文章编号: 1001-2486 (2000) 02-0117-04

主方差分析方法*

胡庆军, 吴　翊

(国防科技大学理学院 湖南 长沙　410073)

摘　要: 该文提出一种对高维随机向量 $X=(x_1, x_2, \cdots, x_p)'_{p\times 1}$ 进行降维处理的实用方法, 其基本思想是利用矩阵的扫描运算, 构造 X 的很少几个综合指标(称为主方差变量) 以反映 X 的统计特性。给出了该方法的理论依据和直观解释以及算法。特别指出, 当变量 X 多重相关性突出时, 该文方法显著地优于主成分分析方法。

关键词: 多元分析; 多重相关性; 主方差变量; 贡献率; 筛选算法; 扫描运算

中图分类号: O212.4　　文献标识码: A

The Method of Principal Variance Analysis

HU Qing-jun, WU Yi

(College of Science, National Univ. of Defense Technology. Changsha, 410073, China)

Abstract: A practical method that reduces the dimensions of a high dimensional random vector $X=(x_1, x_2, \cdots, x_p)'_{p\times 1}$ is put forward. Its fundamental idea is, with the sweep operation of matrix, to structure a few synthetical indexes (called principal variance variables) of X to depict X's statistical feature. The theoretical foundation, audio-visual explanation and algorithm of the method are given. The method is markedly superior to of principal component analysis especially when X has serious multi-correlation.

Key words: multivariate analysis; multi-correlation, principal variance variables; contribution rate; selecting algorithm; sweep operation

在对系统(特别是大型系统) 进行分析或评价过程中, 为了更完备地描述系统, 尽可能不遗漏一些举足轻重的系统特性, 分析人员往往倾向于尽可能周到地选取有关指标。这时, 在系统的指标体系中, 所涉及的指标或变量较多, 同时, 变量之间的多重相关性[1] 会突出。于是, 有关高维随机向量 $X=(x_1, x_2, \cdots, x_p)'_{p\times 1}$ 的统计特性分析或高维变量的观测数据的处理在实际问题中是常见的。

众所周知, 识辨系统在一个低维空间要比在一个高维空间容易得多. 主成分分析[2,3] 是在力保数据信息丢失尽可能少的原则下, 对高维变量空间进行降维处理的有效方法之一, 它具有近 70 年的历史。然而, 当变量 X 多重相关性突出时, 主成分分析可能歪曲真实的数据信息, 一些主成分将会过分地夸大某些因素的作用, 无法客观地反映 X 的统计特性[1]。请看下例:

设系统有两个独立因素 X_{I}、X_{II}, 其中对 X_{I} 用四个变量 x_1, x_2, x_3, x_4 来描述, 且 x_1, x_2, x_3 完全相同, $x_4=0.9x_1$; 而对 X_{II} 仅用一个变量 x_5 来描述。记 $X=(x_1, x_2, x_3, x_4, x_5)'$, 且设

$$V \overset{\mathrm{def}}{=} D(X) = \begin{pmatrix} 1 & 1 & 1 & 0.9 & 0 \\ 1 & 1 & 1 & 0.9 & 0 \\ 1 & 1 & 1 & 0.9 & 0 \\ 0.9 & 0.9 & 0.9 & 0.81 & 0 \\ 0 & 0 & 0 & 0 & 1.2 \end{pmatrix}$$

此处表明, 因素 X_{II} 对系统的作用强于因素 X_{I}。

若用主成分分析, 可得 X 的两个主成分, 第一、二主成分分别为

* 收稿日期: 1999-09-15
第一作者: 胡庆军(1959-), 男, 副教授。

$$z_1 = (x_1 + x_2 + x_3 + 0.9x_4)/\sqrt{3.81},\quad z_1 = x_5$$

且 $D(z_1)=3.81, D(z_2)=1.2$, z_1, z_2 的贡献率分别为 76.0% 及 24.0%。由上可知, z_1 是刻划因素 X_{I} 对系统的作用, z_2 表示了系统的重要特征 X_{II}, 且 z_1 对系统的作用远大于 z_2 的作用。这说明第一主成分明显地夸大了因素 X_{I} 对系统的作用; 另外, 若按 75% 的精度反映该系统, 则仅取第一个主成分 z_1, 而 z_2 被完全忽略掉; 这与系统的实际不相符。

鉴于以上分析, 本文利用矩阵的扫描运算, 根据方差大的分量反映 X 的能力强的原则, 构造的综合指标, 用这些综合指标来反映 X 的统计特性, 以达到对高维变量空间降维处理的目的, 同时又能克服类似于主成分分析中的缺陷。为此, 给出如下预备知识。

1 预备引理

定义 设 $V=(v_{ij})_{p\times p}$, $v_{ii}\neq 0$, 定义一个新的方阵 $B=(b_{kl})_{p\times p}$, 其中

$$b_{ii}=1/v_{ii},\ b_{il}=v_{il}/v_{ii},\ b_{li}=-v_{li}/v_{ii},\ l\neq i$$

$$b_{kl}=v_{kl}-\frac{v_{ki}v_{il}}{v_{ii}},\ k\neq i,\ l\neq i$$

则称由 V 到 B 的这种变换为以 v_{ii} 为枢元的扫描运算(或称为 S 运算), 记为 $S_iV=B$.

引理 **1** 对于矩阵 V, 若以下 S 运算能施行, 则

$$S_iS_iV=V,\quad S_iS_jV=S_jS_iV.$$

对于矩阵 $V=(v_{ij})_{p\times p}$, 若 V_{11} 是 V 的 $i_1,\cdots,i_r$ 行(列)形成的子块矩阵, V_{22} 是 V 其余的 $j_1,\cdots,j_{p-r}$ 行(列)形成的子块矩阵, 而 V_{12}、V_{21} 分别对应于 V 的 $r\times(p-r)$、$(p-r)\times r$ 阶子块矩阵。

引理 **2** 若以下 S 运算能施行, 则 $S_{i_r}\cdots S_{i_2}S_{i_1}V\overset{\mathrm{def}}{=}B$ 的结构(B_{ij} 是 B 对应于 V 的分块)为

$$B_{11}=V_{11}^{-1},\ B_{12}=V_{11}^{-1}V_{12},\ B_{21}=-V_{21}V_{11}^{-1},\tag{1}$$

$$B_{22}=V_{22}-V_{21}V_{11}^{-1}V_{12}.\tag{2}$$

推论 **1** 以 $\mathrm{rk}(V)$ 表示矩阵 V 的秩, 在引理 2 下, 则 $\mathrm{rk}(V)=r+\mathrm{rk}(B_{22})$.

若随机向量 $X=(x_1,x_2,\cdots,x_p)'_{p\times 1}$ 的协方差阵 $D(X)\overset{\mathrm{def}}{=}V=(v_{ij})_{p\times p}$, 且有如前的分块形式, X 的分量的相应分块记为 $X_{(1)}=(x_{i_1},\cdots,x_{i_r})'$, $X_{(2)}$ 为 X 的其余分量, $\mu_{(1)}=E(X_{(1)})$, $\mu_{(2)}=E(X_{(2)})$。

引理 **3** 若 V_{11}^{-1} 存在, 令

$$\begin{pmatrix}Z_{(1)}\\ Z_{(2)}\end{pmatrix}=\begin{pmatrix}X_{(1)}\\ X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}\end{pmatrix}$$

则

$$D\begin{pmatrix}Z_{(1)}\\ Z_{(2)}\end{pmatrix}=\begin{pmatrix}V_{11} & 0\\ 0 & V_{22}-V_{21}V_{11}^{-1}X_{12}\end{pmatrix}$$

推论 **2** 对如上的协方差矩阵 V, 则 $V_{22}-V_{21}V_{11}^{-1}V_{12}\geq 0$(非负定矩阵); 若矩阵 $V_{22}-V_{21}V_{11}^{-1}V_{12}$ 对角元均为零, 则 $V_{22}-V_{21}V_{11}^{-1}V_{12}=0$, 且

$$X_{(2)}=V_{21}V_{11}^{-1}(X_{(1)}-\mu_{(1)})+\mu_{(2)},\ \text{a.s}\tag{3}$$

证明 引理 1、2 的证明[4]略。若 A 为常数矩阵, Y 是随机向量, 则由 $D(AY)=AD(Y)A$ 得引理 3. 由

$$\begin{pmatrix}I & 0\\ -V_{21}V_{11}^{-1} & I\end{pmatrix}\begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}\begin{pmatrix}I & -V_{11}^{-1}V_{12}\\ 0 & I\end{pmatrix}=\begin{pmatrix}V_{11} & 0\\ 0 & V_{22}-V_{21}V_{11}^{-1}V_{12}\end{pmatrix}$$

得推论 1。又由 $V\geq 0$, 则 $V_{22}-V_{21}V_{11}^{-1}V_{12}\geq 0$。且 $V_{22}-V_{21}V_{11}^{-1}V_{12}=0$ 的充要条件是该矩阵的对角元均为零. 再由 $D(Z_{(2)})=0$, 则 $Z_{(2)}=E(Z_{(2)})$, a.s. 而得推论 2。证毕。

引理 **4** 记号如前, 则 $V_{21}V_{11}^{-1}(X_{(1)}-\mu_{(1)})+\mu_{(2)}$ 是 $X_{(2)}$ 在 $X_{(1)}$ 中的投影[3]。从而, $X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}$ 是 $X_{(2)}$ 中扣除 $X_{(1)}$ 的线性部分后的剩余向量。

2 主方差分析思想

注意到, 对于随机向量 $X=(x_1,x_2,\cdots,x_p)'_{p\times 1}$, $D(X)=V$, 主成分分析中是以 $\mathrm{tr}(V)$ 的大小来刻划 X 的变化能力。但当变量 X 多重相关性突出时, $\mathrm{tr}(V)$ 包含了的分量之间的相当一部分的重复信息(这从第 1 节的例子可看出), 这显然是不合理的。如何构造很少几个综合指标以至能刻划高维指标 X 的统计特性? 下面根据方差大的分量反映 X 的能力强的原则, 所选指标应尽量少含 X 的分量之间的重复信息、利用矩阵的扫描运算, 来构造 X 的综合指标, 具体如下:

首先, 在 X 中找方差最大(等价于在矩阵 V 中找最大对角元对应)的分量 x_{i_1}(反映 X 的能力最强), 记 $y_1=x_{i_1}$。对作 S 运算 S_{i_1}, 得矩阵 $S_{i_1}V$(形如(1)、(2)式), 记 $X_{(1)}=x_{i_1}$, $X_{(2)}$ 为 X 的其余分量,

$$D\begin{pmatrix}X_{(1)}\\X_{(2)}\end{pmatrix}=\begin{pmatrix}V_{11}&V_{12}\\V_{21}&V_{22}\end{pmatrix} \tag{4}$$

由引理 4、3 知, 是 $X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}$ 中 $X_{(2)}$ 扣除 $X_{(1)}$ 的线性部分后的剩余向量, $X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}$ 与 $X_{(1)}$ 不相关, 且 $D(X_{(2)}-V_{21}V_{11}^{-1}X_{(1)})=V_{22}-V_{21}V_{11}^{-1}V_{12}$。又由引理 2 知, 矩阵 $-V_{21}V_{11}^{-1}$、$V_{22}-V_{21}V_{11}^{-1}V_{12}$ 均在矩阵 $S_{i_1}V$ 中。

若 y_1 不能完全反映 X(等价于矩阵 $V_{22}-V_{21}V_{11}^{-1}V_{12}$ 的最大对角元大于零), 则在 $X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}$ 中找方差最大(等价于在矩阵 $V_{22}-V_{21}V_{11}^{-1}V_{12}$ 中找最大对角元对应)的分量 $x_{i_2}+\beta_{21}x_{i_1}$, 记为 y_2(y_1,y_2 不相关)。对 $S_{i_1}V$ 作 S 运算 S_{i_2}, 得矩阵 $S_{i_2}S_{i_1}$(形如(1)、(2)式), 仍记 $X_{(1)}=(x_{i_1},x_{i_2})'$, $X_{(2)}$ 为 X 的其余分量, 且有(4)式的形式。若 y_1,y_2 仍不能完全反映 X, 则反复利用引理 2、3、4, 在 $X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}$ 中找方差最大的分量, 对 $S_{i_2}S_{i_1}V$ 作 S 运算, 重复前述过程, …, 如此下去, 若 $\mathrm{rk}(V)=r$, 由推论 1 知, 经过 r 次寻找, 可得到 r 个综合指标 $y_1,y_2,\cdots,y_r$:

$$\begin{cases}y_1=x_{i_1}\\ y_2=x_{i_2}+\beta_{21}x_{i_1}\\ \cdots\\ y_r=x_{i_r}+\beta_{r1}x_{i_1}+\beta_{r2}x_{i_2}+\cdots+\beta_{r,r-1}x_{i_{r-1}}\end{cases} \tag{5}$$

由上可知, $D(y_1)\geq D(y_2)\geq\cdots\geq D(y_r)>0$, $y_1,y_2,\cdots,y_r$ 不相关且完全包含了 X 的信息, $\delta\overset{\mathrm{def}}{=}D(y_1)+\cdots+D(y_r)$ 刻划了 $y_1,y_2,\cdots,y_r$ 反映 X 的能力大小。由于 y_1 反映 X 的能力最大, y_2 次之, …, 本文称 y_k 为 X 的第 k 个主方差变量, $k=1,2,\cdots,r$.

本文目的是用尽可能少的主方差变量 $y_1,y_2,\cdots,y_k(k<p)$ 来反映 X 的统计特性, 类似于主成分分析, 引入量

$$\delta_k\overset{\mathrm{def}}{=}D(y_k)/\delta,\ \delta_k\overset{\mathrm{def}}{=}(D(y_1)+D(y_2)+\cdots+D(y_k))/\delta$$

分别称 δ_k 为第 k 个主方差变量 y_k 对 X 的贡献率, δ_k 为前 k 个主方差变量 $y_1,y_2,\cdots,y_k$ 对 X 的累计贡献率。于是, 贡献率越大, 则对应的主方差变量反映 X 的能力越强, 因此在实用中常略去那些贡献率小的主方差变量。一般而言, 若前 k 个主方差变量的累计贡献率达 $\alpha\overset{\mathrm{def}}{=}75\%$(或由经验确定)以上, 则用 $y_1,y_2,\cdots,y_k$ 反映 p 维指标 $X=(x_1,x_2,\cdots,x_p)$ 的统计特性。

3 求主方差变量的算法

(1) 赋值 $V=(v_{ij})_{p\times p}$, $1\Rightarrow k$, $0\Rightarrow r$, $0\Rightarrow\delta$, 以 I、J 表示指标集, 取 I=ϕ(空集), J=$\{1,2,\cdots,p\}$。

(2) 找 $i_k\in J$, 使 $v_{i_ki_k}=\max\limits_{j\in J}v_{jj}$. 若 $v_{i_ki_k}>0$, 则赋值

$$r+1\Rightarrow r,\ v_{i_ri_r}\Rightarrow d_r,\ \delta+d_r\Rightarrow\delta,\ v_{i_ri_j}\Rightarrow\beta_{rj},\ j=1,2,\cdots,r-1,$$

$$\mathrm{I}\cup\{i_r\}\Rightarrow\mathrm{I},\qquad \mathrm{J}-\{i_r\}\Rightarrow\mathrm{J}.$$

若 $r=p$, 则算法结束, 见注 1; 若 $r<p$, 则对 V 施以 S 运算 S_{i_r}, 得 $S_{i_r}V\Rightarrow V$, 然后, $k+1\Rightarrow k$, 返回到(2); 若 $v_{i_ki_k}=0$, 则算法结束, 见注 1。

注 1 此时由系数$\{\beta_{ij}\}$及指标集 I 得到 X 的 r 个主方差变量((5)式) $y_1,y_2,\cdots,y_r$, 且知 $\mathrm{rk}(V)=r$

及 $\delta = D(y_1) + \cdots + D(y_r)$ 的值。最后又阈值 α 决定在实用中所取主方差变量的个数。

4 注释

(1) 当变量 X 多重相关性突出时, 主方差分析方法效果显著, 且显著地优于主成分分析方法。下面的例子足以能说明这一断言。用主方差分析方法分析第1节中的例子, 可得两个主方差变量 y_1, y_2(其中 $y_1 = x_5, y_2 = x_1$ 系数 $\beta_{21} = 0.$), $\mathrm{rk}(V) = 2$, $\delta = D(y_1) + D(y_2) = 2.2$, 反映 X 的能力大小用数值 2.2 (而不是 $\mathrm{tr}(V) = 5.01$) 刻划, 且 y_1, y_2 的贡献率分别为 $\delta_1 = 54.5\%$, $\delta_2 = 45.5\%$。若按累计贡献率为75%的精度反映 X, 则取两个主方差变量 y_1, y_2, 这样的主方差变量充分反映了 X 的统计特性, 且 y_1 反映 X 的能力强于 y_2, 这与系统的实际十分一致。

(2) 若 $\mathrm{rk}(V) = r < p$, 主方差变量为(5)式, 记 $X_{(1)} = (x_{i_1}, x_{i_2}, \cdots, x_{i_r})'$, $X_{(2)}$ 为 X 的其余分量, 且有(4)式的形式, 由第3节主方差变量的构造方法及推论2知, (3)式成立。若实际用前 $k(< r < p)$ 个主方差变量 $y_1, y_2, \cdots, y_k$ 来反映 X 的统计特性, 仍记 $X_{(1)} = (x_{i_1}, x_{i_2}, \cdots, x_{i_k})'$, $X_{(2)} = (x_{j_1}, x_{j_2}, \cdots, x_{j_{p-k}})$ 为 X 的其余分量, 且有(4)式的形式, $Y \overset{\mathrm{def}}{=} (y_1, y_2, \cdots, y_k)'$, 则可认为 $V_{22} - V_{21}V_{11}^{-1}V_{12}$ 的最大对角元 $v_{i_{k+1}i_{k+1}} \approx 0$。由推论2知, 有类似于(3)式的近似线性关系式

$$X_{(2)} \approx V_{21}V_{11}^{-1}(X_{(1)} - \mu_{(1)}) + \mu_{(2)}, \tag{6}$$

且由(1)式知, (6)式的表示系数 $V_{21}V_{11}^{-1}$ 是 k 次扫描后的矩阵 V 的 $j_1, j_2, \cdots, j_{p-k}$ 行与 $i_1, i_2, \cdots, i_k$ 列对应的子块矩阵(符号相反)。又由(5)式知

$$Y = \begin{pmatrix} 1 & & & \\ \beta_{21} & \ddots & & \\ & & 1 & \\ \beta_{k1} & \cdots & \beta_{kk-1} & 1 \end{pmatrix} X_{(1)} \overset{\mathrm{def}}{=} \beta X_{(1)} \tag{7}$$

则 $X_{(1)} = \beta^{-1}Y$, 故(6)式化为

$$X_{(2)} \approx V_{21}V_{11}^{-1}(\beta^{-1}Y - \mu_{(1)}) + \mu_{(2)}, \quad \text{a.s.} \tag{8}$$

(7)、(8)式说明, 主方差变量 Y 是 $X_{(1)}$ 的线性组合, 而 X 的其余分量 $X_{(2)}$ 可由主方差变量 Y 近似线性表示。另外, 相对于主成分分析中的主成分, 由主方差变量的构造方法及(5)式知, 主方差变量有较明确的解释意义。

(3) 本文方法简单易实行, 计算量少。这是以下原因, 对于 p 个变量 $X = (x_1, x_2, \cdots, x_p)$ 的模型每选入一个主方差变量只要作一次扫描运算, 计算量约 p^2 次乘除法; 若 $\mathrm{rk}(V) = r$, 则可构造 r 个主方差变量, 总共约需 $r \times p^2$ 次乘除法。其次数值计算稳定性好, 主要体现在以最大对角元为枢元作扫描运算。

(4) 若 X 的分量由于量纲(或单位)差别太大, 导致各分量的方差差别大, 类似于主成分分析, 可对 X 的相关矩阵 R 用主方差分析方法求主方差变量, 以达到降维的目的。

(5) 在实际问题中, 若 V 或 R 未知, 则先求相应的估计 $\hat{V}$ 或 $\hat{R}$, 然后对 $\hat{V}$ 或 $\hat{R}$ 采用本文的主方差分析方法求主方差变量。

参考文献:

[1] 任若恩, 王惠文. 多元统计数据分析——理论、方法、实例[M]. 北京: 国防工业出版社, 1997: 115～119.

[2] Hotelling, Harold(1933). Analysis of a complex of statistical variables into principal components[J]. Journal of Educational Psychology, 24: 417～441, 498～520.

[3] 张尧庭, 方开泰. 多元统计分析引论[M]. 北京: 科学出版社, 1982.

[4] 陈希孺, 王松桂. 近代回归分析——原理、方法及应用[M]. 合肥: 安徽教育出版社, 1987: 197～201.