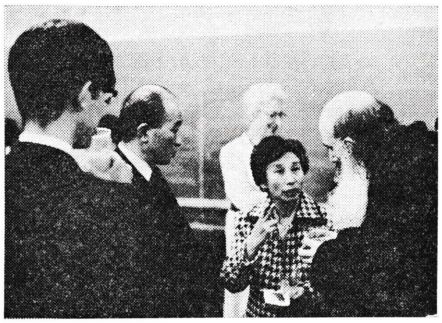
フランスにおけるデータ解析の動向

——Benzécri の数量化法を中心に——

大隅　昇



左から岩坪，林，高倉，Benzécri の各氏

1. まえがき

多くの場合，米国や英国などを中心とするデータ解析の動向については，我々の知識は豊富である．しかし，ヨーロッパ圏のフランス，オランダ，ベルギー，ドイツ，イタリアなどの様子を知る機会にはあまり恵まれていないようである．

だが，オランダでは多次元尺度解析（MDS）などが盛んのようであるし，西独には数値分類やクラスター分析などの研究者も多い．またベルギーには林の数量化法と類似の方法を考え出した J. M. Faverge や，これを踏襲する若手の研究者として G. Karnas らがいる．

ところがフランスとなると，そのデータ解析の動向は日本ではほとんど知られていなかったのではなかろうか．これは勿論筆者の偏った見方かもしれないが，少なくともフランスにおけるデータ解析の発展を知る機会に乏しかったことは誰もが認めることであろう．

ところが数年ほど前から，いわゆる林の数量化法Ⅲ類によく似た手法がフランスにあり，盛んに実際場面に適用を試みている研究者集団があって，その中心人物は Benzécri という研究者であるらしいという噂を耳にするようになった．こうした折も折，日本学術振興会（JSPS）の助成で，日本側から，統計数理研究所の松下嘉米男，西平重喜の両氏，フランス側からはパリ第6大学統計研究所（Institut des Statistique, Université de Paris VI; ISUP）所長の Dugué 教授を中心として，小規模ではあるが日仏の研究者が集まってセミナーを開くことになり，筆者もこれに参加する機会を得た（1978年3月）．このとき，幸いにも噂の人 Benzécri 教授にわずかの時間であるが面会することが出来た．その時の印象では噂通り，かなりの奇人であるとみえた．

その後，同じ年の11月に京都で国際パターン認識学会が開かれたが，そのおり，フランスから自動分類法で最近名前の知られてきた INRIA*（Institut National de Recherche en Informatique et en Automatique）の研究者である E. Diday が来日することを知り接触を図ったところ，偶然にも彼のほうが次の年（1979年）にヴェルサイユで開かれる研究集会（International Symposium on Data Analysis and Informatics）の招待講演者である林知己夫氏のもとへ訪れることとになったため，その機にわずかの時間であったが矢島敬二（日科技研），岩坪秀一（入試センター）などの諸氏をまじえて話し合う機会を得たのである．

この INRIA 主催の集会には林氏を初め，岩坪

* 以前は IRIA (Institut de Recherche D'Informatique et D'Automatique) といったが，最近改称した．

秀一，高倉節子，杉山明子，林文の諸氏と筆者の6人が参加し，ここでフランスのかなりの数の研究者と知り合うことができた．また林知己夫氏の念願であった Benzécri 教授との談合も，Dugué 教授のはからいで実現し，あちらの実状をかなり知ることが出来たわけである（写真は林氏と歓談する Benzécri 教授）．

いうまでもなく Benzécri の仕事がフランスのデータ解析（Analyse des Données）のすべてではないが，あちらの現状を知るには彼を中心とする研究集団の活動を無視するわけにはいかないようで，事実その勢力範囲と人気はかなりのものらしい．これは彼の風貌とそのカリスマ的な挙動に帰因するのかもしれない．

このとき，Diday に再会出来たことはもちろん，MDS の研究で知られる Brouche を初め，Lebart, Fénelon, Saporta, Caussinus, Escoufier 等々数多くの人達と知り合うことができた．彼らをそれとなく観察すると，日本の事情に似てあるクラスター化現象がみられ，やはり気心の知れた仲間同士が集まっているように見えた．

これと並行して，JSPS に申請中であったフランス国立科学研究センター（CNRS）の主任研究者である M. Roux の招へいが実現し，つい最近彼が来日したので，さらに詳しくフランスの実状を知る機会を得た（彼は数値分類法の研究でよく知られている）．

Roux によると，Benzécri は École Normale Supérieure の出身で当年48歳である．また飛行機には絶対乗らないそうで，訪米して Bell 研究所等をたずねたときも，すべて船と汽車で旅行したとのことである．また彼の数少ない英文の報告であるホノルルにおけるパターン認識の研究集会にも本人は出席せず別の研究者をさし向けたという．

前置きが長くなったが Roux の来日により，フランス国内の現状が明らかになってきたので，筆者のわずかの見聞と彼からの情報を合わせてフランスのデータ解析の動向を概観しよう．それと同時にフランスの数量化法の代表として知られる Benzécri 流の方法（Analyse Factorielle des Correspondances*; A. F. C.）の簡単な紹介を試みよう．

2. データ解析の動向

筆者の浅薄な体験で，フランスのデータ解析の実状が十分に把握できるなどとは毛頭考えていないが，とにかく “Analyse des Données” と称されるデータ解析の変遷には，ある意味で日本と類似した事情があるようである．ここでは，とりあえず数量化法を中心に，これと関連させて他の手法の発展を概観しよう．

日本では，1950年代の初めに林により数量化理論（quantification theory）の口火が切られて以来，伝統的な統計解析手法とは趣の異なる質的データ向きの解析手法の発展をみたわけであるが，フランスの事情もこれに似たところがある．

1950年代についてはほとんど情報がないので何ともいえないが，故 Lederman 教授，Dugué 教授らを中心とする伝統的な多変量解析，統計解析が主であったようである．

1962年頃になって初めて，Benzécri が言語学（linguistic）の問題に数量化法の適用を試み，これがこの種の手法の祖とされている．Roux によると60年代の前半は Benzécri の囲りに集まる人もなく，全く独自に研究をすすめていたという．1963年になって，Escoffier 夫人（当時，Miss B. Cordier）によって初めて計算プログラムが作成され，これが現在の発展のきっかけとなった．

その後1965年頃には，Montpellier 大学の Y. Escoufier によりエコロジーの分野の研究に対して多くの多変量解析手法や関連手法の適用が試みられている．この時期に Benzécri はパリ第6大学の統計研究所に移り，Laboratoire des Statistique の責任者として活動を始め現在に至るのである．同じころ Lederman 教授のもとで Bouroche,

* 省略して，Analyse des Correspondances ともいう．しかし，Analyse Factorielle というと，A. F. C. はもちろん主成分分析，因子分析，MDS，クラスター分析，……のいわゆる多変量解析手法の全てを総称するようである．

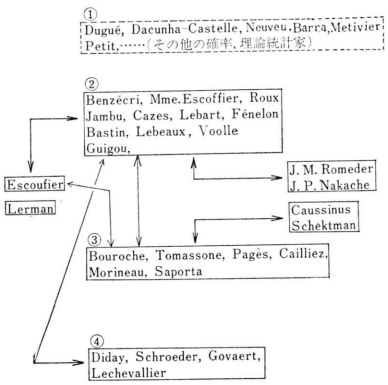
Lebart, Pagès, Saporta といった人達が学生として在籍していたが，教授の死により Benzécri の下で学んだそうである．

また，同じ時期（1966）に Caussinus が多次元分割表に関連した論文を表わしているが，最近注目を集めているこの分野の研究としては，比較的早くにこの仕事を行っている点に注意したい（彼の論文はよく引用される）．

このあと，1965年以降になると，M. Roux, Jambu, Diday といった人達が自動分類（Classification des Automatique）に関した研究を精力的に始めるようになる．さらに60年代後半には J. M. Romeder（現在カナダ），J. P. Nakache（彼は来日の経験がある）らが判別分析法に関連した仕事を始めている．

ところで，筆者が最近入手した文献や目にした資料を整理し，それと同時に Roux から得た情報を参考に，一つの人脈図を作ってみた．もちろん，聞きかじりで試みたことであるから一つの遊びと考えていただきたい．また人の関係がこのようにはっきりと区別できるものでもなかろうが現状を知るある程度の目安にはなるであろう．

図 1



グループ①は伝統的な数理統計，確率論の分野にある人達で，ほかにも多くの研究者がいるであろうが，データ解析とは無縁であるという意味で情報不足である．

第②グループは，Benzécri を中心とするグループで次第に勢力範囲を拡大しつつある一派である．実際，学生や若手の研究者に人気があるという．

③は Bouroche らを中心とするグループで，コンサルティング，調査などの会社を中心に応用分野に広く手を拡げているようである．しかし，一般の多変量解析手法を初め MDS などに精力的に力を注いでおり，日本，米国，オランダなどの研究者とのつながりも密接である（たとえば林，岩坪，Kruskal, Carroll, de Reeuw など）．

また，Diday に代表される INRIA の仲間は分類問題などに盛んに力を投入しているが，彼らは主にパターン認識の分野に密接に関連があり，ここに身をおいて Benzécri グループやその他の人達と接触を図るというやり方のようである．この他自動分類の Lerman や前述の Escoufier らがいるが，これらの人達は，いずれにも属さないで独自の路線をとるフリーランサーのようである．

②と③とは比較的近い関係にあるようで，Lebart, Morineau, Fénelon らは R. Chehessat のペンネームの下に何人かの研究者（というよりも，統計あるいはデータ解析の教育者）の集団を作り主に統計あるいはデータ解析の教育のあり方などを検討しているという．またこれらの研究者の関心分野あるいは関連領域が多岐にわたることはいうまでもない．

大要，こうした流れがみえるわけだが，勿論，これですべてが尽されたわけではない．むしろ大事なことは “Analyse des Données” を新しい観点から把えた若いグループの台頭がみられ，新しい波が起こりつつあるということである．

3. Benzécri の数量化法

結論を先にいうならば，Benzécri の方法は，林の数量化法Ⅲ類に同じ結果を与えるものである．ただ，定式化の過程が若干異なるのでこれを簡単に説明しておこう．

いま，個体数 n，変数 p のデータ行列を考える．そしてこの行列の第 k 変数と第 l 変数との間のクロス表を作る．第 k および第 l 変数のとりう

るカテゴリー値（これを modalité と呼んでいる）をそれぞれ，

k: $\{x_1, x_2, \cdots, x_i, \cdots, x_r\}$

l: $\{y_1, y_2, \cdots, y_j, \cdots, y_c\}$

とする．この $(r\times c)$ 次のクロス表において k が x_i, l が y_j をとる度数を f_{ij} と書く．

このとき，クロス表の各セル内の相対度数（確率）は，

$$p_{ij}=f_{ij}/n \quad (n=\sum_{i,j} f_{ij})$$

である．さらに行および列側の周辺相対度数をそれぞれ，

$$p_{i\cdot}=\sum_{j=1}^{c} p_{ij} \quad (i=1, 2, \cdots, r)$$

$$p_{\cdot j}=\sum_{i=1}^{r} p_{ij} \quad (j=1, 2, \cdots, c)$$

とかく．さらに行和または列和が1になるように条件付き相対度数を考える．

$$p_j{}^i=\frac{p_{ij}}{p_{i\cdot}} \quad (i=1, 2, \cdots, r),$$

$$p_i{}^j=\frac{p_{ij}}{p_{\cdot j}} \quad (j=1, 2, \cdots, c)$$

このとき $p_j{}^i$ は $(c-1)$ 次元内に布置された r 個の点とみることができる．同様に $p_i{}^j$ は $(r-1)$ 次元平面内の分布を与えている．両者は双対であるから，ここでは前者を考えよう．

さていま，変数 k 側に注目してカテゴリー x_i と $x_{i'}$ との間の距離を次の式で表わそう．

$$d^2(x_i, x_{i'})=\sum_{j=1}^{c}\frac{1}{p_{\cdot j}}\left\{\frac{p_{ij}}{p_{i\cdot}}-\frac{p_{i'j}}{p_{i'\cdot}}\right\}^2$$

$$=\sum_{j=1}^{c}\left\{\frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}-\frac{p_{i'j}}{p_{i'\cdot}\sqrt{p_{\cdot j}}}\right\}^2 \quad (1)$$

ここで，

$$z_{ij}=p_{ij}/p_{i\cdot}\sqrt{p_{\cdot j}} \quad (2)$$

と置くと，z_{ij} の平均は，次の式で与えられる．

$$m_j=\sum_{i=1}^{r} p_{i\cdot}z_{ij}=p_{\cdot j}/\sqrt{p_{\cdot j}}=\sqrt{p_{\cdot j}} \quad (3)$$

したがって，y_{ij} の分散共分散行列を $\boldsymbol{V}$ とするとその (j, k) 要素は，

$$v_{jk}=\sum_{i=1}^{r} p_{i\cdot}(y_{ji}-m_j)(y_{ik}-m_k)$$

となるので式(2)，(3)を上に代入して整理すると，

$$v_{jk}=\sum_i \frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}-\sqrt{p_{\cdot j}p_{\cdot k}}$$

$$=\sum_i\left(\frac{p_{ij}-p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\right)\left(\frac{p_{ik}-p_{i\cdot}p_{\cdot k}}{\sqrt{p_{i\cdot}p_{\cdot k}}}\right) \quad (4)$$

を得る．改めて，

$$r_{ij}=(p_{ij}-p_{i\cdot}p_{\cdot j})/\sqrt{p_{i\cdot}p_{\cdot j}} \quad (5)$$

とおき，これを (i, j) 要素とする行列を $\boldsymbol{R}$ とすると，

$$\boldsymbol{V}=\boldsymbol{R}'\boldsymbol{R} \quad (6)$$

となる．$(n\sum_{i,j} r^2{}_{ij}=n\mathrm{tr}\boldsymbol{V}=\chi^2$ はいわゆるピアソンの χ^2 統計量である)

さて問題は正規化データ $p_j{}^i$ または z_{ij} を $(c-1)$ 次元以下の空間内に布置する（$p_j{}^i$ に対する“雲(nuage, cloud)”を与えるという）ことである．ここで変数間の関連性を最も高くするような布置の与え方を考えるということは，式(6)の主成分分析を行うことに他ならない．すなわち，次の方程式

$$(\boldsymbol{V}-\lambda\boldsymbol{I})\boldsymbol{u}=\boldsymbol{0} \quad (7)$$

（ここで $\boldsymbol{I}$ は単位行列，$\boldsymbol{u}$ は固有ベクトル，$\boldsymbol{0}$ はゼロベクトルである）

を解いて固有値 λ とそれに対応する固有ベクトルを算出すればよい．このとき，式(7)の固有値のうちの1つは $\lambda_0=0$ となる．この固有ベクトルは，

$$\boldsymbol{u}_0'=(\sqrt{p_{\cdot 1}}, \sqrt{p_{\cdot 2}}, \cdots, \sqrt{p_{\cdot k}}, \cdots, \sqrt{p_{\cdot c}})$$

となるので λ_0, $\boldsymbol{u}_0$ を除く他の固有ベクトル $\boldsymbol{u}'=(u_1, u_2, \cdots, u_k, \cdots, u_l)$ について，次の関係を満たさねばならない．

$$\boldsymbol{u}'\boldsymbol{u}_0=\sum_{k=1}^{c} u_k\sqrt{p_{\cdot k}}=0$$

このことに注意すると，式(4)の第1の式で $\sqrt{p_{\cdot j}p_{\cdot k}}(\equiv m_{\cdot j}m_{\cdot k})$ の項が消える．こうして，

$$\sum_{k=1}^{c}\left\{\sum_{i=1}^{r}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}\right\}u_k=\lambda u_j \quad (j=1, 2, \cdots, c) \quad (8)$$

を得る．さらに

$$\rho_{ij}=\frac{p_{ij}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\left(\equiv\frac{f_{ij}}{\sqrt{f_{i\cdot}f_{\cdot j}}}\right) \quad (9)$$

とおき，ρ_{ij} を (i, j) 要素とする行列を $\boldsymbol{Q}=(\rho_{ij})$ で表わすと，上の式(8)は次のように表わせる．

$$\boldsymbol{Su}=\lambda\boldsymbol{u},\quad \text{ここで } \boldsymbol{S}=\boldsymbol{Q}'\boldsymbol{Q} \tag{10}$$

この方程式の固有値の最大根はつねに $\lambda_0=1$ となるが，これは式(7)の $\lambda_0=0$ に対応するものでつねに(10)を満たす(u_k として $\boldsymbol{u}_0$ の要素 $\sqrt{p_{ik}}$ を式(8)に代入すれば明らかである)．そこで，この無縁根を除いた次に大きい固有値から順に求め，それを $\lambda_1, \lambda_2, \cdots$ とすると，それぞれに対応する固有ベクトルを算出すればよい．

また，λ_q すなわち第 q 成分におけるカテゴリー x_i に対する数量を $F_q(i)$ で表わすと，

$$F_q(i)=\sum_{k=1}^{c} u_{kq}\left(\frac{p_{ik}}{p_{i.}}\right)\Big/\sqrt{p_{.k}} \tag{11}$$

および第 q 成分ベクトル $\boldsymbol{\varphi}_q$ の要素は，

$$\varphi_{kq}=u_{kq}/\sqrt{p_{.k}} \tag{12}$$

でそれぞれ与えられる．

さて，これでクロス表の列側（表頭）にあるカテゴリーに対して，数量を付与しその布置図を得ることができるわけだが，行側（表側）のカテゴリーに対しても全く同じことがいえる．

すなわち，式(2)あるいは(5)で i と j を入れかえた上で式(6)を $\boldsymbol{V}^*=\boldsymbol{RR}'$ として主成分分析を行えばよい．あるいは，式(9)が i, j の入れかえに対して双対であることに注意して，$\boldsymbol{S}^*=\boldsymbol{QQ}'$ に対して主成分分析を行えばよい．このときの第 q 成分に対する固有ベクトルを $\boldsymbol{v}_q$ で表わすと，前出の $\boldsymbol{u}_q$ との間に，

$$\boldsymbol{v}_q=\boldsymbol{Qu}_q/\sqrt{\lambda_q} \tag{13}$$

の関係がある．したがって式(12)に対応して，第 q 成分ベクトル $\boldsymbol{\phi}_q$ の要素は，

$$\phi_{lq}=v_{lq}/\sqrt{p_{l.}}$$

で与えられる(上の説明で，$\sum_q \lambda_q=\mathrm{tr}\,\boldsymbol{S}$ であり，$\sqrt{\lambda_q}$ は Cor $(\varphi_{kq}, \phi_{lq})$，つまり項目 k, l に与えられた数量の相関係数に相当する．林の方法ではこの λ の最大化を考える)．

以上の準備により，クロス表の表側と表頭のカテゴリーに与えられる数量（得点）を同一の布置

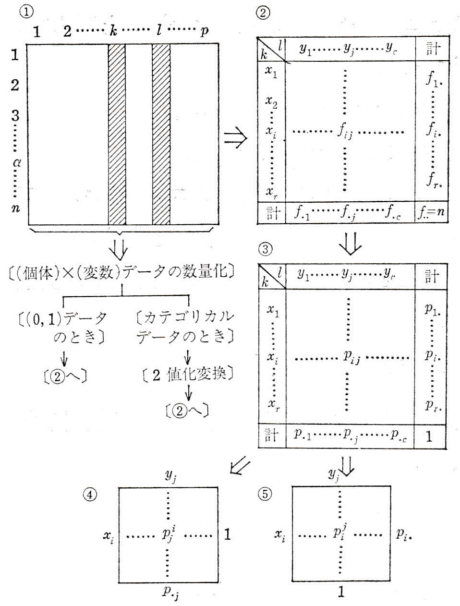


図 2

図の中に置くことが出来る．すなわち，

$$\phi_{lq}=\frac{1}{\sqrt{\lambda_q}}\sum_{j=1}^{c}\left(\frac{p_{lj}}{p_{l.}}\right)\varphi_{jq} \tag{14}$$

または，

$$\varphi_{jq}=\frac{1}{\sqrt{\lambda_q}}\sum_{i=1}^{r}\left(\frac{p_{ij}}{p_{.j}}\right)\cdot\phi_{iq} \tag{14}'$$

の関係を利用すればよい．

さて，上にみるように Benzécri の数量化法は林の数量化法Ⅲ類に同等であることがわかる．また，実際にクロス表に対するⅢ類の適用についても，いろいろの試みがなされている([13], [14])．上に述べたクロス表の立場から眺めた数量化のデータ処理の過程を整理すると図2のように表わせるであろう．

ここで簡単な例により上の過程を検証しておこう（例は Lebart [9] から引用した)．表1にみ

表 1

被験者＼質問	Q_1		Q_2		計
	は い	いいえ	は い	いいえ	
1	1	0	0	1	2
2	1	0	1	0	2
3	1	0	1	0	2
4	0	1	1	0	2
5	0	1	0	1	2
6	1	0	0	1	2
計	4	2	3	3	12

るようにこの例は，データが二値型である．6人の被験者が2つの質問 Q_1, Q_2 に対して「はい」，「いいえ」のいずれかを選択している．これを分割表のように考えても矛盾はないから，このデータ行列に上の方法を適用してみよう．

（手順1） 式(9)から ρ_{ij} を求める．そして式(10)の行列 $\boldsymbol{S}=\boldsymbol{Q}'\boldsymbol{Q}$ を作る．

$$\boldsymbol{S}=\boldsymbol{Q}'\boldsymbol{Q}=\begin{bmatrix} 1/2 & 0 & 1/2\sqrt{3} & 1/2\sqrt{3} \\ 0 & 1/2 & 1/2\sqrt{6} & 1/2\sqrt{6} \\ 1/2\sqrt{3} & 1/2\sqrt{6} & 1/2 & 0 \\ 1/2\sqrt{3} & 1/2\sqrt{6} & 0 & 1/2 \end{bmatrix}$$

（手順2） $\boldsymbol{S}$ の固有値，固有ベクトルを求める．まず固有値は，$\lambda_0=1$, $\lambda_1=\lambda_2=1/2$, $\lambda_3=0$ となる．第1根と4根は除いて，第2，3根に対する固有ベクトルを求めると，

$$\boldsymbol{u}_1'=(0,\ 0,\ 1/\sqrt{2},\ -1/\sqrt{2}),$$

$$\boldsymbol{u}_2'=(1/\sqrt{3},\ -\sqrt{2}/3,\ 0,\ 0)$$

を得る（この例は重根という特別な場合になっている）．

（手順3） 式(12)を使って主成分ベクトルを求める．

$$\boldsymbol{\varphi}_1'=(0,\ 0,\ \sqrt{2},\ -\sqrt{2}),$$

$$\boldsymbol{\varphi}_2'=(1,\ -2,\ 0,\ 0)$$

（手順4） 式(11)から各被験者に対する数量得点を求める．すなわち，$\boldsymbol{\varphi}=(\boldsymbol{\varphi}_1,\ \boldsymbol{\varphi}_2)$ とし，また，$p_{ik}/p_{i.}$ を (i,k) 要素とする行列を $\boldsymbol{R}^*$ で表わすと，$\boldsymbol{R}^*\boldsymbol{\varphi}$ により得点が与えられる．これを求めると，

$$\boldsymbol{R}^*\boldsymbol{\varphi}=\begin{bmatrix} 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \end{bmatrix}$$

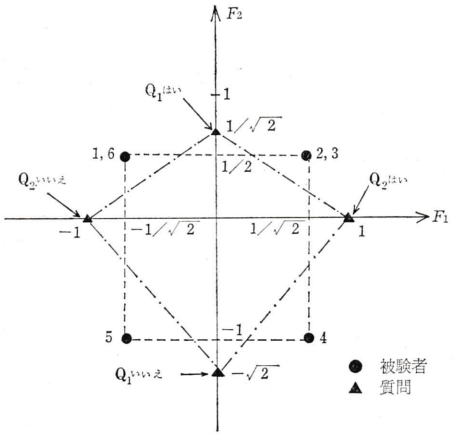
$$\times\begin{bmatrix} 0 & 1 \\ 0 & -2 \\ \sqrt{2} & 0 \\ -\sqrt{2} & 0 \end{bmatrix}=\begin{matrix} F_1(i) & F_2(i) \end{matrix}\begin{bmatrix} -1/\sqrt{2} & 1/2 \\ 1/\sqrt{2} & 1/2 \\ 1/\sqrt{2} & 1/2 \\ 1/\sqrt{2} & -1 \\ -1/\sqrt{2} & -1 \\ -1/\sqrt{2} & 1/2 \end{bmatrix}$$

一方，表側のカテゴリーに対する得点は，式(14)′から簡単に $\sqrt{\lambda_q}\varphi_{jq}$ と変換することで，

$$\begin{bmatrix} 2 & 1/\sqrt{2} \\ 0 & -\sqrt{2} \\ 1 & 0 \\ -1 & 0 \end{bmatrix}$$

と与えられる．こうして被験者と質問の関係は図を見れば明らかであろう．求めた個体と変数の数量化の結果を同時布置図としてプロットすると図3がえられる．

図 3



4. Benzécri 流の数量化法の特徴

技法だけを眺めるならば Benzécri の数量化法は林のそれのⅢ類の変形である．Benzécri らの特色はこの手法を中心に様々の利用の工夫をこらし，実際場面に適用していることである．とくに，こうした適用場面で現われるやっかいな問題点を整理し，技法運用のガイドを与えている点に注目したい．もちろん個別にみると，既に我々にとって周知のことも多い．たとえば，データの事前処理手続きとして，連続量データのカテゴリー化の方法，多値データの2値データへの変換方法，あるいは，いわゆるアイテム・カテゴリー型データの取り扱いなどは我々が日常体験しそのつど苦労してデータ処理を行っていることである．しかしこうした経験則を体系的に整理しようとする態度が，彼らにみられる．

こうした行動の寄りどころの1つは Benzécri のデータ解析に対する考え方にある．彼の主張

は，いわゆる「五つの原理」に要約されている．その第1で，「統計学は確率ではない(Statistics is not probability)」といっている．また，その第2は，「モデルはデータにもとづいて明らかになる，それは創作ではない（Models should come out of data, not out of imagination)」である．以下原理5まで彼独特の意見の展開がみられるのである（[2]参照）．個々の表現はいわゆるデータ解析者にとってきわめて当然のことばかりである．しかしこれを改めて説法するというところが，いかにも Benzécri らしい．

ところで，彼らのデータの把え方を要約すると次のようになる．

1° 分割表（クロス表），多次元分割表

2° (個体)×(変数) 型の多変量構造データ行列

3° (個体)×(個体)，あるいは(変数)×(変数)の形式のデータ

4° 二値型データ

5° 典型的な測定値．これは通常我々が尺度の概念で把えている大部分のデータ，すなわち，名義，順序，区間などの各測定値をさす．

ここで 1°〜3° はデータの形式に関わることであり，その他はデータの種類に関連することである．また，1°〜3° のいずれをも，広い意味で“データ行列”あるいは“データ表”と総称しているようである．

次にデータの加工法である．我々も日常いろいろと工夫しているが，彼等の考え方も大同小異である．しかし，個々の手順に名称をつけて整理している点に注目したい．

たとえば，ある項目（変数）が順序尺度データであるとき，測定値そのものだけ利用するのではなく，その項目のとりうる最大カテゴリー値からみた相補値もデータとしてつけ加える場合を，“重複化 (dédoublement)”という．いまカテゴリーが 1, 2, 3, 4 のとき，測定値が3であると，その相補である 4−3=1 も3の相補データとしてデータ行列上に加えるのである．このとき，3を intensité, 作り出した相補の1を défficience という．データが量的データの場合に適当に区分してカテゴリー化を行うが，このときにも類似したことがおこる．0〜10000 までを 10 等分して，それぞれに 1〜10 をあてる．いま得られたデータが 8200 のとき，コード値は9である．この intensité が9のデータに対して 10−9=1 を défficience として与える．

また，(0, 1) 型データ (descriptive logique) も，「はい—いいえ」の一方だけを採用するときと両者をとり入れてすべてを分析に含める場合がある．さらに多値データを2値データに拡張展開するいわゆるアイテム・カテゴリー型のデータ処理を，“codage en classes (等級，層別のコード化)”と呼びその形式のデータを forme disjonctive complète という．

こうした例にみるように，それぞれを整理するとともに，こうしたデータの，どのような形のものが計算結果としてどう現われるか，換言するならば，入力データに対する出力の布置図の中での各点（個体，変数）の数量の分布を，入力データの形式をあれこれかえて，様々のケースについてその挙動を調べている．彼らは，因子平面上の点の分布のことを“雲 (nuage, cloud)”と呼び，数量化法とは，与えられた原データの個体なり変数なりを新たな空間の中に“雲”として再布置し関連を調べることであるとしている．

ところでデータに付帯する2つの重要な概念がある．その1つは，“homogeneity (等質性)”であり，他の1つは“exhaustivity (完全性)”である．データ行列中に置かれた測定値がすべて同じ性質を備えている場合，それを homogeneous という．たとえば，性×年齢のクロス表では，そのどのセル内のデータ（度数）もすべて性と年齢で特徴づけられており“等質”である．しかし（個体)×(変数) の形で，変数として性，年齢，所得，……と性質の異なる項目が与えられたときには“非等質 (heterogeneity)”という．しかしこの場

合も，個体側に工場名，変数側には地点名だけを取り上げ，A工場は，K, J, L, …地点にある，と考えると等質となる．

exhaustivity は分析の目的に対する母集団とサンプルの構成（あるいは実験の計画）に関連することである．たとえば日本人全体の成年男女の意見を聞くためにそれを考慮した全国標本が作られれば exhaustive である．しかし北海道だけを対象に標本を作ってそれを全体とするのは，non-exhaustive である．別の例として，青少年の実態を知りたいのに，国民全体を対象にサンプリングしたのではおかしなことになる（つまり non-exhaustive).

“等質性”と“完全性”の概念はおよそ上のようなことであるが，これを基礎にして，実際の分析にあたって，等質と非等質あるいは完全性とそうでない場合，これらの対決をはかるのである．

たとえば，日本人全体を代表すると思われる標本を対象とした調査データに対して，アメリカ人に行った同一の質問のデータをそのまま追加して分析を行えば（個体の追加）全体としては heterogeneous である．しかしこれを日本人だけのデータから布置図を設定し，因子の構造を定め，これに対して，アメリカ人のデータを埋め込み同一の布置図の上にアメリカ人の日本人に対する相対的位置づけを図るという場合を考えてみると，これは等質なデータにもとずく一つの対比分析ということになる（後述の élément supplémentaire, supplementary element の発想).

5. 利用法の主な特徴

最後に Benzécri らが考えている技法の運用上の特徴を２つほど挙げておこう．

その１つは，数値例にみたように，データ行列の行側情報と列側情報を“同時布置”するという使い方である．これに類した工夫は，我々も日常行っている．たとえば，変数側の因子ベクトルの布置図の上に，適当なデモグラフィク要因で層別したデータの数量得点の平均値を，ノルムをそろえた上で埋め込み両者の関連を把握するなどがそれである．その意味で斬新さはないが，やはり特色の１つといえよう．ただ，これを解釈する場合に，若干の問題があろうかと思う．

特徴のその２として，“élément supplémentaire”という考え方がある．数量化Ⅲ類を行ったとき，よく体験する例に，少数の特異データ (outlier) のために布置図の中の各点の分布が明確に読みとりにくいという現象がみられる．たとえば，密集した塊が１つあって，それから離れたところに１つないしは数個の点が位置する，というような場合がそれである．つまり，大勢の傾向に対して，布置の構造を歪めるような悪さをする少数のパターンが入っている場合である．このようなとき，その特異値と思われるケースを削除して再分析を行えばもっともらしい結果が得られることを我々は経験的に知っている．Benzécri のやり方は，ここで計算をやめてしまわないで，一旦除いたデータを，すでに求めた布置図の上に再配置するというのである．こういうときの，この再配置の対象となるデータを，élément supplémentaire（追加あるいは補塡データ）と呼ぶのである．このようにまず安定した布置（つまりデータの大勢が示す傾向）をあらかじめおさえておいて，悪さをするものをあとから埋め込むという考え方は大切であろう．

この考え方を押し進めると前述の日本人とアメリカ人の例にみるように，性格の異なるデータの追加を考えて集団間の比較に利用することもできる．また，手法の双対性に注意すればデータ行列の行側，列側のいずれに対しても，データの追加，あるいは削除を適用することが考えられよう．

さらにこの方式を他の分析手法に適用することも考えられる．実際，M. Roux, Jambu らはクラスター分析にこれを利用した例を発表している．

6. むすび

フランスにおける“Analyse des Données”を紹介したわけであるが，これは筆者の目を通してみた断片にすぎない．しかし，日本の事情と比較

すると数量化法に関しては，日本のほうがはるかに普及しているといえそうである．これは解析対象とするデータの性格にもよるのであろう．フランスでは，主にエコロジー，生物学，……といったどちらかといえば計量的データの処理が主である．意見調査データのように大量の質的データの分析では日本の方が一歩進んでいるようである．

Benzécri は数量化法の今後の方向として，多重クロス表の分析（いわゆる n-way データの数量化法），MDS との関連などを挙げているが，前者については既に岩坪，吉澤などによる優れた研究があり，MDS との関連についても，林の体系的な積み上げがある．

筆者はフランスの技法に対してではなく，むしろフランス流のやり方――すなわち，現象の体験や実験を通して，データと技法の接触面で必要となるデータ処理の手続きを大切にするというやり方――に共感を覚える．

参考文献

1) Benzécri, J.-P. (1973): *L'Analyse des Données*, Tom 1 Taxinomie, Tom 2 L'Analyse des Correspondances, Dunod.
2) Benzécri, J.-P. An Introduction to Taxonomical Studies.
3) Bouroche, J.-M. (ed.) (1977): *Analyse des Données en Marketing*, Masson.
4) Cailliez, F., Pagès, J.-P.(1976): *Introduction à l'Analyse des Données*, SMASH (Societe de Mathematiques Appliques et de Science Humaines)
5) Cehessat, R. (1976): *Exercices Commentés de Statistique et Informatique Appliquées*, Dunod.
6) Diday, E. et al. (1979): *Optimization en Classification Automatique*, INRIA.
7) Guigou, J. L. (1977): *Méthodes Multidimensionnelles*: Analyse des Données et Chix à Critères Multiples, Dunod.
8) Jambu, M., Lebeaux, M.-Q. (1978): *Classification Automatique pour l'Analyse des Données*, Tom 1 Méthodes et Algorithmes, Tom 2 Logiciels.
9) Lebart, L., Fénelon, J.-P. (1975): *Statistique et Informatique Appliquées*, 3e, édition, Dunod.
10) Lebart, L., Morineau, A., Fénelon, J.-P. (1979): *Traitement des Données Statistiques*, Dunod.
11) Romeder, J.-M.: *Méthodes et Programmes d'Analyse Discriminante*, Dunod.
12) Saporta, G. (1979): *Theori et Méthodes la Statistique* TECHNIP.
13) 水野欽司 (1978)：クロス集計表の縮約化，「統計情報の地方における多目的利用に関する調査報告書」，全国統計協会連合会．
14) 柳井晴夫，高根芳雄(1977)：多変量解析法，現代人の統計 2，朝倉書店．

（おおすみ・のぼる，統計数理研究所）