*Quim. Nova,* Vol. 35, No. 3, 612-615, 2012

Nota Técnica

# DESENVOLVIMENTO DE UM ALGORITMO PARA IDENTIFICAÇÃO E CORREÇÃO DE *SPIKES* EM ESPECTROSCOPIA RAMAN DE IMAGEM

**Guilherme Post Sabin, André Marcelo de Souza, Márcia Cristina Breitkreitz e Ronei Jesus Poppi***
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13084-971 Campinas – SP, Brasil

Recebido em 25/5/11; aceito em 9/9/11; publicado na web em 8/11/11

DEVELOPMENT OF AN ALGORITHM FOR IDENTIFICATION AND CORRECTION OF SPIKES IN RAMAN IMAGING SPECTROSCOPY. Raman imaging spectroscopy is a highly useful analytical tool that provides spatial and spectral information on a sample. However, CCD detectors used in dispersive instruments present the drawback of being sensitive to cosmic rays, giving rise to spikes in Raman spectra. Spikes influence variance structures and must be removed prior to the use of multivariate techniques. A new algorithm for correction of spikes in Raman imaging was developed using an approach based on comparison of nearest neighbor pixels. The algorithm showed characteristics including simplicity, rapidity, selectivity and high quality in spike removal from hyperspectral images.

Keywords: Raman imaging; spikes filter; algorithm development.

## INTRODUÇÃO

Espectroscopia Raman de Imagem (ERI) combina espectroscopia Raman com tecnologia de imagem digital para fornecer, simultaneamente, informação espacial e espectral sobre uma amostra e, por isso, oferece novas possibilidades para a caracterização de materiais biológicos, farmacêuticos, entre outros. Para a geração da imagem, seleciona-se uma área da amostra, com dimensões espaciais x e y, a qual é dividida em pixels. São então obtidos espectros a cada pixel, gerando um cubo de dados (cubo hiperespectral). Devido ao elevado volume de dados gerados e ao fato de que as respostas espectrais se encontram correlacionadas, métodos de análise multivariada de tratamento de imagens hiperespectrais são comumente aplicados para a identificação dos componentes da amostra e para a efetiva extração de informação sobre a distribuição espacial destes componentes.[1,2]

Detectores de dispositivo de acoplamento de carga, *Charge Coupled Device* (CCD), têm sido utilizados em equipamentos de espectroscopia Raman dispersivos devido à alta sensibilidade e baixo ruído, entretanto, um problema intrínseco ao uso dos detectores de CCD é sua vulnerabilidade a raios cósmicos.[3,4] Raios cósmicos produzem interferências nos espectros, chamadas de *spikes*, que são picos estreitos de intensidade variada. Estes interferentes resultam de raios gama (altamente energéticos) que são capazes de penetrar em diversos materiais e atingir os detectores de CCD. Os sinais resultantes dos *spikes* derivam do excesso de elétrons gerados em um único pixel ou em vários pixels adjacentes no detector sendo caracterizados por:[3] bandas tipicamente estreitas; unidirecionais, sempre positivas em relação à linha de base; padrão aleatório de distribuição temporal e, consequentemente, imprevisibilidade de ocorrência no espaço amostral.[2,5] A identificação e a eliminação de *spikes* presentes nos espectros são de fundamental importância para as aplicações de ERI, porque podem mascarar ou deformar as bandas de interesse químico e limitar o desempenho dos métodos de análise de dados, seja univariada ou multivariada.[2,3,6,7]

Embora em alguns casos a aquisição de imagens pelo monitoramento seletivo de um determinado comprimento de onda seja possível, soluções mais robustas apontam para a utilização de uma abordagem multivariada.[8] Os métodos multivariados mais empregados em análise de imagens hiperespectrais são: análise por componentes principais, *Principal Component Analysis* (PCA),[4] métodos de resolução de curvas multivariadas por mínimos quadrados alternados, *Multivariate Curve Resolution - Alternating Least Squares* (MCR-ALS),[9] mínimos quadrados clássicos, *Classical Least Squares* (CLS)[10] e mínimos quadrados parciais, *Partial Least Squares* (PLS).[11] Na PCA, a presença de *spikes* nos espectros pode causar uma distorção na direção dos eixos das componentes principais em função do aumento da variância em um dado comprimento de onda.[8] Este problema é transferido para o mapa de distribuição dos escores, que mostrará imagens distorcidas. Da mesma forma, os *spikes* podem causar a introdução de falsas variáveis puras em MCR ou, ainda, problemas ligados ao processo de otimização por ALS, que poderão modelar, inadequadamente, o sinal anômalo. No caso do CLS, a calibração e a previsão também serão distorcidas pela presença de *spikes*, já que neste método de calibração direta não existe seleção de informações para correlação com a variável dependente e, consequentemente, não admite calibrações na presença de interferentes, ou seja, a previsão será prejudicada, pois os espectros das amostras serão diferentes da média ponderada dos espectros dos constituintes puros. Embora o PLS possa fornecer calibração na presença de interferentes devido à utilização de informações espectrais que contenham maior correlação com a variável dependente, os *spikes* não podem ser considerados como interferentes previsíveis, uma vez que são ocorrências aleatórias.

Nosso grupo de pesquisa tem diversos projetos em andamento em que ERI e o tratamento multivariado das imagens têm sido amplamente utilizados em estudos de pré-formulação e na investigação de formulações farmacêuticas sólidas (*pellets* e comprimidos) e semissólidas (cremes e pomadas). O surgimento ostensivo de *spikes* tem sido um desafio constante para o sucesso dos projetos, uma vez que comprometem todas as etapas subsequentes de tratamento quimiométrico dos dados. Sendo assim, o desenvolvimento de algoritmos para a remoção de *spikes* tem sido um dos focos principais das pesquisas desenvolvidas.

As abordagens para correção de *spikes* podem variar de acordo com o sentido de observação do conjunto de dados.[2] A análise pode ser realizada em um único espectro percorrendo as variáveis, ou seja, pode-se criar uma janela contendo um determinado número de variáveis, onde é realizada uma análise da intensidade do sinal em um comprimento de onda em função das variáveis adjacentes.[5] Normal-

*e-mail: ronei@iqm.unicamp.br

mente é utilizado um determinado nível de confiança da média das variáveis próximas para detectar um *spike* e um filtro de mediana ou algum tipo de suavização para corrigi-lo. Embora apresente vantagens, este enfoque pode ser pouco sensível à detecção de *spikes*, pois, cada vez que a janela se aproxima de uma banda espectral, o limite de confiança da média é aumentado. Além disto, a deformação dos sinais e redução do número de variáveis no espectro são desvantagens conhecidas destes métodos.

Outra abordagem avalia a presença de *spikes* em uma mesma variável, mas em diferentes espectros, ou seja, a avaliação é feita percorrendo objetos. Assim, é possível minimizar o desvio padrão de um conjunto de espectros, desde que apresentem características espectrais muito semelhantes. A correção do valor anômalo neste caso é feita através da realização de uma replicata da amostra. No entanto, este procedimento pode ser muito demorado ou até mesmo inviável como, por exemplo, em imagens formadas por milhares de espectros. Nestas situações, uma aproximação de uma replicata pode ser feita empregando-se os pixels mais próximos, pois as variações espectrais são bastante sutis quando comparadas em regiões muito próximas na superfície de uma amostra. Assim, esta é uma abordagem rápida e eficiente para correção de *spikes* e pode ser explorada em dados de imagens.

Diversos algoritmos têm sido propostos na literatura para eliminar ou minimizar os efeitos causados pelo surgimento de *spikes* em ERI.[2,4,6] Zhang e Henson[2] mencionam as vantagens e desvantagens de várias abordagens disponíveis na literatura para confrontar o problema de contaminação de espectros Raman por *spikes*, e apresentam um método aplicado à determinação de princípio ativo em baixa dosagem em comprimidos. O método de eliminação de *spikes* utilizado foi baseado na comparação de um pixel com os vizinhos mais próximos proposto anteriormente por Beherend.[6] A base deste método consiste em utilizar a similaridade entre nove espectros distribuídos em um arranjo 3x3 para identificar *spikes* em um pixel central. Este método baseia-se no fato de que a probabilidade de um *spike* proveniente de raios cósmicos aparecer em um mesmo comprimento de onda em um pixel espacialmente adjacente na superfície da amostra é muito baixa, tornando viável a substituição do valor da variável no pixel contaminado por vizinhos mais próximos.

O objetivo deste trabalho foi desenvolver um algoritmo de aplicação simples e rápida para a detecção e eliminação de *spikes* em ERI, a ser aplicado diretamente no cubo de dados. O algoritmo baseia-se na comparação da intensidade do sinal de um determinado pixel, em um dado comprimento de onda, com seus vizinhos mais próximos. O algoritmo foi desenvolvido para obter simultaneamente alta sensibilidade para correção de *spikes* (dados aleatórios) e baixa sensibilidade para correção de bandas espectrais (dados sistemáticos).

## PARTE EXPERIMENTAL

### Instrumentação e estrutura de dados

Os espectros Raman foram coletados em um equipamento Raman Station 400 F da PerkinElmer, equipado com laser de excitação de 785 nm, detector CCD (com área ativa de 1024 x 255 pixels de 26 x 26 µm), resolução de 16 bits, temperatura de operação de -50 °C.

Os resultados mostrados neste artigo são referentes a dois conjuntos de dados de imagens hiperespectrais: (1) semente de soja (resolução espectral de 8 $cm^{-1}$, resolução espacial de 50 µm, 3 exposições de 5 s cada, faixa espectral de 1000 a 3200 $cm^{-1}$); (2) comprimido de carbamazepina (resolução espectral de 4 $cm^{-1}$, resolução espacial de 50 µm, 3 exposições de 3 s cada, faixa espectral de 200 a 3200 $cm^{-1}$). Estas amostras foram selecionadas uma vez que representam desafios ao algoritmo, como similaridade dos sinais analíticos com os *spikes*, deslocamento de linha base e intensidade de fundo. O algoritmo foi escrito usando Matlab 7.0® (Mathworks, Natick, MA) e processado em computador *desktop* equipado com processador Inter Core 2 Duo E7400, 2.80 GHz; memória RAM de 3,24 GB e sistema operacional de 32 bits.

## RESULTADOS E DISCUSSÃO

### Princípio do algoritmo

O algoritmo proposto está baseado na probabilidade de um sinal ser ou não um *spike*, assim como na homogeneidade da composição química entre regiões muito próximas (pixels vizinhos) na amostra. A localização espacial de um pixel, em relação à sua vizinhança, pode ser descrita para três situações distintas: o pixel pode estar localizado em uma região central do cubo apresentando oito vizinhos mais próximos; nas bordas do cubo apresentando cinco vizinhos mais próximos; ou nos cantos do cubo apresentando três vizinhos mais próximos. Estas situações são representadas na Figura 1.

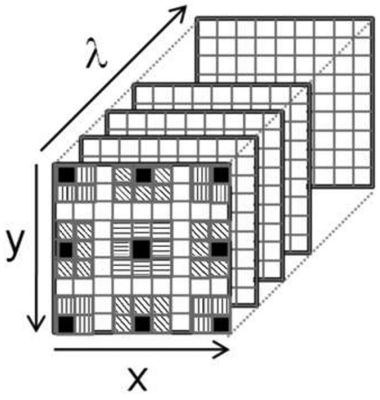


***Figura 1.*** *Localização espacial de um pixel em um cubo de dados: o pixel pode estar localizado em uma região central, nas bordas ou nos cantos do cubo, apresentando oito, cinco ou três vizinhos mais próximos, respectivamente*

O algoritmo foi desenvolvido para seguir uma sequência lógica de avaliações. As três etapas mostradas na Figura 1 são realizadas para cada um dos comprimentos de onda do cubo hiperespectral (fatias do cubo no sentido dos comprimentos de onda). Para cada fatia, a varredura é iniciada pelas colunas de pixels e, em seguida, indo para as linhas. Quando todas as colunas e linhas de uma determinada fatia já foram corrigidas, o algoritmo passa para a análise de outro comprimento de onda. A análise começa pela situação na qual o pixel está rodeado por oito vizinhos, seguindo para a situação na qual está rodeado por cinco e, por último, por três vizinhos mais próximos. A razão desta escolha é que ao chegar à segunda etapa, a maior parte dos *spikes* já foram corrigidos pela primeira etapa e, na última etapa, todos os pixels vizinhos já foram corrigidos pelas duas primeiras.

O procedimento utilizado para identificar e remover um *spike* do pixel central é descrito abaixo:

(1) o pixel central é identificado como sendo s*pike* em função do limite de confiança dos valores de intensidade de espalhamento estabelecido a partir dos pixels vizinhos de acordo com a Equação 1:

$$I_{p_{central}} > \bar{I}_{p_{vizinhos}} + ks_{P_{vizinhos}} \quad (1)$$

onde $I_{P_{central}}$ é a intensidade de espalhamento do pixel central com localização espacial no plano xy da imagem em um dado comprimento de onda λ, $\bar{I}_{P_{vizinhos}}$ é o valor médio de intensidade, descartando-se o

valor extremo para evitar que *spikes* sejam incluídos no cálculo da média; *k* é o fator de abrangência e $s_{P_{vizinhos}}$ é a estimativa de desvio padrão dos pixels vizinhos;

(2) quando o pixel central apresenta um valor acima do limite de confiança, então ele é substituído pela média dos valores de intensidade de espalhamento dos pixels vizinhos, descartados os valores extremos;
(3) na primeira etapa, o pixel central é substituído pela média dos oito vizinhos mais próximos, descartados os valores extremos;
(4) na segunda etapa, o pixel central é substituído pela média dos cinco vizinhos, descartados os valores extremos;
(5) na terceira etapa, o pixel central é substituído pela média dos três vizinhos mais próximos. Neste caso não houve necessidade de descartar valores extremos, pois os vizinhos foram corrigidos nas etapas anteriores;
(6) após a correção de todos os *spikes* presentes, é gerado um cubo de dados binário (onde, posição corrigida = 1 e posição inalterada = 0);
(7) finalmente, o cubo binário é utilizado para retornar aos pontos de correção e realizar melhorias sutis nas variáveis vizinhas no eixo espectral.

Um desafio a ser enfrentado por algoritmos de detecção e remoção de *spikes* é não confundi-los com sinais de linha de base ou variações sistemáticas no espectro Raman e não provocar distorções no perfil das bandas, pois desta forma os espectros perderiam parte de sua informação relevante. O valor do parâmetro k está diretamente relacionado com a detecção de *spikes* nos espectros e por este motivo foi estudado com mais detalhes. A remoção de *spikes,* nas imagens hiperespectrais das amostras, foi avaliada através da variação dos valores de k e da observação dos resultados por meio de gráficos de diagnóstico, os quais serão discutidos adiante. O valor de k a ser utilizado deve ser aquele que retire os *spikes*, sem retirar informações espectrais. Foi observado que o valor ótimo de k variou de acordo com a natureza da amostra, como grau de homogeneidade, intensidade de fundo e formato de bandas. Por outro lado, o valor de k também é influenciado pelas características dos *spikes,* como suas intensidades. Valores altos de k implicam intervalos de confiança grandes, o que permite apenas a retirada de *spikes* de alta intensidade. Um valor de k pequeno implica intervalo de confiança mais estreito, o que permite a retirada de *spikes* de menor intensidade. Entretanto, neste último caso, sinais espectrais podem também ser corrigidos desnecessariamente. Em relação à natureza da amostra, quanto mais heterogênea for a amostra (maior desvio padrão entre os pixels vizinhos), menor deve ser o valor de k para garantir a retirada de *spikes* de menor intensidade.

Pelos motivos expostos acima, o parâmetro k deve ser otimizado pelo usuário de acordo com a sua amostra. Recomenda-se ao usuário que observe três gráficos de diagnóstico: espectros originais, espectros corrigidos e o gráfico da diferença dos dois. O resultado ótimo é alcançado quando a retirada dos *spikes* é completa, sem remoção de informações espectrais. O gráfico da diferença deve mostrar um padrão aleatório sem sinais espectrais. Vale ressaltar que alguns parâmetros instrumentais, como tempo de exposição e número de exposições, também levam a variações de intensidade e número de *spikes* registrados. Quando o tempo de exposição do laser aumenta, o número de *spikes* registrados tende a aumentar. Por outro lado, quando são realizadas muitas exposições por amostra, a intensidade de *spikes* registrados deve diminuir, já que um espectro médio é registrado. Estes parâmetros também podem ser ajustados pelo usuário com o intuito de adequar a intensidade de *spikes* registrados nos espectros.

Para ilustrar os aspectos discutidos acima, calculou-se a diferença entre os espectros originais (contendo *spikes*) e os espectros corrigidos pelo algoritmo, tomando-se os dados do comprimido de carbamazepina como exemplo, com diferentes valores do parâmetro k. Os resultados são mostrados na Figura 2. Na Figura 2A é possível observar que para valores de k = 3, informações espectrais foram corrigidas desnecessariamente. Na Figura 2B observa-se que, quando o valor de k foi aumentado para 5, apenas os *spikes* foram retirados.

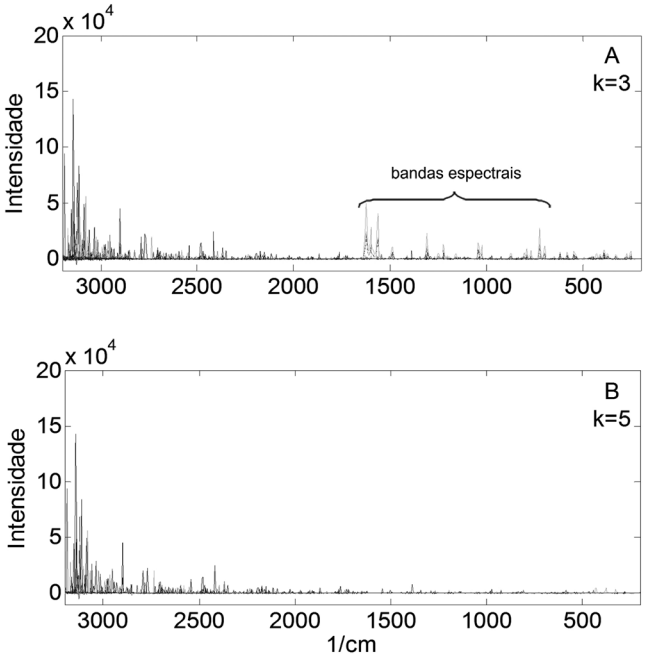


***Figura 2.*** *Espectros Raman da diferença entre os espectros originais da amostra do comprimido de carbamazepina e aqueles corrigidos pelo algoritmo empregando: A) k = 3 e B) k = 5*

A Figura 3 apresenta uma comparação dos espectros da amostra de soja e do comprimido de carbamazepina antes e depois da aplicação do algoritmo, utilizando-se valores de k = 3 e 5, respectivamente. Para a amostra de soja, a utilização de k = 3 foi suficiente para remover os *spikes* sem remoção de picos espectrais. Através da análise desta figura, é possível verificar a qualidade da correção dos *spikes* realizada pelo algoritmo, bem como a sua seletividade na presença de variações de linha base e elevada intensidade de fundo, uma vez que estas informações são preservadas com a aplicação do algoritmo.

Outra maneira de avaliar a qualidade dos dados obtidos após a correção de *spikes* é realizar uma análise de componentes principais (PCA) e avaliar o número necessário de componentes principais para explicar a variância do conjunto de dados. A aplicação da PCA aos conjuntos de dados estudados mostrou que um número maior de componentes principais é requerido para explicar a variância do conjunto de dados contendo *spikes*. Esta observação é explicada pelo fato dos *spikes* causarem distorções na direção dos eixos das componentes principais alterando, portanto, a porcentagem de variância descrita por cada uma delas, uma vez que induzem fontes de variação adicionais no cálculo multivariado. Além disto, uma determinada componente principal pode estar presente exclusivamente para descrever informações relativas a um *spike*, uma vez que o mesmo representa uma fonte de variação muito grande.

Deve ser ressaltado que o algoritmo se baseia na similaridade espectral do pixel central com seus vizinhos mais próximos. Desta maneira, situações nas quais o pixel central apresenta características muito diferentes dos seus vizinhos como, por exemplo, deslocamento de linha base apenas no espectro que contém o *spike*, a correção não poderá ser realizada de forma satisfatória. Neste caso, sugere-se a alteração dos parâmetros instrumentais de tempo de exposição e número de exposições para otimizar a ocorrência/intensidade dos

*spikes*. Além disto, se o sinal analítico estiver presente em apenas um pixel, como é o caso de estudos de *single-molecule* SERS, o mesmo poderia ser confundido com um s*pike* pelo algoritmo. Nestes casos, a resolução espacial pode ser aumentada para que a amostragem de pixels vizinhos se torne mais homogênea ou o número de exposições pode ser diminuído para que as ocorrências de *spikes* sejam mais intensas e possam ser diferenciadas com valores de k maiores, entre outras ações baseadas no conhecimento químico do sistema em estudo. Entretanto, considerando as aplicações usuais do algoritmo proposto, estes são considerados casos raros.

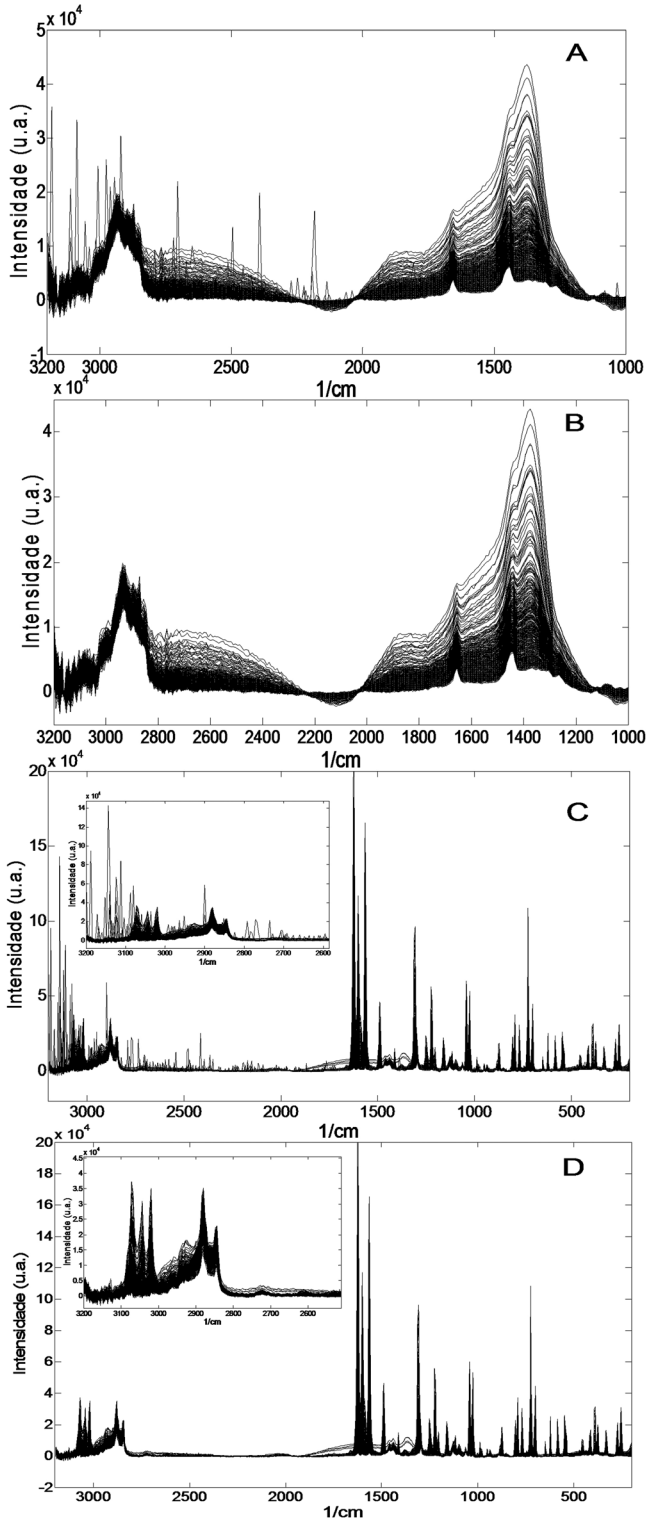


***Figura 3.*** *Exemplos de espectros antes e depois da aplicação do algoritmo: A) espectros da amostra de soja antes da remoção dos spikes; B) espectros da amostra de soja após utilização do algoritmo (k=3); C) espectros do comprimido de carbamazepina antes da remoção dos spikes; D) espectros do comprimido de carbamazepina após utilização do algoritmo (k=5)*

## CONCLUSÃO

A presença de *spikes* inviabiliza o tratamento de dados de imagens hiperespectrais e, por este motivo, deve ser corrigida. Neste trabalho, o algoritmo proposto não requer etapas de pré-processamento como o desdobramento do cubo de dados original ou correção de linha base, o que facilita a sua utilização. Além disso, apresenta vantagens sobre outros algoritmos que são aplicados sobre a matriz desdobrada, como os filtros de mediana, os quais podem deformar as bandas espectrais e reduzem as dimensões originais dos dados. Outro aspecto importante é a correção de espectros de forma seletiva, preservando flutuações de linha base e intensidade de fundo. O fator de abrangência é o único parâmetro ajustável pelo usuário, o qual deverá selecionar o valor a ser utilizado de acordo com os gráficos de diagnósticos. Um bom ponto de partida é considerar um valor de k em torno de 5, podendo-se otimizá-lo de acordo com a natureza da amostra e as características dos *spikes* presente nos espectros. Não é recomendado valores de k abaixo de 3, neste caso a otimização dos parâmetros instrumentais pode ser uma alternativa para a retirada de *spikes* de forma seletiva.

Embora a qualidade das correções seja fundamental, outras características desejáveis também podem ser destacadas neste algoritmo como, por exemplo, a rapidez de execução (aproximadamente 1 min por amostra) e a simplicidade de operação (apenas duas variáveis de entrada: o cubo de dados e o fator de abrangência, k).

## AGRADECIMENTOS

Ao apoio financeiro da CAPES, CNPq e FAPESP.

## REFERÊNCIAS

1. Treado, P. J.; Matthew, P. N. Em *Raman Spectroscopy: From the Research Laboratory to the Process Line*; Lewis, I. R.; Edwards H. G. M., eds.; Marcel Decker: Nova York, 2001, cap. 5.
2. Zhang, L.; Henson, M.; *J. Appl. Spectrosc*. **2007**, *61*, 1015.
3. Zhang, D.; Jallad, K. N.; Ben-Amotz, D.; *Appl. Spectrosc*. **2001**, *55*, 1523.
4. Zhang, D.; Ben-Amotz, D.; *Appl. Spectrosc*. **2002**, *56*, 91.
5. Katsumoto, Y.; Ozaki, Y.; *Appl. Spectrosc*. **2003**, *57*, 317.
6. Behrend, C. J.; Tarnowski, C. P.; Morris, M. D.; *Appl. Spectrosc*. **2002**, *56*, 1458.
7. Ehrentreich, F.; Summchen, L.; *Anal. Chem.* **2001**, *73*, 4364.
8. Gordon, K. C.; McGoverin, C. M.; *Int. J. Pharm*. **2011**, doi:10.1016/j.ijpharm.2010.12.030.
9. Juan, A.; Tauler, R.; Dyson, R.; Marcolli, C.; Rault, M.; Maeder, M.; *Trend. Anal. Chem*. **2004**, *23*, 70.
10. Widjaja, E.; Kanaujia, P.; Lau, G.; Ng, W. K.; Garland, M.; Saal, C.; Hanefeld, A.; Fischbach, M.; Maio, M.; Tan, R. B. H.; *Eur. J. Pharm. Sci*. **2011**, *42*, 45.
11. Henson, M. J.; Zhang, L.; *Appl. Spectrosc*. **2006**, *11*, 1247.