An Improved Data Preprocessing Method in Dynamic Measurement

WU Zhan, *CAI Ping* *

(*Department of Instrument Science and Engineering*, *Shanghai Jiaotong University*, *Shanghai* 200240, *China*)

Abstract: In the data analysis of dynamic process test and measurement, the sampled data need to be pretreated to identify and eliminate the abnormal data caused by accidental interference. An improved algorithm based on Singular Spectrum Analysis is presented. By means of singular spectrum analysis smoothing method, the algorithm establishes a benchmark to identify the abnormal data, and then with the 3σ criterion, the abnormal data is eliminated and replaced by a reasonable one. The feasibility of this algorithm has been validated by practical dynamic process test application.

Key words: dynamic process test and measurement; preprocessing; abnormal data erasing; singular spectrum analysis
EEACC: 7220

一种改进的动态过程测量数据预处理方法

吴 展,蔡 萍*

(上海交通大学测试计量技术及仪器系,上海 200240)

摘 要:在动态过程测量数据分析中,首先必须对测量数据进行预处理,剔除并修正由干扰引起的测量异常值。一种改进的奇异谱分析数据预处理是通过使用奇异谱分析平滑的方法建立判断测量数据中异常值的基准,再通过设立检验准则判断异常值并加以修补,在实际动态过程测量数据预处理应用中效果明显。

关键词:动态过程测量;预处理;坏值剔除;奇异谱分析
中图分类号:TP274.2 **文献标识码**:A **文章编号**:1004-1699(2010)04-0558-04

动态过程测量是指对物体动态变化过程的实时测量,常用于变量间关系复杂的物理现象的研究中。通过对测得数据的非线性拟合等数据处理可以验证或获得反映物体动态变化过程的数学模型。在动态过程测量中,出现机率小但作用强烈的偶发性干扰不可避免,导致出现缺值、奇异值、多值等异常数据[1-2]。异常数据,或称坏值的存在使总体测量的准确度降低,影响从实验数据到动态过程的还原。因此在对数据进行分析处理前必须进行数据预处理,剔除或修正异常数据。由于动态过程本身是变化的,如果对坏值的剔除不当,很有可能把看似异常其实准确反应动态变化过程的数据去掉,造成测量的失真。所以在动态过程测量数据处理过程中,选用合适的数据预处理方法、正确剔除并修正坏值非常重要。

1 现有动态测量数据预处理方法简介

针对动态测量的数据预处理,大都通过人工处理进行,即人为判断出现异常的数据,并逐一将坏值剔除并修补。动态过程测量的数据量大,手动处理效率低下,费时费力。另一方面由于过程量本身具有波动性,人为判断坏值的准确性不高,容易剔除正常值[3]。通过设计合理的算法判断坏值并加以剔除和修正,可以提高数据预处理的效率和可靠性,所以在目前动态测量数据预处理中,越来越多的用到了基于一定算法的计算技术,诸如运用最小二乘法滑动平滑处理测量数据或用奇异谱分析将测量数据进行去噪处理等等。

(1)最小二乘法滑动平滑处理

最小二乘法数据滑动平滑处理的基本假定是:物理系统的惯性及阻尼作用使得整个工作系统的性能变化是缓慢和连续的,传感器获得的测量数据具有一定的连续性,且测量数据整体的统计特性服从正态分布[4]。

基于最小二乘法滑动平滑处理,首先要从测量数据中确定 C 个连续的正确值,当动态变化过程不是很剧烈,同时传感器的非线性不是特别严重时,相邻的这 C 个数据可以认为是线性的,对此 C 个数据进行基于最小二乘法的线性拟合,并用拟合函数的线性外推值对下一个数据进行合理性检验,如果是坏值则给予剔除,由此产生的缺失数据,用线性外推值代替。通常测量数据非线性程度的不同通过改变

收稿日期:2009-11-18 **修改日期**:2009-12-02

上述方法中的 C 值来适应,如采用较小的 C 值算法处理动态测量中非线性更为严重的数据[1]。

该方法对单个出现的坏值具有较好的处理能力,但当出现连续坏值的时候,其判别能力就很快下降,常发生错判和漏判,坏值不能及时剔除。

(2)基于奇异谱分析的去噪平滑处理

奇异谱分析(Singular Spectrum Analysis, SSA)是一种广义的功率谱分析方法,它不受正弦波假定的约束,对信号的识别和描述采用时域性的频域特征分析方式,所以可以更加灵活地对非线性和不稳定的时间序列进行去噪和特征提取的操作,同时可以实现时间序列的动态重构[5-6]。

基于奇异谱分析去噪平滑处理,首先对于长度为 N 的时间序列$\{X_i\}$,$i=1,2,\cdots,N$,用嵌入的方法重构吸引子轨道矩阵 X,对 X 进行奇异值分解可以得到 M 个按不增顺序排列的奇异值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_M \geqslant 0$,记 $S_i = \lg(\lambda_i / \sum_{j=1}^{k} \lambda_j)$,则称 $S_1, S_2, \cdots, S_M$ 为系统的奇异谱,它表示各个状态变量在整个系统中所占能量的相对关系[7]。通常,奇异值中前面的几个值较大,其余的值较小,这些比较大的奇异值对应着信号中的特征成分,比较小的奇异值则对应着信号中的噪声成分。将轨道矩阵对应的前面较大的奇异值保留,后面较小的奇异值人为置零,然后将其反变换即可得到一个新的时间序列$\{x_i'\}$,新序列降低了干扰噪声,对测量数据有平滑效果。

测量数据经过该方法处理后可以得到一条新的数据曲线,它比原始数据曲线要平滑,但同时它几乎已经改变了所有测量点的数值,用这些数据值进行下一步数据处理,显然失去了测量的真实性。

2　改进的奇异谱分析数据预处理

上述几种方法对动态过程数据的处理中都不够理想,基于最小二乘法的滑动平滑处理法虽然考虑剔除坏值,但一边拟合一边剔除的方法使得其判别能力不能始终保持良好,当坏值密集或成片出现时就容易出现错判和漏判,基于奇异谱分析的去噪平滑处理法,没有剔除坏值,而是对整体数据进行了去噪平滑,虽然处理后的数据看上去比原始数据准确,但实际上已经改变了原始数据所包含的信息,不利于下一步的处理。

笔者综合上述两种方法的特点,提出了一种改进的奇异谱分析平滑预处理方法,在提高坏值判断有效性的同时尽量保留正确反映物体变化信息的其它原始数据。首先,建立判断坏值的一个基准,为此需将测量数据进行平滑处理,由 §1 分析可知,通过奇异谱分析处理原始测量数据,能够得到较为平滑的数据曲线,为了进一步加强平滑效果,可将进行一次奇异谱分析去噪平滑处理后得到的数据值作为待处理数据再进行多次迭代奇异谱平滑处理。多次迭代平滑后,得到一条新的曲线,比较经过迭代处理后生成的新的数据曲线与原始数据曲线,可以发现,原始数据中疑似坏值的点,其平滑后的相应点会向正常值靠拢,尤其经过多次迭代平滑后,数值趋于正常值,据此,将迭代后的数据曲线作为坏值判断的基准,设定校验法则,在原始数据中找到坏值并加以剔除和修正,保留正常值。

首先按 §1 中小节(2)所述对原始序列进行奇异谱去噪平滑处理,对时间序列$\{X_i\}$,给定嵌套空间维数 M,$M<N/2$,构建吸引子轨道矩阵

$$\boldsymbol{X}=\begin{bmatrix} x_1 & x_2 & \cdots & x_{N-M+1} \\ x_2 & x_3 & \cdots & x_{N-M+2} \\ \cdots & \cdots & \cdots & \cdots \\ x_M & x_{M+1} & \cdots & x_N \end{bmatrix} \tag{1}$$

对轨道矩阵 $\boldsymbol{X}$ 做经验正交展开,得出 M 个特征向量 $E^k(1 \leqslant k \leqslant M)$,称为经验正交函数,,同时定义第 k 个主分量 p_i^k 为原始序列$\{x_i\}$在 E^k 上的正交投影系数[8]:

$$p_i^k = \sum_{j=1}^{M} x_{i+j} E_j^k, 1 \leqslant i \leqslant M \tag{2}$$

若已知各个主分量和经验正交函数,即可按下式重建各个分量序列:

$$x_i^k = \begin{cases} \dfrac{1}{M}\sum_{j=1}^{m} p_j^k E_j^k & (M \leqslant i \leqslant N-M-1) \\ \dfrac{1}{i}\sum_{j=1}^{i} p_j^k E_j^k & (1 \leqslant i \leqslant M-1) \\ \dfrac{1}{N-i+1}\sum_{j=i-N+M}^{M} p_j^k E_j^k & (N-M+2 \leqslant i \leqslant N) \end{cases} \tag{3}$$

则利用前 K 个主分量重构的序列第 i 个元素 x_i 可表示为:

$$x_i = \sum_{i}^{k} x_i^k \tag{4}$$

在应用奇异谱平滑过程中,需要考虑主分量的截取问题,也就是选择哪些主分量来重构信号. 选用的主分量太少, 会丢失部分信号的特征信息, 而选择的太多, 则又会包含过多的干扰成分,通常根据不同的测量状况由实验者估计选择主分量的值,这样的处理难免带有一定的主观性[9]。因此,可以通过寻求一种判断法则来正确选择有用的主分量,得到更为理想的处理结果。

在对一个时间序列进行奇异谱分析时，只有前面有限个奇异值比较大，这些奇异值反应了信号中的特征成分，而其余的那些较小的奇异值反应的是信号中的噪声成分，因此，信号的奇异谱曲线图中信号有效成分的奇异谱曲线会有一个明显的下降趋势以便过渡到数值较小的噪声成分的奇异谱曲线。

根据以上性质，按式(3)用前 p 个主分量重构信号 x^p，而由剩余的主分量重构的噪声信号为 n^p。此时，由 x^p 构造的像空间的奇异谱曲线应该具有明显的下降趋势，而 n^p 构造的相空间的奇异谱曲线应该是平坦的[10]，比较两者的奇异谱曲线，就可以确定 p，重构出比较理想的降噪后的信号。

根据以上分析，确定奇异谱主分量重构阶次 p 后，对测量数据进行迭代奇异谱分析平滑处理，可以得到较为平滑的数据曲线，计算原始曲线同平滑后曲线的平均距离：

$$\hat{\sigma}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{5}$$

其中，y_i 为原始测量值，$\hat{y}_i$ 为平滑曲线对应值。

若 $|\hat{y}_i-y_i|\geqslant r\hat{\sigma}$（按 $3\hat{\sigma}$ 法则取 $r=3$）[11]，则认为 y_i 是坏值。

对于异常值过多或者部分坏值偏离正常值过大的情况，$\hat{\sigma}$ 会较大，3σ 法则往往不能包括所有的异常值，这时可在剔除异常值后再次对剩下的数据运用 3σ 法则判断可能遗漏的异常值：

$$\hat{\sigma}_{n-s}=\sqrt{\frac{\sum_{i=1}^{n-s}(y_i-\hat{y}_i)^2}{n-s}} \tag{6}$$

式中 s 为上一步中找到的异常值个数。

对剩余的 $n-s$ 个数据继续运用 3σ 法则：

$$|\hat{y}_j-y_j|\geqslant 3\hat{\sigma}_{n-s} \tag{7}$$

根据坏值的分布特点进行多次 3σ 迭代检验，直至找出全部可能的坏值点并加以剔除，再用平滑后曲线对应的数值进行插补，以保证数据整体的完整性。

3 方法验证及分析

图 1 是某一动态过程测量的原始数据[12]，共有 681 个测量值。从图中可以看出，坏值分布于测量的全过程，坏值既有单个出现，又有成片连续出现，具有典型的动态过程数据特点。

分别用上文讨论的三种方法对以上数据进行预处理。在基于最小二乘法的数据滑动平滑处理中，考虑到数值具有较大的变化范围，取 C 为 5，即取 5 点进行线性拟合，并用拟合外推值进行坏值判断，处理结果如图 2 所示。奇异谱分析去噪平滑处理的嵌套系数取 $M=30$，采用前两个主分量进行重构，并进行 3 次迭代平滑处理，处理结果如图 3 所示。改进的奇异谱分析平滑法中，对经过奇异谱平滑迭代处理后的结果进行 3 次 3σ 法则检验，处理结果如图 4 所示。

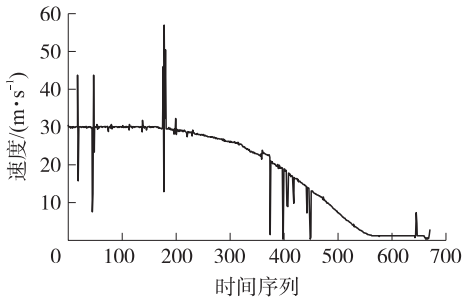

图 1 原始测量数据描点图

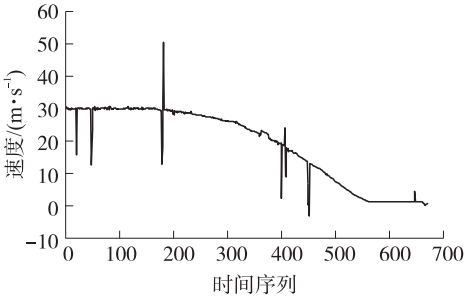

图 2 基于最小二乘法的滑动平滑预处理结果

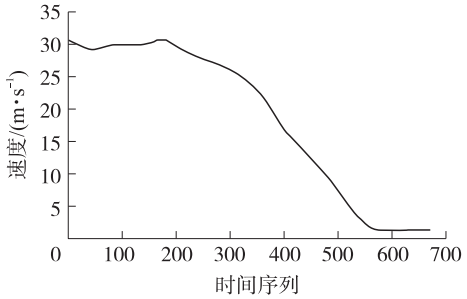

图 3 奇异谱分析去噪平滑预处理结果

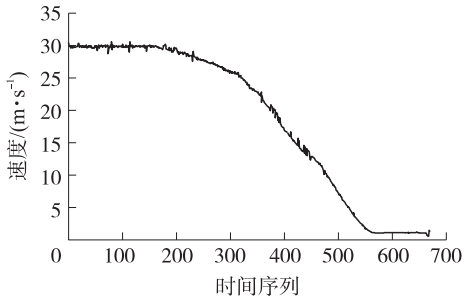

图 4 改进的奇异谱分析平滑预处理结果

从图 2 中可以看出，方法一能够剔除并修正一部分坏值，但效果不是很好，当坏值成片连续出现时，其判断能力很快下降，容易出现错判和漏判。

从图 3 可以看出，方法二可以判断出大部分的坏值，但其在剔除并修正坏值的同时也改变的了许多正确的原始测量数值，虽然处理后的数据看上去

比原始数据更平滑,但由于改变了原始数据,将会影响后续的处理结果。采用改进的奇异谱平滑预处理方法,增强了坏值判断能力。从图 4 中可以看出,该方法处理效果明显优于前两种预处理方法,检验出了全部坏值,并保留了原始测量信息。

为了进一步验证预处理算法的效果,在测量结果中选取 10 个连续的正常测量值,在其中人为制造 4 个坏值,用三种方法分别进行处理,处理结果如表 1 所示。

表 1　各算法坏值剔除及修正效果列表

时间序列点	原始测量值 /(m·s^{-1})	人为破坏值 /(m·s^{-1})	最小二乘法滑动平滑处理 /(m·s^{-1})	奇异谱分析处理 /(m·s^{-1})	改进的奇异谱分析处理 /(m·s^{-1})
452	12.87	12.87	12.87	13.15	12.87
453	12.87	24.454	12.91	13.13	12.82
454	12.72	21.232	21.23	13.08	12.74
455	12.65	3.34	3.34	13.06	12.65
456	12.61	12.61	12.61	13.03	12.61
457	12.48	12.48	12.48	13.01	12.48
458	12.24	12.24	12.24	12.99	12.24
459	12.25	12.25	12.25	12.96	12.25
460	12.15	53.26	12.15	12.94	12.15
461	12.06	12.06	12.06	12.91	12.06

从以上分析我们可以看到,基于最小二乘法滑动平滑预处理对单个出现的坏值处理能力较强,修正误差较小,但对成片连续出现的坏值失去了检验能力,只对成片坏值区域内的第一个坏值有一定的修正能力。奇异谱分析去噪平滑预处理,能准确找到坏值并加以修正,但对坏值的进行修正的同时也改变了坏值相邻的正常测量值的数值,不利于测量数据原始信息的保持和后续的进一步处理。改进后的奇异谱分析处理方法,能准确的剔除各坏值,同时修正效果非常好,无论对单个出现的坏值,还是对成片出现的坏值,实验修正后产生的误差在 0.39 % 以内。

4　结论

讨论了动态过程测量数据预处理的方法和重要性,在论述了传统数据预处理方法和普通的奇异谱分析去噪平滑预处理方法的基础上提出了一种改进的奇异普分析预处理算法,并通过实际动态测试数据对算法进行检验分析,验证了使用新方法进行动态过程测量数据预处理的有效性。

参考文献:

[1] 邝小磊. 动态测量中传感器非线性拟合方法[J]. 传感器技术,2002,21(7):38 - 40.

[2] 何平. 剔除测量数据中异常值的若干方法[J]. 航空计测技术,1995,15(1):19 - 22.

[3] 李明,卢煜,苏振中. 数据预处理中填补空缺值的方法技术[J]. 电脑知识与技术,2009,5(7):1546 - 1548.

[4] 金炯华,杨骅,童宝义. 动态数据系统的异常值智能处理[J]. 南京工学院学报,1988,18(2):55 - 62.

[5] Vautard R, Ghil M. Singular Spectrum Analysis in Nonlinear Dynamic with Application to Paleo-Climatic Series[J]. Physical D, 1989(35): 395 - 424.

[6] Doclo S. Multimicrophone Noise Reduction Using Recursive GSVD-Based Optimal Filtering with ANC Postprocessing Stage[J]. IEEE Trans on Speech and Audio Processing, 2005, 13(1): 53 - 69.

[7] 李亚安,王洪超,陈静. 基于奇异谱分解的水声信号降噪方法研究[J]. 系统工程与电子技术,2007,29(4):524 - 527.

[8] David H Schoellhamer. Singular Spectrum Analysis for Time Series with Missing Data[J]. Geophsical Research Letters, 2001, 28(16): 3187 - 3190.

[9] 江志红,丁裕国. 奇异谱分析的广义性与其应用特色[J]. 气象学报,1998,56(6):736 - 744.

[10] 刘元峰,赵玟. 基于奇异谱分析的降噪方法及其在计算最大 Liapunov 指数中的应用[J]. 应用数学和力学,2005,26(2):163 - 168.

[11] 高燕,林建辉,李翀. 磁悬浮轨道长波不平顺测试数据的预处理及时域分析[J]. 计算机应用,2009,29(1):353 - 355.

[12] 于烨军,蔡萍,吴展. 高速运动物体制动全过程测量[J]. 测控技术,2009,28(4):5 - 9.

吴　展(1984 -),男,浙江宁波人,硕士研究生,研究方向为动态检测技术与自动化装置,wuzhan@ sjtu. edu. cn;

蔡　萍(1963 -),女,江西人,博士生导师,研究方向为测试理论与传感器技术,pcai@ sjtu. edu. cn.