

文章编号: 1000-7598 (2011) 12-3738-05

基于主成分和半参数的大坝变形监测回归模型

许 昌^{1,2}, 岳东杰², 董育烦³, 邓成发⁴

(1. 浙江水利水电专科学校 市政工程系, 杭州 310018; 2. 河海大学 地球科学与工程学院, 南京 210098;
3. 中国水电顾问集团华东勘测设计研究院, 杭州 310014; 4. 浙江省水利河口研究院, 杭州 310020)

摘 要: 主成分分析在一定程度上可以解决大坝变形监测回归模型因子间的复共线性, 然而当提取的主成分信息不充分时, 主成分回归用于大坝安全预测可能失效。提出以主成分分析提取的主成分作为半参数回归的参数分量, 剩余成分和模型误差作为未知的非参数分量对主成分回归进行补偿, 建立一种基于主成分和半参数的大坝变形监测混合回归模型。利用某大坝实测资料进行建模分析, 结果表明该混合模型能克服回归因子间的复共线性, 避免半参数回归补偿最小二乘估计中法矩阵的病态性, 比传统的主成分回归和逐步回归模型具有更好的拟合和预报精度。

关 键 词: 大坝; 复共线性; 主成分分析; 半参数回归; 补偿最小二乘

中图分类号: TV 698.1

文献标识码: A

Regression model for dam deformation based on principal component and semi-parametric analysis

XU Chang^{1,2}, YUE Dong-jie², DONG Yu-fan³, DENG Cheng-fa⁴

(1. Department of Municipal Engineering, Zhejiang Water Conservancy and Hydropower College, Hangzhou 310018, China;

2. School of Earth Sciences and Engineering, Hohai University, Nanjing 210098, China;

3. East China Investigation and Design Institute, China Hydropower Engineering Consulting Group Corporation, Hangzhou 310014, China;

4. Zhejiang Institute of Hydraulics & Estuary, Hangzhou 310020, China)

Abstract: Principal component analysis (PCA) is a solution to the multicollinearity problem of dam regression models. However, uninformative principal components (PCs) may lead to the failure prediction of dam deformation. Thus a hybrid regression model using semi-parametric regression and PCA is proposed, where the PCs with the highest variance are treated as the semi-parametric component; the remaining PCs and model errors are treated as the non-parametric component to be estimated. The hybrid model is tested using the field observations of a dam in China. The result shows the hybrid model can circumvent the multicollinearity of dam causative effects and the ill-conditioned problem in semi-parametric penalized least squares regression. A comparative study with traditional PC regression and stepwise regression demonstrate the superior performance for dam deformation prediction.

Key words: dam; multicollinearity; principal component analysis; semi-parametric regression; compensated least squares

1 引 言

分析大坝施工及运营期间的变形监测资料, 解释坝体变形物理成因, 掌握坝体运行规律和工作状态, 对于大坝安全性评价具有十分重要意义。其中基于经典最小二乘法的统计模型(如多元线性回归和逐步回归)在大坝变形监测资料分析中得到了广泛运用^[1]。然而当库水位、温度和时效因子等自变量间出现复共线性时, 运用经典最小二乘法将无法有效分离这些影响分量^[2]。针对复共线性问题, 人

们提出了岭回归、主成分回归和偏最小二乘回归等方法。景继等^[3]结合某碾压混凝土坝施工期的裂缝资料, 初步分析了回归模型中复共线性产生的原因及其不良影响, 揭示了岭回归在水工监测数据统计分析中的应用价值; 薛桂玉等^[4]探讨了主成分回归在混凝土坝无应力变形分析中的可行性; 李红祥等^[5]提出了基于主成分回归的大坝位移统计模型, 分析了大坝位移的各类影响效应。徐洪钟等^[6]采用偏最小二乘回归建立坝顶水平位移统计模型, 消除了复共线性, 取得了满意结果。上述方法在处理复共线

收稿日期: 2010-06-08

基金项目: 浙江省水利厅科技计划项目 (No. RB1010)。

第一作者简介: 许昌, 男, 1983 年生, 博士, 讲师, 主要从事工程测量和测绘数据处理方面的研究与教学工作。E-mail: xuchang404@163.com

性时取得了一定效果,但依然存在一些问题。例如,岭回归主要通过岭参数来减小均方差,而岭参数的选取问题目前在理论上还未彻底解决^[7];主成分分析则主要通过方差贡献大小从自变量中提取主成分以替代原有自变量,然而在某些情况下剩余成分中含有与主成分同样重要的信息^[8],此时若主成分个数选取不当或主成分包含的信息不充分,主成分回归用于大坝安全预测可能失效,基于主成分分析的偏最小二乘回归同样如此。此外,虽然截断奇异值分解法也可以解决大坝安全监测回归模型因子间的复共线性,但存在着截断参数的选取问题^[9-10],其本质与主成分分析具有等价性^[11]。统计学界在 20 世纪 80 年代提出了半参数模型^[12-13],该模型包含了参数和非参数分量,可同时描述观测量中函数关系确定的误差影响和系统误差等不确定因素,因此,被广泛应用于测绘数据处理中。目前半参数模型的估计方法有很多,如补偿最小二乘估计^[13]、核估计^[14]和样条估计^[15]等,但这些方法的理论研究尚不完善,因此,其实际工程应用并不成熟。

本文从半参数回归补偿最小二乘估计的原理出发,提出一种基于主成分和半参数的大坝安全监测混合回归模型,通过实例分析,指出该混合模型相对于传统主成分回归和逐步回归的优越性。

2 混合回归模型

基于主成分和半参数的混合回归模型(以下简称混合回归模型)其实质就是以主成分分析提取的主成分作为半参数回归的参数分量,以剩余成分和模型误差作为非参数分量进行估计,其计算步骤概括如下:

(1) 标准化原始指标数据。

设原始自变量数据为 $F = (f_{ij})_{n \times m}$, 其中 f_{ij} 为第 i 个样本的第 j 项指标值, n 为样本数, m 为指标数,则可按下式进行标准化^[16]:

$$f_{ij} = \frac{f_{ij} - \bar{f}_j}{\zeta_j}, \quad i=1,2,\dots,n, \quad j=1,2,\dots,m \quad (1)$$

式中: $\bar{f}_j = \sum_{i=1}^n f_{ij} / n$, $\zeta_j = \sqrt{\sum_{i=1}^n (f_{ij} - \bar{f}_j)^2 / (n-1)}$ 。

(2) 提取主成分。

根据式(2)计算指标数据的相关系数矩阵 $U = (u_{jk})_{m \times m}$, 求解该矩阵的特征根和特征向量,最后根据方差贡献率确定主成分的个数^[16]。

$$u_{jk} = \frac{1}{n-1} \sum_{i=1}^n \left[\frac{(f_{ij} - \bar{f}_j)^2}{\zeta_j^2} \right] \left[\frac{(f_{ik} - \bar{f}_k)^2}{\zeta_k^2} \right], \quad j=1,2,\dots,m, \quad k=1,2,\dots,m \quad (2)$$

(3) 建立半参数回归模型。

根据步骤(2)中提取的主成分,建立半参数模型如下^[13]:

$$Y = X\beta + S + \Delta \quad (3)$$

式中: Y 为 n 维观测向量; X 为 $n \times t$ 维主成分矩阵; β 为 $t \times 1$ 维待估参数向量; S 为描述系统误差或模型误差的 n 维未知向量,即非参数分量; Δ 为 n 维服从正态分布 $N(0, \sigma^2)$ 的随机噪声向量。

(4) 确定正规矩阵和光滑参数。

正规矩阵一般反映非参数分量的函数性质,而光滑参数则主要在极化过程中平衡残差和非参数分量^[17-18]。考虑大坝变形观测量均为时间序列,且相邻时刻观测的系统误差并无显著差异,因此,可根据相邻两观测值模型误差之差的平方和来构造正规矩阵 R ^[17-19], 即

$$R = G^T G, \quad G_{n-1,n} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & & \dots & \\ & & & & -1 & 1 \end{bmatrix} \quad (4)$$

光滑参数 α 可采用广义交叉核实(GCV)最小法^[13,18]确定,即

$$GCV(\alpha) = n^{-1} \frac{\sum_{i=1}^n p_i (y_i - \hat{y}_i)^2}{\{1 - n^{-1} \text{tr}[A(\alpha)]\}^2} = \min \quad (5)$$

式中: $A = Q + (I - Q)(P + \alpha R - PQ)^{-1}(P - PQ)$, $Q = X(X^T P X)^{-1} X^T P$; I 为单位矩阵; P 为权矩阵; $\hat{y}_i$ 为 y_i 的估值。

(5) 求解回归系数。

由式(3)可知,其误差方程为:

$$V = X\hat{\beta} + \hat{S} - Y \quad (6)$$

式中: $\hat{\beta}$ 、 $\hat{S}$ 分别为 β 、 S 的估计。

根据 Green 和 Silverman^[13]的补偿最小二乘法原理并结合约束条件 $\Sigma \hat{S} = e^T \hat{S} = 0$, 可估计参数向量 β 和非参数分量 S ^[17]:

$$\begin{bmatrix} \hat{S} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} M^{-1} W Y - \gamma M^{-1} e \\ N^{-1} X^T P (Y - \hat{S}) \end{bmatrix} \quad (7)$$

式中: $N = X^T P X$, $M = P + \alpha R - P X N^{-1} X^T P$, $W = P - P X N^{-1} X^T P$, $\gamma = e^T M^{-1} W Y / (e^T M^{-1} e)$, $e = [1 \ 1 \dots 1]^T$ 。由于 X 为主成分矩阵,因此,在一定程度上避免了半参数回归模型中法矩阵 N 和 M 的病态性问题。

最后将标准化回归系数还原为原始变量的回归系数。

3 实例分析

本文对某混凝土坝 1~89 期的观测资料^[5]进行建模分析,分析过程包括复共线性诊断、主成分分析、位移拟合及预测 4 部分。

3.1 统计模型

坝体在水压力、扬压力和温度等荷载作用下将产生一定位移,考虑该坝体自身的特性,利用引起坝体位移的水压、温度和时效因子建立统计模型^[1]:

$$D = a_0 + \sum_{i=1}^4 a_i (H - H_0)^i + \sum_{j=1}^5 b_j T_j + c_1 (\theta - \theta_0) + c_2 (\ln \theta - \ln \theta_0) \quad (8)$$

式中: D 为位移量; H 为库水位; H_0 为坝底高程; T_j 分别为坝体 5 个不同测点的温度; $\theta = t_n / 100$ 和 $\theta_0 = t_m / 100$ 为时效因子,其中 t_n 为位移观测日至始测日的累计天数, t_m 为建模至始测日的累计天数; a_i 、 b_i 、 c_1 和 c_2 分别为待求的回归系数。

3.2 复共线性诊断

大坝变形监测回归模型自变量间不可避免地出现复共线性,究其原因是水压因子 ($H^1 \sim H^4$) 间存在直接相关性,水位与温度、时效因子之间数值的悬殊性也导致矩阵 $X^T X$ 条件数很大。另外,气温、坝体温度和水量平衡之间也存在某种关联性。本文采用逐步回归中的方差膨胀因子 (VIF) 进行复共线性诊断,结果见表 1。其中,逐步回归中引入和剔除变量的显著性水平分别取 0.05 和 0.10。根据变量 VIF 准则 ($VIF > 10$)^[20]可知库水位、温度和时效因子之间存在严重的复共线性。

表 1 逐步回归变量方差膨胀因子
Table 1 VIF of stepwise regression variables

变量	T引入			T剔除							
	H^4	T_4	$\theta - \theta_0$	H^1	H^2	H^3	T_1	T_2	T_3	T_5	$\ln \theta - \ln \theta_0$
VIF	27	32	2	1 439	3 219	12 806	2	43	2	16	51

表 2 回归系数

Table 2 Regression coefficients

回归系数	变量											常数
	H^1	$H^2/10^{-5}$	$H^3/10^{-7}$	$H^4/10^{-9}$	T_1	T_2	T_3	T_4	T_5	$\theta - \theta_0$	$\ln \theta - \ln \theta_0$	
主成分回归	-0.027	7.403	-2.704	-1.106 0	-0.035	0.295	0.194	0.229	0.193	2.189	2.799	0.713
混合回归	-0.017	4.737	-1.795	-0.7503	-0.185	0.236	0.074	0.183	0.162	1.804	2.405	3.292
逐步回归								0.729		4.512		-5.111

3.3 主成分提取

根据各成分贡献率可以确定主成分的个数为 3, 这些主成分可以解释原有 11 个自变量 98.4% 的信息 (见图 1)。本文利用因子得分系数矩阵和各标准指标变量对主成分进行了提取, 结果见图 2。

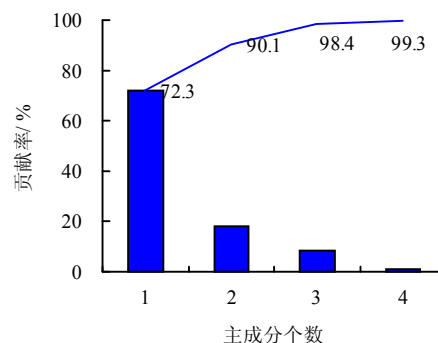


图 1 主成分贡献率

Fig.1 Percentage of variance explained

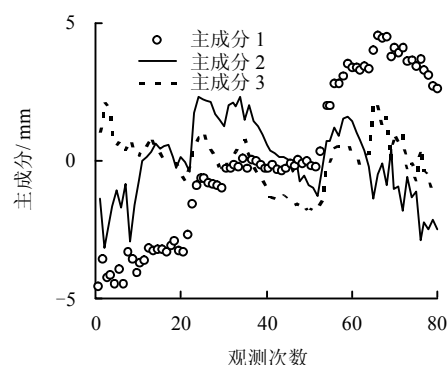


图 2 主成分 1, 主成分 2 和主成分 3

Fig.2 PC1, PC2, and PC3

3.4 回归分析

利用图 2 中 3 个互不相关的主成分作为半参数模型的参数分量, 结合 GCV 函数极小值即光滑参数 (8.4548) 建立混合回归模型, 结果与经典的逐步回归和主成分回归进行对比分析。各回归模型的回归系数见表 2, 模型精度包括复相关系数 R 、残差平方和 RSS 和残差标准差 STD 见表 3, 模型拟合残差曲线见图 3, 半参数回归非参数分量见图 4。

表 3 模型精度
Table 3 Model accuracy

模型	R	RSS/mm^2	STD/mm
主成分回归	0.993	33.15	0.65
混合回归	0.995	10.45	0.36
逐步回归	0.994	25.76	0.57

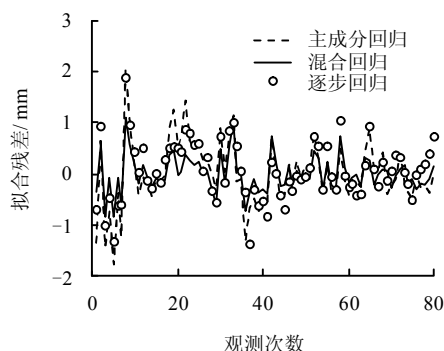


图 3 各模型拟合残差图

Fig.3 Residual curve of different models

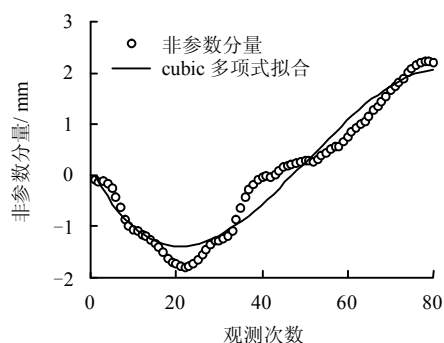


图 4 半参数模型非参数分量

Fig.4 Estimation of non-parametric component

由表 2 可以看出严重的复共线性致使逐步回归中大部分水位因子难以入选, 显然不合理。由表 3 和图 3 可知, 本文提出的混合回归模型不仅保留了原有因子的大部分信息, 还顾及了剩余成分的信息量, 因此, 与传统的主成分和逐步回归模型相比, 该混合模型具有更好的拟合精度。为检验上述 3 种模型的预测效果, 本文采用 9 组实测数据进行预报分析。由于半参数模型非参数分量的离散性, 且事先未知, 因此, 对于补偿最小二乘估计的非参数分量进行预测存在很大难度, 同时成为了半参数回归进行时间序列分析预测的瓶颈问题, 目前尚未见相应研究报道。考虑约束后的非参数分量呈周期性变化^[17], 本文凭经验采用 cubic 多项式函数 (见图 4) 对非参数分量进行拟合外推。从图 5 中位移的预测结果可以看出, 经过半参数模型非参数分量补偿后的混合回归模型比传统的主成分和逐步回归具有更

高的预测精度。此外, 从图 3 和图 5 中的逐步回归结果可以看出, 因子间的复共线性对于回归模型的拟合精度影响并不大, 而在外延预测时可能会出现较大偏差。

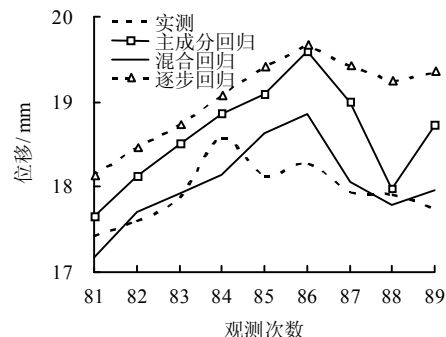


图 5 实测位移及模型预测位移

Fig.5 Observed and statistically predicted displacement

4 结 论

本文将半参数回归补偿最小二乘估计原理和主成分分析引入大坝安全监测资料分析中, 提出一种基于主成分和半参数的大坝安全监测混合回归模型, 通过实例分析, 得出以下结论:

(1) 以主成分作为半参数回归的参数分量, 可以克服回归因子间的复共线性, 避免半参数回归法矩阵的病态性。

(2) 以剩余成分和模型误差作为非参数分量进行补偿, 可以弥补由于主成分信息不充分而导致外延预测失败的不足。因此, 与传统的主成分回归和逐步回归相比, 该混合模型具有更好的拟合与预测能力。

(3) 本文对半参数模型补偿主成分回归作了初步探讨, 而正规矩阵和光滑参数的选取问题, 以及如何寻求非参数分量规律及其蕴含的信息对选用的统计模型加以补偿, 还值得进一步研究。笔者相信随着半参数理论不断完善, 其在大坝监测资料分析领域中的应用将更为广泛。

参 考 文 献

- [1] 吴中如, 沈长松, 阮焕祥. 水工建筑物安全监控理论及其应用[M]. 河海大学出版社, 1990.
- [2] CHRZANOWSKI A, CHEN Y Q, SECORD J M, et al. Problems and solutions in the integrated monitoring and analysis of dam deformations[J]. CISM Journal ACSGC, 1991, 45(4): 547—560.
- [3] 景继, 包腾飞, 谷艳昌, 等. 监测数据共线性问题的岭

- 回归法分析[J]. 水电自动化与大坝监测, 2007, 31(3): 56—59.
- JING Ji, BAO Teng-fei, GU Yan-chang, et al. Analysis of ridge regression for the multi-colinearity of monitoring data[J]. **Hydropower Automation and Dam Monitoring**, 2007, 31(3): 56—59.
- [4] 薛桂玉, 李珍照. 主成分回归在混凝土坝无应力变形测值分析中的应用[J]. 大坝观测与土工测试, 1992, 6(3): 16—21.
- XUE Gui-yu, LI Zhen-zhao. The principal component regression analysis applied to the analysis of the no-stress deformation of the concrete dam[J]. **Dam Observation and Geotechnical Tests**, 1992, 6(3): 16—21.
- [5] 李红祥, 岳东杰, 李立瑞. 基于主成分回归的大坝位移模型[J]. 水电自动化与大坝监测, 2008, 32(5): 61—64.
- LI Hong-xiang, YUE Dong-jie, LI Li-rui. Dam displacement model based on principal component regression[J]. **Hydropower Automation and Dam Monitoring**, 2008, 32(5): 61—64.
- [6] 徐洪钟, 吴中如. 偏最小二乘法回归在大坝安全监测中的应用[J]. 大坝观测与土工测试, 2001, 25(6): 22—25.
- XU Hong-zhong, WU Zhong-ru. Partial least square regression and its application to dam safety monitoring[J]. **Dam Observation and Geotechnical Tests**, 2001, 25(6): 22—25.
- [7] 王振杰, 欧吉坤. 用 L -曲线法确定岭估计中的岭参数[J]. 武汉大学学报(信息科学版), 2004, 29(3): 235—238.
- WANG Zhen-jie, OU Ji-kun. Determining the ridge parameter in a ridge estimation using L -curve method[J]. **Geomatics and Information Science of Wuhan University**, 2004, 29(3): 235—238.
- [8] JOLLIFFE I T. A note on the use of principal components in regression[J]. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, 1982, 31(3): 300—303.
- [9] XU Chang, DENG Cheng-fa. Solving multicollinearity in dam regression model using TSVD[J]. **Geo-Spatial Information Science**, 2011, 14(3): 230—234.
- [10] XU Pei-liang. Truncated SVD methods for discrete linear ill-posed problems[J]. **Geophysical Journal International**, 1998, 135(2): 505—514.
- [11] 吴春国, 梁艳春, 孙延风, 等. 关于 SVD 与 PCA 等价性的研究[J]. 计算机学报, 2004, 27(2): 286—288.
- WU Chun-guo, LIANG Yan-chun, SUN Yan-feng, et al. On the equivalence of SVD and PCA[J]. **Chinese Journal of Computers**, 2004, 27(2): 286—288.
- [12] ENGLE R F, GRANGER C W J, RICE J, et al. Semiparametric estimates of the relation between weather and electricity sales[J]. **Journal of the American Statistical Association**, 1986, 81: 310—320.
- [13] GREEN P J, SILVERMAN B W. Nonparametric regression and generalized linear models: A roughness penalty approach[M]. London: Chapman and Hall/CRC, 1993.
- [14] HÄRDLE W, MÜLLER M. Multivariate and semiparametric kernel regression[M]//Schimek M G. Smoothing and regression: Approaches, computation, and application. New York: John Wiley & Sons, 2000: 357—392.
- [15] SILVERMAN B W. Some aspect of the spline smoothing approach to non-parametric regression curve fitting[J]. **Journal of the Royal Statistical Society: Series B**, 1985, 47(1): 1—52.
- [16] 李艳双, 曾珍香, 张闽, 等. 主成分分析法在多指标综合评价方法中的应用[J]. 河北工业大学学报, 1999, 28(1): 94—97.
- LI Yan-shuang, ZENG Zhen-xiang, ZHANG Min, et al. Application of primary component analysis in the methods of comprehensive evaluation for many indexes[J]. **Journal of Hebei University of Technology**, 1999, 28(1): 94—97.
- [17] 孙海燕, 吴云. 半参数回归与模型精化[J]. 武汉大学学报(信息科学版), 2002, 27(2): 172—174.
- SUN Hai-yan, WU Yun. Semiparametric regression and model refining[J]. **Geomatics and Information Science of Wuhan University**, 2002, 27(2): 172—174.
- [18] 胡宏昌, 孙海燕. 正规矩阵 R 及平滑因子 α 的选取[J]. 测绘工程, 2003, 12(4): 5—8.
- HU Hong-chang, SUN Hai-yan. Choice of the regular matrix R and smoothing parameter α [J]. **Engineering of Surveying and Mapping**, 2003, 12(4): 5—8.
- [19] FISCHER B, HEGLAND M. Collocation, filtering and non-parametric regression[J]. **ZfV**, 1999, 1: 17—24.
- [20] O'BRIEN R M. A caution regarding rules of thumb for variance inflation factors[J]. **Quality and Quantity**, 2007, 41: 673—690.