第 25 卷 第 1 期　　应 用 数 学 学 报　　Vol. 25 No. 1
2002 年 1 月　　ACTA MATHEMATICAE APPLICATAE SINICA　　Jan., 2002

主变量筛选方法

胡庆军　　吴　翊

(国防科技大学理学院系统科学与数学系, 长沙 410073)

摘　要　本文利用矩阵的扫描运算, 提出一种对高维随机向量 $X=(x_1,x_2,\cdots,x_p)'$ 进行降维处理的实用方法 — 主变量筛选方法, 给出了该方法的理论依据、直观解释、算法及数值例子. 该方法是不同于主成分分析法的一种降维方法. 特别, 当变量 X 多重相关性突出时, 本文方法效果显著.

关键词　主变量, 多重相关性, 线性表示, 筛选方法, 扫描运算

1　问题的提出

有关高维随机向量的统计特性分析或高维变量的观测数据的处理在大型工程问题中是常见的, 如判别分析、聚类分析、典型相关分析、回归分析等方法 [1] 均是针对这类问题而提出的实用方法. 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 [1,2] 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. 然而, 主成分分析有三大不足之处: 其一, 计算量大. 这是因为在对随机向量 $X=(x_1,x_2,\cdots,x_p)'$ 的主成分分析中, 必须用到 X 的协方差阵 $D(X)\triangleq V_{p\times p}$ 的全部特征值和对应的特征向量, 这些量通常是对 V 用迭代法计算得到, 这在 p 很大 (如 $p\geq 30$) 时, 其计算量是惊人的, 且当 V "病态" 严重 (即变量 X 多重相关性 [3]x 突出) 时, 其迭代收敛速度慢, 计算误差 (即舍入误差) 可能大, 致使计算结果不可信. 其二, 由于主成分是 X 的分量的线性组合, 这使得在一个问题中所求主成分只有极少数有解释意义, 而多数主成分很难对原变量 X 作出合理的解释, 这是不争的事实. 其三, 当变量 X 多重相关性突出时, 一些主成分将会过分地夸大某些因素的作用 [见本文第 4 节的 3)], 歪曲真实的数据信息, 无法客观地反映原变量 X 的统计特性.

鉴于以上分析, 本文利用矩阵的扫描运算, 从 X 中直接选择部分变量 (称为主变量), 用这些主变量来反映 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.

2　预备知识

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

$$b_{ii}=1/v_{ii},\qquad b_{il}=v_{il}/v_{ii},\qquad b_{li}=-v_{li}/v_{ii},\qquad l\neq i,$$
$$b_{kl}=v_{kl}-\frac{v_{ki}v_{il}}{v_{ii}},\qquad k\neq i,\ \ l\neq i,$$

本文 1999 年 10 月 10 日收到. 2000 年 10 月 2 日收到修改稿.

则称由 V 到 B 的这种变换为以 v_{ii} 为枢元的扫描运算 (或 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 $S_iS_iV=V$, $S_iS_jV=S_jS_iV$.

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, 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 $S_{i_r},\cdots,S_{i_2}S_{i_1}V\triangleq B$ 的结构 (B_{ij} 是 B 对应于 V 的分块) 为

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

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

引理 1, 2 的证明 [4] 略.

若随机向量 $X=(x_1,x_2,\cdots,x_p)'_{p\times1}$ 的期望 $E(X)=\mu$, 协方差阵 $D(X)=V=(v_{ij})_{p\times p}$, 且有如前的分块形式, X 的分量的相应分块记为

$$X_{(1)}=(x_{i_1},\cdots,x_{i_r})',\qquad X_{(2)}=(x_{j_1},\cdots,x_{j_{p-r}})', \tag{3}$$

$$\mu_{(1)}=E(X_{(1)}),\qquad \mu_{(2)}=E(X_{(2)}).$$

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

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

推论 对如上的协方差矩阵 V, 有

1) $V_{22}-V_{21}V_{11}^{-1}V_{12}\geq0$ (非负定矩阵);

2) 若矩阵 $V_{22}-V_{21}V_{11}^{-1}V_{12}$ 对角元均为零, 则

$$V_{22}-V_{21}V_{11}^{-1}V_{12}=0,\qquad X_{(2)}=V_{21}V_{11}^{-1}(X_{(1)}-\mu_{(1)})+\mu_{(2)},\qquad \text{a.s.} \tag{4}$$

证 若 A 为常数矩阵, Y 是随机向量, 则由 $D(AY)=AD(Y)A'$ 得引理 3. 又由 $V\geq0$, 则 $V_{22}-V_{21}V_{11}^{-1}V_{12}\geq0$, 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 $D(Z_{(2)})=0$, 知 $Z_{(2)}=E(Z_{(2)})$, a.s., 而得推论. 证毕.

3 主变量描述及筛选方法

设随机向量 $X=(x_1,x_2,\cdots,x_p)'_{p\times1}$ 的协方差阵 V 的特征值为 λ_i $(i=1,2,\cdots,p)$, 且 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq0$. 我们知道, 在主成分分析中, 主成分是 X 的分量的线性组合 (这使得在一个问题中所求主成分只有极少数有解释意义, 而多数主成分很难对原变量 X 作出合理的解释), 第 i 个主成分的方差为 λ_i, 且各个主成分的方差之和为 $\sum\limits_{i=1}^{p}\lambda_i(=\operatorname{tr}(V)$, 是 X 各分量的方差之和, 即所谓 X 的 "总方差" [1,2]. 但当变量 X 多重相关性严重时, $\operatorname{tr}(V)$ 含有过多的重复信息, 将会导致一些主成分的过分夸大作用 [3].

如何构造很少几个指标以致能刻划高维指标 X 的统计特性? 为了克服主成分分析方法的不足, 本文从 X 的 "总方差" $\operatorname{tr}(V)$ 出发, 按方差大的原则 (方差大的分量反映

X 的能力强), 采用在对角线上选最大值者作枢元 (对角元至多作一次枢元) 对 V 逐次用 S 运算, 从 X 中逐次选择枢元对应的变量 (如枢元在位置 (3,3), 则对应的变量为 x_3). 若经过 r 次 S 运算, 则得到 r 个变量形如 $X_{(1)}$, 并称这 r 个变量为 X 的主变量, 且按枢元选入顺序将第 k 次枢元对应的变量称为 X 的第 k 个主变量, $k=1,2,\cdots,r$. 如何确定 r? 分析如下:

我们知道, 从 $X_{(2)}$ 中扣除 $X_{(1)}$ 的线性部分后是随机向量 $X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}$ (其实, $V_{21}V_{11}^{-1}(X_{(1)}-\mu_{(1)})+\mu_{(2)}$ 是 $X_{(2)}$ 在 $X_{(1)}$ 中的投影 [1]). 由引理 3 知, $X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}$ 与 $X_{(1)}$ 不相关, 且 $D(X_{(2)}-V_{21}V_{11}^{-1}X_{(1)})=V_{22}-V_{21}V_{11}^{-1}V_{12}$, 其 "总方差" $\mathrm{tr}\,(V_{22}-V_{21}V_{11}^{-1}V_{12})$ 的大小刻划了 $X_{(2)}-V_{21}V_{11}^{-1}X_{(1)}$ 还能反映 X 的能力大小, 而且随着主变量的逐次选入, 使 $\mathrm{tr}\,(V_{22}-V_{21}V_{11}^{-1}V_{12})$ 越来越小, 这由 S 运算的定义, 引理 2 及推论可看出. 若记第 i 次枢元为 d_i, $i=1,2,\cdots,r$, 由引理 2, 3 知, 则量 $\sum_{i=1}^{r}d_i$ 类似于 "总方差" $\mathrm{tr}\,(V)$, 它刻划了已入选的主变量 $X_{(1)}$ 反映 X 的能力大小, 且又扣除了 $X_{(1)}$ 的分量之间的重复信息.

于是, 为确定主变量的个数, 引入量

$$\tilde{\delta}=\mathrm{tr}\,(V_{22}-V_{21}V_{11}^{-1}V_{12}),\qquad \delta=\sum_{i=1}^{r}d_i\Big/\Big(\sum_{i=1}^{r}d_i+\tilde{\delta}\Big) \tag{5}$$

及阈值 α (可取 $\alpha\geq 85\%$), 在每次作 S 运算前, 判断: 若 $\delta\geq\alpha$, 则主变量选择结束, 得到形如 (3) 式的 r 个主变量 $X_{(1)}$ 及类似于 (4) 式的近似线性关系式

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

(这里, 当 $\delta\geq\alpha$ 时, 认为 $\mathrm{tr}\,(V_{22}-V_{21}V_{11}^{-1}V_{12})\approx 0$, 则由推论知有 (6) 式). 若 $\delta<\alpha$, 则再以 $V_{22}-V_{21}V_{11}^{-1}V_{12}$ 的最大对角元为枢元作 S 运算, 从 $X_{(2)}$ 中选出一个新的主变量. 如此下去, 至多作 $p-1$ 次 S 运算.

下面给出主变量筛选的具体方法和步骤:

1) 给定阈值 α, 令 $V^{(0)}=(v_{ij}^{(0)})_{p\times p}=V$, 以 $I^{(i)}$, $J^{(i)}$ 表示指标集, 取 $I^{(0)}=\phi$, $J^{(0)}=\{1,2,\cdots,p\}$, 找 $l_1\in J^{(0)}$, 使 $v_{l_1l_1}^{(0)}=\max\limits_{l\in J^{(0)}}v_{ll}^{(0)}$, $v_{l_1l_1}^{(0)}\Rightarrow\delta_1$, $v_{l_1l_1}^{(0)}\Rightarrow d_1$, 进入第 1 步.

2) 第 r ($r=1,2,\cdots$, 至多 $r=p-1$) 步, 对 $V^{(r-1)}$ 施以 S 运算 S_{l_r}, 得到 $V^{(r)}=S_{l_r}V^{(r-1)}$, 形成指标集 $I^{(r)}=I^{(r-1)}\cup\{l_r\}$, $J^{(r)}=J^{(r-1)}-\{l_r\}$, 并赋值 $\sum\limits_{j\in J^{(r)}}v_{jj}^{(r)}\Rightarrow\tilde{\delta}$, $\frac{\delta_1}{\delta_1+\tilde{\delta}}\Rightarrow\delta$. 若 $\delta\geq\alpha$, 则算法结束, 见注 1; 若 $\delta<\alpha$, 则找 $l_{r+1}\in J^{(r)}$ 使 $v_{l_{r+1}l_{r+1}}^{(r)}=\max\limits_{l\in J^{(r)}}v_{ll}^{(r)}$, $\delta_1+v_{l_{r-1}l_{r+1}}^{(r)}\Rightarrow\delta_1$, $v_{l_{r-1}l_{r+1}}^{(r)}\Rightarrow d_{r+1}$, 然后进入第 $r+1$ 步.

注 1 此时 $J^{(r)}=\{j_1,j_2,\cdots,j_{p-r}\}$, 且 $j_1<j_2<\cdots<j_{p-r}$; 而 $I^{(r)}=\{l_1,l_2,\cdots,l_r\}$, 若按从小到大排列, 记为 $i_1<i_2<\cdots<i_r$, X 的分量仍记为 (3) 式, 则在阈值 α 控制下, 得到 r 个主变量 $X_{(1)}$ 及其余变量 $X_{(2)}$ 由 $X_{(1)}$ 表示的近似线性关系式 (6), 且由 (1) 式知 (6) 式的表示系数 $V_{21}V_{11}^{-1}$ 是矩阵 $V^{(r)}$ 的 $j_1,j_2,\cdots,j_{p-r}$ 行与 $i_1,i_2,\cdots,i_r$ 列对应的子块矩阵 (符号相反), 而且 x_{l_1} 是第一个主变量, x_{l_2} 是第二个主变量, $\cdots$, x_{l_r} 是第 r 个主变量. 同时指出, d_1 是第一个主变量 x_{l_1} 的方差, d_2 是第二个主变量 x_{l_2} 扣除 x_{l_1} 的线性部分后的方差, $\cdots$, d_r 是第 r 个主变量 x_{l_r} 扣除 $x_{l_1},\cdots,x_{l_{r-1}}$ 的线性部分后的方差, 且 $d_1\geq d_2\geq\cdots\geq d_r>0$.

4 注释

1) 本文方法简单易实行, 计算量少. 这是因为, 对于 p 个变量 $X=(x_1,x_2,\cdots,x_p)'$ 的模型, 每选入一个主变量只要作一次扫描运算, 计算量约 p^2 次乘除法, 若整个过程选入 r $(r<p)$ 个主变量, 则约需 $r\times p^2$ 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.

2) 本文方法是用所选主变量 $X_{(1)}$ (r 个分量, $r<p$) 来反映原来变量 X (p 个分量) 的统计特性, 以达到对 X 降维处理的目的. 由于 $X_{(1)}$ 是 X 的部分分量. 因此, $X_{(1)}$ 的含义明确, 且有其余变量 $X_{(2)}$ 的近似线性关系式 (6) (见上节注 1). 另外, 该方法是用 $\delta_1=\sum_{i=1}^{r}d_i$ (而不是 $\mathrm{tr}\,(V_{11})$) 来刻划 $X_{(1)}$ 反映 X 的能力大小, 且该量扣除了 $X_{(1)}$ 的分量之间的重复信息 (见上节), 这也就是选取主变量的基本原则. 于是, 阈值 α 可理解为所选主变量 $X_{(1)}$ 包含了原变量 X 的大约 $\alpha\times 100\%$ 的信息. 一般可取 $\alpha\geq 85\%$ 或由经验确定.

3) 在对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, 往往会出现变量多重相关的现象. 当变量 X 多重相关性突出时, 主变量筛选方法效果显著, 下面的例子足以能说明这一断言.

若系统有两个独立因素 X_I, X_{II}, 其中对 X_I 用四个完全相同的变量 x_1,x_2,x_3,x_4 来描述, 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

$$V=D(X)=\begin{pmatrix}1&1&1&1&0\\1&1&1&1&0\\1&1&1&1&0\\1&1&1&1&0\\0&0&0&0&1.1\end{pmatrix},$$

此处表明因素 X_{II} 对系统的作用强于因素 X_I.

若用主变量筛选方法分析该系统, 可得两个主变量 $X_{(1)}=(x_1,x_5)'$ (其中 x_5 是第一个主变量), 反映 X 的能力大小用 $\delta_1(=d_1+d_2=1.1+1)=2.1$ 刻划 (而不是用 $\mathrm{tr}\,(V)=5.1$, 这是因为 $\mathrm{tr}\,(V)$ 包含了分量 x_1,x_2,x_3,x_4 中过多的重复信息). 这样的主变量 $X_{(1)}$ 能充分反映 X 的统计特性, 这与系统的实际十分一致.

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 $y_1=(x_1+x_2+x_3+x_4)/2$, $y_2=x_5$, 且 $D(y_1)=4$, $D(y_2)=1.1$, y_1 的贡献率达 78.4%. 若按 75% 的精度反映该系统, 则仅取第一个主成分 y_1, 而 y_2 被完全忽略掉; 且 y_1 是刻划因素 X_I 对系统的作用, $y_2=x_5$ 表示了系统的重要特征 X_{II}, 而 y_1 对系统的作用远大于 y_2 的作用, 这与系统的实际不相符. 这说明当变量多重相关性突出时, 主成分分析法可能歪曲真实的数据信息.

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 用主变量筛选方法求主变量, 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 $\widehat{V}$ 或 $\widehat{R}$, 然后对 $\widehat{V}$ 或 $\widehat{R}$ 采用主变量筛选方法求主变量.

6) 对于高维变量问题, 若所选主变量仍太多, 可再对主变量用主成分分析法进行再次降维. 按照主变量筛选方法, 主变量的多重相关性远小于原变量 X 的多重相关性, 故上述处理方法将大大削弱了多重相关性对主成分的影响.

7) 在大型工程问题中，主变量筛选方法是一种中间环节，常用在对高维变量或高维观测数据进行降维得到主变量，再对主变量采用相应的数据分析方法，如典型相关分析、判别分析、回归分析等.

5 数值例子

表 1 24 个月 10 个站的降雨量，单位： mm

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	135.7	123.8	108.1	99.0	150.0	140.3	130.7	107.0	80.2	92.4
2	133.9	121.4	104.9	96.0	145.2	136.7	125.3	93.6	125.4	93.9
3	153.6	140.0	120.1	110.2	167.1	157.5	142.8	115.7	110.2	104.7
4	120.5	111.2	97.1	88.7	133.8	125.6	116.0	70.7	140.1	83.3
5	126.0	115.6	98.5	90.3	137.0	129.6	119.1	98.4	110.7	92.0
6	119.4	107.0	93.9	85.6	129.0	121.0	112.4	75.0	140.3	83.8
7	163.0	148.1	127.2	117.0	178.0	167.2	155.1	64.8	160.7	83.9
8	83.8	76.0	65.7	59.7	89.5	85.0	79.2	97.6	135.2	97.2
9	98.5	88.2	80.9	73.1	108.7	100.8	95.3	85.2	185.0	103.0
10	102.5	90.6	82.0	74.4	111.3	103.3	98.1	114.2	138.3	109.4
11	122.1	108.8	94.1	86.7	132.1	123.4	116.2	67.1	141.9	81.4
12	121.8	111.6	96.7	88.6	134.0	125.9	115.9	80.8	129.1	86.6
13	131.2	121.6	101.1	93.6	143.5	136.1	124.5	70.6	145.0	82.2
14	140.8	126.8	109.1	100.7	154.0	143.8	134.1	62.5	179.2	88.1
15	137.1	128.5	110.6	101.0	152.3	144.1	131.2	95.2	162.7	104.8
16	95.1	84.3	74.6	67.8	101.8	95.4	88.6	93.0	153.4	101.5
17	94.9	86.6	71.5	66.3	102.0	96.9	87.4	106.0	127.2	102.9
18	91.3	82.4	71.9	65.6	98.7	92.8	84.4	109.4	106.8	99.0
19	88.8	78.1	70.2	64.2	97.0	89.5	83.3	110.4	117.5	102.1
20	122.4	109.1	97.8	88.9	133.3	124.1	113.7	115.2	109.9	103.9
21	131.9	120.9	102.4	94.4	144.2	136.0	123.1	75.5	171.7	93.7
22	97.2	86.4	75.1	69.6	106.8	98.8	92.6	57.2	150.2	77.1
23	139.1	126.1	108.2	99.5	151.2	142.2	131.7	97.4	149.9	103.2
24	123.2	113.1	100.4	91.1	136.4	127.9	117.1	94.8	186.1	109.6
$\overline{X}$	119.7	108.6	94.2	86.3	130.7	122.7	113.2	89.9	139.9	95.0

为了研究某河流域汛期的降雨趋势，需要该地区的降雨资料. 已知该地区有十个气象站 $1^{\#} \sim 10^{\#}$，其中 $1^{\#} \sim 7^{\#}$ 建在平原， $8^{\#}$ 建在丘陵， $9^{\#}$ 建在山区， $10^{\#}$ 建在山区与丘陵的交界区域. 以 x_i 表示第 i 站每个月的降雨量， $i = 1, 2, \cdots, 10$，记 $X = (x_1, x_2, \cdots, x_{10})'$. 表 1 给出了随机向量 X 的 (样本 $X_1, \cdots, X_{24}$)24 组数据 (最近八年 [每年 5 、 6 、 7 月三个月] 共 24 个月的降雨资料)，表 2 给出了 X 的协方差阵 V 的估计量 $\widehat{V} = \frac{1}{24} \sum_{i=1}^{24} (X_i - \overline{X})(X_i - \overline{X})'$，而 $E(X) \triangleq (\mu_1, \mu_2, \cdots, \mu_{10})'$ 的估计量 $\widehat{\mu} = \overline{X}$，其中 $\overline{X} = \frac{1}{24} \sum_{i=1}^{24} X_i$ (见表 1). 下面研究各气象站降雨量之间的联系及聚集的趋势.

表 2 样本协方差阵 $\widehat{v}$

436.68	411.86	340.08	315.59	485.16	460.59	417.14	−87.15	45.84	−40.58
411.86	390.63	321.29	298.18	458.43	436.08	393.76	−81.61	45.38	−37.48
340.08	321.29	267.88	247.74	379.31	359.67	326.04	−60.17	42.15	−24.99
315.59	298.18	247.74	229.41	351.81	333.70	302.40	−60.03	38.46	−26.01
485.16	458.43	379.31	351.81	540.53	512.89	464.61	−99.83	57.97	−45.04
460.59	436.08	359.67	333.70	512.89	487.32	440.68	−92.89	52.73	−42.30
417.14	393.76	326.04	302.40	464.61	440.68	400.48	−93.03	54.65	−42.29
−87.15	−81.61	−60.17	−60.03	−99.83	−92.89	−93.03	323.32	−267.98	141.01
45.84	45.38	42.15	38.46	57.97	52.73	54.65	−267.98	683.51	−1.18
−40.58	−37.48	−24.99	−26.01	−45.04	−42.30	−42.29	141.01	−1.18	91.24

若用主变量筛选方法，按 $\alpha = 90\%$ 的原则，对 $\widehat{V}$ 经过三次 S 运算得表 3，同时由第 3 节的注 1 知，可得表 4，此时 (5) 式中的 $\delta = \sum_{i=1}^{3} d_i \Big/ \Big(\sum_{i=1}^{3} d_i + \widetilde{\delta} \Big)$ 达到 99.2%. 故所选主变量为 x_9, x_5 和 x_8，依次为第一、二和三个主变量，且由表 4 知其余变量与主变量之间有近似线性关系 (6) 式. 如：

$$x_1 - \widehat{\mu}_1 \approx 0.8986(x_5 - \widehat{\mu}_5) + 0.0005(x_8 - \widehat{\mu}_8) - 0.0090(x_9 - \widehat{\mu}_9), \tag{7}$$

$$x_{10} - \widehat{\mu}_{10} \approx 0.0093(x_5 - \widehat{\mu}_5) + 0.6472(x_8 - \widehat{\mu}_8) + 0.2513(x_9 - \widehat{\mu}_9), \tag{8}$$

其中 $\widehat{\mu}_i$ 如表 1 的最后一行. (7), (8) 式经化简，得

$$x_1 \approx 0.8986x_5 + 0.0005x_8 - 0.0090x_9, \qquad x_{10} \approx 0.0093x_5 + 0.6472x_8 + 0.2513x_9,$$

表 3 主变量顺序及枢元值

顺序 i	1	2	3	4
x_{l_i}	x_9	x_5	x_8	x_2
d_i	683.5	535.6	207.2	1.8

其余变量 x_2, x_3, x_4, x_6, x_7 也有类似的近似线性关系式 (由表 4). 这些结果与实际十分一致：最重要的是 x_9，反映山区的降雨量；其次是 x_5，反映平原地区的降雨量；再次为 x_8，反映丘陵的降雨量；在平原的其余站的情况均可由 x_5 近似线性表示 (这也说明变量 X 多重相关性突出)，而 x_{10} 是 x_8 和 x_9 的近似线性表示，正是反映山区与丘陵的交界区域. 这说明，通过主变量筛选方法可得，仅由 $9^{\#}$, $5^{\#}$ 和 $8^{\#}$ 站的降雨资料就能反映该地区的降雨趋势.

表 4 用主变量表示的近似线性关系

主变量 / 其余变量	x_5	x_8	x_9
x_1	0.8986	0.0005	−0.0090
x_2	0.8498	0.0078	−0.0026
x_3	0.7088	0.0504	0.0213
x_4	0.6544	0.0252	0.0107
x_6	0.9499	0.0047	−0.0016
x_7	0.8550	−0.0260	−0.0028
x_{10}	0.0093	0.6472	0.2513

若用主成分分析该问题，如表 5，可得 X 的主成分分别为 $y_1, y_2, \cdots, y_{10}$，且 y_1 和 y_2 的累计贡献率达 93%，若按 90% 的原则选取主成分，则只取第一和第二个主成分 y_1 和 y_2. y_1 主要是变量 $x_1, \cdots, x_7$ 的线性组合，其贡献率达 72%. 从实际问题看，y_1 过分地夸大了平原站的降雨资料的作用 (主要原因是 $x_1, \cdots, x_7$ 多重相关性突出，其中包含了过多的平原地区降雨量的重复信息)；y_2 主要是 x_8 与 x_9 的线性组合，但解释意义不够明确；而按主成分原则丢弃的第三个主成分 y_3 主要是 x_8, x_9 和 x_{10} 的线性组合，正好反映的是山区、丘陵的降雨趋势，该主成分的贡献率仅为 6.6%，与 y_1 相比，是处于极其次要的地位，这与实际问题不相符.

最后指出，反映 X 变化大小的量，在主变量筛选方法中，是用 $\sum_{i=1}^{10} d_i = 1431$ 来刻

划；在主成分分析方法中，是用 $\sum_{i=1}^{10}\lambda_i = \mathrm{tr}(\widehat{V}) = 3851$ 来刻划. 但从第 3 节的分析及本例中数据知，$\mathrm{tr}(\widehat{V})$ 含有过多的 X 分量之间的重复信息.

表 5 $\widehat{V}$ 的特征值 λ_i 及特征向量

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
	2782.9	806.67	256.07	2.7615	1.2725	0.5989	0.5640	0.1912	4×10^{-14}	2×10^{-16}
	−0.395	−0.051	0.016	0.159	−0.700	−0.234	−0.104	−0.510	0.000	−0.000
特	−0.373	−0.047	0.022	0.602	0.374	−0.286	−0.153	0.130	−0.111	−0.470
	−0.308	−0.038	0.060	−0.556	0.410	−0.344	−0.154	−0.304	0.430	−0.061
征	−0.286	−0.034	0.037	−0.307	0.197	0.024	0.053	−0.162	−0.859	0.121
	−0.440	−0.048	0.022	−0.098	−0.027	0.654	0.434	−0.035	0.176	−0.381
向	−0.417	−0.049	0.023	0.318	0.213	0.116	0.099	0.063	0.185	0.784
	−0.379	−0.032	−0.001	−0.309	−0.330	−0.092	−0.277	0.752	−0.000	−0.000
量	0.099	−0.458	0.704	0.037	0.016	0.300	−0.434	−0.069	−0.000	−0.000
	−0.073	0.878	0.424	0.024	0.004	0.114	−0.168	−0.036	−0.000	−0.000
	0.042	−0.076	0.564	−0.023	−0.084	−0.438	0.669	0.165	0.000	0.000

参 考 文 献

1 张尧庭，方开泰. 多元统计分析引论. 北京：科学出版社，
(Zhang Y T, Fang K T. Introduction to Multivariate Statistical Analysis. Beijing: Science Press, 1982)

2 Hotelling, Harold. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 1993, 24: 417–441, 498–520

3 任若恩，王惠文. 多元统计数据分析理论、方法、实例. 北京：国防工业出版社，1997
(Ren R E, Wang H W. The They, Method and Practice of Mulivariate Statistical Data Analysis. Beijing: Nationl Defence Industry Press, 1997)

4 陈希孺，王松桂. 近代回归分析原理、方法及应用. 合肥：安徽教育出版社，1987
(Chen X R, Wang S G. The Principle, Method and Application of Modern Regression Analysis. Hefei: An Hui Education Press, 1987)

THE METHOD OF SELECTING PRINCIPAL VARIABLES

HU QINGJUN　　WU YI

(*Department of Systems Science and Mathematics, Institute of National University of Defense Technology, Changsha* 410073)

Abstract A practical method, called selecting principal variables' method, that reduces the dimensions of a high dimensional random vector $X = (x_1, x_2, \cdots, x_p)'$, with the sweep operation of matrix, is presented in this paper. The theoretical foundation, audio-visual explanation, algorithm and numerical example of the method are given. The method differs from one of the principal component analysis. Especially,the advantages of the method are marked, while the variables X's multicollinearity being serious.

Key words Principal variables, multicollinearity, linear expression, selecting method, sweep operation