文章编号:1671-6833(2006)02-0117-03

偏最小二乘回归在渗流监控模型中的应用

李宗坤[1], 陈乐意[1], 孙颖章[2]

(1.郑州大学环境与利水学院,河南 郑州 450002;2.水利部小浪底水利枢纽建设管理局,河南 郑州 450000)

摘 要:在渗流监控指标中,库水位之间、库水位与降雨之间存在严重的相关性.利用普通多元线性回归建立渗流监控模型中,监控指标之间存在的多重相关性影响参数估计,扩大模型误差,破坏模型的稳健性.为了克服多重相关性对模型的干扰,引入了能辨别系统信息与噪声的偏最小二乘回归,并编制了程序.算例分析表明,偏最小二乘回归模型所分离出的各个影响分量能对大坝实测变量的变化作出合理的物理成因解释,而且偏最小二乘回归模型的预测能力也远优于普通最小二乘回归模型,前者的预测误差平方和约只有后者的二十分之一.

关键词:偏最小二乘回归;渗流监控模型;原型观测

中图分类号:TH 16;TG 68 **文献标识码**:A

0 引言

一般地,为了更完备地描述和分析系统,尽可能不遗漏一些至关重要的系统特征,往往较周到地选取有关指标,而这样构成的多指标系统常存在严重的多重相关性.土石坝的浸润线高低直接影响边坡稳定,是安全监测的必测项目.在土石坝渗流监测模型中,影响测压管水位主要因素有:库水位、降雨和筑坝材料的渗透时变特性三个方面.由于库水位变化传递到测压管位置还存在一定的滞后效应[1],所以库水位包括前期库水位.在目前常用的普通最小二乘回归(ordinary least-squares regression,OLSR)建模中,当天库水位和前期库水位、降雨之间存在的相关性会严重影响参数估计、扩大模型误差、破坏模型的稳健性.针对最小二乘回归模型存在的不足,本文将偏最小二乘回归法[2](partial least-squares regression ,PLSR)引入土石坝渗流监控模型的建模分析[3].

偏最小二乘回归是一种新型的多元统计分析方法,它于1983年由伍德(S. Wold)和阿巴诺(C. Albano)等人首次提出,它集多元线性回归、典型相关分析和主成分分析的基本功能于一体,将建模预测类型的数据分析方法与非模式的数据认知性分析有机结合起来.与普通最小二乘回归相比,偏最小二乘回归能对数据进行分解和筛选,提取对因变量解释性最强的综合变量,辨识系统中的信息与噪声,从而更好的克服多重相关性在系统建模中的不良作用.

1 建模思路与算法

1.1 建模思路

PLSR[4]的目的是在解释变量空间里寻找某些线性组合,以能更好的解释因变量的变异信息.设有 p 个自变量 $\{x_1, x_2, \cdots, x_p\}$,和 q 个因变量 $\{y_1, y_2, \cdots, y_q\}$.为了研究因变量和自变量的统计关系,观测了 n 个样本点,由此构成了自变量和因变量的数据表 $X = \{x_1, x_2, \cdots, x_p\}_{n\times p}$ 和 $Y = \{y_1, y_2, \cdots, y_q\}_{n\times q}$.PLSR 的基本思路是:分别在 X 和 Y 中提取成分 t_1 和 u_1(t_1 是 $x_1, x_2, \cdots, x_p$ 的线性组合,u_1 是 $y_1, y_2, \cdots, y_q$ 的线性组合).在提取这两个成分时,为了回归分析的需要,有下列两个要求:①t_1 和 u_1 尽可能多地提取各自数据表中的变异信息;②t_1 和 u_1 相关程度能够达到最大.

在第一个成分 t_1 和 u_1 被提取后,PLSR 分别实施 X 对 t_1 的回归以及 Y 对 u_1 的回归.如果回归方程已经达到满意的精度,则算法终止,否则,将利用 X 对 t_1 解释后的残余信息以及 Y 被 u_1 解释后的残余信息进行第二轮的成分提取.如此往复,直到能达到一个较满意的精度为止.若最终对 X 共提取了 m 个成分 $t_1, t_2, \cdots, t_m$,PLSR 将进行

收稿日期:2006-02-09;修订日期:2006-03-11

基金项目:河南省自然科学基金资助项目(511050100)

作者简介:李宗坤(1961—),男,河南南阳人,郑州大学教授,博士,主要从事大坝安全评价方面的研究.

y_k对$t_1, t_2, \cdots, t_m$的回归,然后再表达成y_k对$x_1, x_2, \cdots, x_p$的回归($k=1,2,\cdots,q$).

1.2 交叉有效性(cross validation)识别

在PLSR建模中,究竟选取多少个成分为宜[5],可以考察增加一个新的成分后模型的预测功能是否有明显改进.在单因变量PLSR中,一般采取以下方法:除去某个样本点i的所有样本点组成新的样本,使用$t_1 \sim t_h$个成分拟合一个回归方程,得到y在样本i上的拟合值$\hat{y}_{h(-i)}$.对每一个样本点重复上述计算,定义y的预测误差平方和为$PRESS_h$,有

$$PRESS_h = \sum_{i=1}^{n}(y_i - \hat{y}_{h(-i)}) \tag{1}$$

如果回归方程的稳健性不好,误差很大,它对样本点的变动十分敏感,这种扰动误差作用就会增加$PRESS_h$值.再采用所有的样本点拟合含h个成分的回归方程,得到第i个样本点的预测值$\hat{y}_{h(-i)}$,定义y的误差平方和为SS_h,有

$$SS_h = \sum_{i=1}^{n}(y_i - \hat{y}_{hi}) \tag{2}$$

$PRESS_h$比SS_{h-1}增加了一个成分t_h却含有样本扰动点的误差,如果采用h个成分含有扰动误差的$PRESS_h$能在一定程度上小于采用$(h-1)$成分的SS_{h-1},则认为增加一个成分t_h会使预测的精度明显提高.一般认为:当$PRESS_h/SS_{h-1} \leqslant 0.95^2$时,增加成分$t_h$是有用的,否则,就认为增加新的成分对减小方程的预测误差无明显的改善作用.

1.3 PLSR的算法

首先将数据作标准化处理.X经标准化处理后的数据矩阵记为E_0,Y经标准化处理后的矩阵记为F_0.

(1) 求矩阵$E'_0F_0F'_0E_0$最大特征值所对应的单位特征向量w_1,第一个成分t_1,得

$$t_1 = E_0w_1 \tag{3}$$

$$P_1 = E'_0t_1/\|t_1\|^2 \tag{4}$$

式中:P_1为第一轮提取过程中的回归系数.

$$E_1 = E_0 - t_1p'_1 \tag{5}$$

式中:E_1为E_0被第一轮提取后的残余矩阵.

(2) 重复第一步至第h个成分

$$t_h = E_{h-1}w_h \tag{6}$$

式中:t_h为第h个成分;w_h为矩阵$E'_{h-1}F_{h-1}E_{h-1}$最大特征值所对应的单位特征向量;E_{h-1}为E_0经过$h-1$轮提取后的残余矩阵.

$$P_h = E'_{h-1}t_h/\|t_h\|^2 \tag{7}$$

式中:P_h第h轮提取过程中的回归系数.

$$E_h = E_{h-1} - t_hp_h \tag{8}$$

根据交叉有效性,确定提取h个成分$t_1, t_2, \cdots, t_h$,可以得到一个满意的模型.

2 计算实例

陆浑水库大坝为黏土斜墙坝,上游坝坡1:3.25~1:3.5,下游1:2.25~1:2.75,全长710 m,坝高55 m,总库容13.16亿m^3.大坝于1959年开始兴建,1965年8月基本建成.自建成以来,由于存在洪水标准偏低及西坝肩、坝基渗透稳定问题,未按设计水位正常运用,被列为重点病险库之一.以测压管S1为例,利用matlab编制了PLSR程序.

2.1 统计模型

土石坝浸润线的测压管的实测资料分析表明,其主要受上下游水位、降雨以及筑坝材料的渗流时变特性的影响,即

$$h = \sum_{i=1}^{3}a_{ui}h_{ui} + a_dh_d + \sum_{i=1}^{3}c_ip_i + d_i\theta + d_2\ln\theta \tag{9}$$

式中:h_{u1}, h_{u2}, h_{u3}分别为观测当天、观测日前7 d及前15 d的水位;a_{u1}, a_{u2}, a_{u3}分别为对应h_{u1}, h_{u2}, h_{u3}的回归系数;h_d, a_d为下游水位及对应的回归系数;p_1, p_2, p_3分别为观测当天、观测日前7 d及前15 d的降雨量;c_1, c_2, c_3分别为对应p_1, p_2, p_3的回归系数;θ为蓄水初期开始的天数除以100;d_1, d_2为时效分量的回归系数.

2.2 资料及计算结果分析

模型中9个监控指标的相关系数矩阵见表1.从相关系数矩阵中可以看出,在监控指标之间存在严重的多重相关性,例如$r(h_{u1}, h_{u2}) = 0.993$, $r(h_{u2}, h_{u3}) = 0.995$, $r(d_1, d_2) = 0.998$.这严重背离了线性回归中关于自变量间相互独立的假定,因此普通最小二乘回归的结果也是难以让人信服的.

PLSR和OLSR所得的回归系数见表2.在PLSR结果中,除了上游当日库水位外,上下游水位、降雨均与测压管水位正相关,两个时间分量均与测压管水位负相关.由于该测压管位于斜墙下游并存在滞后效应,因此其水位和下游水位较大的正相关及上游当日库水位微小的负相关是可以接受的.而OLSR得出的回归系数a_{u1}, a_{u3}, c_2, d_1根本不符合物理意义,其结果自然不能接受.PLSR和OLSR的复相关系数分别为0.920,0.935,均为高度相关,虽然前者略低于后者,但回归的目的在于揭示变量间潜在的规律,及时发现工程运

行中的安全隐患,为管理层提供科学决策支持,充分发挥工程效益并减少突发事件的发生,而不是单纯地追求某种统计指标.

表 1　监控指标的相关系数

Tab.1　Correlation coefficients of monitoring indicators

	h_{u1}	h_{u2}	h_{u3}	h_d	p_1	p_2	p_3	d_1	d_2
h_{u1}	1	0.993	0.979	0.549	-0.195	-0.530	0.220	0.510	0.532
h_{u2}		1	0.995	0.609	-0.211	-0.479	0.185	0.565	0.584
h_{u3}			1	0.641	-0.169	-0.465	0.123	0.584	0.604
h_d				1	-0.381	-0.310	-0.173	0.477	0.471
p_1					1	-0.117	-0.182	-0.361	-0.345
p_2						1	0.144	0.031	0
p_3							1	0.196	0.193
d_1								1	0.998
d_2									1

表 2　PLSR 和 OLSR 的回归系数

Tab.2　Regression coefficients of PLSR & OLSR

系数	a_{u1}	a_{u2}	a_{u3}	a_d	c_1	c_2	c_3	d_1	d_1
PLSR	-0.006	0.006	0.014	1.414	0.030	0.011	0.006	-0.002	-0.512
OLSR	-0.523	0.763	-0.221	1.021	0.050	-0.055	0.071	0.070	-6.404

统计回归的一个重要的目的就是预测,因而预测能力是判断一个模型优劣的重要标准.为了比较 PLSR 和 OLSR 的预测能力,选择了 6 个样本分别代入两个模型,其结果见表 3.PLSR 和 OLSR 的预测偏差平方和分别为 0.553 和 10.885,后者是前者的近 20 倍,可见 PLSR 的预测能力远优于 OLSR.由于 OLSR 算法的本身的特点,使其具有较好的拟合能力,但自变量多重相关是其无法逾越的障碍,因此 OLSR 不会有很好的预测能力.

表 3　PLSR 和 OLSR 的预测结果

Tab.3　Prediction results of PLSR & OLSR

样本	1	2	3	4	5	6
实测值	275.92	275.79	275.94	275.67	275.90	276.37
PLSR 回归值	275.82	275.62	275.49	275.74	276.35	276.05
OLSR 回归值	276.20	274.63	275.49	276.17	272.92	276.01

3　结束语

在影响大坝安全监测变量的各类因子之间,往往存在着多重共线性.这种因子之间的多重共线性会导致回归分析的正则方程组出现病态,变量之间的相关性将会严重影响参数估计、扩大模型误差、破坏模型的稳健性.针对普通多元回归难以得到合理的结果,作者应用偏最小二乘回归建立的渗流统计模型,较好地解决许多以往用普通多元回归无法解决的自变量之间的多重共线性问题.工程实例对比分析表明,普通最小二乘回归模型所分离出的各个影响分量无法对大坝实测变量的变化作出合理的物理成因解释,而偏最小二乘回归模型能较好的解释各影响因子对测压管水位的影响.在预测能力上,普通最小二乘回归模型的预测偏差平方和将近是偏最小二乘回归模型的 20 倍,可见,偏最小二乘回归模型的预测能力远优于普通最小二乘回归.偏最小二乘回归模型的结论可靠,整体性强,比目前常用最小二乘回归模型有更广泛的适用性.

参考文献:

[1] 张乾飞,顾冲时,吴中如.基于滞后效应的土石坝渗流监控模型[J].水利学报,2002,(2):85~89.

[2] 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,1999.

[3] 邓念武.偏最小二乘回归在大坝位移资料分析中的应用[J].大坝观测与土工测试,2001,25(6):16~18.

[4] DOYMAZ F. Orthogonal nonlinear partial least - squares regression[J]. Industrial and Engineering Chemistry Research, 2003,42(23):5836~5849.

[5] 杨　杰,胡德秀,吴中如.大坝安全监控模型因子相关性及不确定性研究[J].水利学报,2004,(12):99~105.

(下转第 123 页)

参考文献：

[1] 詹友刚.Pro/E 中文野火版教程[M].北京:清华大学出版社,2004,125～126.

[2] 张淑珍.CAD 系统二次开发方法的研究[J].西北纺织工学院学报,2000,22(3):25～28.

[3] 杨　萍,韩　飞. 基于 PRO/E 软件二次开发复杂钣金件的展开设计[J].工程图学学报,2005,21(5):157～62.

[4] 王艳萍,胡金星.基于 Pro/E 软件的产品三维曲面造型设计方法[J].机械工人·冷加工. 2002,24(3):67～68.

[5] 陈立平,张云清,任卫群,等.机械系统动力学分析及 ADAMS 应用教程[M].北京:清华大学出版社,2005.

Model Analysis and Emulation about the Device of Supercharger Based on Virtual Technologyn

LIU Jun, ZHAO Dong－hui, ZHANG Hui, GAO Xian－kun

(School of Mechanical & Electrical Engineering, Henan Agricultural University, Zhengzhou 450002, China)

Abstract: In this paper, the parametric model of the supercharge device is set up by Pro/E, based on this, the kinematics model is built, and the working device is simulated. First, the connecting rod of the model is done finite element analyzed by the Ansys software. Then with Mechanical/Pro, which is a seamless connected software between Pro/E and Adams, the piston and the connecting rod which constructed by Pro/E are received dynamics analyzed by Adams, and sport curves in many kinds of cases have been got, and every data are reached the actual requirements basically.

Key words: supercharger; kinetic simulation; virtual manufacturing

(上接第 119 页)

Application of Partial Least－squares Regression to Seepage Monitoring Model

LI Zong－kun[1], CHEN Le－yi[1], SUN Ying－zhang[2]

(1. School of Environment & Water Conservancy Engineering, Zhengzhou University, Zhengzhou, 450002, China; 2. Construction Management Bureau of Xiaolangdi Hydraulic Engineering, Ministry of Water Resources, Zhengzhou 450000, China)

Abstract: Among the indicators of seepage monitoring model, there is serous collinearitiy between each water level, water levels and rainfalls. In a seepage monitoring model built by ordinary multilinear regression, the multicollinearity between each monitoring indicator will influence the parameter estimation, enlarge the model error and damage the robustness of model. To avoid multicollinearity's disturbance, partial least－squares regression which can identify system information and noise is introduced into the model, and a program is compiled. It is illustrated by a case that the components of partial least－squares model can give a reasonable physical interpretation to variation of prototype observation data, and predictive power of partial least－squares regression is stronger than ordinary mutilinear regression, the sum square predictive error of former is nearly one of twentieth of the latter.

Key words: partial least－squares regression; seepage monitoring model; prototype observation