文章编号:1004-9037(2014)05-0743-06

基于奇异点检验的 SG 阈值滤波算法

刘晓光 窦曼莉 门晓金 石 春 吴 刚

(中国科学技术大学自动化系工业自动化研究所,合肥,230027)

摘要:SG(Savitzky-Golay)滤波算法是一种去除数字信号中白噪声的有效算法。在实际应用中,这种基于移动窗口的最小二乘算法有一个核心问题有待解决,即如何在保证滤波效果的前提下,尽可能的保留信号中的波峰信息。本文通过理论分析提出了一种 SG 阈值滤波算法,可以对不同特征的信号区间采取不同的 SG 滤波策略,并基于白噪声奇异点检验和迭代算法的思想,提出了一种阈值确定算法,增强了这种 SG 阈值滤波算法的实用性和方便性。

关键词:SG 滤波算法;白噪声;阈值滤波;奇异点检验;迭代算法

中图分类号:TP274.2 **文献标志码**:A

SG Threshold Value Filtering Algorithm Based on Singular Point Detection

Liu Xiaoguang, *Dou Manli*, *Men Xiaojin*, *Shi Chun*, *Wu Gang*

(Institute of Industrial Automation, Department of Automation, University of Science and Technology of China, Hefei, 230027, China)

Abstract: Savitzky-Golay (SG) filtering algorithm is an effective algorithm to remove the white noise in digital signal. The algorithm is based on the least squares algorithm, whose essential concern is how to ensure the filtering effect while remain the information of signals as much as possible. A SG threshold value filtering algorithm is put forward through the theoretical analysis to take different SG filtering strategy in different signal intervals with different characteristics. Then, based on the white noise singular point detection and iterative algorithm, a threshold value determination algorithm is proposed to enhance the practicability and the convenience of the SG threshold value filtering algorithm.

Key words: SG filtering algorithm; white noise; threshold value filtering; singular point detection; iterative algorithm

引 言

Savitzky-Golay(SG)滤波器是一种在时域内基于多项式,通过移动窗口利用最小二乘法进行最佳拟合的方法[1]。SG 滤波算法是一种效果很好的除白噪声的方法[2],在近红外光谱数据采集与处理中普遍被使用。SG 平滑算法的原理参见文献[3]。

在进行近红外光谱信号的采集处理过程中,SG 滤波算法在实际使用时仍有一个问题有待解决,即使得光谱足够平滑的同时又保证其分辨率。这在经典 SG 平滑算法中很难实现。因为算法中平移窗口大小固定,通常导致在变化较为平缓的光谱区域滤波效果较好,而在变化较为剧烈的其他区域导致波峰分辨率下降。为了避免这样的问题,通常需要手动修改平滑窗口大小,这大大降低了 SG 平滑算法在近红外光谱处理中的方便性和实用性。

本文提出一种 SG 阈值滤波算法,该算法可以智能识别光谱图中不含波峰的平缓区域和含有波峰的陡坡区域,在平缓区域使用 SG 滤波算法处理光谱图,在陡坡区域不作 SG 平滑处理[4]。对典型的钨灯和低压汞灯的近红外光谱图的处理结果表

基金项目:国家自然科学基金青年基金(31000672)资助项目。
收稿日期:2012-12-18;**修订日期**:2013-03-03

明,该算法既体现了 SG 良好的白噪声滤除性能又保持了足够的光谱分辨率,具有很好的实用性和应用前景。

1 基于奇异点检验的 SG 阈值滤波

1.1 SG 阈值滤波

定义离散时间序列 $x(k),k=1,2,\cdots,N$ 的一阶差分序列

$$\begin{aligned} y(k)&=x(k)-x(k-1)\\ y(0)&=0,k=1,2,\cdots,N \end{aligned} \tag{1}$$

信号的一阶差分值表征了数据的变化程度,其值越大表明数据波形越陡[5]。对一阶差分值作如下处理

$$\begin{aligned} y(k)&=\begin{cases}0 & |\ y(k)\ |<\varphi\\ y(k) & |\ y(k)\ |\geqslant\varphi\end{cases}\\ k&=1,2,\cdots N \end{aligned} \tag{2}$$

式中:φ 定义为平坦域阈值,定义时间序列 $x(k)$ 的平坦域起点为

$$\begin{aligned} &\{k_{\text{start}}\mid y(k)=0,y(k-1)\neq 0,\\ &y(k+1)=0\}\quad k=1,2,\cdots,N \end{aligned} \tag{3}$$

平坦域终点为

$$\begin{aligned} &\{k_{\text{end}}\mid y(k)=0,y(k-1)=0,\\ &y(k+1)\neq 0\}\quad k=1,2,\cdots,N \end{aligned} \tag{4}$$

于是得到平坦域区间

$$\Psi\in[k_{\text{start}},k_{\text{end}}] \tag{5}$$

$x(k)$ 剩下的区间定义为峰域,峰域中存在有效波峰的可能性较大,应使用窗口宽度较小的 SG 滤波或者不滤波。由式(2)可知,为了确定时间序列的分段区间,必须确定平坦阈值 φ。

1.2 基于 2σ 准则的奇异点检验

由莱特准则可知,若序列 X 服从方差为 σ 的正态分布,则

$$P\{u-2\sigma\leqslant X\leqslant u+2\sigma\}=0.9545 \tag{6}$$

即数据点不在区间$[-2\sigma,2\sigma]$的概率不到 5%[6]。因此可以认为,不在此区间内的数据点均为奇异点,其中标准差 σ 可由贝塞尔公式求得[7]。

正态分布序列的这种统计特性称为 2σ 准则。白噪声服从正态分布,因而可以利用 2σ 准则判断白噪声序列中是否有奇异点存在[8]。对于近红外光谱信号,仪器热噪声占据噪声的主要成分,这导致光谱信号噪声主要为白噪声[9]。

经典 SG 滤波算法采用基于滑动窗口的最小二乘拟合,对于变化缓慢的平坦域信号,通常所采用的 11 点 3 阶拟合可以有比较好地拟合效果,使得残差序列可以被近似地认为是一个白噪声序列,符合 2σ 准则。然而,对于变化剧烈的峰域,在阶数一定的情况下,过宽的窗口将使得拟合效果变差,这种情况下,拟合残差序列中必将出现违反 2σ 准则的奇异点[10]。

图 1,2 充分说明了这种明显的数字特征。图 1(a)显示的是未经滤波的钨灯光谱,对整个光谱做 11 点 3 阶滤波后的谱图如图 1(b)所示,图 1(c)显示的是拟合残差分布。对残差做奇异点检验,显示这种滤波方式下残差奇异点数为 3。与此相对应,对未经滤波的低压汞灯光谱做同样处理得到图 2,对残差做奇异点检验得到奇异点数达到 18。

根据图 1,2 所示,由于经典 SG 滤波算法中未对波峰信号进行保护,导致了滤波后残差序列中出现过多奇异点。SG 阈值滤波为了实现对波峰信号的保护,将信号划分为平坦域和峰域,并对不同区域采用不同窗口宽度的 SG 滤波。判断平坦域和峰域区间的大小十分重要,若峰域区间划分过小将使得波峰信号得不到有效保护。为了判断平坦域区间和峰域区间的划分是否合理,本文提出了一种对平坦域区间信号经 SG 滤波后的残差序列进行奇异点检测的方法。

假设共有 n 个平坦域区间,设为 $\psi_i,i=1,2,\cdots,n$。第 i 个平坦域 ψ_i 的长度为 N_i,平坦域 ψ_i 上

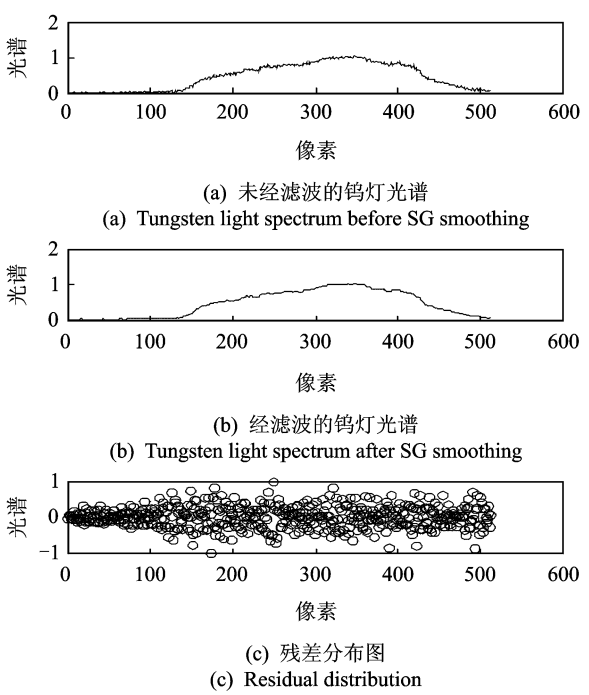
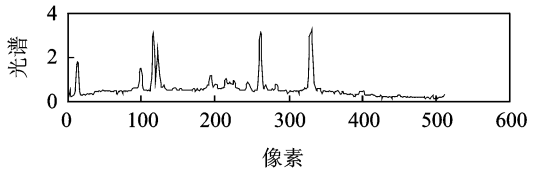
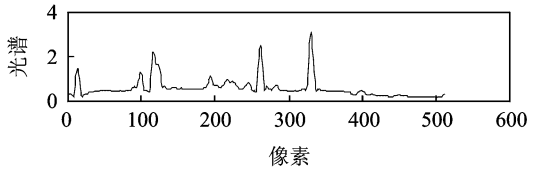


图 1 钨灯光谱 SG 平滑处理前后及残差分布图

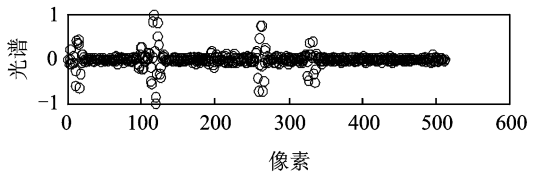
Fig. 1 Tungsten light spectrum before and after SG smoothing and residual distribution



(a) 未经滤波的低压汞灯光谱
(a) Low pressure mercury lamp spectrum before SG smoothing



(b) 经滤波的低压汞灯光谱
(b) Low pressure mercury lamp spectrum after SG smoothing



(c) 残差分布图
(c) The residual distribution

图 2　低压汞灯光谱 SG 平滑处理前后及残差分布图

Fig. 2　Low pressure mercury lamp spectrum before and after SG smoothing and residual distribution

的原始信号序列为 $x_i(k)$，$k=1,2,\cdots,N_i$，经过 SG 滤波算法处理之后的信号序列为 $\hat{x}_i(k)$，$k=1,2,\cdots,N_i$，残差序列为 $\delta_i(k)=\hat{x}_i(k)-x_i(k)$，$k=1,2,\cdots,N_i$。SG 滤波算法中，当阶次一定时，若点数选择合适，则拟合效果良好，此时 $\delta_i(k)$ 应为白噪声，服从正态分布，满足 2σ 准则。

原始的离散时间序列 $x(k)$，$k=1,2,\cdots,N$ 可表示为 $x(k)=\dot{x}(k)+w(k)$，其中 $\dot{x}(k)$ 为真实值，为白噪声，在拟合良好的情况下，子序列 $\delta_1,\delta_2,\cdots,\delta_n$ 的集合形成的新序列 $\delta'=\delta_1\cup\delta_2\cup\cdots\cup\delta_n$ 和原序列满足相同的正态分布特性，可用 2σ 准则对其进行奇异点检验。若 δ' 满足式(7)，则认为残差序列通过奇异点检验。

$$\#\{a\mid a\in\delta',a\notin[-2\sigma,2\sigma]\}\leqslant\gamma \tag{7}$$

式中：$\#$ 表示集合中元素个数，σ 为序列 δ' 的方差，若设 δ' 的长度为 l，算术平均值为 $\overline{\delta'}$，则 σ 由式(8)给出

$$\sigma=\sqrt{\sum_{i=1}^{l}\frac{(\delta'_i-\overline{\delta'})^2}{l-1}} \tag{8}$$

为奇异点数阈值。由式(6)可知

$$\gamma=[0.0455\cdot l] \tag{9}$$

式中：$[\cdot]$ 为取整运算符。

1.3　平坦阈值 φ 的选取

对于给定离散时间序列 $x(k)$，$k=1,2,\cdots,N$，根据式(1)可求得其一阶离散导数序列 $y(k)$，选取较大的平坦阈值 φ 对 $y(k)$ 进行如式(2)所示的处理，并且根据式(3～5)确定 $x(k)$ 的平坦域 ψ_i，$1\leqslant i\leqslant n$，其中 n 为平坦域区间个数。对平坦域内数据点 $x_i(k)$，$1\leqslant i\leqslant n$，$1\leqslant k\leqslant N_i$ 做 SG 平滑处理，其中 N_i 为平坦域区间 ψ_i 中数据个数。经过 SG 平滑后处理后，对应平坦域数据区间内的数据变为 $\hat{x}_i(k)$，则 SG 拟合后的残差序列为

$$\begin{cases}\delta(k)\mid\delta(k)=x_i(j)-\hat{x}_i(j)\\ 1\leqslant k\leqslant n,1\leqslant j\leqslant N_i\\ 1\leqslant k\leqslant\sum_{i=1}^{n}N_i\end{cases} \tag{10}$$

对 $\delta(k)$ 作白噪声奇异点检验，若成功则表明滤去的确实是白噪声；若失败则表明残差序列中含有较多有用信息，这些有用信息的出现是由于阈值设置不合理，导致峰域区间过小，未能起到对波峰信号的良好保护所致。所以，此时应该减小阈值，使得峰域区间扩大，然后重新处理，直到奇异点检验成功为止。初始值取 $y(k)$ 的最大值 y_m，若过大则折半继续，直到满足奇异点检验的相邻平坦阈值相差不大时就认为滤波结束。定义算法结束条件

$$\zeta=\frac{\varphi(j)-\varphi(j-1)}{\varphi(j)}\times100\%\leqslant\mu$$
$$j\geqslant1,\text{且 }\varphi(0)=0 \tag{11}$$

式中：ζ 定义为阈值稳态误差，通常取 $\mu=3\%\sim5\%$，j 为算法迭代次数。

基于白噪声奇异点检验的 SG 平滑算法(SG threshold value filtering based on singular point detection of white noise，WSDSG)的流程如下所示，其中 $x(k)$，$k=1,2,\cdots,N$ 为待滤波的时间序列。

(1)依据式(1)求 $x(k)$ 的一阶差分序列 $y(k)$；

(2)取初始平坦阈值 $\varphi(1)$ 为 $y(k)$ 中最大值 y_m；

(3)依据式(2)对 $y(k)$ 做截断处理，并根据式(3～5)求得平坦域区间；

(4)对(3)中求得的平坦域区间做 SG 平滑滤波；

(5)依据式(10)求取平坦域滤波前后的残差，对该残差做白噪声奇异点检验，若检验成功且 $\varphi(j)=y_m$，$j\geqslant1$ 则算法结束；若检验成功且 $\varphi<y_m$

则取

$$\varphi(j)=\frac{\varphi(j-1)+\varphi(j-2)}{2} \quad j\geqslant 2 \quad (12)$$

式中:$\varphi(0)=0$。校验式(11)是否成立,若成立则算法结束,否则跳至第(3)步;若检验不成功则取

$$\varphi(j)=\frac{\varphi(j-1)}{2} \quad j\geqslant 2 \quad (13)$$

跳至第(3)步;

(6)合适的平坦阈值 $\varphi=\varphi(j)$,$j\geqslant 1$,算法结束,滤波完成。

算法的流程图如图 3 所示。

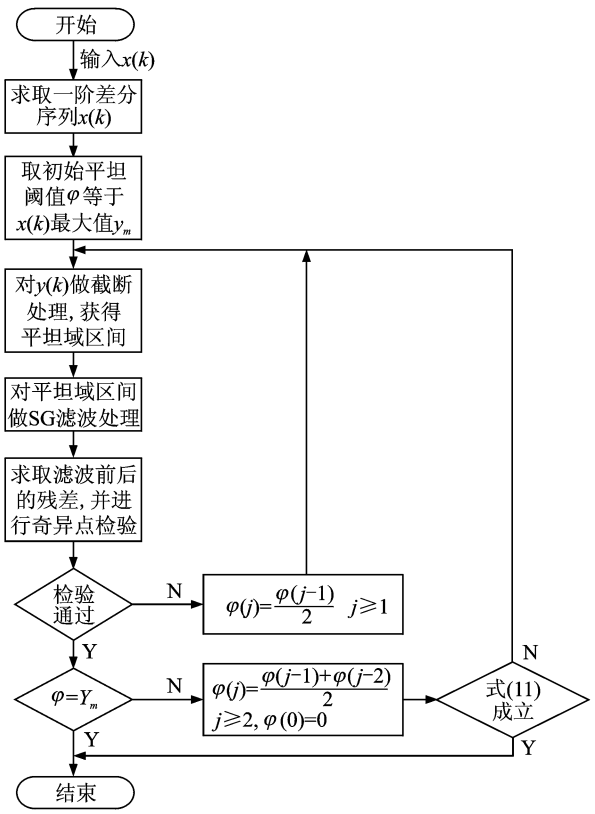


图 3 WSDSG 滤波算法流程图

Fig. 3 Algorithm flowchart of WSDSG filtering

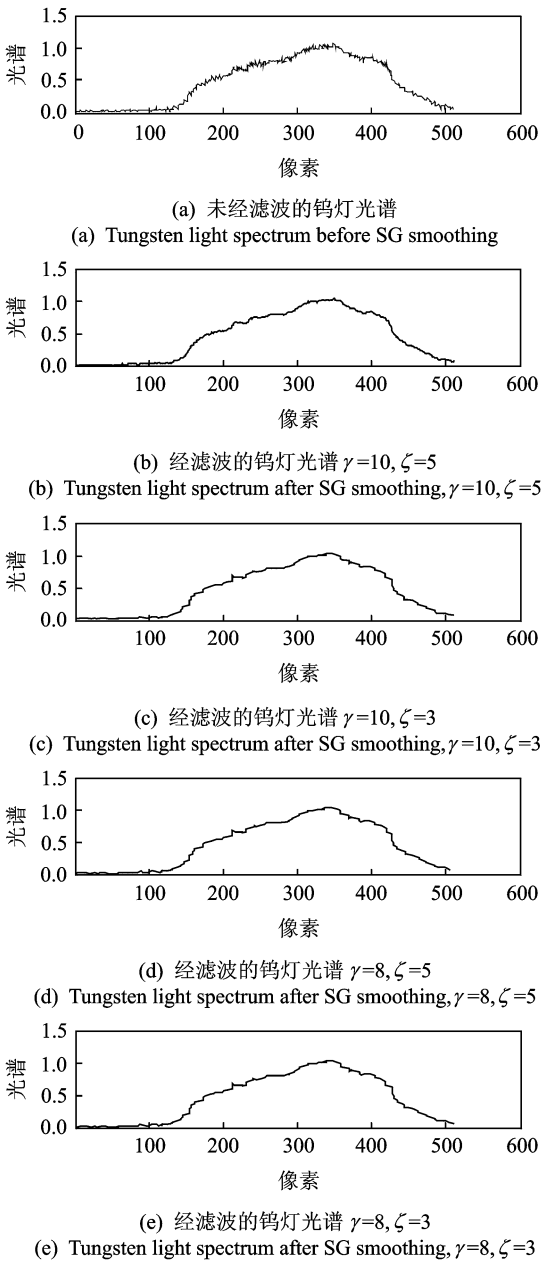


图 4 不同 γ 和 ζ 取值下对钨灯近红外光谱的处理效果图

Fig. 4 Treatment effects of tungsten lamp near infrared spectrum under condition of different γ and ζ

2 算法仿真

算法仿真所使用的近红外光谱的采集仪器为自主研发,探测器采用基于 InGaAs 材料的线性光敏二极管阵列。由于采用了非商用仪器,从而可以采集到未经滤波的光谱数据。仪器采集的钨灯的近红外光谱如图 4(a)所示,该谱图变化趋势缓慢,无明显波峰存在。

图 4(b,c)给出的是算法中令 $\gamma=10$,ζ 分别为 5%和 3%时的光谱处理图。图 4(d,e)给出的是算法中令 $\gamma=8$,ζ 分别为 5%和 3%时的光谱处理图。

从图中可以看出,对于钨灯光谱这种无剧烈变化的光谱图而言,γ 和 ζ 的取值变化对最终滤波效果影响不大。

表 1 是 γ 和 ζ 取不同值时对应的算法迭代次数以及每次迭代所得到的平坦域值变化和白噪声奇异点数检测值。可见,WSDSG 算法将钨灯光谱整体识别为一个平坦域,因而算法迭代 1 次即结束,平坦阈值即为光谱一阶差分序列最大值。这说明,WSDSG 算法对于这类无峰缓坡的处理过程等同于采用经典 SG 滤波算法的处理过程。

表 1　钨灯光谱的 WSDSG 处理过程中，迭代次数以及每次的平坦阈值和奇异点个数变化表

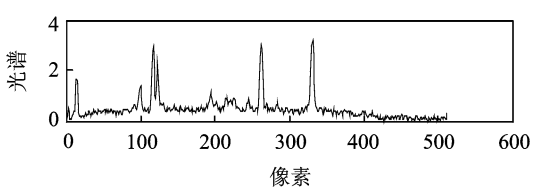
Table 1　Iteration times, threshold value and number change of singular point when processing Tungsten light spectrum with WSDSG algorithm

迭代次数	平坦阈值 $\gamma=10,8$ $\zeta=5\%$	平坦阈值 $\gamma=10,8$ $\zeta=3\%$	奇异点个数 $\gamma=10,8$ $\zeta=5\%$	奇异点个数 $\gamma=10,8$ $\zeta=3\%$
1	0.038	0.038	3	3
	0.038	0.038	3	3
2	结束	结束	结束	结束
	结束	结束	结束	结束

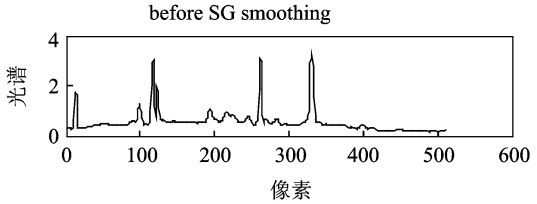
图 5(a)给出的是仪器采集的低压汞灯近红外光谱图，可以看到，该图存在 6 个明显的特征峰，从左到右依次对应的像素点为 13，100，117，123，262，332，每个特征峰的半波宽均很窄，决定了不能采用传统的 SG 滤波算法对该光谱处理，否则必然导致光谱分辨率下降。

图 5(b，c)给出的是算法中令 $\gamma=10$，ζ 分别为 5%和 3%时的光谱处理图。图 5(d，e)给出的是算法中令 $\gamma=8$，ζ 分别为 5%和 3%时的光谱处理图。两图对比中可以看出 $\gamma=10$ 时 WSDSG 算法未能很好地识别出像素序号为 123 的波峰，而当 $\gamma=8$ 时该算法成功识别出了该峰，因而对该峰进行了保护，使得图 5(d，e)比图 5(b，c)的光谱分辨率有所提高。

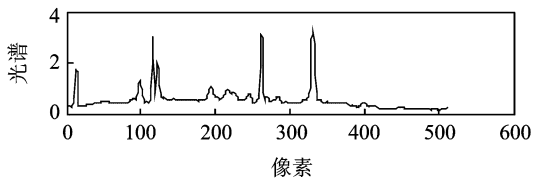
结合表 2 可以看出，对于这类存在剧烈变化的波峰的光谱图，ζ 的取值决定了算法收敛的速度，5%和 3%的收敛误差对应的滤波效果差异并不太明显。对滤波效果有重大影响的是 γ 的取值，γ 值越小，WSDSG 算法得到的平坦阈值越小，对于光谱波峰保护的越完善。然而，由于光谱中非白噪声



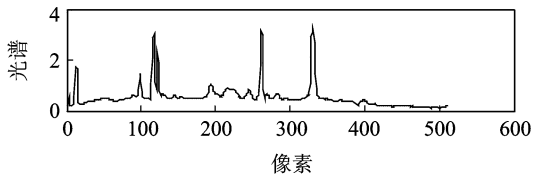
(a) 未经滤波的低压汞灯光谱
(a) Low pressure mercury lamp spectrum before SG smoothing



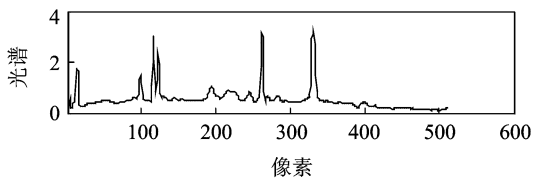
(b) 经滤波的低压汞灯光谱 $\gamma=10,\zeta=5$
(b) Low pressure mercury lamp spectrum after SG smoothing, $\gamma=10,\zeta=5$



(c) 经滤波的低压汞灯光谱 $\gamma=10,\zeta=3$
(c) Low pressure mercury lamp spectrum before SG smoothing, $\gamma=10,\zeta=3$



(d) 经滤波的低压汞灯光谱 $\gamma=8,\zeta=5$
(d) Low pressure mercury lamp spectrum before SG smoothing, $\gamma=8,\zeta=5$



(e) 经滤波的低压汞灯光谱 $\gamma=8,\zeta=3$
(e) Low pressure mercury lamp spectrum before SG smoothing, $\gamma=8,\zeta=3$

图 5　不同 γ 和 ζ 取值下对低压汞灯近红外光谱的处理效果图

Fig. 5　Treatment effects of low pressure mercury lamp spectrum under condition of different γ and ζ

表 2　低压汞灯光谱的 WSDSG 处理过程中，迭代次数以及每次的平坦阈值和奇异点个数变化表

Table 2　Iteration times, threshold value and number change of singular point when processing low pressure mercury lamp spectrum with WSDSG algorithm

迭代次数	平坦阈值 $\gamma=10,8$ $\zeta=5\%$	平坦阈值 $\gamma=10,8$ $\zeta=3\%$	奇异点个数 $\gamma=10,8$ $\zeta=5\%$	奇异点个数 $\gamma=10,8$ $\zeta=3\%$
1	1.207	1.207	18	18
	1.207	1.207	18	18
2	0.603	0.603	10	10
	0.603	0.603	10	10
3	0.905	0.905	10	10
	0.302	0.302	8	8
4	0.754	0.754	10	10
	0.4525	0.453	10	10
5	0.830	0.830	10	10
	0.2262	0.226	6	6
6	结束	0.792	结束	10
	0.339	0.339	8	8
7		结束		结束
	0.283	0.283	8	8
8	0.311	0.311	8	8
9	结束	0.297	结束	8
10		结束		结束

成分的影响，γ 也不能太小，否则将导致算法不能收敛。实际应用中，建议 γ 取 8～10 较为合适。

3 结束语

本文针对经典 SG 算法在滤波过程中不能保证滤波效果的同时兼顾分辨率的问题，提出了 SG 阈值滤波算法，并通过理论分析说明了平坦域阈值在整个滤波算法中的重要性，提出了一种基于白噪声奇异点检验的思想，并运用迭代的方法迅速找到合适平坦阈值的算法，即 WSDSG 算法。仿真实验证明，本算法不但具有很好的白噪声去除性能，还具有优异的波峰保持性能，很好地弥补了传统 SG 滤波算法的不足。本算法在近红外光谱数据的预处理过程中获得了较好的应用，对于 SG 滤波算法在各种降噪领域的推广具有较高的价值。

参考文献：

[1] Chen D L, Chen Y Q, Xue D Y. 1-D and 2-D digital fractional-order Savitzky-Golay differentiator [J]. Signal, Image and Video Processing, 2012, 6(3): 503-511.

[2] 马迎辉，赵刚，陈文针. 高压测试系统中噪声处理的算法研究[J]. 高电压术，2001，27(3)：80-82.
Ma Yinghui, Zhao Gang, Chen Wenzhen. Study of noise reduction for high voltage testing system[J]. High Voltage Engineering, 2001, 27(3): 80-82.

[3] Luo J W, Ying K, Bai J. Savitzky-Golay smoothing and differentiation filter for even number data[J]. Signal Processing, 2005, 85(7): 1429-1434.

[4] 吕振肃，马文. 自适应小波阈值算法在心电信号去噪中的应用[J]. 数据采集与处理，2009，24(3)：313-317.
Lü Zhensu, Ma Wen. Application of adaptive wavelet threshold algorithm in ECG signal denoising [J]. Journal of Data Acquisition and Processing, 2009, 24(3): 313-317.

[5] 陈韬伟，金炜东，李杰. 雷达辐射源信号一阶差分自相关特征提取[J]. 计算机工程与应用，2011，47(26)：143-145.
Chen Taowei, Jin Weidong, Li Jie. Feature extraction of radar emitter signals based on autocorrelation function of first difference[J]. Computer Engineering and Applications, 2011, 47(26): 143-145.

[6] 郭兴明，柯明，肖守中. 3σ 规则和 μ 律阈值法在心音信号去噪中的应用[J]. 计算机工程，2010，36(7)：285-287.
Guo Xingming, Ke Ming, Xiao Shouzhong. Application of 3σ-rule and μ-threshold method in heart sound signal de-noising[J]. Computer Engineering, 2010, 36(7): 285-287.

[7] 朱新岩，史忠科. 一种改进的野值在线预处理 3σ 方法[J]. 弹箭与制导学报，2008，28(6)：69-71.
Zhu Xinyan, Shi Zhongke. An improved 3σ method for outlier on-line pre-processing[J]. Journal of Projectiles, Rockets, Missiles and Guidance, 2008, 28(6): 69-71.

[8] 朱墨子，包鑫. 异常点检测与 Savitzky-Golay 滤波算法在手写系统中的应用[J]. 机电工程，2008，25(8)：8-10.
Zhu Mozi, Bao Xin. Application of outlier detection and Savitzky-Golay filter in handwritten systems[J]. Mechanical & Electrical Engineering Magazine, 2008, 25(8): 8-10.

[9] Xie Sh F, Xiang B R, Yu L Y, et al. Tailoring noise frequency spectrum to improve NIR determinations [J]. Talanta, 2009, 80(2): 895-902.

[10] Krishnan S R, Seelamantula C S. On the selection of optimum Savitzky-Golay filters[J]. IEEE Transactions on Signal Processing, 2013, 61(2): 380-391.

作者简介：刘晓光(1988-)，男，硕士研究生，研究方向：仪器和测控技术、图像处理，E-mail：liuxg88@mail.ustc.edu.cn；窦曼莉(1988-)，女，博士研究生，研究方向：先进控制与优化、汽车电子开发与应用；门晓金((1988-)，男，硕士研究生，研究方向：先进控制与优化、仪器和测控技术；石春(1980-)，男，讲师，研究方向：光谱仪器设计、汽车电子开发与应用；吴刚(1964-)，男，教授，研究方向：先进控制与优化。