A Simple Algorithm for Despiking Raman Spectra

Darren A. Whitaker[1,2] and Kevin Hayes[1,2]

[1]Pharmaceutical Manufacturing Technology Centre (PMTC), Bernal Institute, University of Limerick, Limerick, Ireland
[2]Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland

Abstract

Raman Spectroscopy is a widely used analytical technique, favoured when molecular specificity with minimal sample preparation is required. The majority of Raman instruments use charge-coupled device (CCD) detectors, these are susceptible to cosmic rays and as such multiple spurious spikes can occur in the measurement. These spikes are problematic as they may hinder subsequent analysis, particularly if multivariate data analysis is required. In this work we present a new algorithm to remove these spikes from spectra after acquisition. Specifically we use calculation of modified Z scores to locate spikes followed by a simple moving average filter to remove them. The algorithm is very simple and its execution is essentially instantaneous, resulting in spike-free spectra with minimal distortion of actual Raman data. The presented algorithm represents an improvement on existing spike removal methods by utilising simple, easy to understand mathematical concepts, making it ideal for experts and non-experts alike.

Keywords: Modified Z-Scores; Data Processing; Raman Spectra; Despiking

1 Introduction

Regular practitioners of Raman spectroscopy will be familiar with the problems caused by cosmic spikes. These spurious nuisance spikes typically appear at random positions and present as positive, narrow bandwidth peaks. They arise when a charge-coupled device, used as a detector in modern Raman systems, is struck by an errant high-energy particle. Predominately these are muons but may also be protons or neutrons[1]. Often these particles are caused by genuine cosmic rays (exotic particle produced by exploding supernovae, black holes, *etc.*) but they can also be a result of decay of radioactive atoms present in the locality of the CCD detector. The presence of these cosmic spikes hampers further multivariate data analysis. For example, they cause distortion of the principle component direction in principal component analysis[2], introduce erroneous variables in multivariate curve resolution or regression techniques and can also result in misidentification in classification analysis[3].

It is desirable to be able to automatically identify, reduce and/or remove these spikes from Raman spectra. This becomes even more relevant when processing large mapping datasets prevalent within pharmaceutical research[4,5,6,7]. Methodologies reported in the literature can be broadly separated into three categories: (i) additional acquisition based methods; (ii) methods involving hardware modification; and the category the present work falls into (iii) single-scan correction via filtering or smoothing. In the first category algorithms such as robust summation[8] or upper-bound spectrum[9] methodologies take advantage of the fact the probability of the same pixel experiencing a cosmic spike in successive measurements is very low. In the second category methods such as analyzing the full CCD image[10], division of the spectrograph slit and image curvature correction[11]. Finally in the third category methods such as moving window filtering[12], spike fitting[13], wavelet transforms[14,15,16] and median or polynomial filters[17,18].

In this work we present a despiking algorithm based on the calculation of modified Z scores to locate spikes and a simple moving average filter to remove the located spikes. This algorithm is computationally efficient and inexpensive, accurate and easy to execute and program, and should be of great utility to all users of Raman spectroscopy. Additionally the use of modified Z scores is recommended by the National Institute of Standards and Technology (NIST) as an outlier detection methodology[19] and as such this method should fit easily into regulated industries such as the pharmaceutical industry.

2 Despiking Algorithm

Let $Y_1, \ldots, Y_n$ represent the values of a single Raman spectrum recorded at equally spaced wavenumbers. From this series, form the detrended differenced series $\nabla Y_t = Y_t - Y_{t-1},\ (t = 2, \ldots, n)$. This simple data processing step has the effect of annihilating linear and slow moving curve linear trends, however, sharp localised spikes will be preserved.

Denote the median and the median absolute deviation of the differenced series by $M = \mathrm{median}\{\nabla Y_t\}$ and $\mathrm{MAD} = \mathrm{median}\{|\nabla Y_t - M|\}$ respectively, and define modified Z scores by

$$Z_t = \frac{0.6745 \times (\nabla Y_t - M)}{\mathrm{MAD}}.$$

(The multiplier 0.6745 is included to adjust for asymptotic bias that arises when MAD is calculated from normally distributed data[20,21].) In theory the modified Z scores can be compared with the tabulated tail quantiles from the normal distribution. The criterion $|Z_t| > 3.5$ was proposed as a guideline by the American Society of Quality Control as the basis of an *outlier-labeling* rule with the objective of screening large datasets for observations that are "sufficiently suspect to merit further investigation,"[22]. In this paper, wavenumbers with modified Z scores exceeding $\tau = 6$ in magnitude were flagged as contributing to the formation of an anomalous spike. In practice the scientist will have immediate control over this threshold parameter. Lowering the value of the spike labelling threshold parameter τ will make the algorithm more sensitive to the presence of potential spikes.

Interpolated values $\tilde{Y}_t$ are then obtained at each candidate wavenumber by calculating the mean of its immediate neighbours, specifically $\tilde{Y}_t = \frac{1}{w}\sum_{t-m}^{t+m} Y_t \times \mathbb{I}(Z_t < \tau)$, where $\mathbb{I}(u)$ is an indicator function taking value 1 if the condition u is satisfied and 0 otherwise, and $w = \sum_{t-m}^{t+m} \mathbb{I}(Z_t < \tau)$. This has the effect of excluding the value Y_t itself, and values of Y flagged as contributing to the formation of a spike, from the calculation of $\tilde{Y}_t$. This is desirable because in order to characterise a spike in a Raman spectrum invariably requires a sequence of 2 to 5 inflated values of Y in a row. The width of the moving average neighbourhood is controlled by the parameter m, which was set in our applications to $m = 5$. Finally, in order to accomodate the eventuality of a spike appearing at the first or last wavenumber, the values of Z_1 and Z_n are automatically set to exceed the spike labelling threshold.

3 Case Study

3.1 Experimental

Sample Preparation

Theophylline (99 %, Sigma Aldrich) and Microcrystalline Cellulose (MCC101, Avicel) were blended together in 10 % w/w proportions and compacted into 12 mm diameter tablets using a single rotary punch tablet press.

Instrumental Set-up

Raman spectra were collected from a 12 mm tablet using a LabRAM HR Evolution (HORIBA UK Ltd., Stanmore, UK) spectrometer system. A custom spectrometer control and data acquisition script was written using the VBScripting language to enable mapping of the full tablet surface to be carried out (Figure 1). 407 indivdual spectra were recorded at 500 μm intervals using a 785 nm laser line, 10 x objective, 5 s acquisition time, 100 μm hole diameter in the range 1230 to 1330 cm^{-1}.

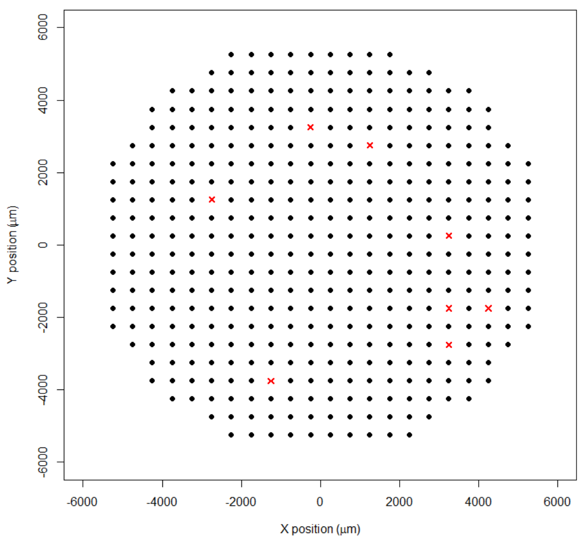

Figure 1: Sites of individual spectra over surface of 12 mm tablet (X denotes subsequently identified spike location).

Data Analysis

Spectral processing and analysis were preformed using the R environment for statistical computing[23]. Custom functions for the calculation of modified Z scoring and annihilation of located spikes were developed and are included as supplementary files to this manuscript. The *hyperSpec* package[24] was used for easy management of spectral data within the R environment.

4 Results and Discussion

The Raman spectra were recorded in the wavenumber region 1230 to 1330 cm^{-1}, in this region three characteristic bands of theophylline are present. The three bands centred at *ca.* 1248, 1286 and 1314 cm^{-1} are assigned to $\nu(C{-}N)_{sym}$[25]. An overlay of all the acquired spectra shows that cosmic spikes are present in the dataset (Figure 2a).

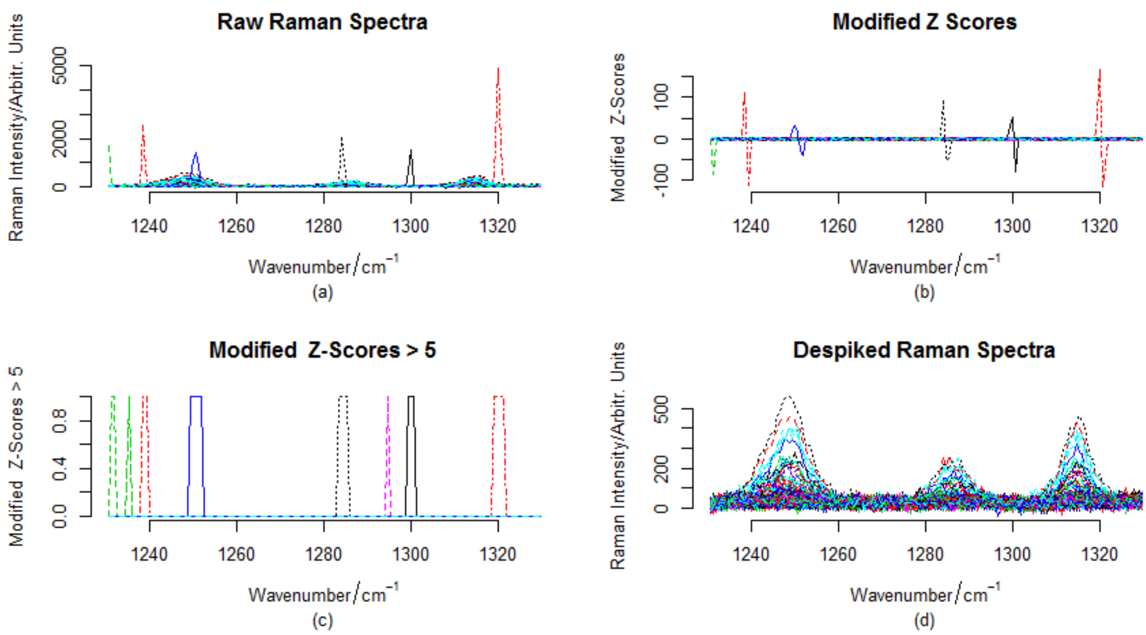

Figure 2: Raman dataset acquired on 12 mm pharmaceutical tablet and results of despiking algorithm.

After calculation of modified Z scores (Figure 2b) and thresholding by setting a suitable value of τ (Figure 2c), the spikes can be removed and smoothed by applying a the moving average filter as described earlier. This results in a corrected dataset where the spikes are removed and the correct signal from the theophylline bands can be easily observed (Figure 2d).

5 Conclusion

We present a new algorithm based on modified Z score outlier detection for the identification and removal of cosmic spikes in Raman spectroscopic data. The algorithm was shown to be effective on a medium sized dataset acquired on a sample of pharmaceutical relevance. The algorithm is sufficiently computationally cheap to be run on almost any computer system and is also platform independent. This makes the algorithm useful for all types of Raman data analysis, including mapping measurements and real-time inline analysis.

Supplementary Information

The example dataset described in this paper and the despiking algorithm are provided free of charge on the publishers website.

Acknowledgements

This work was co-funded by the Pharmaceutical Manufacturing Technology Centre (PMTC) under Enterprise Irelands (EI) - Technology Centres Programme & by the European Regional Development Fund (ERDF) under Ireland's European Structural and Investment Funds Programmes 2014 - 2020

References

[1] Groom, D.. Cosmic rays and other nonsense in astronomical CCD imagers. Experimental Astronomy 2002;14(1):45–55. doi:10.1023/A:1026196806990.

[2] De Groot, P.J., Postma, G.J., Melssen, W.J., Buydens, L.M.C., Deckert, V., Zenobi, R.. Application of principal component analysis to detect outliers and spectral deviations in near-field surface-enhanced Raman spectra. Analytica Chimica Acta 2001;446(1-2):71–83. doi:10.1016/S0003-2670(01)01267-3.

[3] Zhang, L., Henson, M.J.. A practical algorithm to remove cosmic spikes in raman imaging data for pharmaceutical applications. Applied Spectroscopy 2007;61(9):1015–1020. doi:10.1366/000370207781745847.

[4] Potter, C.B., Kollamaram, G., Zeglinski, J., Whitaker, D.A., Croker, D.M., Walker, G.M.. Investigation of polymorphic transitions of piracetam induced during wet granulation. European Journal of Pharmaceutics and Biopharmaceutics 2017;119:36–46. doi:10.1016/j.ejpb.2017.05.012.

[5] Vankeirsbilck, T., Vercauteren, A., Baeyens, W., Van der Weken, G., Verpoort, F., Vergote, G., et al. Applications of Raman spectroscopy in pharmaceutical analysis. TrAC Trends in Analytical Chemistry 2002;21(12):869–877. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0165993602012086`. doi:10.1016/S0165-9936(02)01208-6.

[6] Gordon, K.C., McGoverin, C.M.. Raman mapping of pharmaceuticals. International Journal of Pharmaceutics 2011;417(1-2):151–162. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0378517310009713`. doi:10.1016/j.ijpharm.2010.12.030.

[7] Wartewig, S., Neubert, R.H.. Pharmaceutical applications of Mid-IR and Raman spectroscopy. Advanced Drug Delivery Reviews 2005;57(8):1144–1170. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0169409X0500058X`. doi:10.1016/j.addr.2005.01.022.

[8] Takeuchi, H., Hashimoto, S., Harada, I.. Simple and efficient method to eliminate spike noise from spectra recorded on charge-coupled device detectors. Applied spectroscopy 1993;47(1):129–131.

[9] Zhang, D., Ben-Amotz, D.. Removal of cosmic spikes from hyper-spectral images using a hybrid upper-bound spectrum method. Applied Spectroscopy 2002;56(1):91–98. doi:10.1366/0003702021954269.

[10] Zhao, J.. Image curvature correction and cosmic removal for high-throughput dispersive Raman spectroscopy. Applied spectroscopy 2003;57(11):1368–1375.

[11] Zhang, D., Hanna, J.D., Ben-Amotz, D.. Single scan cosmic spike removal using the upper bound spectrum method. Applied spectroscopy 2003;57(10):1303–1305.

[12] Katsumoto, Y., Ozaki, Y.. Practical algorithm for reducing convex spike noises on a spectrum. Applied Spectroscopy 2003;57(3):317–322. doi:10.1366/000370203321558236.

[13] Hill, W., Rogalla, D.. Spike-Correction of Weak Signals from Charge-Coupled Devices and Its Application to Raman Spectroscopy. Analytical Chemistry 1992;64(21):2575–2579. doi:10.1021/ac00045a019.

[14] Maury, A., Revilla, R.I.. Autocorrelation Analysis Combined with a Wavelet Transform Method to Detect and Remove Cosmic Rays in a Single Raman Spectrum. Applied spectroscopy 2015;69(8):984–92. URL: http://asp.sagepub.com/content/69/8/984.full. doi:10.1366/14-07834.

[15] Ehrentreich, F., Summchen, L.. Spike removal and denoising of Raman spectra by wavelet transform methods. Analytical Chemistry 2001;73(17):4364–4373. doi:10.1021/ac0013756.

[16] Tian, Y., Burch, K.S.. Automatic Spike Removal Algorithm for Raman Spectra. Applied Spectroscopy 2016;70(11):1861–1871. doi:10.1177/0003702816671065.

[17] Phillips, G.R., Harris, J.M.. Polynomial Filters for Data Sets with Outlying or Missing Observations: Application to Charge-Coupled-Device-Detected Raman Spectra Contaminated by Cosmic Rays. Analytical Chemistry 1990;62(21):2351–2357. doi:10.1021/ac00220a017.

[18] Schulze, H.G., Turner, R.F.B.. A fast, automated, polynomial-based cosmic ray spike-removal method for the high-throughput processing of raman spectra. Applied Spectroscopy 2013;67(4):457–462. doi:10.1366/12-06798.

[19] Heckert, N.A., Filliben, J.J.. Exploratory Data Analysis. In: NIST/SEMATECH e-handbook of statistical methods; vol. 1; chap. 1. 2003,URL: http://www.itl.nist.gov/div898/handbook. doi:papers3://publication/uuid/DE51D3F6-8B2C-4EEC-8B48-B598C13EE5F4.

[20] Tukey, J.W.. Exploratory data analysis; vol. 2. Reading, Mass.; 1977.

[21] Hayes, K.. Finite-sample bias-correction factors for the median absolute deviation. Communications in Statistics: Simulation and Computation 2014;43(10):2205–2212. doi:10.1080/03610918.2012.748913.

[22] Iglewicz, B., Hoaglin, D.. Volume 16: How to Detect and Handle Outliers. In: The ASQC Basic References in Quality Control: Statistical Techniques; vol. 16. ISBN 9780873892476; 1993,.

[23] R Development Core Team, . R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria; 2013. URL: http://www.r-project.org.

[24] Beleites, C., Sergo, V.. hyperSpec: a package to handle hyperspectral data sets in R; 2015. URL: http://hyperspec.r-forge.r-project.org.

[25] Gunasekaran, S., Sankari, G., Ponnusamy, S.. Vibrational spectral investigation on xanthine and its derivatives - Theophylline, caffeine and theobromine. Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy 2005;61(1-2):117–127. doi:10.1016/j.saa.2004.03.030.