

文章编号: 1004-0609(2003)03-0743-06

投影寻踪降维方法及其试验数据分析^①

李朝奎^{1, 2}, 朱 庆¹, 赵 杰¹

(1. 武汉大学 测绘遥感信息工程国家重点实验室, 武汉 430079; 2. 湖南科技大学 地球空间信息科学研究所, 湘潭 411201)

摘 要: 针对高维数据分析遇到的困难, 引入投影寻踪降维方法进行变形观测数据分析, 论述了分析原理、分析方法。将投影寻踪降维法应用于桥塔位移观测数据分析, 与偏最小二乘回归法进行了分析效果对比。分析结果表明: 该方法在提取源数据信息与预测方面效果更理想。

关键词: 高维空间; 投影寻踪; 变形观测; 数据分析

中图分类号: P 228; TU 196; U 446

文献标识码: A

变形测量数据分析及其物理解释仍是一个有待进一步深入研究的课题。现有的变形数据分析方法至少存在两个缺陷: 1) 很少顾及多因多果的情形; 2) 没有考虑原因变量的误差。客观地看, 由于变形系统复杂, 人们对系统变形的因果关系又无多少先验信息, 因此在搜集资料时要广纳因果变数。但随着变量的增加, 会产生一些实际难题, 如: 1) 计算量迅速增大; 2) 维数祸根; 3) 统计稳健性变差; 4) 数据的正态性变差。既有的多元统计回归方法大多建立在数据正态分布的假设基础上, 如果偏离假设, 结果就可能产生畸变。针对这些问题, 理论界已经建立一些新的方法, 例如神经网络方法^[1~3]、非参数回归法^[4]、主成分分析法、偏最小二乘回归(PLSR)分析方法^[5]等。这些方法在分析高维数据时, 似乎没有很好地解决上述实际难题。投影寻踪(PP)方法^[6]用一系列岭函数之和来逼近回归函数, 能成功地克服上述困难^[7], 同时减小模型误差。本文作者将该方法首次引入变形测量数据分析、并比较其与偏最小二乘回归分析方法在变形测量数据分析及物理解释中的效果, 由此开辟一条处理高维测量数据的新途径。

1 PP 方法的基本思想

PP(project pursuit)方法是处理高维非正态数据的一种新兴的统计方法, 其基本思想是将高维数据投影到低维子空间, 寻找能反映原高维数据结构或

特征的投影, 从而克服高维数据所带来的“维数祸根”, 同时排除与数据结构和特征无关的或关系甚小的变量干扰。设 $\mathbf{X}$ 是 P 维随机向量, $\mathbf{Y}$ 是 1 维随机变量(如果 $\mathbf{Y}$ 是 m 维随机变量, 则可以采用适当的方法如主成分提取^[8, 9], 将其化为 1 维随机变量), 依据 $(\mathbf{X}, \mathbf{Y})$ 的 n 次观测结果 $(\mathbf{x}_i, y_i) (i = 1, 2, 3, \dots, n)$, 估计回归函数为

$$f(\mathbf{x}) = E(\mathbf{Y} | \mathbf{X} = \mathbf{x}) \quad (1)$$

式中 $E(\cdot)$ 为期望因子。下面给出 PP 方法的具体实现步骤。

1) 选择一个初始的回归模型, 如 $f_0(\mathbf{x}) = C$ 或一元非线性函数等。

2) 寻找一个方向 $\mathbf{a}$, 使得当前的残差 $r_i = y_i - f(\mathbf{x}_i) (i = 1, 2, \dots, n)$ 与投影 $Z = \mathbf{a}^T \mathbf{x}$ 有尽可能大的回归依赖关系, 得到平滑函数 $g_a(Z)$ 。

3) 将模型更新为 $f(\mathbf{x}) = f_0(\mathbf{x}) + g_a(\mathbf{a}^T \mathbf{x})$, 经 m 步迭代后, 回归函数可表示为

$$f(\mathbf{x}) = f_0(\mathbf{x}) + \sum_{j=1}^m g_j(\mathbf{a}_j^T \mathbf{x}) \quad (j = 1, 2, \dots, m) \quad (2)$$

式中 $g_j(\mathbf{a}_j^T \mathbf{x}) = g_{a_j}(\mathbf{a}_j^T \mathbf{x})$ 称为岭函数。这里主要是投影方向 $\mathbf{a}$ 与投影指标的选择。给定 $\mathbf{a}_j, g_j, j < m$, 寻找 $\mathbf{a}_m, g_m$, 使得目标函数

$$\begin{aligned} Q &= \sum_{j=1}^n r_{mj}^2 \\ &= \sum_{j=1}^n [r_{m-1, j} - g_m(\mathbf{a}_m^T \mathbf{x}_j)]^2 = \min \end{aligned} \quad (3)$$

如果给定 $\mathbf{a}$ 后, g_m 的条件期望是

① 基金项目: 霍英东教育基金会青年教师基金资助项目(71017); 国家自然科学基金资助项目(40001017); 湖南省自然科学基金资助项目(02JJY2064) 收稿日期: 2002-06-06; 修订日期: 2002-09-16

作者简介: 李朝奎(1967-), 男, 副教授, 博士, 武汉大学在站博士后。

通讯联系人: 李朝奎, 博士; 0732-8290041; E-mail: chkl_hn@163.com

$$g(\mathbf{a}^T \mathbf{x}) = E(y | \mathbf{a}^T \mathbf{x}) \quad (4)$$

显然这里关键是 $\mathbf{a}$ 的确定。具体方法如下:

1) 任意给定投影方向 $\mathbf{a}$ 。

2) 将 $Z_i = \mathbf{a}^T \mathbf{x}_i$ ($i = 1, 2, \dots, n$) 从小到大排序, 序列记为 $\{Z_i\}$, 相应的当前残差序列记为 $\{r_i\}$ 。

3) 为稳健考虑, 作滑动中位数平滑, 以此剔除孤立离群点

$$r_i = \text{Med}\{r_{i-1}, r_i, r_{i+1}\} \quad (5)$$

4) 对 $\{Z_i\}$ 的 k 个左邻, k 个右邻共 $2k$ 对 (Z_i, r_i) 作局部线性拟合, 并求出残差平方和, 统计出

$$\sigma_i^2 = [VV]/2k \quad (6)$$

5) 对 $\{\sigma_i^2\}$ 作固定宽带的滑动平均

$$\sigma_i^2 = (\sigma_{i-k}^2 + \dots + \sigma_{i-1}^2 + \sigma_{i+1}^2 + \dots + \sigma_{i+k}^2)/2k \quad (7)$$

由 4) 和 5) 可得到局部的平滑估计 σ_i^2 。

6) 对 $\{r_i\}$ 作局部线性拟合, 带宽由 σ_i^2 确定, 得到 $g_a(Z_i)$, 这一步采用较大的邻域宽带以克服较大的局部变异。4)~6) 中不将第 i 个数据本身纳入计算以防过度拟合。

7) 以 r_i 与 $g_a(Z_i)$ 的残差平方和作为 $\mathbf{a}$ 方向的投影指标 $Q(\mathbf{a})$, 直到找出使投影指标达到极小的投影方向 $\mathbf{a}$ 。这里要指出的是: 公式中的 m 由迭代法确定。其原则是第 m 步的投影指标对于第 $m-1$ 步的投影指标的相对增量的绝对值小于或等于某个阈值或迭代次数大于某预定的最大次数即停止迭代。

表 1 塔柱位移观测原型数据

Table 1 Pro-type data of bridge tower observation

No.	Time	Load/t	Point 2701			Point 2702			Point 2601			Point 2602			t/ °C
			dx	dy	dh	dx	dy	dh	dx	dy	dh	dx	dy	dh	
1	3: 00	Free	- 42	29	- 81	- 38	31	- 46	- 19	53	- 54	- 35	- 22	- 62	14. 4
2	5: 40	Middle, 100	- 37	21	- 90	- 27	29	- 42	- 33	29	- 91	- 42	- 21	- 52	14. 2
3	6: 45	Middle, 200	- 26	22	- 91	- 18	21	- 51	- 40	23	- 107	- 56	- 33	- 77	14. 6
4	8: 00	Middle, 250	- 19	27	- 81	- 13	23	- 46	- 47	24	- 95	- 68	- 43	- 81	15. 1
5	9: 00	Middle, 350	2	35	- 77	3	28	- 47	- 52	37	- 97	- 75	- 38	- 75	15. 6
6	10: 20	Middle, 450	23	39	- 75	14	32	- 34	- 66	51	- 78	- 84	- 21	- 58	17. 4
7	11: 30	Middle, 500	27	32	- 79	16	27	- 32	- 76	46	- 96	- 97	- 31	- 64	18. 6
8	19: 30	Middle, 500	42	53	- 72	42	52	- 40	- 42	101	- 92	- 60	61	- 51	16. 6
9	11: 00	Middle, 0	- 10	54	- 80	- 6	58	- 47	28	109	- 71	5	69	- 72	21
10	12: 00	Forth, 100	- 31	43	- 69	- 30	30	- 37	- 8	80	- 61	- 9	34	- 62	21. 6
11	12: 40	Forth, 150	- 7	59	- 71	- 19	54	- 33	26	143	- 38	3	60	- 49	24. 2
12	13: 30	Forth, 250	- 10	57	- 69	- 14	56	- 34	18	131	- 40	- 2	56	- 44	25. 6
13	19: 30	Forth, 500	- 43	30	- 88	- 32	35	- 62	6	68	- 79	- 9	6	- 44	20. 2
14	21: 00	Forth, 500	- 57	16	- 80	- 55	22	- 32	- 36	6	- 75	- 44	4	- 62	21. 2
15	8: 00	Forth, 0	- 34	37	- 78	- 20	33	- 58	- 13	66	- 65	- 37	- 26	- 64	20. 2
16	9: 40	East, 150	- 38	43	- 77	- 27	40	- 41	- 12	55	- 98	- 31	- 21	- 71	22. 4
17	11: 00	East, 500	- 30	53	- 80	- 25	51	- 45	2	83	- 51	- 9	16	- 60	20. 4
18	12: 00	East, 500	- 52	53	- 83	- 41	37	- 43	1	57	- 121	- 18	- 9	- 70	19. 6

The units of dx, dy and dh are mm, and dx, dy and dh are the initial values without load.

2 PP 方法的应用

2.1 试验分析方法

某大桥主桥为双塔双索面预应力混凝土斜拉桥, 采用塔墩固节、塔梁分离的半漂浮体系, 跨径组合为 133 m+ 277 m+ 133 m; 桥塔采用双柱花瓶形塔, 塔高 95. 3 m, 箱形截面, 斜拉索直接锚固于塔壁中心处。根据规范, 通车前所作成桥试验包括静载试验和动载试验。静载试验采集到如下数据:

1) 用全站仪测量各工况下塔顶的水平及竖向位移。测点的平面布置如图 1 所示, 获取的位移数据如表 1 所示。

2) 用连通水准管获取每一工况下桥面的沉降数据。测点布置如图 2 所示, 观测数据如表 2 所示。

3) 记录每一次观测的时间、温度、工况, 数据如表 1 和表 2 所示。

从表 1 和 2 中可以看出: 温度(x_1)、荷载(x_2)、时间效应(x_3)等为原因变量(自变量); x 平面位移(y_1)、 y 平面位移(y_2)、垂直位移(y_3)等为结果变量(因变量)。一个原因可以对多个结果作出贡献, 多个原因可以影响同一结果。图 3 所示为因果关系的网络图。

以表 1 中的点 2701 和表 2 中的 30[#] 测点为例阐述分析方法, 其它数据分析方法与此相同。

1) 对原始数据进行因素(x_i)、指标(y_i)分类,

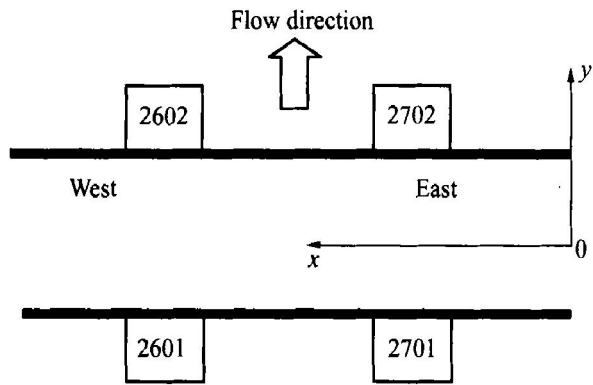


图 1 桥塔位移测点布置平面图
Fig. 1 Distribution of observational points

如表 3 所列; 按式(8)作中心化压缩处理^[10]:

$$x'_i = \frac{(x_i - \bar{x})}{\sigma} \tag{8}$$

2) 依据拉依达准则^[11]对数据进行粗差检验:

$$\sigma = \pm \sqrt{\frac{[d_i d_i]}{n-1}}, \quad d_i = (x_i - \bar{x}), \quad d_i \leq 3\sigma \tag{9}$$

3) 自变量、因变量间的相关系数根据式(10)进行计算, 相关系数如表 4 所列。

$$r_{jk} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{[\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2]^{1/2}} \tag{10}$$

表 2 桥面沉降观测原型数据(表中负号表示下降)

Table 2 Pro-type data of bridge subsidence

No.	Time	t/℃	Load/t	Subsidence of measuring point/mm											
				p1	p7	p14	p21	p30	p31	p32	p39	p48	p53	p57	
1	5:50	13	Middle, 100	1	10	3	- 2	- 31	- 35	- 39	- 17	9	25	15	
2	6:30	14	Middle, 200	2	20	4	- 12	- 78	- 84	- 88	- 36	9	39	21	
3	7:45	15	Middle, 250	2	26	3	- 21	- 107	- 112	- 112	- 45	8	42	22	
4	9:00	15.5	Middle, 350	2	32	4	- 28	- 134	- 140	- 142	- 61	9	54	28	
5	10:30	20	Middle, 450	2	39	4	- 39	- 170	- 178	- 179	- 80	9	65	32	
6	11:30	22	Middle, 500	2	43	4	- 50	- 196	- 203	- 203	- 92	9	68	34	
7	23:00	15	Middle, 500	1	50	3	- 48	- 191	- 199	- 199	- 87	5	74	37	
8	0:40	13	Middle, 200	1	45	2	- 11	- 70	- 77	- 81	- 43	5	70	35	
9	1:40	12	Middle, 0	2	24	3	1	- 4	- 5	- 4	1	6	68	33	
10	11:00	20	Middle, 0	0	3	5	- 3	- 14	- 14	-	-	-	-	-	
11	12:00	22	Forth, 100	1	- 15	5	- 4	- 21	- 23	-	-	-	-	-	
12	13:30	23	Forth, 250	1	- 20	4	- 1	- 24	- 26	-	-	-	-	-	
13	15:20	23	Forth, 300	0	- 32	4	2	- 26	-	-	-	-	-	-	
14	21:00	20	Forth, 300	- 1	- 36	- 2	- 4	- 41	-	-	-	-	-	-	
15	22:00	15	Forth, 100	- 1	- 34	3	11	- 9	-	-	-	-	-	-	
16	8:00	15	Forth, 0	0	- 10	3	- 2	- 19	-	-	-	-	-	-	
17	9:40	18	East, 150	1	5	4	- 5	- 13	- 14	- 13	- 7	10	16	12	

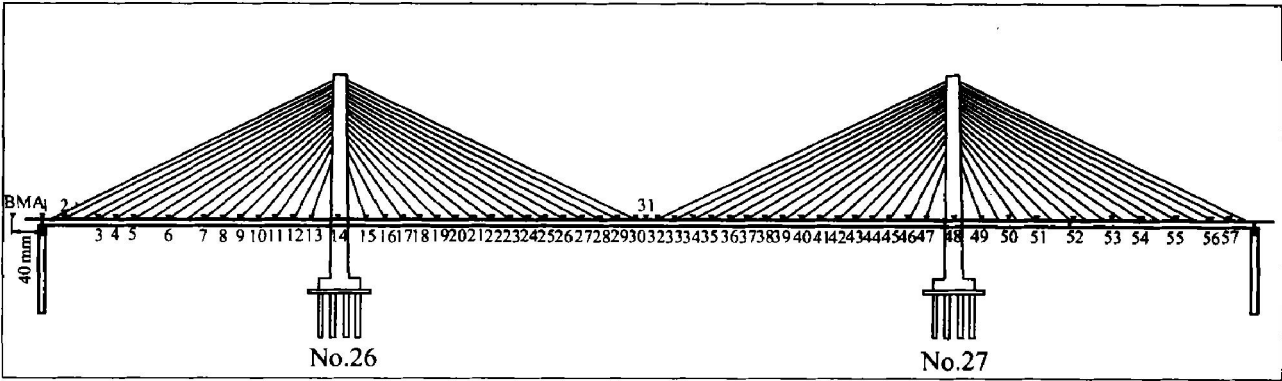


图 2 三通管测点布置图
Fig. 2 Distribution of level points

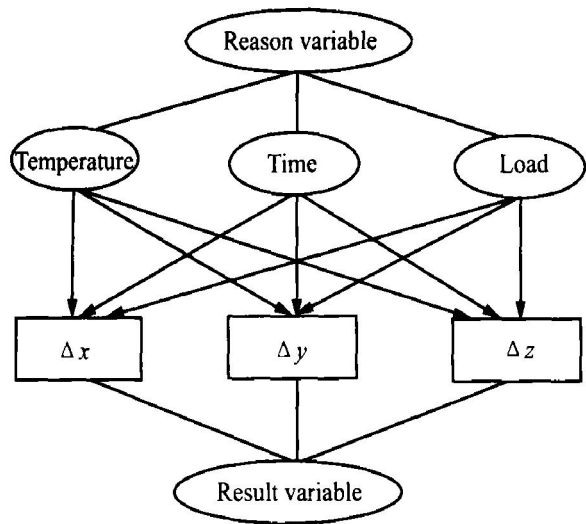


图 3 因果关系的网络图
Fig. 3 Web of reason-result relationships

15), 用于模拟与仿真分析; 第二段数据点占 20% ($n_2=3$), 用于预测对比分析。

6) 对第一段数据分别作 PP 分析及 PLSR 分析, 用 Matlab 及 Excel 软件协助分析计算。

7) 第二段原型数据与预测结果进行对比分析, 统计拟合偏差中误差、拟合相关及预测偏差中误差, 结果如表 5 所列。

2.2 结果分析

本试验的宗旨是测试桥面和塔柱对于不同荷载的敏感性和弹性稳定性。考虑到是新桥, 同时又是钢筋混凝土结构, 因此主要选择了荷载、温度、时效作为自变量因子。为了准确掌握桥的性态, 选择变形最敏感部位—塔顶及桥面 30[#] 测点(见图 2)作为重点试验分析对象。数据及分析表明:

1) 30[#] 测点的沉降(Y_4), 塔柱的纵向位移(Y_1)与荷载(X_2)强相关, 影响规律明显; 塔柱沉降(Y_3), 及横向位移(Y_2)与荷载(X_2)弱相关; 这与结

- 4) 对数据的自变量系统和因变量系统做主成分提取。
- 5) 数据分为二段: 第一段数据点占 80% ($n_1=$

表 3 原始数据的因素、指标分类

Table 3 Classification of factor and index

No.	$X_1/^\circ\text{C}$	X_2/t	X_3	Y_1/mm	Y_2/mm	Y_3/mm	Y_4/mm
1	14.4	0	3:00	-42	29	-81	-31
2	14.2	100	5:40	-37	21	-90	-78
3	14.6	200	6:45	-26	22	-91	-107
4	15.1	250	8:00	-19	27	-81	-134
5	15.6	350	9:00	2	35	-77	-170
6	17.4	450	10:20	23	39	-75	-196
7	18.6	500	11:30	27	32	-79	-191
8	16.6	500	19:30	42	53	-72	-70
9	21	0	11:00	-10	54	-80	-4
10	21.6	100	12:00	-31	43	-69	-14
11	24.2	150	12:40	-7	59	-71	-21
12	25.6	250	13:30	-10	57	-69	-24
13	20.2	500	19:30	-43	30	-88	-26
14	21.2	500	21:00	-57	16	-80	-41
15	20.2	0	8:00	-34	37	-78	-9
16	22.4	150	9:40	-38	43	-77	-19
17	20.4	500	11:00	-30	53	-80	-13
18	19.6	500	12:00	-52	53	-83	-11

表 4 根据式(10)计算的数据相关系数(r_{jk})

Table 4 Factors of correlation(r_{jk}) calculated based on Eqn. (10)

Factor	X_1	X_2	X_3	Y_1	Y_2	Y_3	Y_4
X_1	1.0	0.215 7	0.284 4	0.138 1	0.265 6	0.308 5	0.025 8
X_2	0.215 7	1.0	0.194 4	-0.108 0	0.161 2	0.159 6	-0.015 9
X_3	0.284 4	0.194 4	1.0	0.123 7	0.231 3	0.264 0	0.017 9
Y_1	0.138 1	-0.108 0	0.123 7	1.0	0.107 5	0.107 9	-0.007 1
Y_2	0.265 6	0.161 2	0.231 3	0.107 5	1.0	0.286 7	0.041 8
Y_3	0.308 5	0.159 6	0.264 0	0.107 9	0.286 7	1.0	0.081 6
Y_4	0.025 8	-0.015 9	-0.017 9	-0.007 1	0.041 8	0.081 6	1.0

表 5 PP 方法与 PLSR 结果比较
Table 5 Comparison of PP with PLSR

Factor	Absolute mean error of fit		Fit Correlation		Partial mean error	
	PP	PLSR	PP	PLSR	PP	PLSR
Y_1	± 0.7	± 0.9	0.7	0.5	± 0.3	± 0.3
Y_2	± 0.7	± 0.8	0.6	0.6	± 0.5	± 0.6
Y_3	± 0.8	± 0.9	0.7	0.56	± 0.4	± 0.5
Y_4	± 0.5	± 0.6	0.8	0.8	± 0.3	± 0.4

构理论分析结果完全一致, 且整体结构弹性性能好。

2) 偏最小二乘回归采用主成份分析手段提取自变量集与因变量集的特征信息, 从而建立特征信息之间的线性回归模型, 没有顾及数据之间的非线性特征, 使得拟合与预测效率降低。

3) PP 方法从数据结构出发, 选择数据变异性最大的方向为投影方向, 将非线性或线性数据投影到该方向, 达到降维目的。分析效果较 PLSR 为优。

4) PLSR 与 PP 都采用了降维技术, 分析效果差异表明数据中存在较强的非线性; 表 5 所列数据整体分析结果偏低, 表明数据中有弱相关因素或指标干扰, 例如时效(X_3)、塔柱沉降(Y_3)。

3 结论

1) 投影追踪(PP)法用于建筑物、边滑坡的变形分析与预测是一种探索性研究, 方法较成功地解决了因素、指标之间关系复杂的建模与预测问题。

2) PP 方法将统计数学与计算机科学紧密联系, 具有深刻的理论背景, 为使用一维统计方法解决高维统计问题开辟一条新途径, 但在计算量、高度非线性问题上有待深入研究。

3) PP 方法用于变形分析建模, 不需要对数据的分布做出假定, 也不需要了解因果变量之间的线性或非线性关系, 计算结果精度高, 符合实际。

4) PP 方法采用“审视数据-模拟-预测”的探索性实证分析模式, 有可借鉴意义和相当的应用价值。

REFERENCES

[1] HUANG Quarryi, ZHANG Zheng-lu. Dam deformation prediction by artificial neural network method[A]. Proceedings of Monitoring of Constructions and Local Geodynamic Process [C]. Wuhan: Wuhan University Press, 2001. 123 - 128.

[2] Kropp J. A neural network approach to the analysis of city

system[J]. Applied Geography, 1998, 18(1): 83 - 96.

[3] 白润才, 张普利, 刘景凡. 露天矿边坡稳定性预测的人工神经网络模型[J]. 辽宁工程技术大学学报(自然科学版), 2000(4): 337 - 339.

BAI Run cai, ZHANG Pu li, LIU Jing-fan. The artificial neural network model of forecasting open mining slope stability[J]. Journal of Liaoning Technical University(Natural Science), 2000(4): 337 - 339.

[4] 任若恩, 王惠文. 多元统计数据分析—理论、方法与实例[M]. 北京: 国防工业出版社, 1997. 149 - 158.

REN Ru'en, WANG Hui-wen. Multi-Statistic Data Analysis—Theory, Method and Example[M]. Beijing: National Defence Industry Press, 1997. 145 - 158.

[5] 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. 43 - 59.

WANG Hui-wen. Partial Least Square Regression and Its Applications[M]. Beijing: National Defence Industry Press, 1997. 43 - 59.

[6] 项静恬, 史久恩. 非线性系统中数据处理的统计方法[M]. 北京: 科学出版社, 1997. 278 - 288.

XIANG Jing-gua, SHI Jiu-en. Statistics of Data Processing in the Nonlinear System[M]. Beijing: Science Press, 1997. 278 - 288.

[7] Friedman J H, Stuetzle W. Projection pursuit regression[J]. Journal of American statistical Association, 1981(76): 817 - 823.

[8] 吴有炜. 主成份投影法用于多目标决策与评价[J]. 无锡轻工大学学报, 2001, 20(4): 409 - 415.

WU Your-wei. A principal component projection method for multi-indices decision and evaluation[J]. Journal of Wuxi University of Light Industry, 2001, 20(4): 409 - 415.

[9] 豆俊峰, 邹振扬. 主成份分析在大气环境质量综合评估中的应用[J]. 重庆环境科学, 2001, 23(2): 32 - 33.

DUO Jun-feng, ZOU Zhen-yang. Application of principal component analysis method in comprehensive assessment of atmospheric environmental quality[J]. Environmental Science of Chongqing, 2001, 23(2): 32 - 33.

[10] 李朝奎. 非线性模型空间测量数据处理理论及其应用[D]. 长沙: 中南大学, 2001. 76 - 82.

LI Chao-kui. Theovy and Application of Data Processing in Space of Nonlinear Models[D]. Changsha: Central South U-

- niversity, 2001. 76-82.
- [11] 李朝奎, 黄力民, 曾卓乔, 等. 基于 Monte Carlo 方法的强非线性函数的方差估计[J]. 中国有色金属学报, 2000, 10(4): 613-617.
- LI Chao-kui, HUANG Li-min, ZENG Zhuo-qiao, et al. Variance estimation of intensive nonlinear functions based on Monte Carlo method[J]. The Chinese Journal of Nonferrous Metals, 2000, 10(4): 613-617.

PP method and its application on test data analysis

LI Chao-kui^{1, 2}, ZHU Qing¹, ZHAO Jie¹

(1. State Key Laboratory for Information Engineering in Surveying,

Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China;

2. Research Institute of Geo-spatial Information Science, Hunan University of Science & Technology, Xiangtan 411201, China)

Abstract Aimed at the difficulties coming across during the course of high dimension data analysis, a latest statistic method called as “PP” (project pursuit) method was used. According to PP method, the high dimension data are projected to the low dimension data space, thus the projects can reflect the feature or structure of original high dimension data. By investigating the project in low dimension data space, the effective multi-factors regression analysis on deformation is made, “PP” method is applied to test data regression analysis of new bridge before acceptance. Compared with the results derived from partial least square regression and neural network analysis on test data analysis on new bridge, the “PP” method is proved to be a practical way in some aspects of obtaining original data information and forecasting.

Key words: multi-dimension space; project pursuit; deformation observation; data analysis

(编辑 何学锋)