的な国力」といったような総合的な指標を求めることを考える. そのために, 変量 $x_1, \cdots, x_p$ に対して任意の係数 $a_1, \cdots, a_p$ を用いて次のような線形結合をつくる.

$$z = a_1x_1 + a_2x_2 + \cdots + a_px_p \tag{2.13}$$

このように合成された z を, 主成分と呼ぶ. 主成分は形式的には説明変量の数だけ存在し, 第 1 主成分, 第 2 主成分, $\cdots$, 第 p 主成分と呼ぶ.

主成分の式は p 個の変量 $\boldsymbol{x}$ を「よく代表」していなければならない. そのための基準として次の 4 つの基準が用いられる.

1. 合成変量 z の分散の最大化.

2. 表2.2のデータを p 次元空間の中の n 個の点として表した時, その n 個の点から直線 z に下ろした垂線の長さの二乗和の最小化.

3. 合成変量 z を説明変量, もとの変量 $x_1, \cdots, x_p$ を目的変量として, 回帰式を作った時の残差平方和の合計の最小化.

4. 合成変量 z ともとの変量 $x_1, \cdots, x_p$ との相関係数の二乗和の最大化.

このうち, 1 ～ 3 からは, いずれも $x_1, \cdots, x_p$ の分散共分散行列の固有値問題が得られ, 同じ結果となる. また, 4 からは相関行列の固有値問題が得られる.

2.4.2 主成分分析モデルの定式化

本節では, 2.4.1節で述べた手法 1 の, 合成変量 z の分散の最大化の方法について説明する. p 個の変量から式(2.13)により合成する時, z の分散は,

$$V(z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \sum_{j=1}^{p}\sum_{k=1}^{p} s_{jk}a_j a_k = \boldsymbol{a}'S\boldsymbol{a} \tag{2.14}$$

(s_{jk}はx_jとx_kとの共分散, Sはs_{jk}を(j,k)要素にもつ分散共分散行列を表す)

式(2.13)の z は p 次元空間の中で原点 O からある OZ 方向に z 軸をとることを意味するが, そのとき z 座標のスケールを $x_1, \cdots, x_p$ 軸と同じにとることにすれば, 係数 $a_1, \cdots, a_p$ はそれぞれ直線 OZ の方向余弦 (OZ と $x_1, \cdots, x_p$ 軸となす角を $\theta_1, \cdots, \theta_p$ とすると, $\cos\theta_1, \cdots, \cos\theta_p$) になり,

$$\boldsymbol{a}'\boldsymbol{a} = a_1^2 + a_2^2 + \cdots + a_p^2 = 1 \tag{2.15}$$

を満たす. 従って, 問題は式(2.15)の制約条件のもとで式(2.14)を最大化することになる.
このような問題は, ラグランジュ乗数 λ を用いて,

$$F(\boldsymbol{a}, \lambda) = \boldsymbol{a}'S\boldsymbol{a} - \lambda(\boldsymbol{a}'\boldsymbol{a} - 1) \tag{2.16}$$

を制約なしで最大化する問題に変形される.
式(2.16)を $\boldsymbol{a}$ の各要素で偏微分して 0 とおけば, 次のような S の固有値問題を得る.

$$\frac{1}{2}\frac{\partial F}{\partial \boldsymbol{a}} = (S - \lambda I)\boldsymbol{a} = 0 \tag{2.17}$$

S は対称行列であり, また任意の $\boldsymbol{a}$ に対して $\boldsymbol{a}'S\boldsymbol{a} = V(\boldsymbol{a}'\boldsymbol{x}) \geq 0$ より, 非負値であるので, p 個の実数で非負の固有値 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ を持つ.

各固有値に対する固有ベクトルを $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_p$ とすれば, 式(2.17)より,

$$S\boldsymbol{a}_j = \lambda_j \boldsymbol{a}_j$$

両辺に $\boldsymbol{a}_j'$ を左から掛けると,

$$\boldsymbol{a}_j' S\boldsymbol{a}_j = \lambda_j \boldsymbol{a}_j'\boldsymbol{a}_j = \lambda_j$$

となり, λ_j はちょうど $z_j = \boldsymbol{a}_j'\boldsymbol{x}$ の分散に等しくなる.
従って, 分散を最大にする合成変量は, 最大固有値 λ_1 に対応する固有ベクトル $\boldsymbol{a}_1 = (a_{11}, \cdots, a_{p1})'$ の要素を係数として次のように作ればよい.

$$z_1 = \boldsymbol{a}_1'\boldsymbol{x} = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \tag{2.18}$$

これを, 第1主成分という. 第1主成分 z_1 の分散は λ_1 であり, 分散の最も大きい方向という意味で, p 個の変量をよく代表する合成変量になっている.

第1主成分だけで, もとの p 次元データのばらつきが十分代表されていないときには, 再び式 (2.13) の形の線形結合 z を考える. しかし, このままでは最初の主成分と同じ解になってしまうので, 最初の主成分 z_1 と2番目の主成分 z の相関係数が0となるように2番目の主成分を求める.

$$Cov(z, z_1) = \boldsymbol{a}'S\boldsymbol{a}_1 = \lambda_1\boldsymbol{a}'\boldsymbol{a}_1 = 0 \tag{2.19}$$

すなわち, 式 (2.15) と式 (2.19) の制約条件のもとで, 式 (2.14) の分散を最大化するのである.
2つのラグランジュ乗数 λ, ν を用いると,

$$F(\boldsymbol{a}, \lambda, \nu) \equiv \boldsymbol{a}'S\boldsymbol{a} - \lambda(\boldsymbol{a}'\boldsymbol{a} - 1) - \nu(\boldsymbol{a}'\boldsymbol{a}_1) \tag{2.20}$$

の最大化問題に変形され, これを $\boldsymbol{a}$ の各要素で偏微分して0とおくと, 次のようになる.

$$\frac{1}{2}\frac{\partial F}{\partial \boldsymbol{a}} = S\boldsymbol{a} - \lambda\boldsymbol{a} - \frac{\nu}{2}\boldsymbol{a}_1 = 0 \tag{2.21}$$

この両辺に, 左から $\boldsymbol{a}_1'$ を掛けて式 (2.19) を考慮すると $\nu = 0$ となり, 式 (2.21) は第1主成分を求めたのと同じ固有値問題, 式 (2.17) に帰着する.

最大固有値 λ_1 に対応する固有ベクトル $\boldsymbol{a}_1$ は, 既に第1主成分に用いられているので, 今度は2番目に大きい固有値 λ_2 に対応する固有ベクトル $\boldsymbol{a}_2 = (a_{12}, \cdots, a_{p2})'$ の要素を係数として次の式を合成する.

$$z_2 = \boldsymbol{a}_2'\boldsymbol{x} = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \tag{2.22}$$

これを, 第2主成分という. この第2主成分の分散は λ_2 である.

以下同様にして, 分散が $\lambda_3, \cdots, \lambda_p$ となるような第3主成分, $\cdots$, 第 p 主成分を求めることが出来る.

2.4.3 寄与率・累積寄与率

以上のようにして求められた主成分の分散と, もとの変量の分散の間には

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = s_{11} + s_{22} + \cdots + s_{pp} \tag{2.23}$$

のような関係がある.

$$\lambda_k / \sum_{j=1}^{p} \lambda_j \tag{2.24}$$

$$\sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{p} \lambda_j \tag{2.25}$$

ここで, 式 (2.24) を第 k 主成分の寄与率といい, その主成分がどの程度モデルの特性を説明しているかを表す. また, 式 (2.25) を 第1 ～ 第k 主成分の累積寄与率といい, k 個の主成分によるモデルの特性の説明力の合計を表している.
もし, $(m+1)$ 番目以下の固有値が0に近ければ, 第1 ～ 第m 主成分だけで, もとの変量のばらつきの大部分を説明出来ることになる.

2.4.4 主成分得点

得られた主成分について，各サンプルの(標準化した)変量の値を代入して計算される得点が主成分得点である．
l 番目のサンプルの第 k 主成分の主成分得点 f_{lk} は，以下の式で求める．

$$f_{lk} = \sum_{i=1}^{m} z_{ki} x'_{il} \tag{2.26}$$

ある程度の主成分の意味付けが出来ているときには，この得点を算出することによって，各サンプルの特徴を知ることが出来，分類することも出来る．さらに，主成分得点とサンプルの持つ他の情報とをつき合わせることにより，主成分の意味の確認や新たな知見が得られる場合があるなど，外部分析が可能になる．

2.4.5 主成分分析の並列化

主成分分析のプログラムでは，まず 2.2.2節で示した方法で基本統計量を求める．
そして，分散共分散行列を用いて解くか，相関行列を用いて標準化されたデータとして解くかをユーザが選択する．その行列を用いてヤコビ式を解き，得られた値から寄与率，累積寄与率，主成分得点を求める．尚，ヤコビ式は各 PU がそれぞれ計算し，主成分得点の計算などは各 PU が分担したデータについて並列に計算した．
よって，最初に行列やデータの割り当てを行った後は，計算途中には通信を行っていない．
尚，本研究の性能評価実験では，どちらの手法で解くかは予めプログラム中で与えておき測定した．
主成分分析の全体の流れを図 2.6に示す．

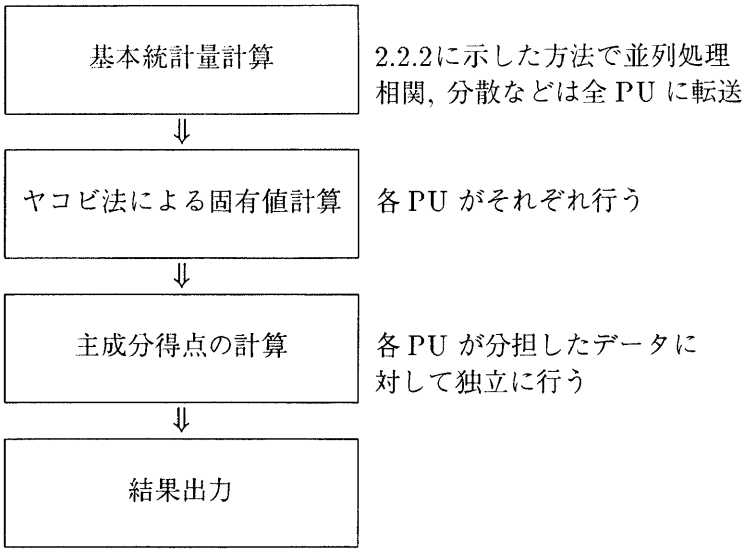

図 2.6: 主成分分析の並列処理の流れ

2.5 判別分析

2.5.1 判別分析とは

判別分析とは, p 変量の観測値に基づいて目的とする判別などを行う処理のことである. 例えば, 色々な検査結果からいくつかの病気のうちのどれであるかを判別したり, がくや花弁など色々な要素から, その植物の種類を判別したりということに用いられる. いくつかの手法があるが, 本研究では線形判別関数を用いた多群の判別と, 群間の相違を正準変量という少数個の変量を用いて表現した正準判別分析とを行った. [12], [13]

2.5.2 線形判別関数を用いた判別分析

観測値と各群の距離が最小になるような群に判別する処理のこと.
具体的には, どの群に属するか分からない観測値 $x = (x_1, \cdots, x_p)'$ を g 個の群のどれかに判別するとすると, 観測値 x と各群の重心 (平均ベクトル) $\mu^{(k)}, (k = 1, \cdots, g)$ との間のマハラノビスの汎距離

$$\Delta^2_{(k)} = (x - \mu^{(k)})'\Sigma^{-1}(x - \mu^{(k)}) \tag{2.27}$$

が最小となるような群に判別することである.
未知のベクトルと分散共分散行列に対する推定値として

$$\bar{x}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)} \tag{2.28}$$

$$S = \frac{1}{n-g}\sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_i^{(k)} - \bar{x}^{(k)})(x_i^{(k)} - \bar{x}^{(k)})' \tag{2.29}$$

($x_i^{(k)}$: k群のi番目の観測ベクトル, n : 総個体数)

を用いるとすると,

$$D_{(k)}{}^2 = (x - \bar{x}^{(k)})'S^{-1}(x - \bar{x}^{(k)}) \tag{2.30}$$

を最小にする群 k に判別することにする. 上式を最小にする k を求めることは, 次のような線形関数 $u_k(x)$ を最大にする k を求めることと同じになる.

$$u_k(x) = x'S^{-1}\bar{x}^{(k)} - \frac{1}{2}\bar{x}^{(k)'}S^{-1}\bar{x}^{(k)} \tag{2.31}$$

よって, この線形判別関数を用いた判別分析の判別ルールは, 観測値 x に対して $u_k(x), (k = 1, \cdots, g)$ を計算し, 最大値 $u_k(x)$ を与える k に判別するということになる.

2.5.3 正準判別分析

正準判別分析とは, 一つの観測値 x を特定のどれかの群に判別するという観点よりも, 群間の相違 (群ごとの傾向) を正準変量 (canonical variate) と呼ばれる少数個の変量を用いて, 出来るだけ明確に表現することを目的とする.
方法としては, p 個の変量 $x_1, \cdots, x_p$ に対して任意の係数 $(a_1, \cdots, a_p)$ を用いて,

$$z = a_1x_1 + \cdots + a_px_p \tag{2.32}$$

のような線形結合 z を作り，この値によって判別する．
係数 $a_1, \cdots, a_p$ が与えられると，n 個の個体のそれぞれに対して，合成変量

$$z_i^{(k)} = a_1 x_{1i}^{(k)} + \cdots + a_p x_{pi}^{(k)}, (k = 1, \cdots, g; i = 1, \cdots, n_k) \tag{2.33}$$

を計算することが出来る．この $\{z_i^{(k)}\}$ の変動を表す平方和は次のように分解される．

$$\underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(z_i^{(k)} - \bar{z})^2}_{\text{総平方和}S_T} = \underbrace{\sum_{k=1}^{g} n_k(\bar{z}^{(k)} - \bar{z})^2}_{\text{群間平方和}S_B} + \underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(z_i^{(k)} - \bar{z}^{(k)})^2}_{\text{群内平方和}S_W} \tag{2.34}$$

$$(\bar{z}^{(k)} : k\text{群の平均}, \bar{z} : \text{総平均})$$

z により g 個の群がよく判別されるということを，群間平方和 S_B が総平方和 S_T に対して大きくなることと考え，相関比

$$\eta^2 = S_B/S_T \tag{2.35}$$

を最大にするように係数 $a_1, \cdots, a_p$ を定める．
それはまた，群間平方和 S_B と群内平方和 S_W の比

$$\lambda = S_B/S_W \tag{2.36}$$

を最大化することと同じである．
λ を a の各要素で偏微分してゼロとおくと，一般固有値問題が得られる．求める係数を $a = (a_1, \cdots, a_p)'$ として，最大固有値 λ_1 に対応する固有ベクトル $a_1 = (a_{11}, \cdots, a_{p1})$ の要素を用いれば良い．このようにして得られた線形結合

$$z_1 = a_{11} z x_1 + \cdots + a_{p1} x_p \tag{2.37}$$

を第 1 正準変量と呼ぶ．
この後，2 番目以降の正準変量を求める場合には，新たな z は z_1 と無相関になるように計算する．
そして，得られた $a_2 = (a_{12}, \cdots, a_{p2})'$ の要素を用いて線形結合

$$z_2 = a_{12} z x_1 + \cdots + a_{p2} x_p \tag{2.38}$$

を作れば良い．この z_2 を第 2 正準変量と呼ぶ．
同様にして，必要ならば第 3, 第 4 の正準変量を求めることが出来る．

2.5.4 判別分析の並列化

判別分析では, データの群ごとに基本統計量などを求め, その値を利用して解析を行うので, データの群内での処理が多い. そこで, 図2.7のように, ある群のデータを分割して各ノードに割り当てると, 却って計算効率が悪化すると考えた.
よって, 各ノードへのデータの割り当て方法は, 図2.8のように, データ群ごとにまとめてそれぞれのノードに割り当てた. それぞれのノードで基本統計量などは計算し, 必要な数値は通信を行ってやり取りをした.
判別分析の処理全体の流れを図2.9に示す.

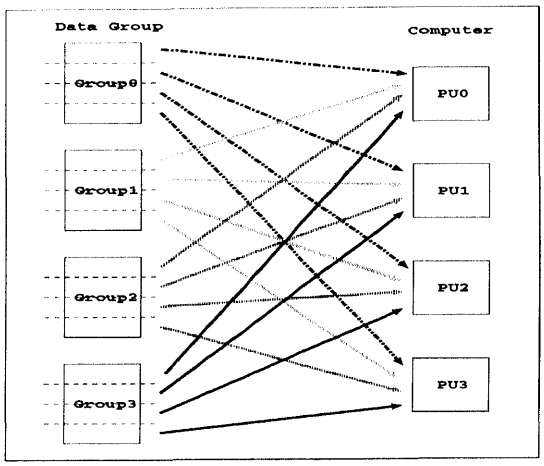

図 2.7: データの各ノードへの割り当て方法 2

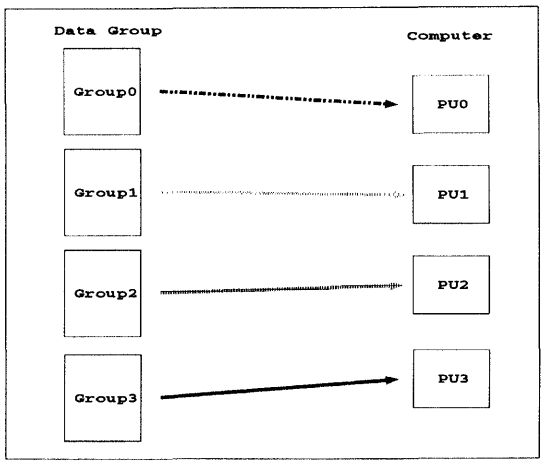

図 2.8: データの各ノードへの割り当て方法 3

[問題点]

この方法の問題点は, データの群数が少ない場合には, 並列計算機の全ノードを使用しない可能性があり, 並列処理効率が著しく低下することになるということである. しかし, 本研究では大量のデータ処理を行うということを目的としているので, その点は問題にならないと考えた.

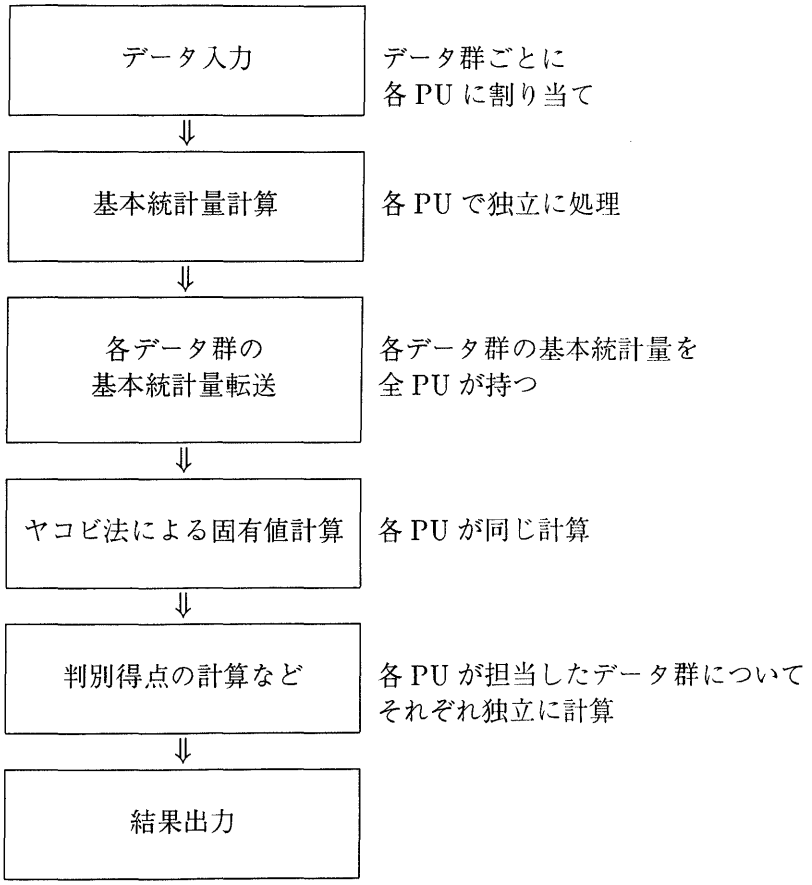

図 2.9: 正準判別分析の並列処理の流れ

2.5.5 並列処理での利点

従来の判別分析では, ユーザが結果を見たり予想したりして, 処理をしながら説明変量を加えたり除いたりしていた.
そのため, ユーザの先入観 (予見) によって結果が左右されることがあった.
本研究では, 並列計算機のパワーを生かして, 全ての説明変量の組合せについてしらみ潰しに処理するようにし, これらの処理を各ノードで分担するようにした. p 変量の時, 処理する組合せの個数は, 以下に示す式のようになる.

$$\sum_{i=1}^{p} {}_pC_i$$

今, 20 変量のデータを解析するとした場合,

$$\sum_{i=1}^{20} {}_{20}C_i = 1048575\text{通り}$$

の組合せについて解析することになる. [17]

また, 予め閾値を与えておくことにより, 著しく判別結果の悪いものは自動的に削除するようにした.

これらの処理により, ユーザの負担を出来るだけ軽くしながら, 全ての変量の組合せについての解析を, 洩れなく行うことが可能になった.

2.6 数量化理論 I 類

2.6.1 数量化理論 I 類とは

数量化理論 I 類とは, 質的な要因 (カテゴリカルデータ) に関する情報に基づいて, 量的に測定された外的基準 (目的変量) の値を, 説明あるいは予測するための処理である. 回帰分析において, 説明変量が定性的な形で与えられた場合に相当する. [12], [15], [18]
例えば, 性別, 職業, 居住地域などから収入を説明 (予測) するなどである.

処理方法は以下の通りである. 図 2.10の 1 番目に示すようなデータテーブルを考える. ここで, 2 番目に示すように, それぞれのアイテム・カテゴリに点数 (数量) を設定する. そして, 3 番目に示すように各サンプルの反応しているアイテム・カテゴリの点数を合計する. この合計した点数と, 各サンプルの外的基準の値 (y) が出来るだけ近くなるようにアイテム・カテゴリに与える点数を定めるのである.

2.6.2 数量化理論 I 類の並列化

数量化理論 I 類の処理の流れは, 図 2.11に示す通りである. データは 1 群なので, 図 2.2のようにサンプル数が均等になるように各ノードにデータを割り当てた. また, プログラム中の行列に関わる計算部分は, データを行ごとに各ノードに割り当てて計算した.
この解析で計算する行列の次元数は, 全アイテムのカテゴリ数の合計である. つまり, カテゴリ数 5 のアイテムが 100 あれば, 500 次元の行列となり, 並列化の効果が期待出来ると考えた.
図 2.11の中で, テーブル作成部分, 正規方程式を解く部分, スコアを求める部分, 逆行列の計算部分を並列に処理している.

Outside Value	Item	1			2			3		
	Category	1	2	3	1	2	3	1	2	3
y_1			∨		∨					∨
y_2		∨				∨		∨		
⋮			⋮			⋮			⋮	
y_n				∨		∨			∨	

⇓

Outside Value	Item	1			2			3		
	Category	1	2	3	1	2	3	1	2	3
y_1			∨		∨					∨
y_2		∨				∨		∨		
⋮			⋮			⋮			⋮	
y_n				∨		∨			∨	
Category score		0.1	0.2	0.3	-0.1	-0.2	-0.3	1.0	2.0	3.0

⇓

Outside Value	Item	1			2			3			Sample score
	Cate.	1	2	3	1	2	3	1	2	3	
y_1			∨		∨					∨	3.1
y_2		∨				∨		∨			0.9
⋮			⋮			⋮			⋮		⋮
y_n				∨		∨			∨		2.1
Category score		0.1	0.2	0.3	-0.1	-0.2	-0.3	1.0	2.0	3.0	

図 2.10: 数量化理論 I 類の処理手順

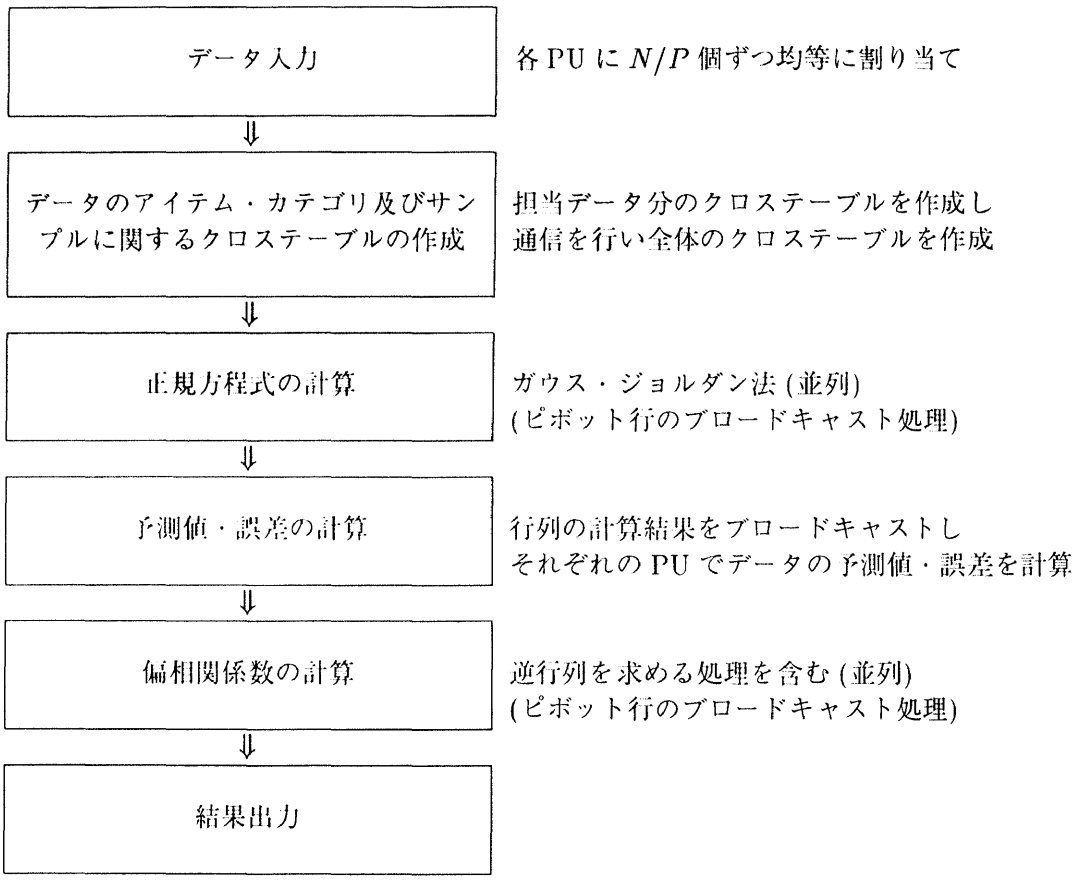

図 2.11: 数量化理論 I 類の並列処理の流れ

2.7 数量化理論 II 類

2.7.1 数量化理論 II 類とは

数量化理論 II 類とは, 質的な要因 (カテゴリカルデータ) に基づいて, 質的な外的基準 (目的変量) を予測あるいは判別するための方法である. 判別分析で, 要因データが質的に与えられた場合と考えることも出来る. [12], [15], [18]

処理方法としては, m 個の質的変量 (アイテム) の各カテゴリに

$$x_{jk} \qquad (j = 1, \cdots, m; k = 1, \cdots, c_j \quad (c_j \text{は第} j \text{変量のカテゴリ数}))$$

なる数値を与え, 各サンプルごとに反応したカテゴリの数値の和として合成変量 α を定義する. この時, 数量化理論 II 類では, 外的基準によるサンプルのグループを最もよく判別するという立場から, 同じグループに属するサンプル間では α の値が類似し, 異なったグループ間では α の値が異なるように, つまり外的基準に対する合成変量 α の相関比が最大となるように各カテゴリに数値を与える. (図 2.12)
このように, それぞれのカテゴリアイテムの点数 (数値) を求めて, 未知のデータを予測したり分類したりするのが数量化理論 II 類の目的である.

Outside Value	Sample No.	Item 1			Item 2			Item 3			Sample score (α)
		1	2	3	1	2	3	1	2	3	
1	1	∨					∨		∨		1.8
	2		∨		∨				∨		2.1
	⋮		⋮			⋮			⋮		⋮
	n_1			∨		∨		∨			1.1
2	1		∨			∨				∨	3.0
	2	∨			∨				∨		2.0
	⋮		⋮			⋮			⋮		⋮
	n_2			∨		∨		∨			1.1
⋮	⋮		⋮			⋮			⋮		⋮
k	1			∨			∨			∨	3.0
	2		∨			∨				∨	3.0
	⋮		⋮			⋮			⋮		⋮
	n_k			∨		∨			∨		2.1
Category score		0.1	0.2	0.3	-0.1	-0.2	-0.3	1.0	2.0	3.0	

(α が, k 個のグループごとに近い数値になるように, Category score を調節する.)

図 2.12: 数量化理論 II 類のデータテーブルとスコアの設定

2.7.2 数量化理論 II 類の並列化

図 2.13は数量化理論 II 類の処理の流れである. 要因のアイテム・カテゴリに付与する点数の計算をする部分の中に, 逆行列を求める計算があるが, この部分は並列化した. 何故なら, この計算を行う行列の次元数が,

次元数 = カテゴリ数の合計 − アイテム数

となるからである.
例えば, 100 アイテムで各アイテムが 10 カテゴリ持つデータだとすると, 次元数は,

$$10 \times 100 - 100 = 900$$

となり, 並列化の効果が見込めると判断した.
逆に, 固有値・固有ベクトルを求める部分は, 並列化しないで各ノードで同じ計算をすることにした. それは, この計算を行う行列の次元数は,

次元数 = 外的基準となるアイテムのカテゴリ数 − 1

となり, 上と同条件の場合, $10 - 1 = 9$ となり並列化の効果が見込めないと判断したからである.
数量化理論 II 類の並列処理の流れを、図 2.13に示す.

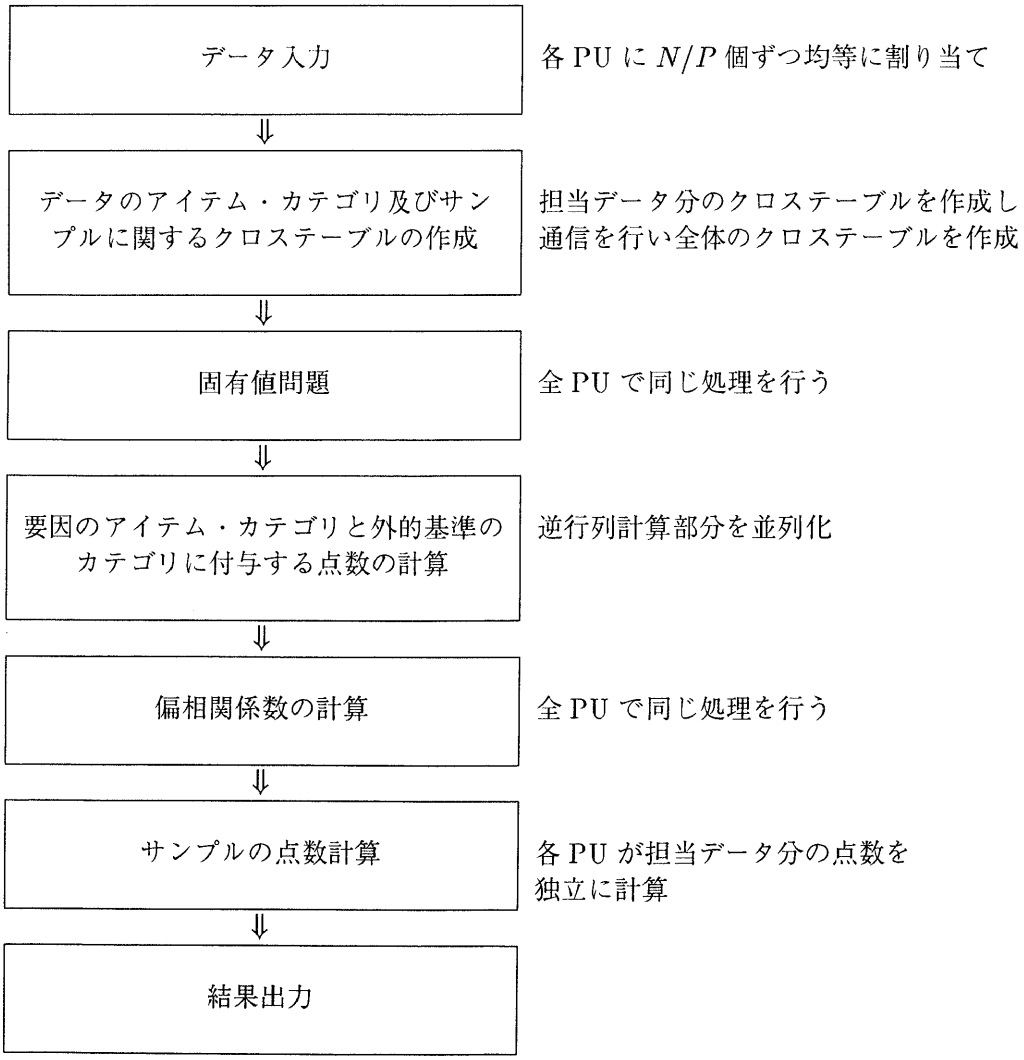

図 2.13: 数量化理論 II 類の並列処理の流れ

2.8 クラスタ分析

2.8.1 クラスタ分析とは

クラスタ分析とは, 一群の集団の中で互いに似たものを集め集落 (クラスタ) を作り, 集団を分類しようとする手法である. 各サンプル間の類似度を求め, それを手がかりにしてクラスタを形成していき, そこから図 2.14のような樹形図 (デンドログラム) を構成することが目的である. [12], [19], [20]
解析の手順は, 以下に示す通りである.

1. 各サンプルどうしがどの程度類似しているかを類似係数によって求める.
2. 求めた類似度により, 類似マトリクスを生成する.
3. 類似マトリクスの中で, 最も類似度の高いサンプルどうしを融合し, 一つのクラスタにまとめる.
4. まとめて新たに生成したクラスタと他のクラスタとの間で新たに類似度を計算し, 類似マトリクスを再生成する.
5. 以上の手順を, 最終的にクラスタが一つにまとまるまで繰り返す.

尚, 本研究では類似係数として, 平均ユークリッド距離[2]を用いた.

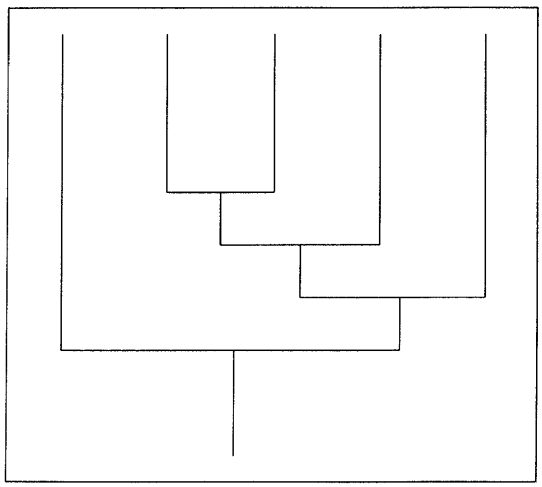

図 2.14: デンドログラム

[2][平均ユークリッド距離]

$$d_{jk} = (\sum_{i=1}^{m}(X_{ij} - X_{ik})^2/m)^{\frac{1}{2}}$$

2.8.2 クラスタ分析の並列化

まず計算量の多い類似度を計算する部分の並列化を行った. また, それぞれのPUにおいて属性の除き方を変えて, 一度に複数パターンのクラスタ分析を行った. これにより, どの属性がどの程度クラスタリングに影響を与えているのかを見ることが可能になる. また, 人間が属性の選択をせずに複数パターンのクラスタ分析が同時に行えることにより, ユーザの予見によらない結果の発見を助けることが出来る. 尚, クラスタ分析の並列化は, 筑波大学第3学群工学システム学類の山野 尚大氏の協力により行った. [19]

属性に対するクラスタ分析

前述の属性を除く処理において, 除く属性は属性自身に対するクラスタ分析を行って決定した. これは, クラスタ分析を行った結果同じクラスタに分類された属性というのは性質の似た属性であり, 同時に除くことが出来ると考えたからである. 以下に例を示す.

[例] A ～ Z まで26種類の属性を持つデータ
この属性に対するクラスタ分析を行って, 8クラスタに分けたとする.

Cluster1: C, D, Y

Cluster2: B, I, N, O, Z

⋮

Cluster8: A, P

Cluster1に属する属性C, D, Yは似た性質を持つということなので, データから属性C, D, Yを除いてクラスタ分析を行う. (図2.15)
同様に, 各Clusterに属する属性を除いたクラスタ分析を, それぞれのPUで同時に行うことにより, 8種類の異なった性質の分析を同時に行うことが出来る.

Sample	Attribute							
	A	B	C	D	·····	X	Y	Z
1	1.2	0.05	100	2.8	·····	91	1200	0.3
2	1.5	0.09	120	3.2	·····	88	1800	0.4
⋮	⋮	⋮	⋮	⋮	·····	⋮	⋮	⋮
n	2.3	0.01	95	1.9	·····	98	1500	0.2

⇓

Sample	Attribute							
	A	B	E	F	·····	W	X	Z
1	1.2	0.05	75	0.02	·····	12.1	91	0.3
2	1.5	0.09	63	0.05	·····	18.3	88	0.4
⋮	⋮	⋮	⋮	⋮	·····	⋮	⋮	⋮
n	2.3	0.01	80	0.01	·····	13.5	98	0.2

図 2.15: クラスタ分析におけるデータマトリクスの変更例

2.9 開発環境

本研究では，開発したシステムを様々な並列計算機システム上で実行するという汎用生の確保のために，通信ライブラリとしてMPIを用いた．また，パソコン上での開発実行環境として，LAMを用いた．

2.9.1 MPI

MPIはMessage Passing Interfaceを意味し，メッセージ通信のプログラムを記述するために広く使われる標準に発展することを目的に，MPI Forumにより策定された．MPI Forumには40以上の組織が参加しており，1992年11月より，メッセージ通信用の標準ライブラリセットの定義および検討がなされている．メッセージ通信の標準を確立することの主要な利点は，移植性と使い易さである．[21]

2.9.2 LAM

本研究のパソコン・クラスタで用いたMPIの開発・実行環境は，Ohio Supercomputer Centerで開発されたLAM(Local Area Multicomputer)である．LAMはネットワーク接続されたコンピュータのための並列処理環境及び開発システムであり，拡張モニタリング，デバッグツールなどによってサポートされている．LAMでは，ネットワーク接続された複数台のコンピュータを，あたかも一台の並列計算機のように扱うことが出来る．[22]
(尚，現在LAMのサポート・開発は，University of Notre DameのLaboratory for Scientific Computingで行われている．)