主成分分析在一类水泥沙观测数据建模中的应用

张超

河海大学理学院，江苏南京（210098）

E-mail：keke.5219@163.com

摘 要:本文首先介绍主成分回归分析是将回归模型中有严重复共线性的变量进行因子提取，得到正交的因子变量，然后对因子变量进行回归模型的建立。通过对一类黄河复杂水泥沙试验数据介绍复共线性的判别方法，并利用主成分回归分析方法建立回归模型。从数据拟合的效果和模型的稳定性方面与普通的最小二乘回归模型进行对比，显示了主成分分析在复共线性条件下建立回归模型的优越性。从而解决了由于复共线而造成病态回归方程的问题。

关键词：主成分分析;复共线性;回归分析

中图分类号：O29

1 引言

在实际数据回归建模过程中，为了全面分析问题，往往涉及众多有关的变量，但是，变量太多不但会增加计算的复杂性，而且也给分析问题和解决问题带来困难。一般来说，虽然每个变量都提供了一定的信息，但其重要性有所不同。实际上，在很多情况下，众多变量都有一定的相关关系，人们希望利用这种相关关系对这些变量加以改造，用为数较少的新变量来反映原变量所提供的大部分信息，通过对新变量的分析，达到解决问题的目的。主成分分析的基本方法是通过构造原变量的新型组合来产生一系列不相关的新变量，从中选出少数几个新变量并使它们含有尽可能多的原变量带有的信息，从而使得用这几个新变量代替原变量来解决问题成为可能。因此，利用主成分分析可以很好的解决回归分析中复共线问题。

水情预报中高精度预报模型的建立，有赖于对历史资料能建立高精度的拟合模型。黄河下游强烈游荡严重淤积，水，沙，河道等影响水位的因素，即使在同一汛期的不同时段上也常差别很大，这使得黄河水位过程特征非常复杂。国内为对黄河这方面的研究甚少[1-4]。本文选取黄河花园口及其下游夹河滩两断面 92 年汛期相应水沙过程资料，按特征点对应采集自变量与相应因变量的值，首先用最小二乘法建立回归模型，然后用主成分回归模型，从模型拟合优度和稳定性方面进行对比分析。所用数据来源于黄河水利委员会水文局（由河海大学袁永生教授提供），分析所用 92 年数据是黄河之典型的高含沙类洪水。

2 评价一个回归模型好坏的一般标准

（1）残差平方和：样本的残差平方和的表达式为 $e_i=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，即观测值与预测值的差的平方和。一个好的拟合方程，其参差平方和越小越好。

（2）调整的复测定系数 R^2： $R^2=1-\frac{SSE/(n-p-1)}{SST(n-1)}$。其中 $SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，这是拟合方程可解释的变异平方和，自由度 $df=p$； $SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$,这是原始数据 y_i 的

总变异平方和，自由度 $df=n-1$；$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，这是残差平方和，自由度 $df=n-p-1$。一般来讲，调整的复测定系数在 0～1 之间，越接近 1，表明拟合直线的优良程度越好。

（3）预测误差平方和 $PRESS$：$e_i=\sum_{i=1}^{n}(y_i-\hat{y}_{(-i)})^2$，$\hat{y}_{(-i)}$ 表示：除去第 i 个样本，用剩余的 $n-1$ 个样本做回归模型，再将去掉的第 i 样本代入模型，得到 $\hat{y}_{(-i)}$。这个标准主要用于判定回归方程的稳健性好不好。也有可以防止选取过拟合的回归模型。

（4）预测的 P R^2：P R^2= $1-\frac{PRESS}{SST}$。$PRESS$ 和 SST 的含义在（1）（3）中。这个准则和（3）的意义差不多，具体数值介于 0～1 之间，越接近 1 越好，对过拟合的回归模型，具有很好的识别能力。

3 主成分回归分析应用背景

3.1 从协方差矩阵求解主成分

1． 引论:设矩阵 $A^T=A$。将 A 的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 依大小顺序排列，不妨设 $\lambda_1>\lambda_2>\cdots\lambda_p$，$\gamma_1,\gamma_2,\cdots,\gamma_p$ 为矩阵 A 各特征值的标准正交特征向量，则对任意向量 x，有

$$\max_{x\neq 0}\frac{x^TAx}{x^Tx}=\lambda_1,\quad \min_{x\neq 0}\frac{x^TAx}{x^Tx}=\lambda_p \tag{1}$$

证明见[5]。

2. 结论[6][7]：设随机向量 $X=(X_1,X_2,\cdots,X_p)^T$ 的协方差矩阵为 Σ，$\lambda_1>\lambda_2>\cdots>\lambda_p$ 为 Σ 的特征值，$\gamma_1,\gamma_2,\cdots,\gamma_p$ 为矩阵 A 各特征值对应的标准正交特征向量，则第 i 主成分为：

$$Y_i=\gamma_{1i}X_1+\gamma_{2i}X_2+\cdots+\gamma_{pi}X_p,\qquad (i=1,2,\cdots,p)$$

此时

$$\mathrm{var}(Y_i)=\gamma_i'\Sigma\gamma_i=\lambda_i$$

$$\mathrm{cov}(Y_i,Y_j)=\gamma_i'\Sigma\gamma_j=0,\ (\ i\neq j\ )$$

证明[6]：由引论知，对于任意常向量 u，有

$$\max_{u\neq 0}\frac{u^T\Sigma u}{u^Tu}=\lambda_1$$

又 γ_i 为标准正交特征向量，于是

$$\gamma_i^T\gamma_j=\begin{cases}1, i=j\\ 0, i\neq j\end{cases}$$

且 $$\gamma_i^T \Sigma \gamma_i = \sum_{k=1}^{p} \lambda_k \gamma_i^T \gamma_k \gamma_k^T \gamma_i = \lambda_i$$

令 $u_i = \gamma_i$，则有

$$\max_{u \neq 0} \frac{u^T \Sigma u}{u^T u} = \lambda_1 = \frac{\gamma_1^T \Sigma \gamma_1}{\gamma^T \gamma} = \mathrm{var}(Y_1) = \lambda_i \sigma_{ij} = 0 \text{。}$$

由以上结论，我们把 $X_1, X_2, \cdots, X_p$ 的协方差矩阵 Σ 的非零特征 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ 对应的标准化特征向量 $\gamma_1, \gamma_2, \cdots, \gamma_p$ 分别作为系数向量，$Y_1 = \gamma_1^T X$，$Y_2 = \gamma_2^T X$，$\cdots$，$Y_p = \gamma_p^T X$ 分别为随机向量 X 的第一主成分，第二主成分，$\cdots$，第 p 主成分。Y 的分量 $Y_1, Y_2, \cdots, Y_p$ 依次是 X 的第一主成分，第二主成分，$\cdots$，第 p 主成分的充分必要条件是：

(1) $Y = u^T X$，$u^T u = I$，即 u 为 p 阶正交阵；

(2) Y 的各分量之间各不相关；

(3) Y 的 p 个分量是按方差大小排列。

于是随机向量 X 与随机向量 Y 之间存在下面的关系：

$$Y = u^T X = \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{bmatrix} \bullet X = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \gamma_1^T \\ \gamma_2^T \\ \vdots \\ \gamma_p^T \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \qquad (2)$$

3. 两个定义：

(1) 称 $a_k = \dfrac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$ $(k = 1, 2, \cdots, p)$ 为第 k 个主成分 Y_k 的方差贡献率，称 $\dfrac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$ 为主成分 $Y_1, Y_2, \cdots, Y_m$ 的累积方差贡献率。

进行主成分分析的目的之一是减少变量的个数，所以一般不会取 p 个主成分，通常以所取 m 使得累计贡献率达到 85%以上为宜，即 $\dfrac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq 85\%$ 。

(2) 第 k 个主成分 Y_k 与原始变量 X_i 的相关系数 $\rho(Y_k, X_i)$ 称为因子负荷量。

因子载荷量是主成分解释中非常重要的解释依据，因子负荷量的绝对值大小刻画了该主成分的主要意义及其成因。

4．主成分的性质[5]

性质 1 Y 的协方差阵为对角阵 $\Lambda=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & \lambda p\end{bmatrix}$。

性质 2 记 $\Sigma=(\sigma_{ij})_{p\times p}$，有 $\sum_{i=1}^{p}\lambda_i=\sum_{i=1}^{p}\sigma_{ij}$

性质 3 $\rho(Y_k,X_i)=u_{ij}\sqrt{\lambda_k}/\sqrt{\sigma_{ii}}$，$k,i=1,2,\cdots,p$

性质 4 $\sum_{i=1}^{p}\rho^2(Y_k,X_i)\sigma_{ii}=\lambda_k$

性质 5 $\sum_{k=1}^{p}\rho^2(Y_k,X_i)=\frac{1}{\sigma ii}\sum_{k=1}^{p}\lambda_k u_{ki}{}^2=1$

3.2 从相关矩阵出发求解主成分

考虑如下的数学变换：

$$令\ Z_i=\frac{X_i-u_i}{\sqrt{\sigma_{ii}}},i=1,2,\cdots,p \qquad (3)$$

式中，u_i 与 σ_{ij} 分别表示变量 X_i 的期望与方差。于是有

$E(Z_i)=0$，$\mathrm{var}(Z_i)=1$

令 $\Sigma^{1/2}=\begin{bmatrix}\sqrt{\rho} & 0 & \cdots & 0\\ 0 & \sqrt{\rho} & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \sqrt{\rho}\end{bmatrix}$，于是，对原始变量 X 进行如下标准化：

$$Z=(\Sigma^{1/2})^{-1}(X-u) \qquad (4)$$

经过标准化后，显然有

$$E(Z)=0$$

$$\mathrm{cov}(Z)=(\Sigma^{1/2})^{-1}\Sigma(\Sigma^{1/2})^{-1}=\begin{bmatrix}1 & \rho_{12} & \cdots & \rho_{1p}\\ \rho_{12} & 1 & \cdots & \rho_{2p}\\ \vdots & \vdots & & \vdots\\ \rho_{1p} & \rho_{2p} & \cdots & 1\end{bmatrix}=R$$

由于上面变换过程，原始变量 $X_1,X_2,\cdots,X_p$ 的相关阵实际上就是对原始变量标准化后的协方差阵，因此，由相关矩阵出发求主成分的过程与从协方差矩阵出发求主成分的过程实际上是一致的。如果用 λ_i，γ_i 分别表示相关阵 R 的特征值与对应的标准正交特征向量，此时求得的主成分与原始变量的关系式为：

$$Y_i=\gamma_i^T Z=\gamma_i^T(\Sigma^{1/2})^{-1}(X-u) \qquad (5)$$

3.3 由样本求主成分

由于总体协方差Σ与相关阵通常是未知的，于是需要通过样本数据来估计。设有n个样品，每个样品有p个指标，这样得到np个数据，原始矩阵资料如下：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}，S=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{ki}-\bar{x}_i)^T，\bar{x}_i=\frac{1}{n}\sum_{k=1}^{n}x_{ki}，i=1,2,\cdots,p$$

$R=(r_{ij})_{p\times p}$，$r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$，其中，$S$为样本协方差阵，用它来估计总体协方差阵；$R$为样本相关矩阵，为总体相关矩阵的估计。由从总体求解主成分的过程，可以容易的推导出有样本求主成分的过程。

3.4 主成分回归

当我们对一组由解释变量组成的矩阵完成主成分的提取，得到从第一主成分到最后一个主称分：$\gamma_1,\gamma_2,\cdots,\gamma_p$，根据累积方差贡献率，选取前$m$($m<p$)个主成分$\gamma_1,\gamma_2,\cdots,\gamma_m$。由于这$m$个新变量彼此不相关，这时再与因变量$Y$做一般的最小二乘回归，就这$m$个新变量而言，避免了多重共线性。

4 主成分回归建模

4.1 复共线性诊断

（1）自变量间的简单相关系数

表1　92年自变量和因变量之间的简单相关系数表

$r(.)$	x_1	x_2	x_3	x_4	y
x_1	1.00000	0.81112	0.97546	0.94426	0.96796
x_2		1.00000	0.91998	0.70532	0.86321
x_3			1.00000	0.89865	0.97337
x_4				1.00000	0.91417
y					1.00000

（2）容许值：$TOL_{(1)}=1.7272\times10^{-5}$，$TOL_{(2)}=5.3891\times10^{-5}$，$TOL_{(3)}=7.692\times10^{-4}$，$TOL_{(4)}=0.0827$

（3）方差膨胀因子：$VIF_1=57897$，$VIF_2=18556$，$VIF_3=1.3\times10^5$，$VIF_4=12.098$

（4）特征根：

表2　92年自变量矩阵的特征值和条件数

特征值	条件数

3.3793e-006	1.0751e+006
0.045347	80.1178
0.32154	11.2991
3.6331	1

由以上分析，可以看出自变量间存在多重共线性。

4.2 普通最小二乘回归建模

用普通最小二乘回归建立线型模型：

$$y=-1790+19.94x_1+24.75x_2-0.2648x_3+1.004e-4x_4 \quad (6)$$

4.3 主成分回归模型

（1）利用 SPSS 软件[8]进行主成分分析，由于原始数据量纲差别很大，故首先对数据标准正交化，见公式（2.4）。我们从相关系数矩阵出发求主成分。首先得到总方差分析表 2.5.3。可以看出，第一个主成分占总方差的比例为 90.857%，故只选取第一个主成分建立模型。第一个成分

$$\gamma_1=-0.515x_1^*-0.472x_2^*-0.522x_3^*+0.489x_4^*。 \quad (7)$$

表3 总方差分析表

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.634	90.857	90.857	3.634	90.857	90.857
2	.321	8.028	98.885			
3	.045	1.115	100.000			
4	3.24E-006	8.10E-005	100.000			

（2）用 y 与 γ_1 做普通的最小二乘回归：

$$y=74.17444-0.22337\gamma_1$$

$$=0.1150356x_1^*+0.1054306x_2^*+0.1165991x_3^*-0.1092279x_4^* \quad (8)$$

γ_1 和常数项的 t 检验和 F 检验的 P 值都小于 0.01。

将（8）还原为原变量的形式为：

$$y=26.8348312+0.1582536x_1+0.3212929x_2+0.001448545x_3+7.63\times10^{-5}x_4 \quad (9)$$

（3）两种回归模型比较

表 4 两种回归模型比较

	普通最小二乘回归	主成分回归
残差平方和	0.09359895	0.1114712
绝对残差最大值	0.1417367	0.1295927
预测误差平方和	0.438940	0.2005685

预测的测定系数	0.8712	0.94166

（4）就 92 年数据：从残差平方和看，普通最小二乘回归模型要比主成分回归模型好一些。但从衡量模型稳定系性的其它三个指标可以看出，主成分回归模型要比普通最小二乘回归模型更稳定一些。这就是说，在用最小二乘拟合回归方程时，出现的过拟合现象。实际上，即使变量之间存在多重共线性，最小二乘拟合的单样本数据还是不错的，但稳定性就很差了。本例很好的说明了这一点。

5 总结

本文给出了主成分分析与回归分析相结合的一般思路，总结出一般主成分回归模型建立的一般步骤，并对黄河花园口及其下游夹河滩两断面 92 年汛期相应水沙过程资料数据进行 OLS 建模和主成分建模，发现利用主成分建立的模型更加稳定，有效的防止了过拟合。在分析黄河下游水位预测方面提供了一个新的解决思路。

参考文献

[1] 袁永生，时政华，朱庆平.改进的多元方差分析用于黄河水位过程研究[J].水利学报,2003,(11):48-53
[2] 胡汝南,张优礼,李世东等.黄河下游变动河床洪水位预报方法的探讨[A].水位预报论文集[C].北京:水利电力出版社，1985：76-83.
[3] 麦乔威，赵亚南，番贤弟等.黄河下游水沙特征及河道冲於规律的研究[A].科学研究论文集[C].郑州:河南科学技术出版社,1990:100~146.
[4] 袁永生.冲於河道相应水位过程中的非线性分析[D].江苏南京:河海大学,2002:38-39.
[5] 李伟明.多元描述统计分析方法[M].上海:华东师范大学出版社,2000,118-142.
[6] 何晓群.多元统计分析[M].北京:中国人民大学出版社,2008,152-173.
[7] 陆旋.实用多元统计分析[M].北京:清华大学出版社,2000:284-428
[8] SPSS for Windows 统计分析（第三版）[M].北京:电子工业出版社,2006:477-489.

The application of Principle Component Analysis in a kind of modeling complicated water-sand data

zhangchao
college of science,Hohai University,Nanjing(210098),China

Abstract

In this paper,Principle Component Analysis is intoduced to draw the new factor variables from collinearity variables in the regression equation.So we can obtain the orthogonal factor variables,then use factor variables to regression ananlysis.Through a kind of complicate sand data,introduce different ways to judge collinearity and introduce how to use factor analysis to establish a good regression equation. Through the goodness of fit and modelstability, Principle Component Regression show more advanced than Ordinary Least Square Regrossion.Then,the problem of ill-condition equlation resule in collinearity is to be solved.

Keywords：*nprincipal Component Analysis;Collinearity;Regression Analysis*