# Automated Principal Component–Based Orthogonal Signal Correction Applied to Fused Near Infrared–Mid-Infrared Spectra of French Olive Oils

**ARTICLE** *in* ANALYTICAL CHEMISTRY · SEPTEMBER 2009

Impact Factor: 5.64 · DOI: 10.1021/ac900538n · Source: PubMed



**4 AUTHORS:**

Peter B Harrington
Ohio University
**180** PUBLICATIONS **2,186** CITATIONS
SEE PROFILE

Jacky Kister
French National Centre for Scientific Resea…
**36** PUBLICATIONS **387** CITATIONS
SEE PROFILE

Jacques Artaud
Aix-Marseille Université
**65** PUBLICATIONS **868** CITATIONS
SEE PROFILE

Nathalie Dupuy
Aix-Marseille Université
**103** PUBLICATIONS **1,380** CITATIONS
SEE PROFILE

*Anal. Chem.* **2009,** *81,* 7160–7169

# Automated Principal Component-Based Orthogonal Signal Correction Applied to Fused Near Infrared−Mid-Infrared Spectra of French Olive Oils

**Peter de B. Harrington,*,† Jacky Kister,‡ Jacques Artaud,‡ and Nathalie Dupuy‡**

*OHIO University Center for Intelligent Chemical Instrumentation, Department of Chemistry and Biochemistry, Clippinger Laboratories, Athens, Ohio 45701-2979, and ISM², UMR 6263, Equipe AD²EM, Université Paul Cézanne, Case 451, 13397 Marseille Cedex 20 France*

**An approach for automating the determination of the number of components in orthogonal signal correction (OSC) has been devised. In addition, a novel principal component OSC (PC-OSC) is reported that builds softer models for removing background from signals and is much faster than the partial least-squares (PLS) based OSC algorithm. These signal correction methods were evaluated by classifying fused near- and mid-infrared spectra of French olive oils by geographic origin. Two classification methods, partial least-squares-discriminant analysis (PLS-DA) and a fuzzy rule-building expert system (FuRES), were used to evaluate the signal correction of the fused vibrational spectra from the olive oils. The number of components was determined by using bootstrap Latin partitions (BLPs) in the signal correction routine and maximizing the average projected difference resolution (PDR). The same approach was used to select the number of latent variables in the PLS-DA evaluation and perfect classification was obtained. Biased PLS-DA models were also evaluated that optimized the number of latent variables to yield the minimum prediction error. Fuzzy or soft classification systems benefit from background removal. The FuRES prediction results did not differ significantly from the results that were obtained using either the unbiased or biased PLS-DA methods, but was an order of magnitude faster in the evaluations when a sufficient number of PC-OSC components were selected. The importance of bootstrapping was demonstrated for the automated OSC and PC-OSC methods. In addition, the PLS-DA algorithms were also automated using BLPs and proved effective.**

Measurement data are frequently plagued with background variations especially for large scale studies and for complex samples. Accurate background correction is a difficult yet important problem for data analysis. There are many approaches to correcting backgrounds or removing baseline variations. The key idea is to remove unwanted or irrelevant variances from the measurement data. Because background correction is a key first step for data processing, improper corrections can generate errors that propagate through the modeling process and result in inaccurate predictions. In many cases, background correction, such as polynomial curve-fitting, the practice is more a time-consuming art than a science. Adjustments to parameters such as polynomial order and window size can have a pronounced effect on the corrected data. Furthermore, spurious artifacts can be introduced that may cause problems later during model building and interpretation steps. With the modern trend of making analytical measurements of more samples that are complex, automated methods of model building and evaluation are essential. Embedding automated chemometric methods so that they are transparent to the user into analytical instrumentation is a key step toward the design of an intelligent chemical instrument.

Quality assurance of food products and safety is a growing concern as the global economy expands. Food from foreign nations may not conform to the same hygiene and safety standards as the consumer nation. Furthermore, varieties of foods from different cultivars, geographic regions, and ages will vary in quality and price. There is a need to develop methods to ensure the veracity of food labels and detect adulterated or misidentified samples, whether accidental or deliberate. The Mediterranean basin is the most important virgin olive oil (VOO) producing area in the world. VOO comes from very many olive varieties, which are generally related to a soil that confers specific sensory and chemical properties. These characteristics are developed by attribution by national or European organizations of mark of quality as the Registered Designation of Origin (RDO) in France or the Protected Designation of Origin (PDO) in Europe. In France, eight RDOs were recently created. The RDOs are regulated by specific articles and conditions that rule all of the official channels from growth to fabrication. Characterization of VOO based on the chemical composition is related to various families of compounds present in VOO: fatty acids, triacylglycerols, phenols, tocopherols, sterols, aroma, pigments, etc. The chemical composition is generally determined by chromatographic techniques (gas chromatography (GC) and high-performance liquid chromatography (HPLC)).[1] However, these techniques are relatively slow and require complex sample preparation (extraction, separation). Increased traceability, especially for food-processing products, requires fast and powerful analytical methodology to solve this problem. The determination of origin and VOO authenticity has thus been the subject of numerous studies using

* Corresponding author. E-mail: Peter.Harrington@OHIO.edu.
† OHIO University Center for Intelligent Chemical Instrumentation.
‡ Université Paul Cézanne.

(1) Aparicio, R.; Aparicio-Ruiz, R. *J. Chromatogr., A* **2000**, *881*, 93–104.



10.1021/ac900538n CCC: $40.75 
Published on Web 08/05/2009

a variety of spectral techniques ($^1$H nuclear magnetic resonance (NMR), $^{13}$C NMR, fluorescence, ultraviolet (UV), and infrared (IR)). The main advantages of IR spectroscopy to the gold standard chromatographic methods are that sample preparation is simpler and spectral acquisition is faster. IR spectroscopy in combination with multivariate analysis was developed for three important VOO characteristics, prediction of peroxide value,[2] free fatty acid,[3] acidity[4] and cis/trans fatty acids[5,6] authentication, characterization, and detection of adulteration,[7−9] and determination of geographic origin, namely, France,[10] Italy,[11−13] Spain,[14,15] and Turkey.[16]

Instrumental methods provide a more efficient and effective means for evaluating food samples (i.e., olive oils) than organoleptic evaluation and are less subjective because the results do not depend on subconscious cues or the moods of the expert. Two spectroscopic methods of evaluating olive oils are near-infrared spectroscopy (NIRS)[17] and mid-infrared spectroscopy (MIRS).[18] Spectra from these two methods may be combined so that the data are fused and the information content is increased as long with the size of the data.

For high-resolution methods, such as nuclear magnetic resonance,[16] or data fusion such as MIR-NIR, a common problem is that the data is underdetermined in that there are more variables than objects. This case presents a problem for all supervised chemometric methods in that models exist for fitting the data perfectly by modeling insignificant variations in the data set (i.e., noise). Therefore, data compression is important to improve the efficacy and also the efficiency. Smaller data sets require less computational power. Many classification algorithms will scale as the square of the number of variables. For data with a large number of variables, even with modern advances in computation, the geometric increase of load may exceed the computational capacity of the computer with respect to memory and speed. Data compression is as important as ever to analytical chemists.


(2) Vandevoort, F. R.; Ismail, A. A.; Sedman, J.; Dubois, J.; Nicodemo, T. *J. Am. Oil Chem. Soc.* **1994**, *71*, 921–926.
(3) Bertran, E.; Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Montoliu, I. *J. Am. Oil Chem. Soc.* **1999**, *76*, 611–616.
(4) Inon, F. A.; Garrigues, J. M.; Garrigues, S.; Molina, A.; de la Guardia, M. *Anal. Chim. Acta* **2003**, *489*, 59–75.
(5) Bendini, A.; Cerretani, L.; Di Virgilio, F.; Belloni, P.; Bonoli-Carbognin, M.; Lercker, G. *J. Food Qual.* **2007**, *30*, 424–437.
(6) Sinelli, N.; Cosio, M. S.; Gigliotti, C.; Casiraghi, E. *Anal. Chim. Acta* **2007**, *598*, 128–134.
(7) Christy, A. A.; Kasemsumran, S.; Du, Y. P.; Ozaki, Y. *Anal. Sci.* **2004**, *20*, 935–940.
(8) Lai, Y. W.; Kemsley, E. K.; Wilson, R. H. *Food Chem.* **1995**, *53*, 95–98.
(9) Yang, H.; Irudayaraj, J.; Paradkar, M. M. *Food Chem.* **2005**, *93*, 25–32.
(10) Mignani, A. G.; Ciaccheri, L.; Cimato, A.; Attilio, C.; Smith, P. R. *Sens. Actuators, B: Chem.* **2005**, *111*, 363–369.
(11) Di Bella, G.; Maisano, R.; La Pera, L.; Lo Turco, V.; Salvo, F.; Dugo, G. *J. Agric. Food Chem.* **2007**, *55*, 6568–6574.
(12) D'Imperio, M.; Mannina, L.; Capitani, D.; Bidet, O.; Rossi, E.; Bucarelli, F. M.; Quaglia, G. B.; Segre, A. *Food Chem.* **2007**, *105*, 1256–1267.
(13) Casale, M.; Casolino, C.; Ferrari, G.; Forina, M. *J. Near Infrared Spectrosc.* **2008**, *16*, 39–47.
(14) Perez, M. M.; Yebra, A.; Melgosa, M.; Bououd, N.; Asselman, A.; Boucetta, A. *Grasas Y Aceites* **2003**, *54*, 392–396.
(15) Lopez-Feria, S.; Cardenas, S.; Garcia-Mesa, J. A.; Valcarcel, M. *Talanta* **2008**, *75*, 937–943.
(16) Rezzi, S.; Axelson, D. E.; Heberger, K.; Reniero, F.; Mariani, C.; Guillou, C. *Anal. Chim. Acta* **2005**, *552*, 13–24.
(17) Downey, G.; McIntyre, P.; Davies, A. N. *Appl. Spectrosc.* **2003**, *57*, 158–163.
(18) Tapp, H. S.; Defernez, M.; Kemsley, E. K. *J. Agric. Food Chem.* **2003**, *51*, 6110–6115.


There also is an interest in fuzzy and soft pattern recognition methods because they can be used to avoid the overfitting problem with underdetermined data. Because these methods model variances of the measurements, when unwanted background variances dominate, the modeling of weaker signals may be obscured. Thus, it is crucial to remove these background variances for soft modeling methods to work effectively. If these background variances are removed prior to compression, then more efficient compressions may also be achieved.

Orthogonal signal correction[19] (OSC) is one of many methods for background correction. The method is supervised in that the dependent variables are used to help find variances that are uncorrelated to the dependent variables. Thus, OSC models must always be built with model-building or model-building data and these same models evaluated with the prediction data set.

This article presents an approach for automating signal correction algorithms and presents a new signal correction approach alluded to in a previous publication.[19] The two methods of signal correction are evaluated using a fuzzy rule-building expert system and discriminant partial least-squares.

## THEORY

**Orthogonal Signal Correction.** For many analytical measurements correction of baseline fluctuations is a critical step in preprocessing. In spectroscopic measurements, the common approach is to calculate first and second derivatives to correct baseline drift, but these calculations will decrease the signal-to-noise ratios and the results depend on the choice of the filter width or window size or the order of the polynomial used in the correction. OSC[19] was devised by Wold et al. which removed variances that are orthogonal or unrelated to the properties of interest or the dependent variables. The approach of removing variation that is orthogonal is an application of the multivariate net analytical signal concept[20] formalized by Lorber et al. for the determination of a component in a mixture, except in this case the baseline fluctuations and noise are considered as other components.

The first step in the correction is to mean center the data by subtracting the averages of the independent $\mathbf{X}$ and dependent $\mathbf{Y}$ variables from their respective matrices.

The key idea is to define a subspace from the $m \times p$ matrix of dependent variables $\mathbf{Y}$ as

$$\mathbf{H} = \mathbf{Y}(\mathbf{Y}^\mathbf{T}\mathbf{Y})^{-1}\mathbf{Y}^\mathbf{T} \tag{1}$$

for which $m$ is the number of model-building spectra, $p$ is the number of properties, and $\mathbf{H}$ is an $m \times m$ matrix of the space defined by the columns of $\mathbf{Y}$. $\mathbf{H}$ can be considered the Hat matrix of the dependent variables $\mathbf{Y}$. The product is defined by

$$\hat{\mathbf{X}} = \mathbf{H}\mathbf{X} \tag{2}$$

for which $\hat{\mathbf{X}}$ is an $m \times n$ matrix of the estimated spectra that correlate with the dependent variables. When $\mathbf{Y}$ is a binary or bipolar matrix that encodes the class designations, the matrix $\hat{\mathbf{X}}$ comprises the


(19) Wold, S.; Antti, H.; Lindgren, F.; Ohman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175–185.
(20) Lorber, A.; Faber, K.; Kowalski, B. R. *Anal. Chem.* **1997**, *69*, 1620–1626.

corresponding class mean for each row object. Thus, the other uncorrelated variances can be removed by subtraction of the estimate $\hat{\mathbf{X}}$ from the matrix $\mathbf{X}$.

$$\mathbf{X_o} = \mathbf{X} - \hat{\mathrm{X}} = (\mathbf{I} - \mathbf{H})\mathbf{X} \tag{3}$$

for which $\mathbf{X_o}$ is the $m \times n$ data matrix comprising the variations that are not related to the properties. For classification, $\mathbf{X_o}$ is the ANOVA-PCA residual matrix that comprises the within class variations. The class mean has been subtracted from each data object.

A basis could be constructed directly from $\mathbf{X_o}$, but the problem was how to accommodate prediction objects. For this approach, Wold et al. wrote that it was "unworkable" for new prediction objects and left it for others to pursue.[19] The PC-OSC approach calculates an orthogonal basis of $\mathbf{X_o}$ by singular value decomposition

$$\mathbf{X_o} = \mathbf{USV^T} \tag{4}$$

for which two orthogonal matrices of eigenvectors $\mathbf{U}$ and $\mathbf{V}$ and a diagonal matrix of singular values $\mathbf{S}$ are obtained. Because the background variances arise from the correlation of variables, the $\mathbf{V}$ eigenvectors will form a basis that can be used for correction.

The process is very straightforward as

$$\hat{\mathbf{X}}_\mathbf{p} = (\mathbf{X_pV})\mathbf{V^T} \tag{5}$$

The spectra located in the rows of $\mathbf{X_p}$ are projected onto the orthogonal vectors $\mathbf{V}$ to yield a set of scores. These scores are then used with the transposed $\mathbf{V}$ eigenvectors to reconstruct the background components of the spectral matrix. Remember, the $\mathbf{V}$ eigenvectors were computed from a data matrix that had the analytical signal removed, so the reconstructed spectra will have no signal but comprise the systematic background fluctuations.

The prediction data are easily corrected as given below

$$\mathbf{X_c} = \mathbf{X_p} - \hat{\mathbf{X}}_\mathbf{p} \tag{6}$$

for which $\mathbf{X_c}$ is the matrix of background corrected spectra and $\hat{\mathbf{X}}_\mathbf{p}$ is the matrix of background estimates. The number of components in $\mathbf{V}$ controls the degree of the background subtraction similar to the number of components in OSC.

In the PLS approach, inside the NIPALS algorithm the scores are orthogonalized against $\mathbf{Y}$ using an iterative approach in the NIPALS algorithm. The iterative PLS approach is much slower than the new approach given above. However, it has an important consequence in that it finds components that maximize the covariance between the subspace that is orthogonal to the properties and the data matrix $\mathbf{X}$ so that the components are more efficient at removing unrelated variations and fewer components are required to correct the data. This difference in efficiency also appears when PLS is compared to principal component regression.

The orthogonal signal corrections were enhanced so that they would determine the optimum number of components. The approach used the projected difference resolution[21] (PDR) method combined with bootstrapped Latin partitions[22,23] to estimate the number of components within the OSC methods.

The PDR method is used for measuring the separation of classes in the multivariate data space and gives a result that is analogous to chromatographic resolution so it is easy to interpret. The resolutions between all combinations of pairs of classes are calculated and the geometric mean of the classes is used for optimizing the number of components. The first step is to calculate the projections of the objects in two classes onto the mean difference vector between the two classes as

$$p_i = (\bar{\mathbf{x}}_\mathbf{A} - \bar{\mathbf{x}}_\mathbf{B})\mathbf{x_i^T} \tag{7}$$

for which $p_i$ is the $i$th data object projected onto the difference vector calculated from classes A and B. Now a resolution measure is calculated from the scalar projections as

$$\mathrm{Rs(A, B)} = \frac{|\bar{p}_\mathrm{A} - \bar{p}_\mathrm{B}|}{2(s_\mathrm{A} + s_\mathrm{B})} \tag{8}$$

for which the resolution between classes A and B is measured by the absolute value of the difference between the averages of the class A and B projections divided by twice the sum of the standard deviations about each class mean. This simple calculation is fast and is a good indicator for when data preprocessing steps are improving the class separations. Large values are desirable, and a resolution value greater than 1.5 indicates that the classes are resolved from each other.

The PDR measures can be bootstrapped as well. All PDR values were reported as averages with 95% confidence intervals across 10 bootstraps. Each bootstrap randomly removed one object from each class. Because there are six classes for the data set, six spectra were removed from each bootstrap calculation.

Because these signal correction methods are supervised, one must be aware of the risk of overfitting the data. Within the OSC and PC-OSC functions, the data are split into subsets, so that one set is used for building the orthogonal basis sets and the other set is used for determining the number of components.

The number of components or basis vectors depends on the compositions of both the model building and test set, so it is judicious to use bootstrapped Latin partitions.[23] In this work, the data were randomly split into two sets of equal size and the Latin partitions maintain the same distribution of classes between the two sets. Because there were 10 bootstraps and 2 Latin partitions, there were 20 PDR values measured for each component number. The averages and 95% confidence intervals were calculated across the 20 BLP PDRs. The maximum average PDR is determined from the average PDRs with respect to component numbers. Components that occur after the maximum average PDR are omitted from consideration. The confidence interval at the maximum is subtracted from the maximum to determine a threshold. The components from the first to the component with

(21) Cao, L. Nonlinear Wavelet Compression Methods for Ion Analyses and Dynamic Modeling of Complex Systems. Dissertation, OHIO University, Athens, OH, 2004, p 180.
(22) Wan, C. H.; Harrington, P. D. *Anal. Chim. Acta* **2000**, *408*, 1–12.
(23) Harrington, P. D. B. *TrAC, Trends Anal. Chem.* **2006**, *25*, 1112–1124.

the lowest average PDR that is above the threshold value are used for the correction. For the results reported in this paper, all orthogonal signal corrections partitioned the data into two equally sized subsets using Latin partitions and 10 bootstraps, with the exception for the cases that had no bootstrapping of the signal correction optimization and the cases for which the numbers of components were configured manually.

The background basis is constructed from the model-building data. The basis is used to reconstruct the backgrounds form the prediction set. After removal of the reconstructed background, the PDR is measured. The PDR is measured while the number of components is incremented. A set of PDRs with respect to component number will be obtained for each Latin partition (i.e., 2) within the 10 bootstraps. There were 20 PDR profiles with respect to component number that are used to calculate an average profile and the corresponding confidence interval. The maximum of the PDR profile is determined. A threshold is obtained by subtracting the 95% confidence interval from the maximum value, and the smallest component number that yields a PDR larger than this threshold will define the number of components to be used in the orthogonal signal correction.

The PDR measure is applied to each prediction set separately. It is possible to reconstitute the full data set with the two background corrected prediction sets. However, there is a correlation with respect to background correction that occurs when two or more sets of corrected prediction data are combined into a single data set and the PDR is measured. This phenomenon is not understood and will be investigated further.

**Preprocessing.** Preprocessing was applied separately to the IR and NIR data before fusing the spectra together. Multiplicative scatter correction[24] (MSC) was applied and followed by normalizing each IR and NIR spectrum to unit length. The spectra were concatenated by combining the NIR and IR spectra from each corresponding oil sample to form a fused data object. Lastly, the mean spectrum was subtracted from each fused data object. This process was only applied to the model-building spectra. The averages calculated from the MSC and calculated after the normalization were stored, so that these averages could be used to correct the objects in the prediction set. This precaution is important to avoid any bias during the evaluations.

The prediction objects were preprocessed by applying MSC using the average of the model-building set. The spectra were normalized to unit vector length, and the average of the normalized and corrected model-building set was subtracted from each prediction object.

**Classification.** Two classification methods were compared. The first was a fuzzy rule-building expert system[25] (FuRES), and the second was partial least-squares-discriminant analysis[26,27] (PLS-DA). Both methods have been well described in the literature. It should be noted that FuRES is a soft classifier and applied directly to NIR data will not perform well unless the baseline variations are removed, hence, application of orthogonal signal correction methods.

(24) Geladi, P.; MacDougall, D.; Martens, H. *Appl. Spectrosc.* **1985**, *39*, 491–500.
(25) Harrington, P. B. *J. Chemom.* **1991**, *5*, 467–486.
(26) Frank, I. E.; Kowalski, B. R. *Anal. Chim. Acta* **1984**, *162*, 241–251.
(27) Kowalski, B. R. *Chemometrics, Mathematics, and Statistics in Chemistry*; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1984.

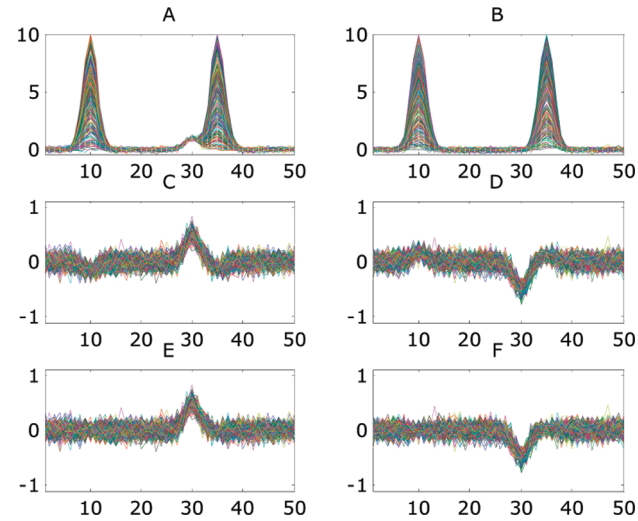


**Figure 1.** Overdetermined synthetic data set with random background peaks located at 10 and 35. Positive signals (peak at 30) are located on the left side, and null samples are on the right side. The top row (A and B) are the uncorrected data objects. The middle row (C and D) are the OSC corrected data with two components, and the bottom row (E and F) are the PC-OSC corrected data with two components. The components were determined automatically. The axes for parts C−F are identical.

**Validation.** All the classifications were evaluated using 10 bootstrapped Latin partitions with the number of partitions equal to 3. The 3 sets of prediction results were combined for each bootstrap and then averaged across the 10 bootstraps. The prediction sets were never used during any of the model building steps including the orthogonal signal correction. ANOVA was used to compare the FuRES and PLS models.

## EXPERIMENTAL SECTION

**Synthetic Data Sets.** Three simple synthetic data sets were constructed to evaluate the signal correction methods. The first set is overdetermined with 600 objects and 50 variables, while the second set is underdetermined with 50 objects and 600 variables. Many evaluations were run of the methods that all yielded similar results, but three exemplary sets are included to demonstrate the results obtained with the fused data are not fortuitous for a single data set.

For the overdetermined data set, a single Gaussian peak at a location of 30 of the 50 points with an amplitude of unity and a standard deviation of 2 points was created for the first 300 objects and is missing for the subsequent 300 objects that represent the null signal. To create a varying background, two spurious Gaussian peaks were added at points 10 and 35. These peaks had the same widths of a standard deviation of 2, but their amplitudes varied using a random uniform deviates from 0−10 for all 600 objects. Noise in the form of a standard normal deviates with a standard deviation of 0.1 was added to all data points. The positive and negative data objects can be observed in parts A and B of Figure 1, respectively.

The underdetermined data set was generated in a similar fashion. A signal represented by a single Gaussian peak at location 360 with a standard deviation of 20 points and amplitude of unity was create for the first 25 objects. The same procedure was used to create the varying background with two Gaussian peaks that were located at 120 and 420 with standard deviations of 20 points. The amplitudes of these two Gaussians varied using uniform

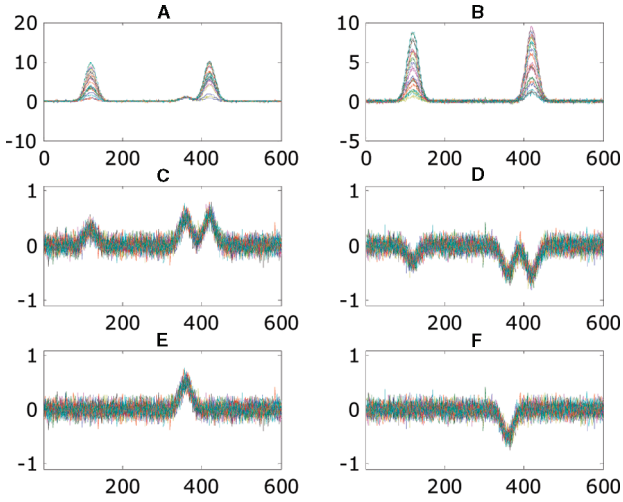


**Figure 2.** Underdetermined synthetic data set with random background peaks located at 120 and 420. Positive signals (peak at 360) are located on the left side, and null samples are on the right side. The top row (A and B) are the uncorrected data objects. The middle row (C and D) are the OSC corrected data with two components, and the bottom row (E and F) are the PC-OSC corrected data with two components. The components were determined automatically. The axes for parts C−F are identical.

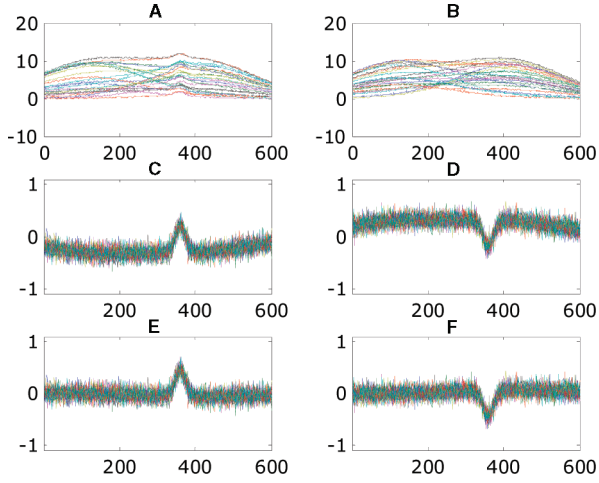


**Figure 3.** Underdetermined synthetic data set with random broad background peaks located at 120 and 420. Positive signals (peak at 360) are located on the left side and null samples are on the right side. The top row (A and B) are the uncorrected data objects. The middle row (C and D) are the OSC corrected data with two components, and the bottom row (E and F) are the PC-OSC corrected data with 2 components. The components were determined automatically. The axes for parts C−F are identical.

random deviates 0−10. Once again, a normal random deviate with a standard deviation of 0.1 was added to each data point in the data set. The positive and negative data objects can be observed in parts A and B of Figure 2, respectively.

The third data set was designed in an identical procedure as the underdetermined set described in the paragraph above except that the two background Gaussian peaks that were located at 120 and 420 had much larger standard deviations of 200 points. The positive and negative data objects can be observed in parts A and B of Figure 3, respectively.

**Table 1. Number of Olive Oil Samples for Each Geographic Region**

| registered designation of origin | number of samples |
|---|---|
| Aix-en-Provence | 104 |
| Haute-Provence | 50 |
| Nice | 57 |
| Nîmes | 39 |
| Nyons | 49 |
| Vallée des Baux de Provence | 112 |

**Virgin Olive Oil Samples.** A total of 411 commercial virgin olive oil samples (Table 1) were obtained from the French Inter-Professional Olive Oil association (AFIDOL), Aix-en-Provence, France, and from the Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes (DGCCRF) laboratory, Marseille, France. Sampling was carried out during five successive crops (2003/2004−2007/2008). The RDOs are principally made up of primary and secondary cultivars and also local and old varieties. “Haute-Provence”, “Nice”, “Nimes”, and “Nyons” are made up of one unique principal cultivar. Aix-en-Provence and Vallée des Baux have up to three or four principal cultivars of which at least two do not have their proportions specified.

**Near Infrared Spectroscopy.** FT-NIR spectra were recorded with a Nicolet Antaris FT-NIR spectrometer (Thermo Fisher Scientific, Inc., Waltham, MA) interfaced to a personal computer. Oil samples were filled into a 2 mm path length quartz cell directly sampled from the bottle without any chemical treatment. All the spectra were digitized at 4 $cm^{-1}$ resolution between 4500 and 10 000 $cm^{-1}$ using the integrated Thermo Nicolet software version 2.1. Coaddition of phase-corrected interferograms of 10 scans was performed for each spectrum. A reference spectrum was recorded before each sample measurement of the empty cell. The spectra were collected at a temperature of 18.0 °C.

**Mid-Infrared Spectroscopy.** Mid-Infrared spectra of each VOO sample were obtained using a Nicolet Avatar FT-IR spectrometer (Thermo Fisher Scientific, Inc., Waltham, MA) equipped with a DTGS detector, an Ever-Glo source, and a KBr/germanium beam splitter. Samples were deposited without preparation on an attenuated total reflection (ATR) cell equipped with a diamond crystal. Spectra were digitized from 600 to 4000 $cm^{-1}$ at a nominal resolution of 4 $cm^{-1}$. A total of 64 scans were coadded for each spectrum. The clean ATR plate was sampled to furnish a background spectrum before each sample. After each measurement, the ATR plate was cleaned by scrubbing with ethanol solution, which made it possible to dry the ATR. Cleanliness was verified by collecting a background spectrum and comparing it to the other background spectra. The spectra were collected at a temperature of 18.0 °C.

All evaluations were run under the Windows XP Pro x64 SP2 (Microsoft, Redmond, WA) operating system and in MATLAB 2008b x64 (The MathWorks, Natick, MA). The computer was a home-built Intel Core I7 CPU at 3.0 GHz with 12 GB of DDR3 random access memory.

## DISCUSSION OF RESULTS

**Synthetic Data.** The algorithms were tested using three synthetically generated data sets. The first set is the overdeter-

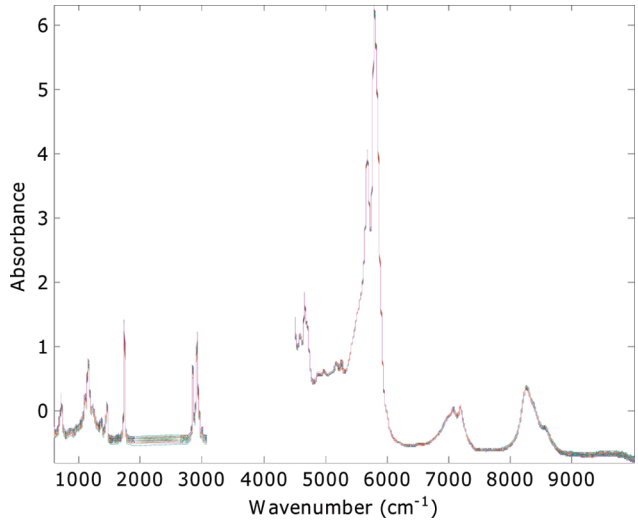


**Figure 4.** The 411 olive oil fused (IR 599−3081 $cm^{-1}$; NIR 4491−9999 $cm^{-1}$) absorbance spectra.

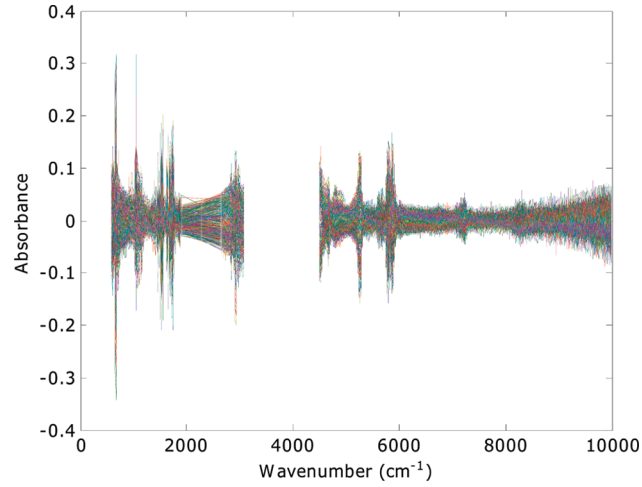


**Figure 5.** The 411 olive oil fused (IR 599−3081 $cm^{-1}$; NIR 4491−9999 $cm^{-1}$) absorbance spectra after preprocessing by multiplicative signal correction, normalization to unit vector length, and subtracting the mean followed by data fusion.

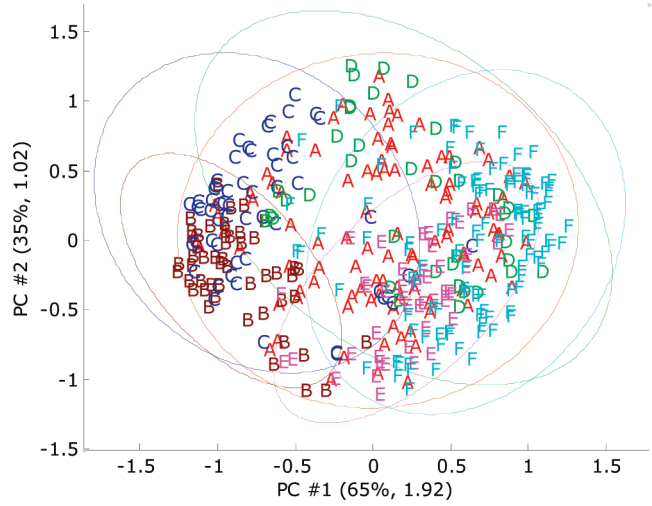


**Figure 6.** Principal component score plot of the fused preprocessed spectra: A, Aix-en-Provence; B, Haute-Provence; C, Nice; D, Nîmes; E, Nyons; and F, Vallée des Baux de Provence. The ellipses indicate 95% confidence intervals that define each class.

mined data set and comprised 600 objects and 50 variables. The top row of Figure 1 gives the uncorrected positive and negative spectra that contain or lack the analytical peak, respectively. For this case, both signal correction methods calculated the expected number of background components of 2, because there were two independently varying peaks that contribute to the background. In Figure 1C−F, a negative analytical peak is obtained in the corrected OSC and PC-OSC spectra of the negative objects because both methods rely on removing the mean or average from the data. The mean of the data set is a constant background that does not contribute any information to the measurement.

The underdetermined case is much more interesting because with the improved resolving power of modern instrumentation, data sets often have many more variables than objects. In this case, there are 50 objects each with 600 variables. The top row of Figure 2 gives the 25 positive and negative objects on the left and right, respectively. Both methods determined that there were two components.

In Figure 1C,D, after careful examination of the OSC corrected objects, one can see artifacts located at positions 10 and 35, where the background variations occurred. These artifacts are more prominent in Figures 2 and 3.

Although aesthetically displeasing, the artifacts will not affect later modeling, and note that the systematic variations are removed. Especially for data that are underdetermined, chance correlations may exist between the analytical peaks and the background fluctuations. These random correlations are included in the OSC models. A synthetic prediction set was calculated in the same manner as before except with different random deviates for the noise and background components. Signal correction of the prediction sets yielded similar trends with baseline artifacts appearing in the OSC objects. However, these artifacts in the OSC corrected spectra may be problematic when one attempts to interpret the model and make causal inferences.

**Fused Olive Oil Spectra.** The spectra before preprocessing are given in Figure 4 and after preprocessing in Figure 5. The unprocessed fused spectra had average PDR values and 95% confidence intervals across 10 bootstraps were 0.177 ± 0.002 for the minimum resolution value and 0.431 ± 0.001 for the geometric mean. One can see the improvement that arises from preprocessing: 0.502 ± 0.009 and 0.740 ± 0.003, for the minimum and geometric means, respectively. The principal component spectral scores from the preprocessed data are given in Figure 6. The 95% confidence intervals of the class averages are given as ellipses.

The classifiers were applied directly to the preprocessed spectra. For PLS-DA, the model-building data was split into two equal sizes using Latin partitions. Each subset was used once for prediction and once for model-building. The sum of the squared prediction errors was calculated with respect to the number of latent variables for each prediction set and stored. The procedure was bootstrapped five times. The errors were averaged with respect to the 10 evaluations, and the number of latent variables was selected that yielded the lowest average error.

For FuRES and PLS-DA, the principal component compression was applied to the model-building data. The prediction data obtained were projected onto the same components that were calculated from model-building data. Initial studies examined the

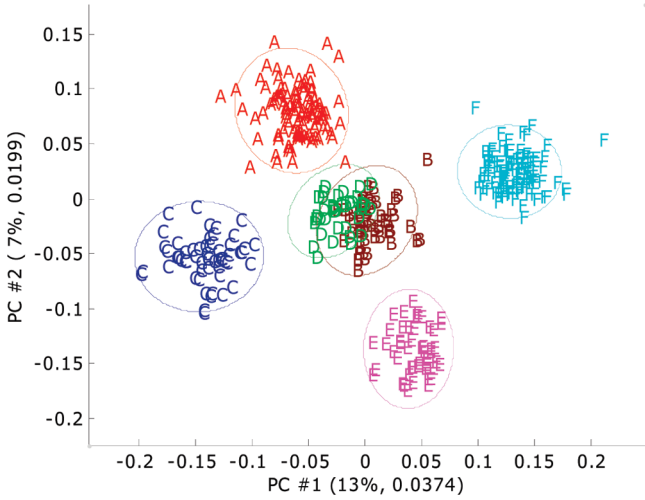


**Figure 7.** Principal component score plot after PC-OSC with 40 components: A, Aix-en-Provence; B, Haute-Provence; C, Nice; D, Nîmes; E, Nyons; and F, Vallée des Baux de Provence. The ellipses are the 95% confidence intervals about the class averages.

effect of the number of components, and the classification trees constructed between 60 and 120 components were constant. FuRES has similar temperature constraints as a temperature constrained cascade correlation network (TCCCN). Chen et al. has shown that the TCCCN models are robust and are not influenced by additional components.[28] FuRES models exhibited the same property so the number of principal components was set to 60 for compression which accounted for 27% of the variance. This 27% proportion is a relatively small amount of variance, so by using OSC or PC-OSC a larger proportion of signal would be retained with the 60 components. For example, by using PC-OSC with 40 components to correct the data in Figure 7, the PDR increased to 3.03 ± 0.02 and 4.544 ± 0.007 for the minimum and geometric resolution values, respectively, so that all classes are resolved. Note that the scores of classes B and D appear to be overlapped when projected onto the subspace defined by the two principal components in Figure 7 but are resolved in the higher dimensional data space. The advantage of the PDR is that this limitation of two- and three-dimensional renderings of PC score plots is overcome, and the separation is quantified in the full data space and not limited to the two dimensions that are displayed in Figure 7.

To automate the classical OSC and PC-OSC algorithms, the bootstrap Latin partition method was used to optimize the number of components so that the average PDR of the prediction sets was maximized. The model-building data was passed into the function and split into two subsets. Figures 8 and 9 give the PDR with respect to component number for the OSC and PC-OSC methods, respectively. In Figure 8, the OSC is much more efficient at removing uncorrelated variations from the model-building set than the PC-OSC method in Figure 9; however, this efficiency has a cost in the ability to correct the prediction set. With PC-OSC, a much larger number of components are required to obtain the same PDR value for the model-building set, and the improvement of the PDR for the prediction set is significantly greater. The PC-OSC is a softer or more variance-based approach to signal correction and therefore being less efficient allows the correction to be fine-tuned and yields improved predictions. The extra components remove embedded noise from the model-building data allowing faster convergence and better models when they are constructed later.

The next study compares the PDR with respect to the basis set number with the entire data set. Once again, 10 bootstraps were used with 2 Latin partitions. The 20 average PDRs are compared with respect to component number in Figure 10. For 60 components, the OSC spectra yielded a PDR of 1.8 ± 0.1 and the PC-OSC spectra yielded a PDR of 3.1 ± 0.1, so the effect of the PC-OSC on prediction data is significantly better once the optimum number of components has been obtained. With respect to the PDR, additional components do not cause problems for either of the OSC methods. The extra components remove uncorrelated noise from the model-building data which could benefit model construction in terms of faster convergence and the resulting classifiers in terms of better generalization of the models.

In terms of computation, the PC-OSC method is an order of magnitude faster than the OSC method. The difference is that OSC uses the nonlinear iterative partial least-squares algorithm (NIPALS) and the PC-OSC method uses singular value decomposition (SVD) that is computationally efficient and numerically stable.

The olive oil spectra were classified by geographical region. The geographical regions and number of samples from each region are given in Table 1. The classification was evaluated by using 10 bootstraps and 3 Latin partitions. As a reminder, three Latin partitions will divide the spectra into three mutually exclusive prediction sets so that each spectrum is contained once and only once and the class distributions will be proportional among the three prediction sets. Two of these prediction sets are combined into a model-building set, and the other is used for prediction. This procedure was accomplished three times per bootstrap so that each prediction set was used once for prediction and twice for model building.

The OSC models were built only with the model-building data using the automated procedure described in the Theory section. The corrected model-building data were then used to calculate 60 principal components to compress the data set, which is required by the FuRES method. The OSC corrected prediction sets were then projected onto the principal components from the model-building set to accomplish the predictions. The FuRES and PLS-DA methods were evaluated side by side in that the same model-building and prediction sets were used to compare between the classification methods. Five modes of classification were conducted. The first compared automated PLS1-DA and PLS2-DA algorithms with no signal correction or principal component compression as a control method. Then optimized PLS1-DA and PLS2-DA were run as positively biased controls. The third through fifth studies evaluated no OSC, OSC, and PC-OSC, respectively.

The classification results of these studies are given in Table 2. The first two columns give the PLS-DA classifications using a separate PLS model to classify each class or geographic region (PLS1-DA) and one PLS model to classify all geographic regions PLS2-DA. These PLS methods used no principal component compression. The same bootstrap Latin partition approach was used to select the number of latent variables in the PLS models

(28) Chen, P.; Harrington, P. B. *Appl. Spectrosc.* **2008**, *62*, 133–141.

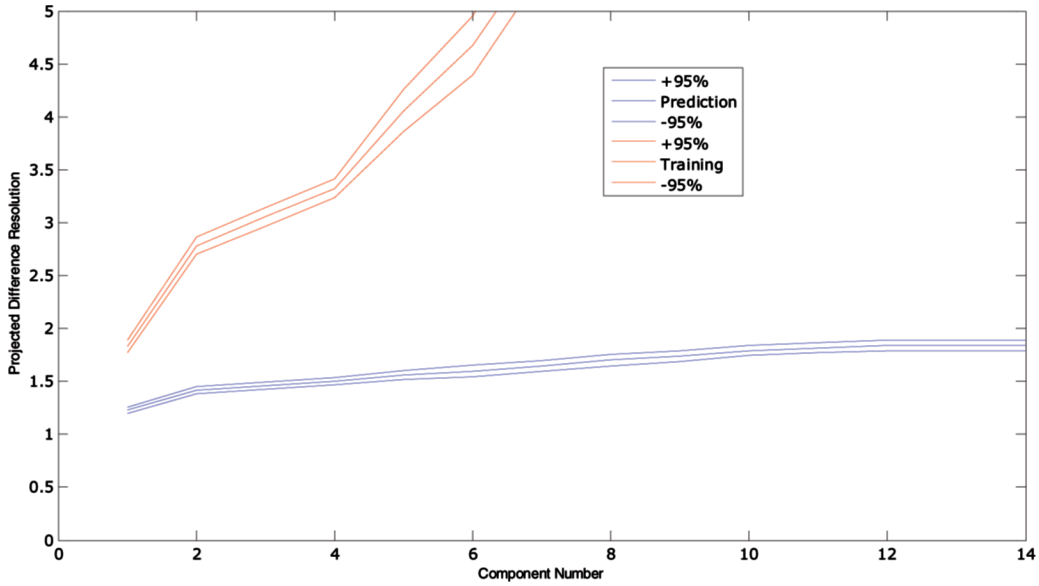


**Figure 8.** The effect of number of OSC components on the training or model-building set (red) and the prediction set (blue) average PDRs and 95% confidence intervals from 10 bootstraps and 2 Latin partitions.

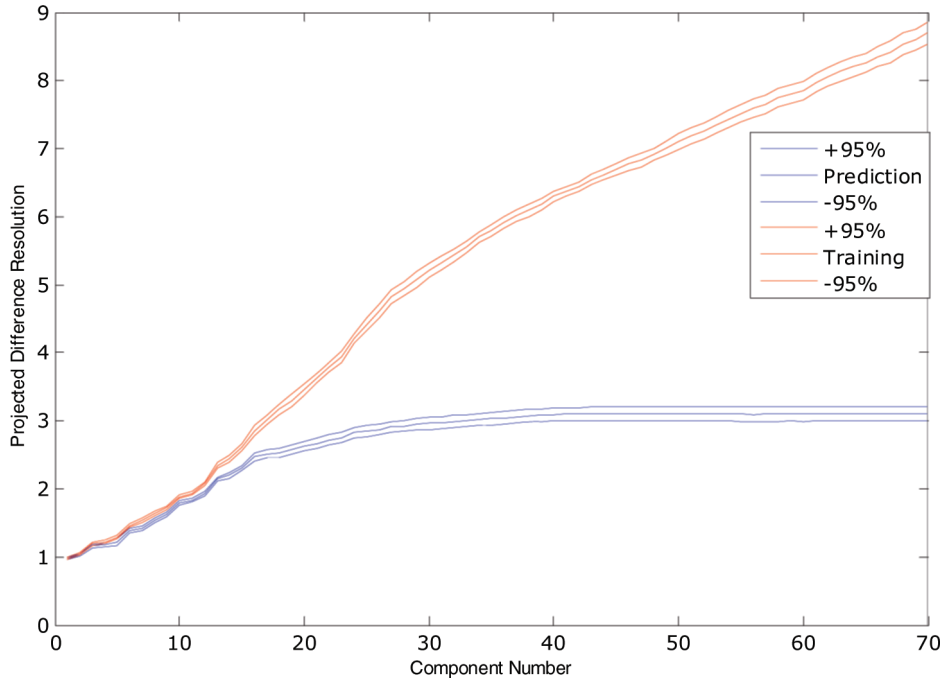


**Figure 9.** The effect of number of PC-OSC components on the model-building or training set (red) and the prediction set (blue) average PDRs and 95% confidence intervals from 10 bootstraps and 2 Latin partitions.

by splitting the model-building data in two subsets and pooling the prediction results. The number of components was selected that yielded the lowest prediction error sum of squares for the pooled results.

In Table 2, the results indicate that the PLS-DA methods applied directly to the data are effective but also computationally inefficient. The prediction results for the automated PLS-DA algorithms agree perfectly with the biased results that were obtained by selecting the number of latent variables with the lowest prediction error. In these studies, bootstrapping was never used. These biased classification accuracies are used as references to estimate the upper bound for the prediction accuracy.

PLS2-DA had a marginal loss in prediction accuracy when the data were compressed to 60 components. It was hoped that by using OSC, a basis set of the 60 components would yield improved predictions, and the results improved so they were perfect. However, the OSC correction had no effect on the FuRES predictions. The PC-OSC method did not diminish the PLS2-DA predictions but significantly improved the FuRES predictions.

A further test was to compress the PC-OSC corrected data using 120 components to investigate the effect of component number on predictions. The results were 410.3 ± 0.8 and 410.9 ± 0.2 for FuRES and PLS-DA, respectively. The FuRES results did

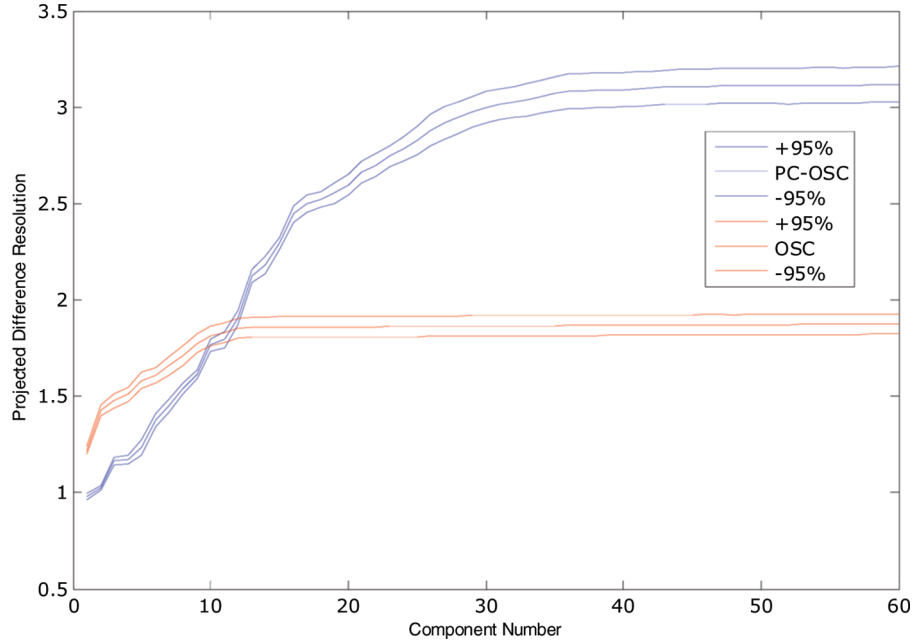


**Figure 10.** A comparison of the average PDR and 95% confidence intervals from 10 bootstraps and 2 Latin partitions of the corrected prediction sets using PC-OSC (blue) and OSC (red).

**Table 2. Comparison of the Average Number of Correct Classifications of Olive Oils to Geographic Regions for 411 Spectra in the Evaluations for 10 Bootstraps and 3 Latin Partitions[a]**

| no compression | | | compressed onto 60 PCs | | | |
|---|---|---|---|---|---|---|
| | no OSC | best | | uncorrected | OSC | PC-OSC |
| PLS1-DA | 411 ± 0 | 410.9 ± 0.2 | FuRES | 397 ± 2 | 400 ± 2 | 410.1 ± 0.5 |
| PLS2-DA | 411 ± 0 | 410.9 ± 0.2 | PLS2-DA | 410.6 ± 0.4 | 411 ± 0 | 411 ± 0 |
| ANOVA2 | ND | ND | | SD | SD | SD |
| interaction | ND | ND | | SD | SD | SD |
| CPU time | 59:10:33 | 19:06:02 | | 2:44:36 | 93:35:16 | 11:23:37 |

[a] Two-way analysis of variance was applied with interaction to the prediction results for comparing the two rows above. ND is no difference, and SD is significant difference. Results are presented for uncompressed and spectra that were compressed by projection onto 60 principal components. CPU time is given in h:min:s.

not degrade significantly when the number of principal components was doubled. The CPU time was 40:07:24 (h:min:s), so clearly it is an advantage to reduce the number of variables and an advantage to preprocess the data with a signal correction method.

In addition, the evaluations were rerun but this time the number of bootstraps in the self-optimizing signal corrections were set to 1. Therefore, the model-building data was divided into two Latin partitions. Each subset was used to construct a signal correction model and applied to the other as the prediction set. The correct spectra were recombined, and the PDR was measured of the reconstituted set. These results are reported in Table 3.

To examine the effect of overfitting by OSC methods, the signal correction methods were used with manually configured components. OSC used 12 and 40 components, and PC-OSC used 40 and 60 components. The results of the evaluation are given in Table 4. The number of components was ascertained by visual inspection of Figures 8 and 9, respectively. One can see that excessive components do not influence either signal correction method detrimentally. For the PC-OSC method, it is interesting that the CPU time for using 40 components is longer than the time for 60 components. The time difference of modeling with different numbers of components for the PC-OSC method is insignificant because a full SVD calculates all the components, and the time difference of generating the reconstructed background from 40 and 60 components is small in comparison to the PLS-DA and FuRES calculations. A plausible explanation for the time decrease is that by having more components in the scatter correction model, uncorrelated noise is removed from the model-building data set thereby causing faster convergence for PLS-DA and FuRES.

**Table 3. Effect of the Self-Optimizing OSC Routines without Bootstrapping[a]**

| compressed onto 60 PCs one bootstrap | | |
|---|---|---|
| | OSC | PC-OSC |
| FuRES | 389 ± 3 | 398 ± 1 |
| PLS2-DA | 410.4 ± 0.6 | 410.7 ± 0.3 |
| ANOVA2 | SD | ND |
| interaction | SD | SD |
| CPU time | 9:52:33 | 3:33:16 |

[a] The number of components maximized the PDR of the combined set of corrected spectra. The classifiers were evaluated with 10 bootstraps and 3 Latin partitions. ND is no difference and SD is significant difference.

**Table 4. Effect of Extra Components on OSC and PC-OSC Models[a]**

| | | compressed onto 60 PCs | | |
|---|---|---|---|---|
| | OSC | OSC | PC-OSC | PC-OSC |
| components | 12 | 60 | 40 | 60 |
| FuRES | 398 ± 4 | 401 ± 3 | 410.6 ± 0.4 | 410.7 ± 0.3 |
| PLS2-DA | 410.9 ± 0.2 | 410.8 ± 0.3 | 410.8 ± 0.4 | 410.9 ± 0.2 |
| ANOVA2 | SD | SD | ND | ND |
| interaction | SD | SD | ND | ND |
| CPU time | 2:56:50 | 11:31:38 | 2:36:27 | 2:05:31 |

[a] The classifiers were still evaluated with 10 bootstraps and 3 Latin partitions. CPU time is given in h:min:s. ND is no difference and SD is significant difference.

In summary, the PC-OSC method resulted in significant time-savings without compromising prediction accuracy. The FuRES method requires good background correction to remove unwanted variances. The PLS2-DA method works well without background correction but should compression be required to speed up the computational efficiency either scatter correction method works equally well. There is no need to spend much time optimizing the number of components because once an adequate number of components are used, the prediction error converges to a constant value.

## CONCLUSIONS

Background correction is important for soft or fuzzy methods of pattern recognition because variances in the independent variable space are important for model building. For data compression by the principal component projection, removal of unwanted variances can improve efficiency by allowing a smaller basis set of components to span more characteristic variances, thus improving speed.

A novel method was developed for signal correction. This principal component based method PC-OSC is not as efficient as the PLS version of OSC. However, it is computationally efficient (an order of magnitude faster to implement), which is useful for bootstrapping and statistically valid results. The inefficiency with regard to spanning background variances is also an advantage because the method allows fine-tuning of the background correction especially when an automated procedure is used. The PC-OSC method is a softer approach. Lastly, correlated background components do not appear in the signal corrected spectra with the PC-OSC method that may confound data interpretation and peak assignments.

Both OSC and PC-OSC methods were optimized by using the PDR metric. The metric was measured using bootstraps to create general models for background subtraction that would not affect the prediction accuracy. The results used 10 bootstrapped Latin partitions that compared favorably to not bootstrapping at all. The model-building data set was partitioned and evaluated inside the signal correction functions. If the removal of uncorrelated noise is desired, the number of components may be determined by the maximum of the averaged PDR values. If the OSC should only be removing correlated background components, then a threshold can be built using statistical confidence intervals calculated at the number of components that yielded the maximum PDR value and can be used to determine signal correction models with fewer components.

Both PLS1-DA and PLS2-DA algorithms were automated as well by using 10 bootstraps with 2 Latin partitions to calculate the average prediction error with respect to component number. The number of components that corresponded to the minimum average prediction error proved to build reliable models.

## ACKNOWLEDGMENT

The OHIO University Faculty Fellowship Leave program and The French Scientific Research National Center (CNRS) are thanked for funding Professor Harrington's visit to France. Yao Lu, Xiaobo Sun, Weiying Lu, and Zhanfeng Xu are thanked for their helpful comments. The authors are grateful to Christian Pinatel and Carole Fusari (Centre Technique de l'Olivier, Aix-en-Provence, France) and Oswin Galtier for the technical assistance.

Received for review March 13, 2009. Accepted July 17, 2009.

AC900538N