Modeling and Simulation 建模与仿真, 2016, 5(4), 198-204
Published Online November 2016 in Hans. http://www.hanspub.org/journal/mos
http://dx.doi.org/10.12677/mos.2016.54025

Hans汉斯

Principal Basis Analysis and Application in Feature Selection of Water Quality Data

Hui Zou[1,2], Zhihong Zou[1*], Xiaojing Wang[1]

[1]School of Economics and Management, Beihang University, Beijing
[2]School of Science, China Agricultural University, Beijing
Email: huizou@cau.edu.cn, *zouzhihong@buaa.edu.cn

Received: Nov. 11th, 2016; accepted: Nov. 27th, 2016; published: Nov. 30th, 2016

Abstract

With the increasing emphasis on the environment and the improvement of monitoring technology, there appear more and more multivariate data in which the variable sets have multi-collinearity problem. The water quality data of Taizi River belong to this kind of data. In order to avoid the limitation of the traditional method, the principal basis analysis method based on the Gram-Schmidt transform is used to the feature selection of the water quality data of the Taizi River. This method selects information effectively from the large-scale variable set with the minimal loss of original information. Meanwhile, this method can exclude all redundant variables and reduplicate information. Furthermore, it can obtain a mini-dimensional orthogonal basis. Using the measurement of the net information content ratio of the selected features, it is effective to select the representative water quality monitoring variables. It is conducive to the improvement of water quality monitoring work and the experimental results indicate the effectiveness of this method.

Keywords

Gram-Schmidt Transform, Principal Basis, Variable Selection

主基底分析方法及在水质监测指标筛选中的研究

邹　辉[1,2]，邹志红[1*]，王晓静[1]

*通讯作者。

文章引用: 邹辉, 邹志红, 王晓静. 主基底分析方法及在水质监测指标筛选中的研究[J]. 建模与仿真, 2016, 5(4): 198-204. http://dx.doi.org/10.12677/mos.2016.54025

[1]北京航空航天大学经济管理学院，北京
[2]中国农业大学理学院，北京
Email: huizou@cau.edu.cn, *zouzhihong@buaa.edu.cn

收稿日期：2016年11月11日；录用日期：2016年11月27日；发布日期：2016年11月30日

摘　要

随着人们对环境的日益重视和监测技术的提高，水质监测中出现了越来越多变量相关的多变量数据。其中，太子河水质数据属于数据相关的多变量数据。由于传统方法的局限性，本文利用基于Gram-Schmidt变换的主基底分析方法进行太子河水质数据的监测指标筛选工作。这种方法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，可以有效地选择具有代表性的水质监测变量。有利于对水质监测工作进行科学合理的改进。数值实验表明，使用Gram-Schmidt变换的主基底分析方法对太子河水质数据进行分析是有效的。

关键词

Gram-Schmidt变换，主基底，变量筛选

1. 引言

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。这样形成的数据集具有高维度，变量中的冗余信息较多。很多时候自变量集合中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。很多的多元统计方法比如多元线性回归、Fisher判别、聚类分析等都具有变量筛选的功能。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都会受到影响。在数据降维方面也有许多非常经典的方法，例如主成分分析、因子分析、典型相关分析等。已经被广泛应用于环境问题中高维数据的降维和分类[1]-[8]。这些方法的共同特点是依照某种最优化原则，在原变量 $x_1, x_2, \cdots, x_p$ 中提取成分 $F_1, F_2, \cdots, F_m, m < p$，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[9]提出的改良主成分分析方法。该方法通过变量筛选，找到原始变量集合的一个子集，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。此外，一种简便可靠的基于Gram Schmidt过程的主基底分析方法被提出来并应用于指标筛选[10]。该方法可以消除重叠信息的所有冗余变量，在信息的损失降低到最少的前提下，得到一组正交基，可以更有效地进行变量筛选。

本文中，将应用这种方法进行太子河水质数据的指标筛选，更有效地选择具有代表性的变量。

2. 基于 Gram-Schmidt 变换的变量筛选方法

2.1. 主基底分析方法

用于指标筛选的主基底分析方法就是基于主成分分析的 Gram-Schmidt 过程[10]。它的原理如下：

定义 1：记 $x_1, x_2, \cdots, x_p$ 是秩为 $s(s \le p)$ 的向量集，$z_1, z_2, \cdots, z_s$ 是经过 Gram-Schmidt 过程后得到的正交

向量。如果满足条件$Var(z_1) \geq Var(z_2) \geq \cdots \geq Var(z_s)$，并且$\sum_{j=1}^{s} Var(z_j)$达到最大值，则$z_1, z_2, \cdots, z_s$就是相应于原始变量的主基底。

从 Gram-Schmidt 过程和定义 1 得到，最大方差法可以被用于得到一组变量的主基底，具体计算过程如下：

1) 对向量$x_1, x_2, \cdots, x_p$进行标准化；

2) 令$z_1 = x_k$，要求满足

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

不妨假设$k=1$，那么$z_1 = x_1$；

3) 对于剩下的变量$x_2, x_3, \cdots, x_p$，继续分别和z_1进行 Gram-Schmidt 变换，得到一组相应的变量：

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

4) 对于$z_2^{(1)}, z_3^{(1)}, \cdots, z_p^{(1)}$，选择这组变量中方差最大的一个记为$z_2$，that is

$$Var(z_2) = \max_{j=2,3,\cdots,p} \left\{ Var\left(z_j^{(1)}\right) \right\} \tag{3}$$

不妨假设$z_2 = z_2^{(1)}$，也就是x_2是第二个进行 Gram-Schmidt 变换的向量；

5) 对于剩下的变量$x_3, x_4, \cdots, x_p$，分别和z_1, z_2进行 Gram-Schmidt 变换。得到一组相应的变量：

$$z_j^{(2)} = x_j - \frac{x_j^{\mathrm{T}} z_1}{z_1^{\mathrm{T}} z_1} z_1 - \frac{x_j^{\mathrm{T}} z_2}{z_2^{\mathrm{T}} z_2} z_2, \quad j = 3, \cdots, p \tag{4}$$

6) 对于$z_3^{(2)}, \cdots, z_p^{(2)}$，选择具有最大方差的变量记为$z_3$，即

$$Var(z_3) = \max_{j=3,\cdots,p} \left\{ Var\left(z_j^{(2)}\right) \right\} \tag{5}$$

7) 重复以上的过程，得到一组相互正交的向量$z_1, z_2, \cdots, z_s$。

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：$x_1, x_2, \cdots, x_p$是秩为$s(s \leq p)$的一组向量。经过 Gram-Schmidt 变换后得到主基底$z_1, z_2, \cdots, z_p$，主基底的信息总量$\sum_{j=1}^{s} Var(z_j)(j = 1, 2, \cdots, s)$就是原始向量$x_1, x_2, \cdots, x_p$的信息总量。

定义3：对任意的$z_k(k = 1, 2, \cdots, s)$，定义它所携带的“净信息含量比”为

$$R_{N1} = \frac{Var(z_h)}{\sum_{j=1}^{s} Var(z_j)} \tag{6}$$

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。

2.2. 主基底分析和主成分分析的比较

设原始变量集合为$x_1, x_2, \cdots, x_p$，则该数据集合的总信息量为$\sum_{j=1}^{s} Var(x_j)(j = 1, 2, \cdots, p)$，经主成分分析计算后得到的主成分集合为$F_1, F_2, \cdots, F_p$，并且有$Var(F_1) \geq Var(F_2) \geq \cdots \geq Var(F_p) \geq 0$。如果最终保留了$m$个成分，则原始变量的总信息量被分解为两个部分，这可以表达为

$$\sum_{j=1}^{p} Var\left(x_j\right) = \sum_{j=1}^{m} Var\left(F_j\right) + \sum_{j=m+1}^{p} Var\left(F_j\right) \tag{7}$$

其中，$\sum_{j=1}^{m} Var\left(F_j\right)$ 为保留的信息；$\sum_{j=m+1}^{p} Var\left(F_j\right)$ 为被删除的信息。

由于有 $F_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{pj}x_p\ (j = 1, \cdots, p)$，所以很显然，主成分分析不能解决变量筛选的问题。而且，变量之间的多重相关性和冗余信息还会对主成分分析的计算结论产生重要的影响。

与之相反，主基底分析中的信息分解与选择则采用了完全不同的一种形式。对于一组秩为 $s\left(s \le p\right)$ 的变量集合 $x_1, x_2, \cdots, x_p$，经过主基底分析后，得到s个主基底变量 $z_1, z_2, \cdots, z_s$。为了记号方便起见，这里不妨设与 $z_1, z_2, \cdots, z_s$ 对应的关联变量分别为 $x_1, x_2, \cdots, x_s$。于是，其信息分解方式可以表达为

$$\sum_{i=1}^{s} Var\left(x_i\right) = Var\left(x_1\right) + \sum_{i=2}^{s} Var\left(x_j\right) + \sum_{i=s+1}^{p} Var\left(x_j\right) = \sum_{j=1}^{s} Var\left(z_j\right) + \sum_{i=2}^{s}\sum_{j=1}^{i-1} r^2\left(x_i, z_j\right) + \sum_{i=s+1}^{p} Var\left(x_i\right)$$

其中，$\sum_{j=1}^{s} Var\left(z_j\right)$ 为保留信息，也为全部的净信息；$\sum_{i=s+1}^{p} Var\left(x_j\right)$ 为被删除变量的信息。特别值得关注的部分是 $\sum_{i=2}^{s}\sum_{j=1}^{i-1} r^2\left(x_i, z_j\right)$，这里是在由 $x_1, x_2, \cdots, x_p$ 得到 $z_1, z_2, \cdots, z_s$ 的过程中，逐渐被剥离的重叠信息的总和。式(19)说明，通过Gram-Schmidt过程不但可以完整地删除所有的冗余变量，而且还可以有效地拆分和去除 $x_1, x_2, \cdots, x_p$ 之间多重相关性的冗余信息。

3. 案例研究

3.1. 研究区域及数据

太子河发源于抚顺，全长约413 km，流域面积约13,883 km^2。太子河辽阳段位于太子河中游，经本溪市进入辽阳市境内，入口与参窝水库相接，出口进入鞍山境内，境内流程142.8 km，流域面积约4000 km^2，约占全市总面积的85%。太子河辽阳段及其支流汤河、北沙河和柳壕河与参窝水库和汤河水库构成了辽阳地表水监测体系。太子河辽阳段干流上共设有3个监测断面，分别为入市断面参窝坝下断面、国控断面下王家桥断面和出市断面下口子断面。本文采用太子河流域的一个监测点参窝坝下2012年月度监测数据，选取了13个变量，x_1——流量(m^3/s)，x_2——水温(℃)，x_3——pH，x_4——电导率(MS/m)，x_5——溶解氧(mg/l)，x_6——高锰酸盐指数(mg/l)，x_7——五日生化需氧量(mg/l)，x_8——氨氮(mg/l)，x_9——石油类(mg/l)，x_{10}——挥发酚(mg/l)，x_{11}——总磷(mg/l)，x_{12}——氟化物(mg/l)，x_{13}——阴离子表面活化剂(mg/l)。原始数据见表1。

3.2. 水质监测指标的筛选

将主基底分析方法应用于太子河的水质数据，进行监测指标的筛选。

首先，对于原始变量做标准化处理，之后，选取 $z_1 = x_k$，使得

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

从表 2 中对应的最大数值可知，$z_1 = x_{10}$。

通过计算这 13 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。

然后分别将 $x_1, x_2, \cdots, x_9, x_{11}, x_{12}, x_{13}$ 与 z_1 做 Gram-Schmidt 变换，并计算 $z_i^1\left(i = 1,2,\cdots,9,11,12,13\right)$ 的方差，见表 3。由表 3 中对应的最大数值可知，$z_2 = x_3$，所以 x_3 是主基底采用的第 2 个变量。重复以上过程，直到所有的变量都经过 Gram-Schmidt 变换变成相互正交的变量。

Table 1. Original data
表 1. 原始数据表

变量	数值
x_1	(14.01, 8.32, 16.00, 15.60, 36.52, 96.59, 168.00, 66.00, 27.36, 32.00, 75.30, 72.00)
x_2	(1, 1, 1, 12, 14, 15, 15, 18, 22, 15, 10, 4)
x_3	(8.13, 8.11, 7.68, 8.00, 8.33, 8.12, 8.05, 7.79, 7.90, 8.22, 8.10, 8.00)
x_4	(48.3, 48.0, 50.3, 47.6, 43.7, 44.8, 45.3, 40.4, 39.6, 36.1, 33.5, 32.1)
x_5	(9.8, 9.8, 9.2, 10.4, 11.3, 8.0, 6.2, 7.5, 5.7, 6.3, 7.5, 8.3)
x_6	(2.35, 2.60, 2.31, 2.68, 3.16, 2.68, 2.89, 2.69, 2.32, 2.33, 2.47, 1.87)
x_7	(1.0, 2.5, 2.2, 2.6, 2.1, 1.0, 2.4, 2.6, 1.0, 1.0, 1.0 , 1.0)
x_8	(0.747, 0.543, 2.980, 1.958, 1.332, 0.863, 0.988, 0.688, 1.028, 1.399, 0.977, 0.535)
x_9	(0.06, 0.06, 0.06, 0.08, 0.07, 0.08, 0.08, 0.07, 0.06, 0.07, 0.06, 0.09)
x_{10}	(0.001, 0.001, 0.001, 0.001, 0.001, 0.0005, 0.0005, 0.0005, 0.0005, 0.0005, 0.0005, 0.0005)
x_{11}	(0.025, 0.068, 0.060, 0.048, 0.043, 0.040, 0.036, 0.070, 0.056, 0.005, 0.050, 0.020)
x_{12}	(0.504, 0.653, 0.795, 0.565, 0.553, 0.517, 0.377, 0.274, 0.200, 0.265, 0.255, 0.233)
x_{13}	(0.025, 0.063, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025, 0.025)

Table 2. Values of $\sum_{j=1}^{p} r^2\left(x_i, x_j\right)$

表 2. $\sum_{j=1}^{p} r^2\left(x_i, x_j\right)$ 的数值

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$\sum_{j=1}^{p} r^2\left(x_i, x_j\right)$	2.556	2.589	1.782	3.680	3.514	2.185	2.860	2.212
变量	x_9	x_{10}	x_{11}	x_{12}	x_{13}			
$\sum_{j=1}^{p} r^2\left(x_i, x_j\right)$	1.856	4.528	2.450	4.345	2.134			

Table 3. Variance of $z_i^1\left(i=1,3\sim13\right)$

表 3. $z_i^1\left(i=1,3\sim13\right)$ 的方差

变量	z_1^1	z_2^1	z_3^1	z_4^1	z_5^1	z_6^1	z_7^1
$Var\left(z_i^1\right)$	0.578	0.648	0.995	0.453	0.241	0.941	0.797
变量	z_8^1	z_9^1	z_{11}^1	z_{12}^1	z_{13}^1		
$Var\left(z_i^1\right)$	0.813	0.886	0.941	0.305	0.873		

计算每个变量的净信息含量 R_{NI}，得到主基底中各维变量 $z_i\left(i=1,\cdots,13\right)$ 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 $z_i\left(i=1,\cdots,13\right)$ 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，$z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8$ 的累积净信息含量比达到 96.096%。所以在太子河水质数据的研究中，选取 8 个原变量 x_{10} ——挥发酚(毫克/升)，x_3 ——pH，x_9 ——石油类(毫克/升)，x_6 ——高锰酸盐指数(毫克/升)，x_{13} ——阴离子表面活化剂(毫克/升)，x_8 ——氨氮(毫克/升)，x_1 ——流量(m^3/s)，x_4 ——电导率(MS/m)，x_{11} ——总磷(毫克/升)，便可以 96.096%的精度代表所有原变量集合中的净信息量。

Table 4. R_{NI} of variables
表 4. 每个变量的 R_{NI} 值

变量	$z_1(x_{10})$	$z_2(x_3)$	$z_3(x_9)$	$z_4(x_6)$	$z_5(x_{13})$	$z_6(x_8)$	$z_7(x_1)$	$z_8(x_4)$
R_{NI}/%	17.519	17.4340	15.148	14.395	14.257	7.781	5.329	4.234
变量	$z_9(x_{11})$	$z_{10}(x_2)$	$z_{11}(x_7)$	$z_{12}(x_{12})$	$z_{13}(x_5)$			
R_{NI}/%	1.707	1.625	0.271	0.261	0.041			

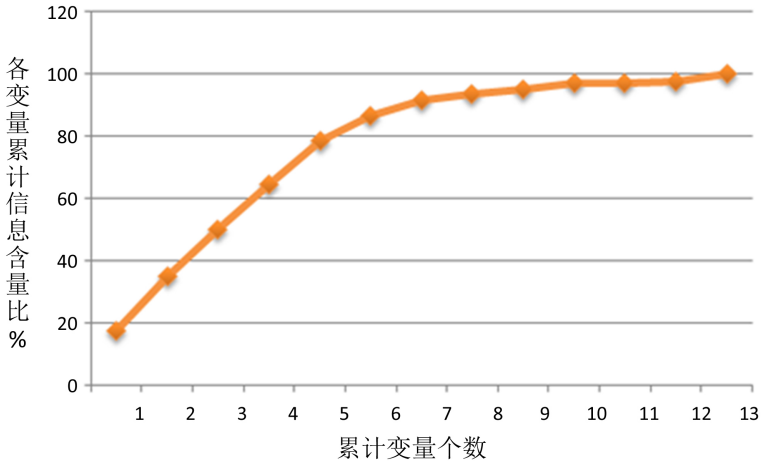

Figure 1. The cumulative net information content ratio
图 1. 累计净信息含量比

4. 结论

在实际工作中，数据分析人员往往要面对变量数目庞大的数据.本文利用基于 Gram-Schmidt 变换的一种主基底分析方法进行太子河水质数据的监测指标筛选工作。有效地利用主基底分析方法对大规模变量集中的信息进行筛选，并且通过对所选基底的“净信息含量比”的测度从 13 个监测指标中选择具有代表性的 8 个水质监测变量。有利于对水质监测工作进行科学合理的改进。

基金项目

国家自然科学基金(No. 51478025)资助。

参考文献 (References)

[1] Shrestha, S. and Kazama, F. (2007) Assessment of Surface Water Quality Using Multivariate Statistical Techniques: A Case Study of the Fuji River Basin, Japan. *Environmental Modelling & Software*, **22**, 464-475. http://dx.doi.org/10.1016/j.envsoft.2006.02.001

[2] Kowalkowski, T., Zbytniewski, R., *et al*. (2006) Application of Chemometrics in River Water Classification. *Water Research*, **40**, 744-752. http://dx.doi.org/10.1016/j.watres.2005.11.042

[3] Wang, X., Lu, Y., *et al*. (2007) Identification of Anthropogenic Influences on Water Quality of Rivers in Taihu Watershed. *Journal of Environmental Sciences*, **19**, 475-481. http://dx.doi.org/10.1016/S1001-0742(07)60080-1

[4] Juahir, H., Zain, S.M., *et al*. (2011) Spatial Water Quality Assessment of Langat River Basin (Malaysia) Using Environmetric Techniques. *Environmental Monitoring and Assessment*, **173**, 625-641. http://dx.doi.org/10.1007/s10661-010-1411-x

[5] Venkatesharaju, K., Somashekar, R.K., *et al*. (2010) Study of Seasonal and Spatial Variation in Surface Water Quality of Cauvery River Stretch in Karnataka. *Journal of Ecology and the Natural Environment*, **2**, 1-9.

[6] Singh, K.P., Malik, A., *et al*. (2005) Water Quality Assessment and Apportionment of Pollution Sources of Gomti River (India) Using Multivariate Statistical Techniques—A Case Study. *Analytica Chimica Acta*, **538**, 355-374. http://dx.doi.org/10.1016/j.aca.2005.02.006

[7] Zhou, F., Liu, Y., *et al*. (2007) Application of Multivariate Statistical Methods to Water Quality Assessment of the Watercourses in Northwestern New Territories, Hong Kong. *Environmental Monitoring and Assessment*, **132**, 1-13. http://dx.doi.org/10.1007/s10661-006-9497-x

[8] Wang, Y., Liu, C., *et al*. (2013) Spatial Pattern Assessment of River Water Quality: Implications of Reducing the Number of Monitoring Stations and Chemical Parameters. *Environmental Monitoring and Assessment*, **186**, 1781-1792. http://dx.doi.org/10.1007/s10661-013-3492-9

[9] Tanaka, Y. and Mori, Y. (1997) Principal Component Analysis Based on a Subset of Variables: Variable Selection and Sensitivity Analysis. *American Journal of Mathematical and Management Sciences*, **17**, 61-89. http://dx.doi.org/10.1080/01966324.1997.10737430

[10] 王惠文, 仪彬, 叶明. 基于主基底分析的变量筛选[J]. 北京航空航天大学学报, 2008, 34(11): 1288-1291.