

DOI:10.3876/j.issn.1000-1980.2011.03.016

主成分分析在 GPS 高程拟合中的应用

岳东杰¹ 李立瑞² 郭培闪¹

(1.河海大学地球科学与工程学院,江苏 南京 210098;2.中国电力工程顾问集团中南电力设计院,湖北 武汉 430071)

摘要:为了合理确定多项式的阶数,克服参数间的相关性,提高 GPS 高程拟合的精度,提出采用主成分分析法,对相同的数据点分别采用常规多项式拟合与主成分分析拟合.苏通大桥首级施工控制网的 GPS 高程拟合计算结果表明,主成分分析法受参数个数与相关性影响小,计算结果稳定,而且能有效降低数据冗余度,是一种比较好的拟合方法.

关键词:多项式拟合;GPS 高程拟合;主成分分析法;苏通大桥

中图分类号: TU198+.2 **文献标志码:** A **文章编号:** 1000-1980(2011)03-0323-04

GPS 测量以精度高、速度快、操作简单等诸多优点在各行各业得到广泛的应用.由于 GPS 测量直接得到的高程为相对于 WGS-84 椭球的大地高,而实际应用的高程为相对于大地水准面的正常高,为此必须进行高程转换.通常采用一定的数学模型对测区的高程异常进行拟合,进而结合 GPS 测量的成果,求出待测点的正常高^[1].在各种拟合法中,多项式拟合法以其计算简单、易于编程实现而得到广泛应用.即在小区域内,将似大地水准面看成曲面或平面,将高程异常 ξ 表示为平面坐标 (x, y) (或大地坐标 (B, L)) 的多项式函数 $f(x, y)$,通过区域内的已知控制点(大地高和正常高均为已知),构造误差方程,采用最小二乘法确定模型中的参数,从而确定测区的似大地水准面形状,求出其余各点的高程异常.其数学模型为

$$\xi = f(x, y) + \varepsilon \quad (1)$$

式中: $f(x, y)$ ——以平面坐标表示的似大地水准面函数; ε ——拟合误差; ξ ——高程异常. $f(x, y)$ 的多项式形式为

$$f(x, y) = a_0 + a_1x + a_2y + a_3x^2 + a_4xy + a_5y^2 + a_6x^3 + a_7y^3 + \dots \quad (2)$$

对于 n 个已知点,写成误差方程的形式为

$$V = BX + L \quad (3)$$

其中

$$V = -(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T \quad L = (-\xi_1, -\xi_2, \dots, -\xi_n)^T$$
$$B = \begin{bmatrix} 1 & x_1 & y_1 & x_1y_1 & x_1^2 & y_1^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & y_n & x_ny_n & x_n^2 & y_n^2 & \dots \end{bmatrix} \quad X = (a_0, a_1, a_2, a_3, a_4, a_5, \dots)^T$$

上述模型在实际应用中存在多项式阶数(拟合参数个数)的确定问题,以往是根据经验确定.实践验证表明并不是阶数越高、参数越多越好;其次,当拟合参数过多时,系数之间会产生很大的相关性,导致法方程系数病态,最小二乘解不稳定.针对以上问题,本文尝试采用主成分分析法,根据携带信息的多少逐步选取独立的主成分进行回归计算,克服参数相关性的影响,提高计算结果的稳定性.

1 主成分分析法原理及其计算步骤

主成分分析最早是由美国统计学家皮尔逊在 1901 年的生物学理论研究中引入的,它从多指标分析出发,运用统计分析原理与方法提取少数几个彼此不相关的综合性指标而保持其原指标所提供的大量信息的一种统计方法,也是考察多个变量间相关性的一种多元统计方法,研究如何通过原来变量的少数几个线性组

合来解释原来变量绝大多数的信息.具体地说,就是将高维空间的问题转化到低维空间去处理,使问题变得比较简单、直观,而且这些较少的综合指标之间互不相关,又能提供原有指标的绝大部分信息.目前主成分分析方法已经广泛地应用于图像识别、图像分析、管理决策、回归分析、模式识别、时间序列预测等各个领域^[2-6].

设 $X = (x_1, x_2, \dots, x_p)$ 是 p 维随机向量, $E(X) = \mu$, $Cov(X) = Q$. 主成分分析的思想就是将这 p 维向量 $(x_1, x_2, \dots, x_p)$ 综合成尽可能少的 q 维综合向量 $F = (f_1, f_2, \dots, f_q)$ ($q < p$), 而且要求这些新的综合向量 $(f_1, f_2, \dots, f_q)$ 既能充分反映原来 p 维向量 $(x_1, x_2, \dots, x_p)$ 所反映的信息, 又互不相关^[7-8].

考虑到 $f_i = t_{i1}x_1 + t_{i2}x_2 + t_{i3}x_3 + \dots + t_{ip}x_p$ ($i = 1, 2, \dots, q$), 记为矩阵形式 $f_i = X t_i$, 其中 $t_i = (t_{i1}, t_{i2}, t_{i3}, \dots, t_{ip})^T$, $\|t_i\| = 1$.

要使 f_1 携带最多的信息, 即要求 f_1 的方差达到最大值. f_1 的方差可为

$$V(f_1) = \frac{1}{N} \|f_1\|^2 = t_1^T Q t_1 \quad (4)$$

式中: N —— X 的样本数; Q —— X 的协方差阵. 而当 X 中的变量均是标准化后变量时, Q 就是 X 的相关系数矩阵.

为求 $V(f_1)$ 的最大值 $\max_{\|t_1\|=1} t_1^T Q t_1$, 构造拉格朗日函数, 记 λ_1 是拉格朗日系数, 令

$$\varphi = t_1^T Q t_1 - \lambda_1 (t_1^T t_1 - 1) \quad (5)$$

对 φ 求关于 t_1 和 λ_1 的偏导, 并令其为零, 有

$$\frac{\partial \varphi}{\partial t_1} = 2Q t_1 - 2\lambda_1 t_1 = 0 \quad (6)$$

$$\frac{\partial \varphi}{\partial \lambda_1} = -(t_1^T t_1 - 1) = 0 \quad (7)$$

由式(6)得

$$Q t_1 = \lambda_1 t_1 \quad (8)$$

由此可见, t_1 是 Q 的一个标准化特征向量, 它对应的特征值为 λ_1 . 由 $V(f_1) = t_1^T Q t_1 = t_1^T (\lambda_1 t_1) = \lambda_1 t_1^T t_1 = \lambda_1$, 所以 t_1 对应的特征值 λ_1 应取到最大值, t_1 为最大特征值对应的特征向量, 这里称 λ_1 为第一主轴, $f_1 = t_1^T X$ 为第一主成分^[9-10].

依次可以推出其他主轴及其对应的主成分 $f_2, f_3, \dots, f_q$:

$$\begin{cases} f_2 = t_{12}x_1 + t_{22}x_2 + t_{32}x_3 + \dots + t_{p2}x_p \\ f_q = t_{1q}x_1 + t_{2q}x_2 + t_{3q}x_3 + \dots + t_{pq}x_p \end{cases} \quad (9)$$

对主成分进行线性拟合回归, 回归方程为

$$Y_i = b_0 + b_1 f_1 + b_2 f_2 + \dots + b_n f_n + \varepsilon_i \quad (10)$$

满足

$$\sum_{i=1}^n (Y_i - b_0 - b_1 f_1 - b_2 f_2 - \dots - b_n f_n)^2 = \min \quad (11)$$

利用最小二乘法求出系数. 通常根据一定的累计贡献率(如 90%)来决定回归因子的个数 n .

归纳起来主成分分析的计算步骤如下:

a. 为了防止数值差别过大和量纲不同影响方程组的求解, 对数据进行标准化处理

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (12)$$

式中: $\bar{x}_j$ —— x_j 的样本均值; s_j ——样本标准差.

b. 计算标准化后矩阵 X^* 的协方差阵 Q (Q 也是 X 的相关系数阵).

c. 求 Q 的特征值 $\lambda_1 > \lambda_2 > \dots > \lambda_q$ 及其对应的特征向量, 计算所有的主成分 $f_i = t_i^T X$ ($i = 1, 2, \dots, q$).

d. 对 $f_1, f_2, \dots, f_q$ 进行线性拟合回归, 得到主成分分析回归方程 $y = b_0 + b_1 f_1 + b_2 f_2 + \dots + b_q f_q$;

e. 将式(9)代入主成分分析回归方程, 得到用原始自变量因子表示的方程 $y = b'_0 + b'_1 x_1 + b'_2 x_2 + \dots + b'_p x_p$.

2 实例计算分析

为了验证主成分分析在 GPS 高程拟合上的适用性,选取苏通大桥首级施工控制网中的 16 个控制点(A1 ~ A13 为位于江堤上的点, S1 ~ S3 为位于江中间的点)进行计算分析. 为了避免由于拟合点的选取对检验结果的影响,各种计算方案均选择相同的 11 个均匀分布的控制点作为拟合点,位于拟合点区域内的 5 个点作为检核点. 图 1 为苏通大桥区域高程异常等高线图. 可见该区域高程异常面相对平缓,略有起伏^[11].

根据检核点高程异常的已知值 ξ_i 与拟合值 ξ'_i 之差 $\Delta\xi_i$, 计算拟合的外符合精度 $\sigma = \pm \sqrt{\frac{1}{n} \sum_{i=1}^n \Delta\xi_i^2}$ 其中 n 为检核点个数.

首先对数据进行多项式拟合,分别选取式(2)中 3 个、4 个、6 个、8 个参数进行拟合计算,计算结果列于表 1. 各检核点的已知值与拟合值之差以及外符合精度见表 2.

表 1 2 种方法计算的参数比较

参数 个数	多项式拟合								主成分分析拟合							
	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7
3	9.791	-0.024	0.030						9.791	-0.024	0.030					
4	9.791	-0.024	0.030	-0.001					9.791	-0.024	0.030	-0.001				
6	9.780	-0.024	0.032	-0.001	0.013	-0.002			9.817	-0.017	0.021	0.011	-0.022	-0.008		
8	9.793	-0.006	0.039	-0.005	0.001	-0.040	0.038	-0.015	9.815	-0.010	0.021	0.008	-0.023	-0.001	-0.005	-0.009

从表 2 可见,若直接采用多项式拟合,采用 3 参数、4 参数结果相对较好,随着参数个数的增加,或者幂次的增加,精度越来越低. 这是由于参数个数的增加,后面因子与前面各因子只是简单平方、立方关系,从而导致系数矩阵存在较大的相关性,影响拟合结果.

对数据进行主成分分析,在不同的参数个数下(取累计贡献率满足 99% 的前几个因子),计算出的参数列于表 1,各检核点的已知值与拟合值之差以及外符合精度见表 3.

表 2 不同参数下多项式拟合结果与已知结果比较

Table 2 Comparison between fitting results by polynomial model and given ones under different parameters							mm
参数个数	$\Delta\xi_1$	$\Delta\xi_2$	$\Delta\xi_3$	$\Delta\xi_4$	$\Delta\xi_5$		σ
3	-7.1	-8.3	6.7	-6.3	6.2		7.0
4	-8.6	-9.3	7.1	-6.6	6.5		7.7
6	-18.2	-8.7	11.2	-5.4	8.8		11.3
8	-28.6	-13.0	3.2	0.2	1.5		14.1

表 3 不同参数下主成分分析拟合结果与已知结果比较

Table 3 Comparison between fitting results by principal component analysis and given ones under different parameters							mm
参数个数	$\Delta\xi_1$	$\Delta\xi_2$	$\Delta\xi_3$	$\Delta\xi_4$	$\Delta\xi_5$		σ
3	-7.1	-8.3	6.7	-6.3	6.2		7.0
4	-8.6	-9.3	7.1	-6.6	6.5		7.7
6	5.0	-8.0	2.6	5.0	6.2		5.6
8	8.3	-8.8	4.8	3.9	6.8		6.8

从表 1 可见,相同参数个数时,两种方法拟合系数有很大不同,这是由于主成分分析顾及了因子之间的相关性.

表 3 与表 2 比较可见,主成分分析较好地改善了因子之间相关性对结果的影响,具有较好的抗相关能力,而且从拟合精度方面均有较大提高. 3 参数、4 参数法二者结果一致,说明采用平面与双线性曲面拟合模型不足于描述该区域大地水准面的形状,这与图 1 显示的结果一致.

从以上计算结果可以得出以下结论:(a)多项式拟合法的精度受参数个数影响较大,但并不是参数越多,精度越高. 随着参数的增多,参数的相关性增强,多余观测数目相应减少,导致拟合精度降低,而且当因子次数过大时,还会使模型出现一些峰值,对真实情况造成扭曲.(b)主成分分析法受参数个数影响较小,因为主成分分析法本身能够很好地处理数据的相关性,降低数据的相关度.(c)在算例分析的过程中,主成分个数的选取是按 99% 的累计贡献率来取舍的,计算过程中,一般取 3 个主成分分量就能满足要求,有效地降低了数

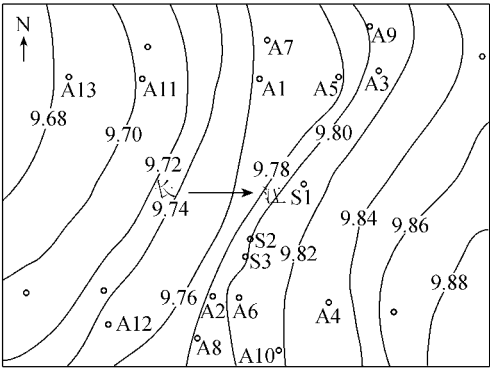


图 1 苏通大桥区域高程异常等高线(单位:m)
Fig. 1 Contour map of height anomaly of Su-Tong Bridge area(unit : m)

据的维数。 d 主成分分析法虽然受参数选取多少的影响较小,但选取参数过多过少都不合适,参数个数过少会导致模型对真实世界的反应程度不够,参数个数过少又会使计算量急剧增加。在对算例进行计算试验过程中发现,选择4~7个参数或者二次多项式时较好,建立的模型精度较高而且计算量又不大。

3 结 语

在进行多项式拟合时,多项式的阶数或参数个数的选取直接影响拟合的结果与精度,参数过少或阶数过低,尽管可以避免参数相关,但可能不足以反映实际情况,参数过多又容易受到数据相关性的影响,造成结果不稳定。主成分分析根据携带信息的多少逐步选取主成分,可以规避数据相关性的影响,因此受参数个数影响小,计算结果稳定,而且能有效降低数据冗余度,是一种比较好的拟合方法。

参考文献:

- [1] 岳东杰,黄腾. GPS高程的抗差拟合估计[J]. 河海大学学报:自然科学版, 1999, 27(6): 90-93. (YUE Dong-jie, HUANG Teng. Anti-error fitting of GPS height[J]. Journal of Hohai University: Natural Sciences, 1999, 27(6): 90-93. (in Chinese))
- [2] 李红祥. 偏最小二乘回归及其在大坝安全监控中的应用[D]. 南京: 河海大学, 2009.
- [3] 张鹏. 多指标综合评价理论与方法问题研究[D]. 厦门: 厦门大学, 2000.
- [4] 李磊,金菊良,梁忠民. 核主成分分析的参数优化及其在水资源可持续利用综合评价中的应用[J]. 水利水电科技进展, 2009, 29(4): 36-38. (LI Lei, JIN Ju-liang, LIANG Zhong-min. Parameters optimization of kernel principal component analysis and its application in evaluation of water resources sustainable utilization[J]. Advances in Science and Technology of Water Resources, 2009, 29(4): 36-38. (in Chinese))
- [5] 余祖德,周明友,陈俊芳. 基于模糊主成分分析法的供应商评价与选择决策[J]. 重庆交通学院学报, 2006, 25(3): 111-115. (YU Zu-de, ZHOU Ming-you, CHEN Jun-fang. Appraisalment and selection decision of suppliers based on the method of fuzzy principal component analysis[J]. Journal of Chongqing Jiaotong University, 2006, 25(3): 111-115. (in Chinese))
- [6] 胡平香,张鹰,王进华. 基于主成分融合的盐田水体遥感分类研究[J]. 河海大学学报:自然科学版, 2004, 32(5): 519-522. (HU Ping-xiang, ZHANG Ying, WANG Jin-hua. Principal component fusion-based classification of salt field water bodies with remote sensing technique[J]. Journal of Hohai University: Natural Sciences, 2004, 32(5): 519-522. (in Chinese))
- [7] 何秀丽. 多元线性模型与岭回归分析[D]. 武汉: 华中科技大学, 2005.
- [8] 何晓群. 现代统计分析与方法[M]. 北京: 中国人民大学出版社, 2007: 281-315.
- [9] 李艳军. 多元线性模型回归系数的主成分估计[D]. 长春: 东北师范大学, 2006.
- [10] 曾鹏,王绍棣. 多变量时间序列的主成分分析及估计[J]. 南京邮电学院学报, 1996, 16(4): 112-116. (ZENG Peng, WANG Shao-di. The main component analysis and estimation of multi variable time series[J]. Journal of Nanjing Institute of Posts and Telecommunications, 1996, 16(4): 112-116. (in Chinese))
- [11] 邬昱昆,岳东杰,杨柏宁. GPS水准在苏通大桥工程建设中的应用研究[J]. 测绘通报, 2007(2): 58-60. (WU Yu-kun, YUE Dong-jie, YANG Bai-ning. On the application of GPS leveling in the Sutong Bridge[J]. Bulletin of Surveying and Mapping, 2007(2): 58-60. (in Chinese))

Application of principal component analysis in GPS height fitting

YUE Dong-jie¹, LI Li-rui², GUO Pei-shan¹

- (1. College of Geosciences and Engineering, Hohai University, Nanjing 210098, China;
2. Central Southern China Electric Power Design Institute of CPECC, Wuhan 430071, China)

Abstract: In order to rationally determine the order of polynomial, to overcome the correlativity among parameters and to raise the fitting precision of GPS height, it was put forward that the conventional polynomial fitting and the principal component analysis were respectively employed for the same data point by means of the principal component analysis method. The calculated results of GPS height fitting for the first stage construction control network of Su-Tong Bridge indicate that the principal component analysis method is little affected by the number and the correlativity of parameters and has stable computational results. It can effectively reduce the data redundancy, thus it is a good method for fitting.

Key words: polynomial fitting; GPS height fitting; principal component analysis; Su-Tong Bridge