Combinação de Modelos Obtidos por Regressão no Domínio Wavelet

Luiz Alberto Pinto
Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica
12228-900, São José dos Campos, SP, Brasil
E-mail: luizpt@ita.br

Roberto Kawakami Harrop Galvão
Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica
12228-900, São José dos Campos, SP, Brasil
E-mail: kawakami@ita.br

***Resumo**: A Transformada Wavelet é uma ferramenta útil para pré-processamento e compressão de dados em calibração multivariada. Contudo, a sua aplicação requer a escolha da wavelet e do número de níveis de decomposição a serem empregados, o que pode não ser uma tarefa simples. Este artigo propõe uma abordagem alternativa, que consiste em combinar modelos calibrados a partir de diferentes decomposições wavelet do conjunto de dados. Como ilustração, apresenta-se um estudo de caso envolvendo a determinação da densidade de gasolina por espectroscopia no infravermelho. Os resultados indicam que a abordagem proposta é uma alternativa vantajosa ao uso de uma única decomposição wavelet para construção do modelo.*

Exposição do Problema

A Transformada Wavelet (TW) tem sido utilizada, em diversos ramos da ciência, para aplicações tais como processamento de imagens [3], classificação de padrões [18], análise e compressão de sinais [9]. Na área de Quimiometria [1,29], em particular, a TW tem sido aplicada em problemas de filtragem [13], compressão de sinais [28], correção de linhas de base [22], classificação [11] e calibração multivariada [10,20], entre outros.

O problema de calibração multivariada em Quimiometria consiste em construir um modelo matemático que permita a determinação de propriedades físicas e/ou químicas de uma amostra a partir de medidas espectrais (tipicamente de intensidade de emissão, absorção ou reflexão) registradas em diferentes comprimentos de onda [23]. Aplicações incluem, por exemplo, a determinação de metais em ligas de aço [21], a análise da composição de fórmulas farmacêuticas [7] e a predição de propriedades de combustíveis [5].

No contexto de calibração multivariada a TW tem sido utilizada para compressão de conjuntos de dados para construção de modelos por Análise de Componentes Principais [27] ou Mínimos-Quadrados Parciais [26]. Como alternativa, algoritmos de seleção de variáveis podem ser empregados na seleção de um subconjunto apropriado de coeficientes wavelet para calibração de modelos por Regressão Linear Múltipla [6,8].

Diferentemente da Transformada de Fourier, que envolve apenas funções de base do tipo seno e cosseno, a TW requer algumas escolhas por parte do usuário. Caso a decomposição wavelet seja efetuada com o uso de um banco de filtros [1,25], tais escolhas tipicamente se referem à família (Daubechies, Coiflet ou Symlet, por exemplo [25]) e comprimento dos filtros, bem como ao número de níveis de decomposição. Embora inúmeros trabalhos tenham sido publicados sobre a aplicação da TW em calibração multivariada, tais escolhas ainda constituem um problema em aberto. Em geral, os trabalhos se referem a resultados prévios obtidos para conjuntos de dados similares [10,26]. Alternativamente, alguns autores optam por testar diferentes combinações de filtros wavelet e níveis de decomposição, escolhendo a mais apropriada com base em métricas de desempenho para os modelos obtidos [2,20].

Neste contexto, o presente artigo propõe uma nova abordagem, que consiste em combinar modelos obtidos a partir de diferentes decomposições wavelet. Como ilustração, apresenta-se um estudo de caso envolvendo a determinação da densidade de gasolina por espectroscopia no infravermelho. Os resultados da abordagem proposta são comparados com os obtidos ao se empregar os modelos individuais usados na combinação. Para tal, consideram-se métricas associadas ao erro de previsão e à sensibilidade a ruído.

Formulação Matemática da Transformada Wavelet

A transformada wavelet no caso contínuo, TWC [19,25], pode ser definida por

$$TWC(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right)dt \text{ , com } a \neq 0, \qquad (1)$$

onde $1/\sqrt{|a|}$ é um termo de normalização de energia [25], $x(t)$ é o sinal sob análise, e a função ψ(t) é denominada "wavelet-mãe". A função $\psi\left(\frac{t-b}{a}\right)$ corresponde à wavelet-mãe dilatada/contraída pela escala a, e transladada pelo fator b [19,25]. A magnitude de TWC(a, b) está associada ao conteúdo espectral do sinal $x(t)$ em torno de $t = b$. Pequenos valores de escala a correspondem a altas freqüências, enquanto, grandes escalas correspondem a baixas freqüências.

A versão discreta da transformada wavelet [19,25] pode ser obtida a partir da TWC pela discretização dos parâmetros a e b, na forma $a = a_0^{\,j}$ e $b = k\, b_0\, a_0^{\,j}$, sendo j, k valores inteiros e $a_0 > 1$, $b_0 > 0$. Fazendo $a_0 = 2$ e $b_0 = 1$, obtém-se uma discretização diádica [25]. Neste caso, pode-se obter a transformada wavelet discreta (TWD) de forma computacionalmente eficiente através do uso de de bancos de filtros digitais, como será visto na próxima seção.

Implementação da Transformada Wavelet empregando Banco de Filtros

A TWD pode ser implementada de forma eficiente através de um algoritmo de banco de filtros proposto por Mallat [17, 25]. A Figura 1 ilustra um banco de filtros para cálculo da TWD com dois níveis de decomposição.

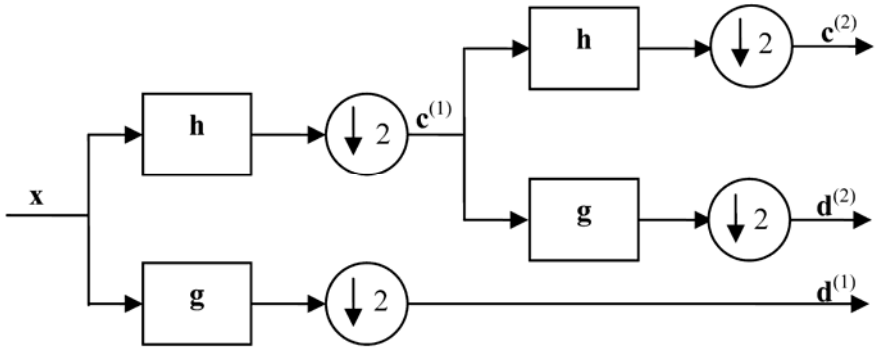

Figura 1: Banco de filtros para implementação da TWD em dois níveis de resolução.

A estrutura básica do banco de filtros consiste em um par de filtros passa-baixas (**h**) e passa-altas (**g**), seguido por uma operação de subamostragem diádica. As saídas subamostradas dos filtros passa-baixas e passa-altas denominam-se, respectivamente, coeficientes de aproximação e detalhe, representados na Figura 1 por $\mathbf{c}^{(1)}$, $\mathbf{c}^{(2)}$ e $\mathbf{d}^{(1)}$, $\mathbf{d}^{(2)}$, respectivamente. As operações de filtragem/subamostragem podem ser reaplicadas aos coeficientes de aproximação até o número L de níveis de decomposição especificado pelo analista. O resultado final da decomposição

consiste nos coeficientes de aproximação no último nível L, em acréscimo aos coeficientes de detalhe nos níveis 1 a L.

Como os filtros **h** e **g** são tipicamente de comprimento finito, cada coeficiente de aproximação ou detalhe corresponde a um segmento do sinal original. Esta característica de localização espacial é uma das principais vantagens da Transformada Wavelet sobre a Transformada de Fourier.

Método Proposto

O método proposto baseia-se na abordagem de combinação de modelos. Dentre as vantagens da aplicação de estratégias para combinar modelos, citam-se [30]: *(i)* Encontrar o melhor modelo para um problema específico não é uma tarefa simples porque, geralmente, os fenômenos envolvidos não são totalmente conhecidos ou são de modelagem complexa. *(ii)* Modelos combinados muitas vezes produzem melhores resultados que os modelos individuais. *(iii)* A estratégia de combinar modelos pode reduzir o consumo de tempo e recursos computacionais envolvidos na modelagem. Com base nessas premissas, no que segue apresenta-se um método para combinação linear de modelos calibrados no domínio wavelet.

O método proposto envolve duas etapas. Na primeira etapa, de Geração de Modelos, modelos individuais são obtidos para cada configuração (filtros e níveis de decomposição) considerada para o banco de filtros wavelet. Para isso, emprega-se neste trabalho o algoritmo de regressão por passos (*stepwise regression*) [12]. O n-ésimo modelo assim obtido é da forma

$$\hat{y}^{(n)} = w_0 + \sum_{k=1}^{K} w_k t_k \tag{2}$$

em que t_k, $k = 1, ..., K$ correspondem aos coeficientes wavelet de aproximação e detalhe dos dados de entrada e $w_0, w_1, ..., w_K$ denotam os coeficientes de regressão.

Na segunda etapa, de Combinação, os modelos individuais são linearmente combinados com pesos iguais. A saída do modelo resultante é dada por (4) sendo N o número de configurações consideradas para o banco de filtros.

$$\hat{y} = \frac{1}{N} \sum_{n=1}^{N} \hat{y}^{(n)} \tag{3}$$

Neste trabalho, três estratégias de combinação de modelos são testadas. Na “estratégia 1”, são combinados modelos obtidos usando os mesmos filtros wavelet, mas com diferentes níveis de decomposição. Na “estratégia 2” são combinados modelos obtidos no mesmo nível de decomposição, mas com filtros wavelet diferentes. Na “estratégia 3”, são combinados modelos obtidos variando-se tanto os filtros wavelet quanto os níveis de decomposição.

Exemplo de Aplicação

O método proposto foi aplicado em um problema de determinação da densidade de amostras de gasolina por espectroscopia no infravermelho. Para isso, foi utilizado um conjunto de 104 amostras com 6443 variáveis descritoras, que correspondem a medidas de absorbância nos correspondentes comprimentos de onda. A Figura 2a apresenta os espectros desse conjunto. Como se pode observar, os espectros apresentam flutuações de linha de base, que foram parcialmente corrigidas com a aplicação de um filtro diferenciador de Savitzky-Golay [4] com polinômio de segunda ordem e janela de 11 pontos. Os espectros derivativos resultantes, que serão utilizados na fase de modelagem, estão apresentados na Figura 2b.

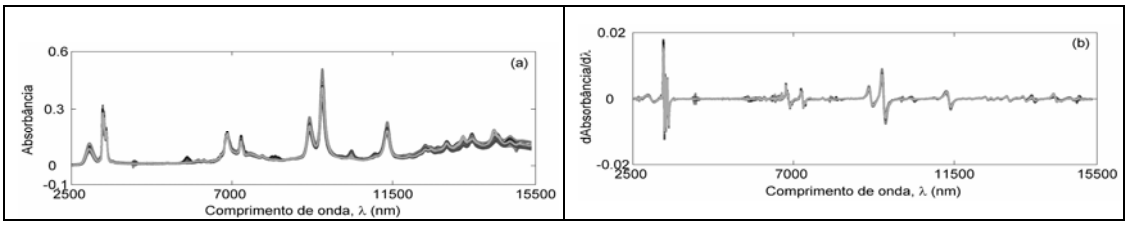

Figura 2: (a) Espectros originais. (b) Espectros derivativos.

O algoritmo de Kennard-Stone [15,16] foi aplicado para selecionar 70% das amostras, a serem empregadas para fins de modelagem. As 30% restantes foram reservadas como um conjunto de teste para avaliar e comparar o desempenho dos modelos obtidos. Para reduzir o custo computacional na fase de modelagem, aplicou-se aos coeficientes wavelet um procedimento de compressão preliminar que descartou os de menor valor absoluto, retendo 99% da variância explicada.

Os modelos foram calibrados aplicando o algoritmo de regressão por passos (*stepwise regression*), com valor α de 0.01 para inclusão e exclusão de variáveis [12]. Como os filtros wavelet mais comumente empregados em calibração multivariada são das famílias Daubechies (db), Symlet (sym) e Coiflet (coif) [2,10,20,26], modelos foram construídos utilizando os filtros db2 – db10, coif1 – coif5 e sym4 – sym8, totalizando 19 filtros diferentes. Para amenizar efeitos de borda, foi empregada uma extensão constante a partir do início e fim do sinal a ser decomposto [25]. Em cada caso, o número de níveis de decomposição foi variado de um até um valor máximo *L*. O nível máximo *L* foi tomado como sendo aquele para o qual pelo menos um coeficiente wavelet ainda se referia ao sinal original (sem extensão). Os modelos individuais foram combinados conforme as estratégias 1, 2 e 3 mencionadas na seção anterior. Todos os cálculos foram desenvolvidos usando o pacote de wavelets do Matlab® 6.5 R13.

Os modelos resultantes foram analisados empregando duas métricas, a saber: (i) valor RMSEP (*root-mean-square error of prediction*) obtido no conjunto de teste e (ii) norma 2 do vetor de coeficientes de regressão **b** no domínio original dos dados. O RMSEP é definido como

$$RMSEP = \sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}(\hat{y}_i - y_i)^2} \qquad (4)$$

em que $\hat{y}_i$, y_i correspondem aos valores predito e esperado para a densidade (em g/cm^3) da *i*-ésima amostra do conjunto de teste (i = 1, ..., N_t). Por sua vez, o vetor **b** pode ser obtido convertendo-se modelos da forma (1) para o domínio original do sinal **x**. Como discutido em [14, 24], $\|\mathbf{b}\|_2$ é um indicador da sensibilidade do modelo a ruído nos dados de entrada.

Resultados

A Figura 3 mostra os resultados obtidos empregando-se a "estratégia 1" (combinação de modelos obtidos usando os mesmos filtros wavelet, com diferentes níveis de decomposição), representados por círculos. Para comparação, apresentam-se também os resultados médios obtidos pelos modelos individuais, representados por quadrados. Vale salientar que as médias aqui consideradas são calculadas para cada filtro wavelet, tomando-se diferentes níveis de decomposição. Como se pode observar, a estratégia proposta permite obter melhores resultados, em termos das métricas adotadas. Adicionalmente, um resultado ainda melhor é proporcionado pela estratégia 3 (combinação de modelos obtidos variando-se tanto os filtros wavelet quanto os níveis de decomposição), conforme indicado pela estrela no gráfico.

O gráfico da Figura 4 mostra o resultado das combinações pela “estratégia 2” (combinação de modelos obtidos no mesmo nível de decomposição, mas com filtros wavelet diferentes). Aplicando os mesmos critérios da análise anterior, observa-se que o desempenho dos modelos resultantes é superior ao resultado médio obtido com os modelos individuais. Neste caso, o desempenho do modelo médio obtido com a “estratégia 3” é comparável ao melhor resultado geral.

Conclusões

Este trabalho propôs um método para combinar modelos obtidos por regressão no domínio wavelet, empregando diferentes configurações de bancos de filtros. Tal abordagem é uma alternativa à construção de modelos utilizando escolhas específicas de filtros wavelet e níveis de decomposição.

Para ilustração, considerou-se um estudo de caso envolvendo a determinação da densidade de gasolina por espectroscopia no infravermelho. Os resultados foram avaliados em termos da capacidade preditiva e sensibilidade a ruídos dos modelos obtidos. Ao final, constatou-se que o desempenho dos modelos gerados pela abordagem proposta foi superior ao resultado médio obtido com modelos individuais (isto é, modelos construídos empregando-se bancos de filtros específicos).

Pesquisas futuras poderiam contemplar critérios para ponderação dos modelos individuais na combinação. Para isso, poderiam ser utilizados pesos que refletissem o desempenho de cada modelo individual.

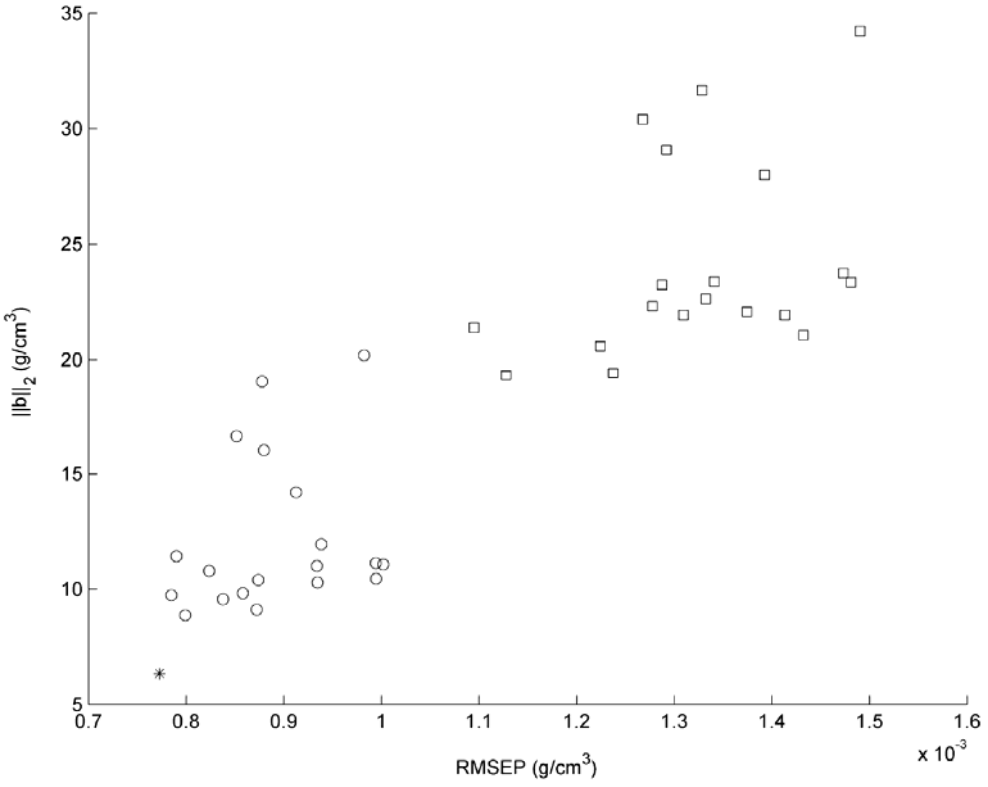

Figura 3: Resultados da combinação de modelos pela “estratégia 1” (O) e “estratégia 3” (*), em comparação com os resultados médios obtidos por modelos individuais (□).

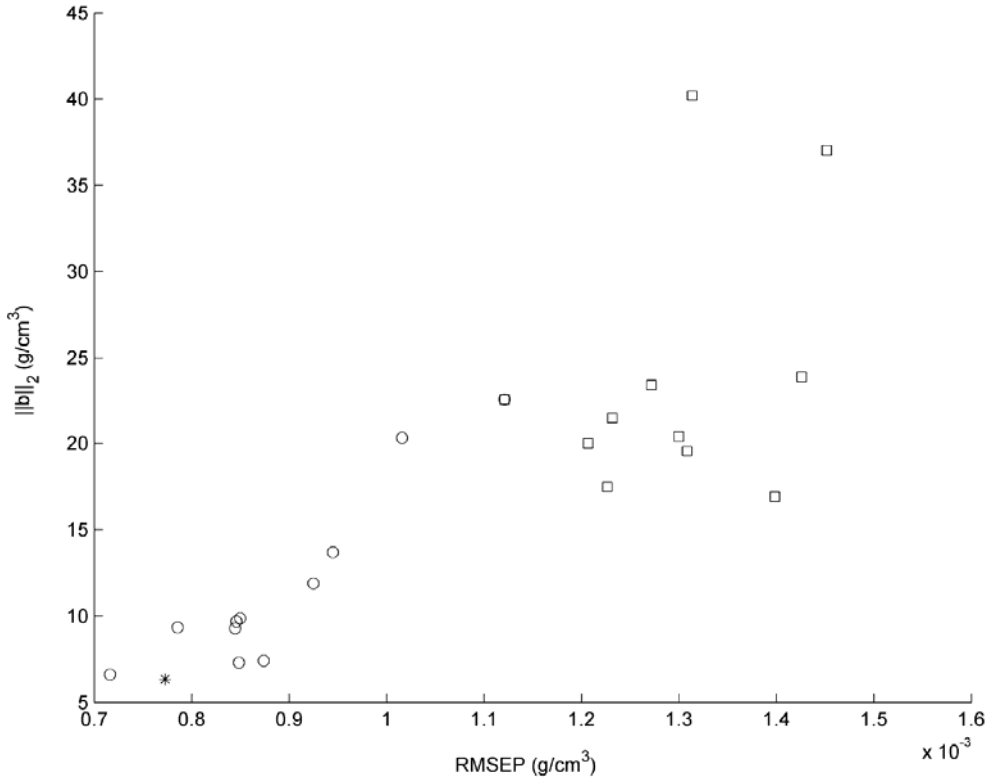

Figura 4: Resultados da combinação de modelos pela “estratégia 2” (O) e “estratégia 3” (*), em comparação com os resultados médios obtidos por modelos individuais (□).

Referências

1. B. K. Alsberg; A. M. Woodward; D. B. Kell, An introduction to wavelet transform for chemometricians: A time frequency approach, Chemom. Intell. Lab. Syst., vol. 37, pp. 215-339, (1997).
2. B. K. Alsberg; A. M. Woodward; M. K. Winson; J. J. Rowland; D. B. Kell, Variable selection in wavelet regression models, Anal. Chim. Acta, vol. 368, pp. 29-44, (1998).
3. M. Antonini; M. Barlaud; P. Mathieu; I. Daubechies, Image coding using wavelet transform, IEEE Transactions on Image Processing, vol. 1, pp. 205-230, (1992).
4. K. R. Beebe; R. J. Pell; B. Seasholtz, “Chemometrics – A Practical Guide”, Wiley, New York, 1998.
5. G. Bohacs; Z. Ovadi; A. Salco, Prediction of gasoline properties with near infrared spectroscopy, Journal of Near Infrared Spectroscopy, vol. 6, pp. 341-348, (1998).
6. P. J. Brown; T. Fearn; M. Vannucci, Bayesian wavelet regression on curves with application to a spectroscopic calibration problem, J. Am. Stat. Assoc., vol. 96, pp. 398-408, (2001).
7. P. Chalus; S. Walter; M. Ulmschneider, Combined wavelet transform-artificial neural network use in tablet active content determination by near-infrared spectroscopy, Anal. Chim. Acta, vol. 591, pp. 219-224, (2007).
8. C. J. Coelho; R. K. H. Galvão; M. C. U. Araújo; M. F. Pimentel; E. C. Silva, A solution to the wavelet transform optimization problem in multicomponent analysis, Chemom. Intell. Lab. Syst., vol. 66, pp. 205-217, (2003).
9. I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Transactions on Information Theory, vol. 36, pp. 961-1005, (1990).
10. I. E. Díez, J. M. Saiz; C. Pizarro, OWAVEC: a combination of wavelet analysis and an orthogonalization algorithm as a pre-processing step in multivariate calibration, Anal. Chim. Acta, vol. 515, pp. 31- 41, (2004).

11. D. Donald; D. Coomans; Y. Everingham; D. Cozzolino; M. Gishen; T. Hancock, Adaptive wavelet modeling of a nested 3 factor experimental design in NIR chemometrics, Chemom. Intell. Lab. Syst., vol. 82, pp. 122-129, (2006).
12. N. R. Draper; H. Smith, "Applied Regression Analysis", Wiley, New York, 1998.
13. R. K. H. Galvão; H. A. D. Filho; M. N. Martins; M. C. U. Araújo; C. Pasquini, Sub-optimal wavelet denoising of coaveraged spectra employing statistics from individual scans, Anal. Chim. Acta, vol. 581, pp.159-167, (2007).
14. J. H. Kalivas, Pareto calibration with built-in wavelength selection, Anal. Chim. Acta, vol. 505, pp. 9-14, (2004).
15. K.R. Kanduc; J. Zupan; N. Majcen, Separation of data on the training and test set for modeling: a case study for modeling of five colours properties of a white pigment, Chemom. Intell. Lab. Syst., vol.65, pp. 221-229, (2003).
16. R. W. Kennard; L. A. Stone, Computer aided design of experiment, Technometrics, vol. 11, pp.137-148, (1969).
17. S. G. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 11, pp. 674-693, (1989).
18. Y. Mallet; D. Coomans; J. Kautsky; O. De Vel, Classification using adaptive wavelets for feature extraction, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 1058-1066, (1997).
19. Matlab 6.5 User's Guide, The Mathworks; Natick, MA, USA.
20. B. M. Nicolai; K. I. Theron; J. Lammertyn, Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple, Chemom. Intell. Lab. Syst., vol. 85, pp. 243–252, (2007).
21. M. F. Pimentel; B. B. Neto; M. C. U. Araújo; C. Pasquini, Simultaneous multielemental determination using a low-resolution inductively coupled plasma spectrometer/diode array detection system, Spectrochimica Acta B., vol. 52, pp. 2151-2161, (1997).
22. X. Shao; W. Cai; Z. Pan, Wavelet transform and its application in high performance liquid chromatography (HPLC) analysis, Chemom. Intell. Lab. Syst., vol.45, pp. 249-256, (1999).
23. D. A. Skoog; F. J. Holler; T. A. Nieman, "Princípios de Análise Instrumental", Bookman, Porto Alegre, 2002.
24. F. Stout; M. R. Baines; J. H. Kalivas, Impartial graphical comparison of multivariate calibration methods and the harmony/parsimony tradeoff, J. Chemometrics, vol. 20, pp. 464-475, (2006).
25. G. Strang; T. Nguyen, "Wavelet and Filter Banks", Cambridge Press", Wellesley, 1996.
26. J. Trygg; S. Wold, PLS regression on wavelet compressed NIR spectra, Chemom. Intell. Lab. Syst., vol. 42, pp. 209–220, (1998).
27. F. Vogt; M. Tacke, Fast principal component analysis of a large data set, Chemom. Intell. Lab. Syst., vol. 59, pp. 1–18, (2001).
28. B. Walczak; D. L. Massart, Wavelet packet transform applied to a set of signals: A new approach to the best-basis selection, Chemom. Intell. Lab. Syst., vol. 38, pp. 39-50, (1997).
29. B. Walczak, "Wavelet in Chemistry", Elsevier Science, New York, 2000.
30. S. R. Waterhouse, "Divide and Conquer: Pattern Recognition using Mixture of Experts", Doctorate's Thesis, University of Cambridge, Cambridge, England 1997.