第 14 章

高度な解析法についての概説

14.1 主成分分析

n 個体のサンプルがあって，それぞれについて，p 個の変数 $x_1, x_2, ..., x_p$ の観測値が得られているとする。一般に，p 個の変数の情報を全部一度に考えて n 個体の情報を把握することは難しい。そこで考えられるのが，p 個の変数を，もっと少ない数の，互いに独立な主成分 (principal component) で表せないかということである。

いま，主成分 $\xi_1, \xi_2, \ldots, \xi_p$ を考え，これらを x の一次関数で表すことにする。つまり，

$$\xi_i = \sum_{j=1}^{p} l_{ij} x_j$$

として，p^2 個の適当な係数 l_{ij} を見つけることを考える。各 x_j をそれぞれの平均からの偏差として測れば，どの x_j も n 個体についての和はゼロになり，したがって ξ_i の和もゼロになる。ここで p 個の ξ は互いに無相関であるとする。すなわち

$$E(\xi_i \xi_j) = E(\{\sum_{k=1}^{p} l_{ik} x_k \sum_{m=1}^{p} l_{jm} x_m\})) = 0 \qquad (i \neq j)$$

とする。これだけではまだ $p(p+1)/2$ 個の自由度が残っているので，この変換を直交変換であると条件付ける，すなわち

$$\sum_{k=1}^{p} l_{ik} l_{jk} = 0 (i \neq j), = 1 (i = j)$$

とすれば，符号の付け替えの自由度を加味しても有限組の解が得られることになる (数学的な解は行列の固有値と固有ベクトルを求めることによって得られるが，普

通はコンピュータソフトにやらせるので説明は省略する)。より詳しくは，ケンドール (1981) を参照されたい。

この新しい変数 ξ は主成分と呼ばれる。ξ は，もとの変数 x が正規分布に従うなら互いに独立である。行列の固有値の大きさの順に $\xi_1, \xi_2, \ldots, \xi_p$ と番号をつけると，これらは順に第 1 主成分，第 2 主成分，...，第 p 主成分と呼ばれる。第 1 主成分は，あらゆる一次関数の中で可能な最大の分散をもつ。第 2 主成分は第 1 主成分と無相関な一次関数の中で可能な最大の分散をもつ。このようにして主成分を決めると，それぞれの固有値の，固有値の和に対する割合を使って，それぞれの主成分が全変動の何パーセントを説明するかを表すことができる。それを主成分の寄与率と呼ぶ。普通は，たくさんの変数から少数（たとえば 2 つとか 3 つ）の主成分だけを使って全変動の 80% が説明できる，のように使う。

本当はこんなに少数のデータに使うような分析法ではないのだが，前章の例題で使ったデータについて，R を使って主成分分析をしてみる。まず，`library(mva)` として多変量解析ライブラリを呼び出しておく必要がある。ついで，`mat<-matrix(c(CAR1990,TA1989,DIDP1985),nrow=47)` として `res<-princomp(mat); summary(res)` とすれば，下表が得られる。

	Comp.1	Comp.2	Comp.3
Standard deviation	21.1842224	11.7510982	1.897637799
Proportion of Variance	0.7600359	0.2338654	0.006098678
Cumulative Proportion	0.7600359	0.9939013	1.000000000

この結果から，第 1 主成分の寄与率が 76%，第 2 主成分までの累積寄与率が 99% で，取り上げた 3 つの変数のばらつきは，ほぼ完全に 2 つの直交する主成分に分解できることがわかった。そこで，各都道府県の第 1 主成分 （得点）と第 2 主成分 （得点）を図示するには，`biplot(res,xlabs=PREF)` とすれば，図 14.1 が得られる。

14.2 因子分析

思想は逆だが，数学的には因子分析は，主成分分析に良く似ている。つまり，p 個の観測された変数 x があるときに，個々の x が m 個 ($m < p$) の潜在因子の線型結合と誤差によって表されると考える。たくさんの変数を，別の少数の変数の線型結合によって表すことによって情報を集約する方法論である。R では `factanal` という最尤法で因子分析を行う関数があるが，3 つの変数を 2 つの因子で説明することはできず，1 つの因子しか想定できない（上の例題のデータで `factanal(mat,2)` とする

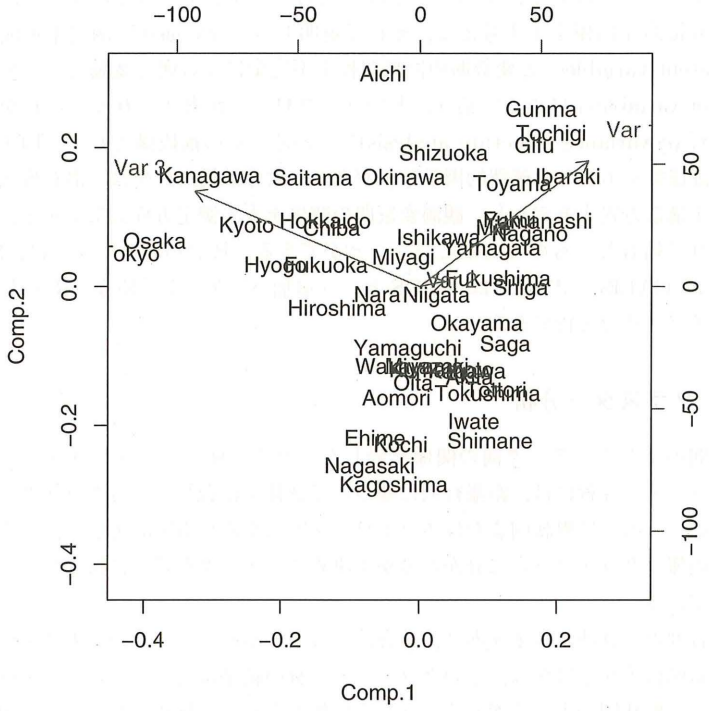

図 14.1 都道府県別の交通事故件数，人口集中地区割合，世帯当たりの自家用車保有台数についての主成分分析の結果

とエラーが出る)。`factanal(mat,1)` とすると，第 1 因子の因子負荷量は 1.764 であり，寄与率は 0.588 である。このことは，取り上げた 3 つの変数は，共通の潜在因子によって約 59% 説明されるということを意味する。少数の主成分また因子による累積寄与率を最大にするために varimax 回転や promax 回転を行うことがあるが，R ではこれらの関数も用意されている。

因子分析は，観測された変数 (observed variables; 観測変数) 間の関係が，実は測定不可能な構成概念 (construct)，すなわち因子 (factor) との関係によって説明されると捉えるモデルであるということもできる。しかし因子分析には，観測変数間の関係は，因子との関係においてしか説明できないし，因子間の因果関係を論じることが

できないし，仮説検証ができないという欠点がある。そこで，測定不可能な因子間の関係もあるだろうけれど，すべてをそれで説明しようとするのではなくて，観測変数間の直接的な関係をまず考えて，それで説明しきれない部分を測定不可能な潜在変数 (latent variables) と変数間の因果関係を不完全にする偶然変動としての誤差変数 (error variables) によって補う，というアプローチが考えられる。これが共分散構造分析 (covariance structure analysis)[*1]である。その前段階として，因子分析に仮説検証機能を追加した確認的因子分析がある。共分散構造分析は，潜在変数間の関係を表す構造方程式モデルと，観測変数間の関係を表す測定方程式モデルを，誤差変数を入れて結合したものであるということができる。統計パッケージでは，SAS では PROC CALIS，SPSS では AMOS という追加パッケージを使う。R でも sem というライブラリで実行できる。

14.3 クラスター分析

変数間のでなく，データ間の関係を表したいときに使うのがクラスター分析である。クラスター分析には，距離行列に基づいて個体を結合しながらクラスターを積み上げていく（出力は樹状図またはネットワーク図になる）階層的手法と，予めいくつくらいの塊（クラスター）に分かれるかを決めて，データを適当に振り分ける非階層的手法がある。

距離行列の計算法にも多々あり，結合法にも多々ある。いくつかの方法でやってみて，樹状図に差がなければ，そのクラスター分析の結果は安定していて，信頼できるといえる。樹状図が大きく変わるようなら信頼できない。解釈としては，変数が足りないために，個体間の関係が十分にわからないと考える。例としては，R で，先ほどのデータを読んで mva ライブラリを呼び出した後で，

```
mat <- matrix(c(CAR1990,TA1989,DIDP1985),nrow=47)
dis <- dist(mat,method="euclidean")
clus <- hclust(dis)
op<-par(ps=8)
plot(clus,PREF,xlab="",ylab="",sub="")
par(op)
```

とすれば，樹状図 (dendrogram とか tree とかいう）が図 14.2 のように描ける（次ページ参照)。

[*1] 共分散構造解析と訳すこともある。

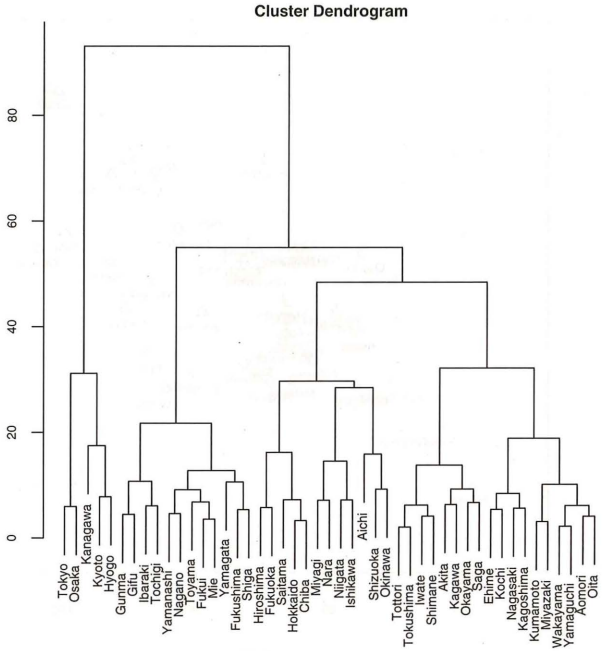

図 14.2 都道府県別の交通事故件数，人口集中地区割合，世帯当たりの自家用車保有台数についての階層的クラスター分析結果

R ではデフォルトの距離の計算法はユークリッド距離（要するに差の二乗和），クラスター結合法は，完全連結法 (complete linkage) である。クラスター分析の結果は見やすいが，解釈には主観が入りがちである。ちなみに山口は和歌山ともっとも近いようである。

非階層的手法の k-means 法の R での実行例も示しておく。5 つのクラスタを仮定すると，データを読んで mva ライブラリ呼び出した後で，次のプログラム

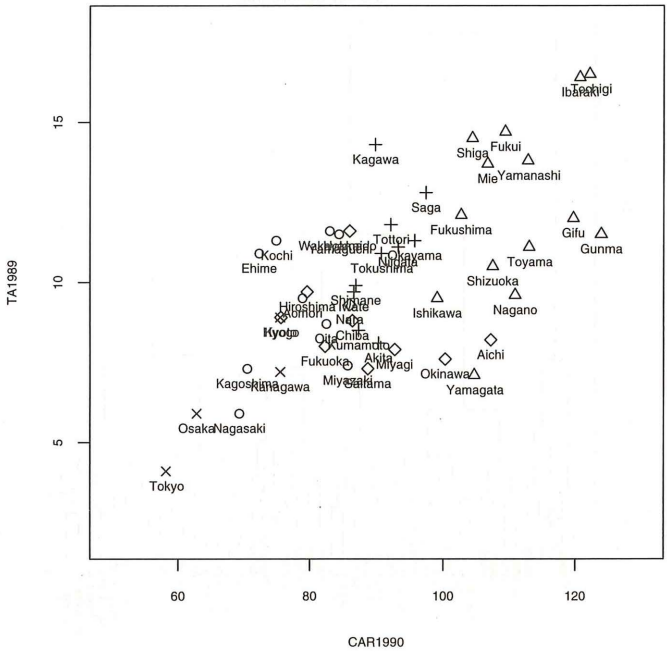

図 14.3 都道府県別の交通事故件数，人口集中地区割合，世帯当たりの自家用車保有台数についての非階層的クラスター分析（**kmeans** 法）の結果

```
clus5 <- kmeans(mat,5)
op<-par(ps=7)
plot(CAR1990,TA1989,pch=clus5$cluster,xlim=c(50,130),
ylim=c(2,18))
text(CAR1990,TA1989,paste(PREF),pos=1)
par(op)
```

によって図 14.3 が得られる。