基于主基底分析的变量筛选

王惠文 仪 彬 叶 明

(北京航空航天大学 经济管理学院, 北京 100191)

摘 要: 利用 Gram-Schmidt 变换,提出一种主基底分析方法.解释并证明了 Gram-Schmidt 变换所删除的信息量.给出"主基底"的定义及构造方法,并提出"净信息含量比"的概念,用以测度所选基底包含的信息.该方法能在原始数据信息损失尽可能小的前提下,排除所有的冗余变量以及变量集合中的重叠信息,得到一个正交的主基底,从而更有效地对大规模变量集合中的信息进行筛选.多角度的理论分析指出,主基底在尽可能多地携带原始变量信息的同时,还可保证样本点间的相似性改变最小.实际案例分析说明了该方法的合理性和有效性.

关 键 词: Gram-Schmidt 变换; 变量筛选; 数据降维; 主基底

中图分类号: O 212

文献标识码: A **文 章 编 号**: 1001-5965(2008)11-1288-04

Variable selection based on principal basis analysis

Wang Huiwen Yi Bin Ye Ming

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: A principal basis analysis method based on Gram-Schmidt process was proposed. Information deleted by Gram-Schmidt process was explained and proved. Principal basis and introduced its construction method was defined. The concept of net information ratio was also put forward to measure the information retained in principal basis. This method selects information effectively from the large-scale variable set while excludes all the redundant variables and reduplicate information, on the promise that the loss of original information is minimized to obtain a mini-dimensional orthogonal basis. From different perspectives, the principal basis can not only retain original information to the utmost extent, but also ensure least similarity changes between elements. Case study was addressed to validate the rationality and effectiveness of this method.

Key words: Gram-Schmidt process; variable selection; dimension reduction; principal basis

在许多研究领域的建模与数据处理工作中,为了不遗漏重要的信息,研究人员往往倾向于选择数量较多的初始变量.这样形成的变量集合规模庞大,变量中的冗余信息较多.由于自变量集合中往往包含严重的多重相关性,如果直接采用所有这些变量进行建模,还将对分析结论造成不良影响.事实上,在多元线性回归、Logistic 回归、Fisher 判别等方法中都有变量筛选的方法.然而从理论上看,当变量集合中存在严重的多重相关性时,这些方法的计算精度和参数检验都会受到损害.

在数据降维方面也有许多非常著名的工作,例如主成分分析、因素分析、典型相关分析,以及偏最小二乘回归[1]、分段逆回归[2]等.这些方法的共同特点是依照某种最优化原则,在原变量 $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_p$ 中提取成分 $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, …, $\boldsymbol{F}_m$ ($m<p$),然后再利用这些成分进行相关的分析工作.但是,由于每一个成分都是原始变量的线性组合,所以这些方法都没有变量筛选的功能.一个突破性的研究进展是文献[3]提出的改良主成分分析方法(M. PCA, Modified Principal Component Analysis).该方法通过变量筛选,找到原始变量集合的一个

收稿日期: 2007-11-08

基金项目: 国家自然科学基金资助项目(70771004, 70521001, 70531010)

作者简介: 王惠文(1957—),女,辽宁大连人,教授, wanghw@vip.sina.com.

子集,使得筛选后的变量尽可能地在表现主成分的结构方面与原变量有很相似的效果.

为了处理多元线性回归中的变量筛选问题,文献[4]提出一种基于 Gram-Schmidt 变换的变量筛选方法.该方法还可以推广到其他"有导师"类的统计模型(即存在因变量的模型).

在本文中,将针对"无导师"类的统计分析(即没有因变量,所有变量的地位都是同等的),讨论大规模变量的筛选方法.提出一种新的变量筛选方法,即利用 Gram-Schmidt 变换来构造变量集合的主基底,以保留最有代表性的变量,进行统计分析.

1　Schmidt 变换及其相关性质

本节主要为提出论文研究方法做理论准备.首先介绍 Gram-Schmidt 变换;然后说明 Gram-Schmidt 变换所删除的信息量.

1.1　Gram-Schmidt 变换

定理 1　任意一组线性无关的变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_s$,总可以经过 Gram-Schmidt 变换把它们变换成正交变量 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_s$. Gram-Schmidt 变换公式见式(1).

$$\left.\begin{aligned}
\boldsymbol{z}_1&=\boldsymbol{x}_1\\
\boldsymbol{z}_2&=\boldsymbol{x}_2-\frac{\boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}\boldsymbol{z}_1\\
\boldsymbol{z}_3&=\boldsymbol{x}_3-\frac{\boldsymbol{x}_3^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}\boldsymbol{z}_1-\frac{\boldsymbol{x}_3^{\mathrm{T}}\boldsymbol{z}_2}{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2}\boldsymbol{z}_2\\
&\vdots\\
\boldsymbol{z}_s&=\boldsymbol{x}_s-\sum_{k=1}^{s-1}\frac{\boldsymbol{x}_s^{\mathrm{T}}\boldsymbol{z}_k}{\boldsymbol{z}_k^{\mathrm{T}}\boldsymbol{z}_k}\boldsymbol{z}_k
\end{aligned}\right\}\tag{1}$$

定理 1 的证明见文献[5].根据定理 1,通过 Gram-Schmidt 变换可以得到空间的一组基底.因此,不难证明推论 1 的结论.

推论 1　对于任意一组秩为 $s(s\leqslant p)$的变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$,对它们做 Gram-Schmidt 正交变换后,得到 $\boldsymbol{z}_1,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_p$.这其中,必有 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_s$ 是相互直交的,而 $\boldsymbol{z}_{s+1}=\cdots=\boldsymbol{z}_p=0$.

从推论 1 可以看出,Gram-Schmidt 变换有两个功能:①将变量集合中的信息进行正交分解;②排除 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 中的冗余变量(即被变换成 0 的那些变量).

1.2　测度经 Gram-Schmidt 变换删除的信息量

在统计数据分析方法中,$\boldsymbol{x}_j$ 的变异程度通常可用方差 $\mathrm{Var}(\boldsymbol{x}_j)=s_j^2$ 来测量.若变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 经过标准化处理(即每一个变量的均值为 0,方差为 1),则标准化后的全部变量的变异信息量为 $\sum_{j=1}^{p}\mathrm{Var}(\boldsymbol{x}_j)=p$.然而,如果在变量集合中存在多重相关性,则 $\sum_{j=1}^{p}\mathrm{Var}(\boldsymbol{x}_j)=p$ 所表述的总变异信息量中就会包含着冗余信息.定理 2 将给出被 Gram-Schmidt 变换所删除的冗余信息量的测度公式,从而指出 Gram-Schmidt 变换是如何剥离 $\boldsymbol{x}_k$ 与 $\boldsymbol{z}_j(j=1,2,\cdots,k-1)$之间的公共信息的.

定理 2　若变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 是标准化的,经过 Gram-Schmidt 变换后,得 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_p$.变量 $\boldsymbol{z}_k(k>1)$所包含的变异信息量为

$$\mathrm{Var}(\boldsymbol{z}_k)=\mathrm{Var}(\boldsymbol{x}_k)-\sum_{j=1}^{k-1}r^2(\boldsymbol{z}_j,\boldsymbol{x}_k)\tag{2}$$

证明　对标准化变量 $\boldsymbol{x}_k$,由于 $\mathrm{Var}(\boldsymbol{x}_k)=1$,即 $\frac{1}{n}\|\boldsymbol{x}_k\|^2=1$,所以 $\|\boldsymbol{x}_k\|^2=n$.根据式(1),有

$$\boldsymbol{x}_k=\boldsymbol{z}_k+\sum_{j=1}^{k-1}\frac{\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{z}_j}{\|\boldsymbol{z}_j\|^2}\boldsymbol{z}_j\tag{3}$$

由于 $\boldsymbol{z}_k$ 与 $\boldsymbol{z}_j(j=1,2,\cdots,k-1)$相互正交,由式(3)可以推出:$\frac{1}{n}\|\boldsymbol{x}_k\|^2=\frac{1}{n}\|\boldsymbol{z}_k\|^2+\frac{1}{n}\sum_{j=1}^{k-1}\left(\frac{\boldsymbol{x}_k\boldsymbol{z}_j}{\|\boldsymbol{z}_j\|^2}\right)^2\|\boldsymbol{z}_j\|^2=\frac{1}{n}\|\boldsymbol{z}_k\|^2+\sum_{j=1}^{k-1}r^2(\boldsymbol{z}_j,\boldsymbol{x}_k)$.　**证毕**

根据式(2),有:

$$\mathrm{Var}(\boldsymbol{x}_k)-\mathrm{Var}(\boldsymbol{z}_k)=\sum_{j=1}^{k-1}r^2(\boldsymbol{z}_j,\boldsymbol{x}_k)$$

从上式可以明显看出,经 Gram-Schmidt 变换后,被删除的冗余信息量恰好等于 $\sum_{j=k}^{k-1}r^2(\boldsymbol{z}_j,\boldsymbol{x}_k)$.

2　主基底的定义及其构造方法

根据第 1 节的讨论,本节提出变量集合的主基底定义,并给出主基底的构造方法.

定义 1　一组秩为 $s(s\leqslant p)$的变量集合 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$,对它们做 Gram-Schmidt 变换后得到 s 个相互正交的变量 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_s$.如果有 $\mathrm{Var}(\boldsymbol{z}_1)\geqslant\mathrm{Var}(\boldsymbol{z}_2)\geqslant\cdots\geqslant\mathrm{Var}(\boldsymbol{z}_s)$,并且 $\sum_{j=1}^{s}\mathrm{Var}(\boldsymbol{z}_j)$取到极大值,则称 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_s$ 为原变量集合所对应的主基底.

按照 Gram-Schmidt 变换的过程和定义 1,不难给出主基底的构造方法.对于一组秩为 $s(s\leqslant p)$的变量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$,可以采用"最大方差法"得到它的主基底.其计算步骤如下.

1) 将 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 做标准化处理;

2) 选择 $z_1=x_k$,使得

$$\sum_{j=1}^{p} r^2(x_k,x_j)=\max_{i=1,2,\cdots,p}\sum_{j=1}^{p} r^2(x_i,x_j)$$

为了记号方便起见,这里不妨设 $k=1$,即 $z_1=x_1$;

3) 分别将剩余变量 $x_2,x_3,\cdots,x_p$ 与 z_1 做 Gram-Schmidt 变换,得到备选的变量集合

$$z_j^{(1)}=x_j-\frac{x_j^{\mathrm{T}}z_1}{z_1^{\mathrm{T}}z_1}z_1 \qquad j=2,3,\cdots,p$$

4) 在 $z_2^{(1)},z_3^{(1)},\cdots,z_p^{(1)}$ 中,选方差最大的备选变量作为 z_2,即 $\mathrm{Var}(z_2)=\max\limits_{j=2,3,\cdots,p}\{\mathrm{Var}(z_j^{(1)})\}$,不妨设 $z_2=z_2^{(1)}$,即 x_2 为第 2 个做 Gram-Schmidt 变换的变量;

5) 对于剩余的变量 $x_3,x_4,\cdots,x_p$,分别将它们与 z_1,z_2 做 Gram-Schmidt 变换,得到备选的变量集合 $z_j^{(2)}=x_j-\frac{x_j^{\mathrm{T}}z_1}{z_1^{\mathrm{T}}z_1}z_1-\frac{x_j^{\mathrm{T}}z_2}{z_2^{\mathrm{T}}z_2}z_2$,$j=3,\cdots,p$;

6) 在 $z_3^{(2)},\cdots,z_p^{(2)}$ 中,选择方差最大的备选变量作为 z_3,即 $\mathrm{Var}(z_3)=\max\limits_{j=3,\cdots,p}\{\mathrm{Var}(z_j^{(2)})\}$;

7) 重复以上过程,直至经 Gram-Schmidt 变换后,得到 s 个相互正交的向量 $z_1,z_2,\cdots,z_s$.

从最大方差法的计算过程可以看出,在排除所有冗余信息后,主基底将携带最多的原始数据信息,即有 $\sum\limits_{j=1}^{s}\mathrm{Var}(z_j)\to\max$. 结合第 1 节的分析还容易证明,按照以下 2 种最优化原则构造的主基底与最大方差法的计算结果是等价的.

1) 相似性改变最小的原则. 如果以点点间的距离来描述样本点之间的相似性,可以证明在上述算法中,把"每次选择方差最大的备选变量"改换成"每次选择使样本群点间的相似性改变最小的备选变量",则所生成的主基底和"最大方差法"是一致的(类似的证明可参考主成分分析中的性质证明,见文献[6]).

2) 直交性最佳的原则. 由于原始变量 $x_1,x_2,\cdots,x_p$ 都是标准化的,根据式(2)有

$$\mathrm{Var}(z_k)=\mathrm{Var}(x_k)-\sum_{j=1}^{k-1}r^2(z_j,x_k)=1-\sum_{j=1}^{k-1}r^2(z_j,x_k)$$

可见,当 $\mathrm{Var}(z_k)\to\max$ 时,必然有 $\sum\limits_{j=1}^{k-1}r^2(z_j,x_k)\to\min$.

因此,在第 $k(k<s)$ 步的计算中,如果每次都选择与 $z_1,\cdots,z_{k-1}$ 相关性总体达到最小的变量 x_k 进行 Gram-Schmidt 变换,则所得到的主基底和"最大方差法"也是一致的. 这种优化原则被称为"直交性最佳的原则".

根据主基底中信息量的特点,下面给出净信息总量和净信息含量比的概念.

定义 2 一组秩为 $s(s\leqslant p)$ 的变量集合 $x_1,x_2,\cdots,x_p$ 经 Gram-Schmidt 变换后,得到主基底 $z_1,z_2,\cdots,z_s$. 称 $z_1,z_2,\cdots,z_s$ 的信息总量 $\sum\limits_{j=1}^{s}\mathrm{Var}(z_j)(j=1,2,\cdots,s)$ 为原变量集合 $x_1,x_2,\cdots,x_p$ 的净信息总量.

定义 3 对于任意的 $z_k(k=1,2,\cdots,s)$,定义它所携带的"净信息含量比"为

$$R_{\mathrm{NI}}=\frac{\mathrm{Var}(z_h)}{\sum\limits_{j=1}^{s}\mathrm{Var}(z_j)} \tag{4}$$

根据累计的净信息含量比,就可以最终确定在应用分析中所需要的主基底的维数.

3 案例研究

本节采用案例分析来验证主基底分析方法的变量筛选效果. 某项关于我国教育事业的研究在分析各年教育水平状况时选取了 9 个变量,它们分别是 x_1—人均 GDP(元),x_2—财政收入(亿元),x_3—财政收入增长率(%),x_4—国家财政科研支出(亿元),x_5—科技三项费用(亿元),x_6—科学支出(亿元),x_7—科研基建费(亿元),x_8—高校数(所),x_9—高校教师数(万人). 共选取了 1992 年到 2005 年 14 年的数据,原始数据见表 1.

通过计算这 9 个变量的相关系数矩阵可以发现,在自变量集合间存在非常严重多重相关性,其中有 26 对变量之间的相关系数超过 0.90.

下面对该变量集合进行筛选分析. 首先,对原始变量做标准化处理. 之后,选取 $z_1=x_k$,使得

$$\sum_{j=1}^{p} r^2(x_k,x_j)=\max_{i=1,2,\cdots,p}\sum_{j=1}^{p} r^2(x_i,x_j)$$

从表 2 中对应的最大数值可知,$z_1=x_2$. 分别将 $x_1,x_3,x_4,x_5,x_6,x_7,x_8,x_9$ 与 z_1 做 Gram-Schmidt 变换,并计算 z_i^1,$i=1,3,4,5,6,7,8,9$ 的方差,见表 3. 由表 3 中对应的最大数值可知,$z_2=z_3^1$,所以 x_3 是主基底采用的第 2 个原始变量.

进一步地,分别将 $x_1,x_4,x_5,x_6,x_7,x_8,x_9$ 与 z_1,z_2 一道做 Gram-Schmidt 变换,在 z_j^2 中取 z_3 使其满足 $\mathrm{Var}(z_3)=\max\{\mathrm{Var}(z_j^2)\}$,重复以上过程,直到所有原始变量都经过 Gram-Schmidt 变换变为相互正交的变量. 计算每个变量的净信息含量 R_{NI},得到主基底中各维变量 z_i,$i=1,2,\cdots,9$ 的净信息含量比 R_{NI}(见表 4). 该表的第 1 行中,

括号内的变量为各个 z_i 所对应的原始变量.

表 1　原始数据表

变量	数　值
x_1	(2 311.1,2 998.4,4 044.0,5 045.7,5 845.9,6 420.2,6 796.0,7 158.5,7 857.7,8 621.7,9 398.1,10 542.0,12 335.6,14 040.0)
x_2	(3 483.4,4 349.0,5 218.1,6 242.2,7 408.0,8 651.1,9 876.0,11 444.1,13 395.2,16 386.0,18 903.6,21 715.3,26 396.5,31 649.3)
x_3	(10.6,24.9,20.0,19.6,18.7,16.8,14.2,15.9,17.1,22.3,15.4,14.9,21.6,19.9)
x_4	(189.3,225.6,268.3,302.4,348.6,408.9,438.6,543.9,575.6,646.1,816.2,975.5,1 095.3,1 111.3)
x_5	(89.4,106.6,114.2,136.0,155.0,190.0,189.9,272.8,277.2,359.6,398.6,416.6,484.0,609.7)
x_6	(57.2,65.6,87.9,96.9,109.7,127.1,151.9,168.1,189.0,223.1,269.9,300.8,335.9,389.1)
x_7	(24.6,34.0,36.1,38.0,48.6,42.7,47.3,52.9,61.5,63.4,70.0,111.1,95.9,112.5)
x_8	(1 053,1 065,1 080,1 054,1 032,1 020,1 022,1 071,1 041,1 225,1 396,1 552,1 731,1 792)
x_9	(38.8,38.8,39.6,40.1,40.3,40.5,40.7,42.6,46.3,53.2,61.8,72.5,85.8,96.6)

注:表中各变量取值为 1992 年～2005 年共 14 年的数据.

表 2　$\sum_{j=1}^{p} r^2(x_i,x_j)$ 的数值

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
$\sum_{j=1}^{r} r^2(x_i,x_j)$	7.36	7.67	1.14	7.57	7.51	7.62	7.28	7.02	7.36

表 3　$z_i^1\ (i=1,3\sim9)$ 的方差

变量	z_1^1	z_3^1	z_4^1	z_5^1	z_6^1	z_7^1	z_8^1	z_9^1
$\mathrm{Var}(z_i^1)$	0.034	0.982	0.026	0.013	0.008	0.082	0.125	0.064

表 4　每个变量的 R_{NI} 值

变量	$z_1(x_2)$	$z_2(x_3)$	$z_3(x_8)$	$z_4(x_7)$	$z_5(x_4)$
R_{NI}/%	44.98	44.17	5.47	3.50	0.82
变量	$z_6(x_1)$	$z_7(x_5)$	$z_8(x_6)$	$z_9(x_9)$	
R_{NI}/%	0.61	0.32	0.10	0.03	

注:第 1 行括号内的变量为各个 z_j 所对应的原始变量.

由表 4 得到的各变量累计净信息含量比,见图 1. 可见,经过 Gram-Schmidt 正交变换后,z_1,z_2,z_3,z_4 的累积净信息含量比达到 98.13%. 所以在该项教育水平的研究中,选取 4 个原变量 x_2 为财政收入(亿元),x_3 为财政收入增长率(%),x_8 为高校数(所)x_7 为科研基建费(亿元),便可以 98.13% 的精度代表所有原变量集合中的净信息量.

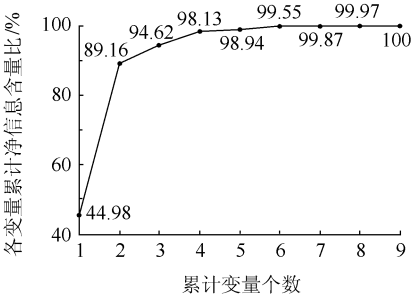

图 1　累计净信息含量比

4　结 束 语

在实际工作中,数据分析人员往往要面对变量数目庞大的大规模数据. 本文利用 Gram-Schmidt 变换,提出一种主基底分析方法. 该方法在"无导师"类的统计分析研究中,能够在原数据信息损失尽可能小的前提下,排除所有的冗余变量以及变量集合中的重叠信息,有效地对大规模变量集中的信息进行筛选,从而得到一个标准正交的主基底. 并且,通过对所选基底的"净信息含量比"的测度,可以更有效地选择具有代表性的变量. 所提方法的信息分解与筛选过程清晰,操作简单易行,便于推广使用.

参考文献(References)

[1] Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method[M]. Heidelberg: Springer-Verlag, 1983

[2] Cook R D, Weisberg S. Sliced inverse regression for dimension reduction[J]. Journal of the American Statistical Association, 1991, 86(414): 328－332

[3] Yutaka Tanaka, Yoichi Mori. Principal component analysis based on a subset of variables: variable selection and sensitivity analysis[J]. American Journal of Mathematical and Management Sciences, 1997, 17(1－2): 61－89

[4] 王惠文,陈梅玲,Gilbert Saporta. 基于 Gram-Schmidt 过程的回归模型及其在刀具磨损预报中的应用[J]. 北京航空航天大学学报,2008,34(6):729－733
Wang Huiwen, Chen Meiling, Gilbert Saporta. Gram-Schmidt regression and application in cutting tool abrasion prediction [J]. Journal of Beijing University of Aeronautics and Astronautics, 2008, 34(6): 729－733 (in Chinese)

[5] Jain S K, Gunawarddena A D. Linear algebra: an interactive approach[M]. Beijing: China Machine Press, 2003

[6] 王惠文. 偏最小二乘回归方法及其应用[M]. 北京:国防工业出版社,1997:93－107
Wang Huiwen. Partial least-squares regression-method and applications[M]. Beijing: National Defense Industry Press, 1997: 93－107(in Chinese)

(责任编辑:吴　琼)