Hindawi Publishing Corporation
Journal of Automated Methods and Management in Chemistry
Volume 2011, Article ID 323196, 6 pages
doi:10.1155/2011/323196

Research Article

Automatic Discrimination of the Geographical Origins of Milks by Excitation-Emission Fluorescence Spectrometry and Chemometrics

Lu Xu,[1] De-Hua Deng,[1] Chen-Bo Cai,[2] and Hong-Wei Yang[2]

[1] *College of Chemistry and Chemical Engineering, Anyang Normal University, Anyang 455002, Henan Province, China*
[2] *Department of Chemistry and Life Science, Chuxiong Normal University, Chuxiong 675000, China*

Correspondence should be addressed to De-Hua Deng, ddh@aynu.edu.cn

Received 20 May 2011; Accepted 13 June 2011

Academic Editor: Lu Yang

This paper presents the automatic discrimination of geographical origins of milks from Western Yunnan Plateau areas and eastern China by excitation-emission fluorescence spectrometry and chemometrics. Genuine plateau milks ($n = 60$) and milks from eastern China ($n = 89$) are scanned in the regions of 180–300 nm for excitation and 200–800 nm for emission. Different options of data analysis are investigated and compared in terms of their performance in discriminating milks of different geographical origins: (1) two-way partial least squares discriminant analysis (PLSDA) based on excitation and emission spectra, respectively; (2) two-way PLSDA based on fusion of excitation and emission spectra; (3) three-way PLSDA based on excitation-emission matrix spectra. The two-way PLSDA methods with excitation spectra, emission spectra, and fusion of excitation and emission spectra correctly classify 91.3%, 88.6%, and 95.3% of the milk samples, respectively; while the total accuracy of three-way PLSDA is 96.0%. The results demonstrate the two-way data combining excitation and emission spectra are sufficient to characterize and identify the plateau milks. Considering both model accuracy and the analytical time required, two-way PLS-DA with fusion of excitation and emission spectra is recommended as a reliable and quick method to discriminate plateau milks from ordinary milks.

1. Introduction

Recently China has witnessed several food crises, among which one of the most serious being the adulteration of milk products with melamine [1]. Fake and shoddy food products are more than a matter of commercial fraud but also invoke considerable concerns about public safety and interest. Consumers are increasingly demanding food products with conditions of production that are friendlier to the environment and/or warrant the product quality from a sensory, nutritional, or safety point of view [2].

For milks, some conditions of production such as geographical zones or cow grass feeding are known to confer specific organoleptic and nutritional qualities to the milk products [3–5] and thus provide an added value to the product and justify its higher price. In China, the mainstream milk manufactures and their sources of raw materials are located in the heavily populated eastern areas. Milk production in these areas might be influenced by various adverse conditions, such as potential environmental problems caused by rapid industrialization, the quality uncertainty during purchase, and storage of raw material [1]. On the contrary, Western Yunnan Plateau areas (about 2,000 m altitude), located in the southwest of China, has a unique geographical position and a sparse population. The place also enjoys a temperate climate with plenty of rainfall and sunshine. All the above factors contribute to the high quality of plateau milks, including rich nutrition, particular flavor, and more reliable safety guarantee [5, 6]. Moreover, the output of milks in Western Yunnan plateau is much lower than that of eastern China; therefore, it is attractive to falsely denote the origins of milk products for manufacturers, and it is necessary to develop quick and reliable methods for discrimination of milk origins.

Traditional methods for discrimination of food origins depend on chemical component analysis and sensory analysis. Because many food products like milks are highly complex chemical systems, the cost of a thorough analysis

of chemical components is often prohibitive. Moreover, the quality of milks usually cannot be sufficiently characterized just by the contents of a single or a few components. Sensory analysis is an expert-dependent technique and is thought as a reliable method for the purpose of food authentification, but it suffers the disadvantages of high cost and lack of objectivity. Compared with traditional methods, the combination of various spectrometry (such as near infrared [7–9] and fluorescent spectrometry [10]) and chemometric methods has provided promising alternative approaches for food control [11]. In spectroscopic analysis, the chemistry of the complicated samples can be characterized by the measured multivariate spectra and then multivariate statistical methods are used to extract information concerning food quality. Some advantages of spectrometry analysis are (1) no or less sample preparations are required; (2) the analysis time is largely reduced compared with traditional methods, so it is very suitable to analyze batch samples; (3) it is a nondestructive analysis method and can be used for online analysis; (4) when combined with chemometrics, it provides an automatic and quick analysis method for food control.

Among various spectroscopic techniques, fluorescent spectrometry is widely available in analytical labs, and its high sensitivity to a wide array of potential analytes makes it a powerful tool for food analysis [10]. For milk products, different fluorescent bands can be attributed to the differences in compositions (fluorescent analytes such as aromatic amino acids, nucleic acids, and tryptophan) and properties (e.g., antioxidant activity and acidity) of samples. This forms the basis for fluorescent analysis of milks of different kinds and sources. With the development of chemometric data fusion and multiway techniques like parallel factor analysis (PARAFAC) [12] and multiway partial least squares (PLS) [13], excitation-emission fluorescent spectrometry has been increasingly used in food analysis [11]. Compared with traditional excitation and emission fluorescence data, excitation-emission matrix data not only provides much more information, but also enables more options of data fusion and analysis methods.

This paper presents a case study of automatic discrimination of plateau milks from ordinary milks by fluorescent spectrometry and chemometrics. Different options of data fusion and analysis are investigated: (1) two-way partial least squares discriminant analysis (PLSDA) [14] based on the traditional excitation and emission spectra, respectively, (2) two-way PLSDA based on fusion of excitation and emission spectra and (3) three-way PLSDA [11, 13] based on excitation-emission matrix data. The objective is to develop a quick and yet reliable analysis method to distinguish the plateau milks from the milks produced in eastern China areas. More details of the work will be presented later.

2. Experimental and Methods

2.1. Preparation of Milk Samples and Fluorescent Spectrometric Analysis. A set of 60 pure and authentic milk samples from Western Yunnan Plateau area are collected from domestic markets. The samples consist of three brands, including *Ouya* (20), *Laisier* (20), and *Butterflyspring* (20). Another 89 milk samples from eastern China areas are collected of five mainstream brands including *Mengniu* (20), *Yili* (20), *Guangming* (20), *Wandashan* (17), and *Wangzai* (12). All the milk samples are produced by pasteurising technology and stored in a cool, dark area before spectrometry analysis.

The fluorescent spectra are measured on an MC-960 fluorescence spectrophotometer by Shanghai Xianke Instrument Co., Ltd. A trial experiment demonstrates the pure milk should be diluted to reflect the absorption characteristics in excitation spectra. Then the excitation-emission matrix data are measured with no further preprocessing of milk samples except a dilution of 1 : 500 with distilled water. The scanned excitation and emission wavelength regions are 180–300 nm (with an interval of 5 nm) and 200–800 nm (with an interval of 1 nm), respectively. Therefore, for each sample, a 25-by-601 excitation-emission matrix is obtained for each sample. A typical fluorescent matrix data set is shown in Figure 1.

2.2. Two-Way Partial Least Squares Discriminant Analysis (PLSDA). If each sample is described by a vector, for example, the multiwavelength emission spectra measured with the maximum excitation wavelength, one can obtain an $n \times p$ matrix $\mathbf{X}$ (a two-way data set) including p wavelength variables for n samples. For two-class problems, $\mathbf{X}$ contains samples from two different classes. A vector $\mathbf{y}$ $(n\times 1)$ contains the category variable of each sample in $\mathbf{X}$, for example, an element of 1 for class A and -1 for class B. The objective is to predict the class of new samples based on $\mathbf{X}$ and $\mathbf{y}$. The above problem can be solved by two-way PLSDA.

PLSDA is a classification method based on partial least squares (PLS) regression. As the key method in chemometrics, PLS has been widely used to solve various regression problems. The goal of PLS is to find a set of orthogonal latent variables that are the linear combinations of the original $\mathbf{X}$ variables, where the covariance between the latent variables and $\mathbf{y}$ is maximized under some constraints

$$\begin{aligned} &\max \quad (\mathbf{Xw})^{\mathrm{T}}\mathbf{y}, \\ &\text{subject to} \quad \mathbf{w}^{\mathrm{T}}\mathbf{w} = 1, \qquad \mathbf{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}_j = 0, \\ &\qquad\qquad \forall i \neq j,\ 1 \le i \le A,\ 1 \le j \le A, \end{aligned} \tag{1}$$

where A is the number of latent variables and $\mathbf{w}$ is the $p \times 1$ weighting vector of original $\mathbf{X}$ variables

$$\mathbf{t} = \mathbf{Xw}. \tag{2}$$

The above objective function can be solved by the Lagrange multiplier method. After all the A latent variables have been calculated, $\mathbf{y}$ is related to $\mathbf{X}$ by A latent variables

$$\mathbf{y} = \mathbf{Tq}, \tag{3}$$

where $\mathbf{T}$ $(\mathbf{T} = \mathbf{XW})$ contains A latent variables in its columns and $\mathbf{W}$ contains the corresponding A weighting vectors. Regression coefficients $\mathbf{q}$ can be solved by least squares regression

$$\mathbf{q} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y}. \tag{4}$$

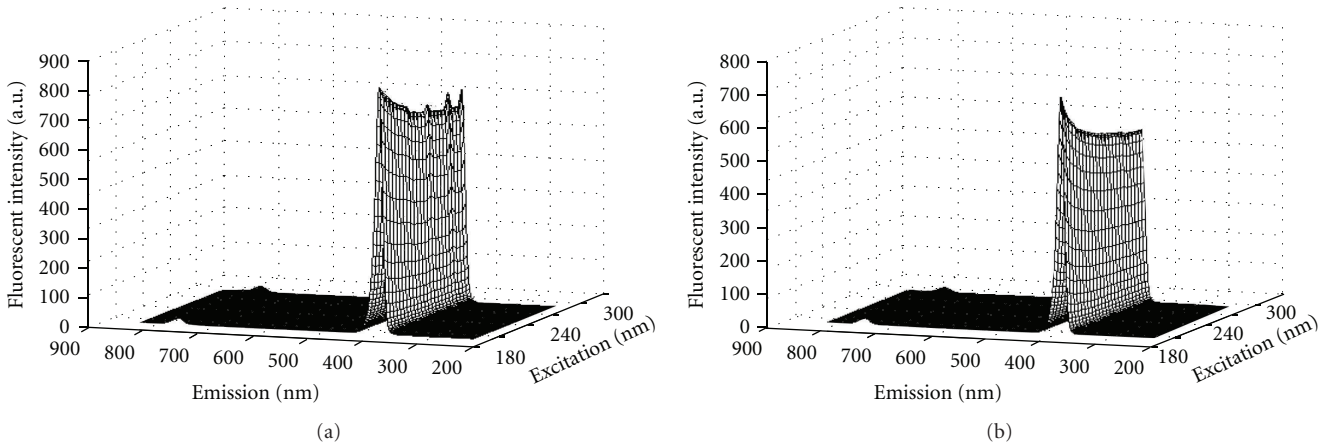

Figure 1: A typical fluorescence excitation-emission matrix for milk samples from (a) plateau area and (b) eastern China areas.

Then, $\mathbf{y}$ is related to $\mathbf{X}$ by PLS regression coefficients $\mathbf{b}$ ($\mathbf{b} = \mathbf{Wq}$) as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}, \tag{5}$$

where $\mathbf{e}$ is the error vector, and the dependent variable $\mathbf{y}_{\text{un}}$ of unknown samples can be predicted from the corresponding predictor variables $\mathbf{X}_{\text{un}}$

$$\mathbf{y}_{\text{un}} = \mathbf{X}_{\text{un}}\mathbf{b}. \tag{6}$$

For PLSDA, instead of a set of continuous values, $\mathbf{y}$ contains a binary vector of 1 and -1 (or 1 and 0) denoting class A and B, respectively. A predicted value of a new sample above 0 means the sample is predicted to belong to class A by the model and vice versa.

2.3. Three-Way PLSDA. Here, a brief introduction to three-way PLS will be given. If each sample is described by a matrix, for example the fluorescent excitation-emission spectra, one can obtain an $n \times p_1 \times p_2$ cubic matrix $\mathbf{X}$ for n samples including fluorescent intensities scanned at p_1 excitation wavelengths and p_2 emission wavelengths. A three-way data set is shown in Figure 2.

Three-way PLS is an extension of two-way PLS to tackle three-way data. Three-way PLS maximizes the covariance between a latent variable $\mathbf{t}$ ($n \times 1$) and $\mathbf{y}$. The score of sample i ($i = 1, 2, \ldots, n$) in ($i = 1, 2, \ldots, n$)$\mathbf{t}$ can be calculated as

$$\mathbf{t}_i = \mathbf{w}_{1i}{}^T \mathbf{X}_{i..} \mathbf{w}_{2i}, \tag{7}$$

where $\mathbf{w}_{1i}(p_1 \times 1)$ and $\mathbf{w}_{2i}(p_2 \times 1)$ are weighting vectors for latent variable j ($j = 1, 2, \ldots, A$), $\mathbf{X}_{i..}(p_1 \times p_2)$ is a matrix containing the fluorescent excitation-emission data for sample i. The weighting vectors can be deduced by unfolding the cubic matrix and solving an eigenvalue problem [13]. When the A latent variables for three-way PLS are obtained, three-way PLSDA can be performed as in (3)–(6).

2.4. Monte Carlo Cross-Validation (MCCV) [15, 16]. For discriminant models based on two-way and three-way partial least squares, an important problem is to select the number of latent variables or determine the model complexity. Including too few latent variables will lose some useful information in the data structure and fail to classify the samples sufficiently, while a model with too much complexity will include the class-uncorrelated data variances and have a bad prediction performance. Therefore, a well-established cross-validation method, MCCV [15, 16], is used to determine the complexity of classification models.

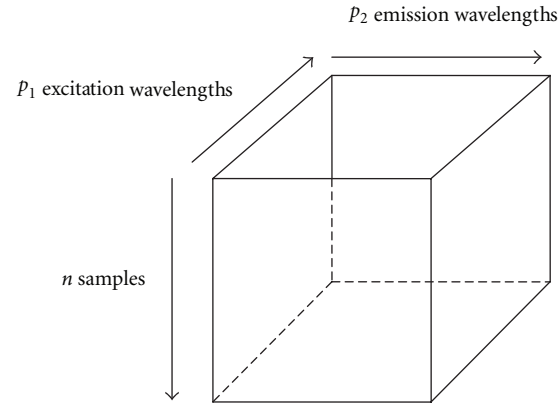

Figure 2: The $n \times p_1 \times p_2$ three-way matrix data for n samples scanned at p_1 excitation wavelengths and p_2 emission wavelengths.

MCCV is originally proposed and used to reduce the risk of selecting too many PLS components [15] and then corrected for model errors estimation [16]. By multiple resampling and excluding certain percent of training samples, MCCV has been proved to be an effective method to estimate model complexity [17]. With a predefined model complexity, the root mean square error of MCCV (RMSEMCCV) can be calculated as

$$\text{RMSEMCCV} = \sqrt{\frac{1}{S \times n_v} \sum_{i=1}^{S} ||\mathbf{y}_{S(i)} - \hat{\mathbf{y}}_{S(i)}||^2}, \tag{8}$$

where S and n_v are the resampling time and size of left-out samples, respectively; $\mathbf{y}_{S(i)}$ and $\hat{\mathbf{y}}_{S(i)}$ represent the reference

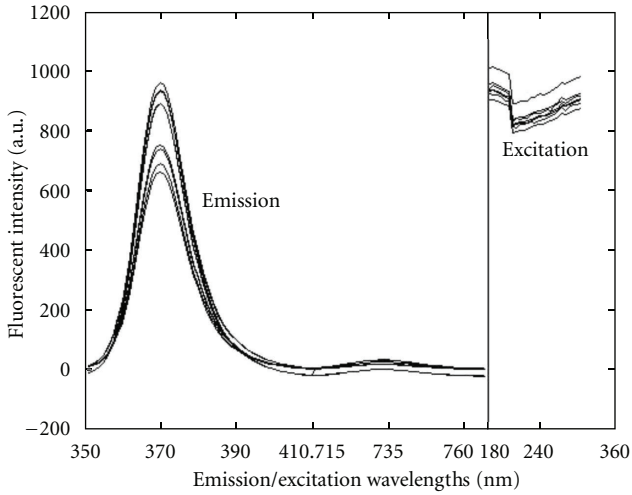

Figure 3: Fusion of excitation and emission spectra for two-way PLSDA.

and predicted values of left-out samples during the ith resampling, respectively. The number of PLS components is selected to get the lowest RMSEMCCV value. The percent of left-out samples can be adjusted according the size of training set.

3. Results and Discussion

To remove the baselines, all the data are corrected by subtracting the spectra matrix of distilled water. Moreover, to reduce the computational burden, wavelength channels that have no significant signals compared with backgrounds (signals of water) are eliminated. For two-way methods, the emission spectra and fusion of excitation and emission spectra are demonstrated in Figure 3. Both the excitation and emission spectra are selected to have the maximum fluorescent intensities.

To make the data analysis and comparison of model performances reliable, potential outliers must be removed. With the 149 milk samples, robust PCA [18] is performed, and no outliers are detected. To select the representative training and test samples for model building and validation, Kennard and Stone (KS) algorithm [19] is used to split the samples into a representative training set and a test set. The KS algorithm selects the set of training samples that covers the overall sample domain based on their distance (Euclidean distance) from each other. For the four models, the KS algorithm is performed on the two-way fusion data as shown in Figure 3(b). Therefore, a training set of 80 samples (40 genuine plateau milks + 40 nonplateau milks) and a test set of 69 samples (20 genuine plateau milks + 49 nonplateau milks) are obtained.

Two-way PLSDA models are developed with excitation spectra, emission spectra, and fusion of excitation and emission spectra, respectively. Three-way PLSDA is built on the excitation-emission matrix data. Considering the size of training set is not very large, for all the four models, MCCV with 20 percents of left-out samples is used to determine the number of PLS components and the sampling time is 100. The results of different models are listed in Table 1.

Table 1: The classification results of different models for milk samples.

Models	Two-way PLSDA[1]	Two-way PLSDA[2]	Two-way PLSDA[3]	Three-way PLSDA
Model complexity	4	4	5	5
Number of misclassified[4]	7/6	7/10	4/3	2/4
Total accuracy rate	91.3%	88.6%	95.3%	96.0%

[1]Two-way PLSDA with excitation spectra,
[2]Two-way PLSDA with emission spectra,
[3]Two-way PLSDA with fusion of excitation and emission spectra,
[4]The number of misclassified samples for training/predicting.

Seen from Table 1, the two-way PLSDA with excitation spectra and emission spectra has an accuracy of 91.3% (136/149), and 88.6% (132/149), respectively, which is much inferior to those of the other two methods. This can be partially explained by the insufficiency of chemical information carried by pure excitation and emission spectra, because the differences between the excitation or emission spectra of the different milk samples are very subtle. On the other hand, for two-way PLSDA with data fusion and three-way PLSDA with matrix data, both the numbers of PLS components are 5, which can be attributed to the similarity of information contained in the data. Moreover, the error rate for the two models is 4.7% (7/149) and 4.0% (6/149), respectively, indicating that the performance of two-way PLSDA with data fusion is comparable to that of three-way PLSDA with matrix data. The detailed results obtained by two-way PLS with data fusion are further shown in Figure 4, where the numbers of misclassified samples for training and prediction are 4 and 3, respectively. Seen from Figure 4, the two-way PLSDA model is sufficiently trained, and no overfitting has been found, because the prediction results are equally well compared with training results.

4. Conclusions

In order to achieve automatic identification of genuine plateau milk samples, excitation-emission fluorescence matrix spectra are measured, and different data analysis and fusion methods are investigated. The results demonstrate that two-way PLS with pure excitation or emission spectra are not very sufficient to classify the milk samples, while two-way PLSDA with fusion of emission and excitation spectra and three-way PLSDA with matrix data are effective in distinguishing milk samples of different geographical origins.

Compared with three-way PLSDA, the two-way PLS with data fusion has some advantages. Firstly, the measurement of full excitation-emission fluorescence matrices is time consuming, especially when the sample size is large or in case of batch samples. Conversely, the measurement of an excitation and emission spectra is much more convenient. Secondly, while the three-way PLSDA is a somewhat complex

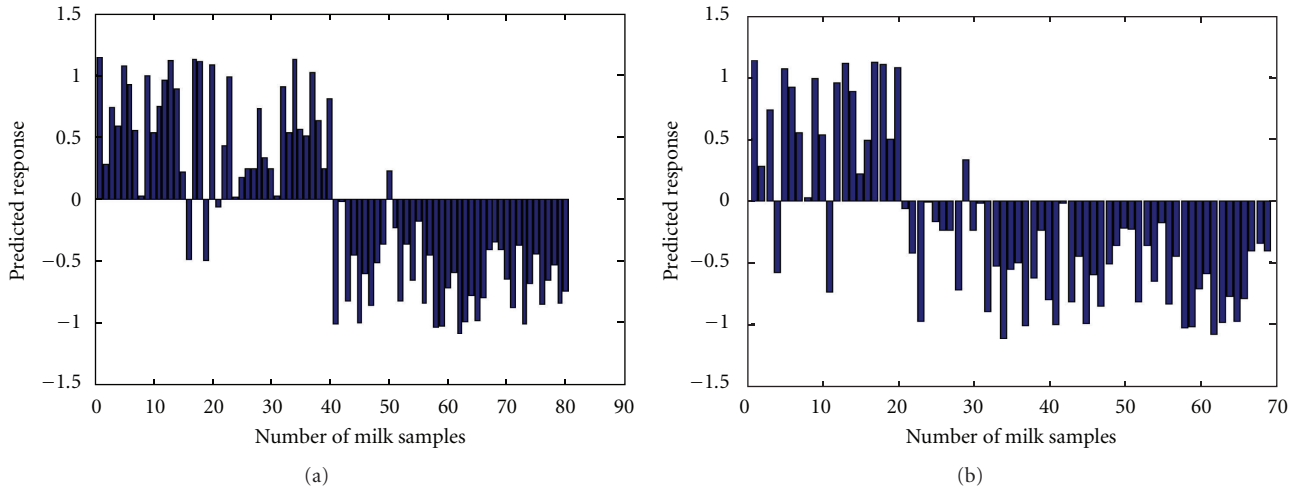

Figure 4: The training (a) and prediction (b) results by two-way PLSDA with fusion data. Samples 1–40 (a) and 1–20 (b) are genuine plateau milks; Samples 41–80 (a) and 21–69 (b) are nonplateau milks.

mathematical tool for routine use, two-way PLSDA is a well-established and easy-to-use tool in chemometrics. Therefore, two-way PLSDA with data fusion of emission and excitation spectra is recommended as a quick and reliable method for authentication of plateau milks. Our future work will be focused on quantitative analysis of milk quality parameters by fluorescence spectrometry and chemometrics.

Acknowledgment

This work is financially supported by the Basic Research Plans on Natural Science of the Education Department of Henan Province (nos. 2008A150001 and 2011A430001).

References

[1] ACS, "Chinese Baby-Food Crisis Widens," 2008, http://www.nhlbi.nih.gov/meetings/workshops/cardiorenal-hf-hd.htm.

[2] E. Engel, A. Ferlay, A. Cornu et al., "Relevance of isotopic and molecular biomarkers for the authentication of milk according to production zone and type of feeding of the cow," *Journal of Agricultural and Food Chemistry*, vol. 55, no. 22, pp. 9099–9108, 2007.

[3] J.-B. Coulon, A. Delacroix-Buchet, B. Martin, and A. Pirisi, "Relationships between ruminant management and sensory characteristics of cheeses: a review," *Lait*, vol. 84, no. 3, pp. 221–241, 2004.

[4] B. Martin, I. Verdier-Metz, S. Buchin, C. Hurtaud, and J.-B. Coulon, "How do the nature of forages and pasture diversity influence the sensory quality of dairy livestock products?" *Animal Science*, vol. 81, no. 2, pp. 205–212, 2005.

[5] A. Lucas, C. Agabriel, B. Martin et al., "Relationships between the conditions of cow' s milk production and the contents of components of nutritional interest in raw milk farmhouse cheese," *Lait*, vol. 86, no. 3, pp. 177–202, 2006.

[6] M. A. De La Fuente and M. Juarez, "Authenticity assessment of dairy products," *Critical Reviews in Food Science and Nutrition*, vol. 45, no. 7-8, pp. 563–585, 2005.

[7] A. Alishahi, H. Farahmand, N. Prieto, and D. Cozzolino, "Identification of transgenic foods using NIR spectroscopy: a review," *Spectrochimica Acta A*, vol. 75, no. 1, pp. 1–7, 2010.

[8] H. Huang, H. Yu, H. Xu, and Y. Ying, "Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: a review," *Journal of Food Engineering*, vol. 87, no. 3, pp. 303–313, 2008.

[9] D. Toher, G. Downey, and T. B. Murphy, "A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies," *Chemometrics and Intelligent Laboratory Systems*, vol. 89, no. 2, pp. 102–115, 2007.

[10] J. Sádecká and J. Tóthová, "Fluorescence spectroscopy and chemometrics in the food classification-a review," *Czech Journal of Food Sciences*, vol. 25, no. 4, pp. 159–173, 2007.

[11] L. Munck, L. Nørgaard, S. B. Engelsen, R. Bro, and C. A. Andersson, "Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance," *Chemometrics and Intelligent Laboratory Systems*, vol. 44, no. 1-2, pp. 31–60, 1998.

[12] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 149–171, 1997.

[13] R. Bro, "Multiway calibration. Multilinear PLS," *Journal of Chemometrics*, vol. 10, no. 1, pp. 47–61, 1996.

[14] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[15] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.

[16] Q.-S. Xu, Y.-Z. Liang, and Y.-P. Du, "Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration," *Journal of Chemometrics*, vol. 18, no. 2, pp. 112–120, 2004.

[17] S. Gourvénec, J. A. Fernández Pierna, D. L. Massart, and D. N. Rutledge, "An evaluation of the PoLiSh smoothed regression and the Monte Carlo Cross-Validation for the determination

of the complexity of a PLS model," *Chemometrics and Intelligent Laboratory Systems*, vol. 68, no. 1-2, pp. 41–51, 2003.

[18] M. Hubert, P. Rousseeuw, and T. Verdonck, "Robust PCA for skewed data and its outlier map," *Computational Statistics and Data Analysis*, vol. 53, no. 6, pp. 2264–2274, 2009.

[19] R. W. Kennard and L. Stone, "Computer aided design of experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.