Vitale, R.; Bevilacqua, M.; Bucci, R.; Magrì, A.; Magri, A.; Marini, F. (2013). A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics. Chemometrics and Intelligent Laboratory Systems. 121:90-99. doi:10.1016/j.chemolab.2012.11.019.

A RAPID AND NON-INVASIVE METHOD FOR AUTHENTICATING THE ORIGIN OF PISTACHIO SAMPLES BY NIR SPECTROSCOPY AND CHEMOMETRICS

Raffaele Vitale, Marta Bevilacqua, Remo Bucci, Andrea D. Magrì, Antonio L. Magrì, Federico Marini*

Dept. Chemistry, University of Rome "La Sapienza", Rome, Italy.

***Corresponding author:**

dr. Federico Marini

Dept. of Chemistry

University of Rome "La Sapienza"

P.le Aldo Moro 5

I-00185 Rome

Italy

Tel +39 06 4991 3680

Fax +39 06 4457 050

e-mail: fmmonet@hotmail.com

A RAPID AND NON-INVASIVE METHOD FOR AUTHENTICATING THE ORIGIN OF PISTACHIO SAMPLES BY NIR SPECTROSCOPY AND CHEMOMETRICS

Raffaele Vitale, Marta Bevilacqua, Remo Bucci, Andrea D. Magrì, Antonio L. Magrì, Federico Marini

Dept. Chemistry, University of Rome "La Sapienza", Rome, Italy.

Abstract

In this study, near-infrared spectroscopy coupled to chemometrics is used to build an analytical protocol to authenticate the origin of pistachio nuts (Pistacia vera L.), a high value-added food product.

In particular, 483 samples from six different origins (Sicily, India, Iran, Syria, Turkey and U.S.A.) were analyzed by NIR spectroscopy. Spectra were recorded on half seeds cut longitudinally in reflectance mode. Spectral data were then processed by chemometrics to build classification models by SIMCA and PLS-DA. The discriminant approach resulted in classification accuracies higher than 90% for most of the classes. On the other hand, SIMCA built class-models with high sensitivity and specificities, the only exception being the two categories Turkey and Iran, whose heterogeneity resulted in a poorer specificity (anyway higher than 80%). In particular, the results obtained for the samples coming from Bronte (Sicily), the only PDO pistachio production in Europe – 95.5% non error rate in PLS-DA, 90% sensitivity and 97% specificity in SIMCA, as evaluated on the external test set – are very promising from the viewpoint of the authentication of this product.

In general, the results show that the coupling of NIR spectroscopy to chemometric classification techniques can be a valuable tool for tracing the origin of pistachio nuts, providing a reliable authentication in a rapid, relatively cheap and non invasive way.

Keywords: Pistachio (*Pistacia vera L.*) nuts, Near Infrared Spectroscopy (NIR), Classification, Partial Least Squares-Discriminant Analysis (PLS-DA), Soft Independent Modeling of Class Analogies (SIMCA).

1. Introduction

Pistachio (*Pistacia vera L.*) is a nut having peculiar organoleptic characteristics. It is widely consumed as a raw or toasted snack or ingredient of many desserts, ice cream, cakes, pastry and for the production of some sausages such as mortadella [1]. The genus *Pistacia L.* is a member of the

Anacardiaceae family and consists of at least 11 species. Among these species, *Pistacia vera* is the only cultivated and economically important one [2]. It can grow in dry and hot areas and under saline conditions [3]. Because of its marked resistance to extreme environmental (pedoclimatic and hydrologic) conditions, it is cultivated in Europe and Asia on soils that are unsuitable for other fruit crops [4].

The *Pistacia vera* tree is native to arid zones of central and west Asia [5]. Nowadays, only a few major growing areas exist worldwide: the principal pistachio-producing countries of the world are, in order, Iran, USA (California), Turkey, Syria, but to lesser extent, other countries, such as, Italy and India [2,3,6], cultivate pistachios as well [7].

In Italy, only a single variety (Bianca) is grown [4,8] and its cultivation is concentrated mainly in Bronte, an area around the Etna volcano, where the lava and climate allow the production of an intensely green nut with a very aromatic taste that is highly prized on the international markets. Italian production is very poor in comparison to Asian and American ones; however, it is compensated by the very high quality of the final products [5,9]. Moreover, fee rates and national laws on commodities of each producing country vary dramatically [7,10]. So, pistachios variation in quality, food safety (*e.g.*, contamination by aflatoxins), import/export fees, legal implications, and financial concerns makes determining the country of origin for pistachios important to protect the consumers against potential fraud [7,10,11]. Moreover, given the fundamental economic implications of any fraud, not only the consumers but also pistachio producers and traders are moved to discover objective chemical techniques that can confirm food labels identifying geographic indications [7,10]. As a consequence, there is the need to develop analytical procedures, which can provide a reliable authentication of the geographical origin of this product.

In this framework, some works have already been published in the literature concerning the possibility of differentiating the geographical origin of samples coming from various producing countries, using different chemical indices and analytical techniques. For example, Dyszel & Pettit used the triglycerol profile determined by HPLC and the areas of some DSC peaks to discriminate nuts coming from California from the Iranian and Turkish ones [11]. Furthermore, Anderson & Smith proposed the use of stable isotope analysis to distinguish pistachio samples from the three main growing areas (USA, Asia and Mediterranean countries) [10,12]. Other researches from different groups hypothesized that the variations in the fatty acid composition of pistachio nuts, determined by various techniques (HPLC, GC, DSC, NMR), could be related to the different geographical origin of the product [1,7,11]. Lastly, differences in the profiles of inorganic anions, organic acids, and in color among pistachios of different origins and varieties have also been reported in the literature [4,12].

On the other hand, there is a great number of researches where the problem of assessing the authenticity of a wide range of other food commodities, and in particular the problem of tracing their geographical origin, is tackled and solved by the use of near infrared spectroscopy (NIR) coupled with the application of chemometric classification methods for data processing. The possibility of using this spectroscopic technique to address problems connected to the authentication of foodstuff has attracted extensive attention by scientists due to its being rapid, relatively cheap and non-polluting, characteristics which perfectly fit the concept of "green analytical chemistry" [13]. Moreover, in many cases, the use of NIR spectroscopy allows the operator to analyze samples without the need to perform any previous chemical or physical treatment [14]. In the framework of the authentication of food origin, these methods have already been successfully applied, for example, to discriminate geographical origin of olive oils [15-19], meat [20], cheese [21], and honey [22]. However, to our knowledge this approach has never been tried before to trace the origin of pistachio nut samples.

Therefore, the aim of the present work is to investigate the possibility of using NIR spectroscopy coupled to chemometric classification methods to build a rapid, relatively cheap and non-invasive analytical procedure for the authentication of the geographical origin of pistachio samples and, in particular, for the recognition of the PDO samples from Bronte (Italy). To this purpose, an experimental setup allowing the spectroscopic determinations to be carried out directly on the nuts, without any sample pretreatment steps was designed. Furthermore, from a data processing standpoint, both a discriminant and class-modeling chemometric approaches (by means of the algorithms PLS-DA and SIMCA, respectively) were used. Special care was also taken in the choice of the suitable method of signal spectral pretreatment prior to the construction of the models.

2. Materials and methods

2.1 Samples

Pistachio nut samples from 6 different countries – the four main producers (Iran, USA, Turkey, and Syria) and two of the smaller but still relevant ones (India and Italy) – were collected and analyzed. In particular, the Italian samples were all coming from the Protected Designation of Origin (PDO) "Pistacchio Verde di Bronte" (Bronte, Sicily). In all cases, samples were obtained from different sources and suppliers, chosen to be as representative as possible of the different production areas. Pistachio nuts were stored in a refrigerator at 4 °C and protected from light until the day prior to analysis, to prevent any kind of surface modification and photodegradation of their molecular constituents. In total, 483 pistachio samples - 41 from Bronte (Italy), 41 from India, 121 from Iran, 40 from Syria, 120 from Turkey, 120 from USA (California) - were analyzed.

122

123 *2.2 Acquisition of NIR spectra*

124 For the acquisition of spectra a Nicolet 6700 FT-NIR instrument (Thermo Scientific Inc., Madison,
125 WI), equipped with a tungsten-halogen source and an InGaAs detector, was used. The signals were
126 recorded between 10000 and 4000 cm^{-1}, collecting 82 scans at a nominal resolution of 4 cm^{-1}. All
127 the spectra were acquired at room temperature in the interval on individual pistachio nuts, without
128 any further sample treatment, in reflectance mode, through the use of an integrating sphere (Thermo
129 Scientific Inc., Madison, WI). Operationally, each nut was split in two by a longitudinal cut, so that
130 a flat surface was produced on both halves. Two different NIR spectra were then recorded on each
131 half nut, aligning the pistachio first parallel and then perpendicular to the axis of the optical slit of
132 the integrating sphere, and the four spectra corresponding to a single sample (two halves at two
133 orientations) were averaged prior to the successive elaboration. The data were then exported from
134 Omnicare Suite software (Thermo Fisher Scientific Inc., Waltham, MA) as ASCII files, which were
135 then imported into MATLAB (release R2011a, The MathWorks Inc., Natick, MA), for the
136 successive chemometric analysis. In the data analytical stage, 7 different signal pre-processing
137 techniques were evaluated and compared: MSC (Multiplicative Scatter Correction) [23,24],
138 detrending [25], first and second derivatives, computed according to the Savitzky-Golay method (15
139 points window and third-degree interpolating polynomial) [26], and the combinations of MSC with
140 each of the other three; the possibility of no pretreatment was taken into account, too.

141

142 *2.3 Statistical data analysis*

143 Since the aim of this study is to develop a method to predict the geographical origin of pistachio
144 nuts and, in particular, to build a traceability model for the PDO "Pistacchio verde di Bronte", the
145 measured NIR data were processed by statistical pattern recognition techniques. In particular, two
146 different techniques were chosen, PLS-DA [27,28] and SIMCA [29,30], as examples of
147 discriminant and class-modeling approaches, respectively. Discriminant techniques focus on the
148 differences between samples coming from different classes and operate by dividing the hyperspace
149 of the variables in as many regions as the number of available categories, while class-modeling
150 techniques are rather focused on the similarities among samples of the same class than on the
151 differences among the classes and act by defining the category space of one class at a time.

152

153 *2.3.1 Partial least squares-discriminant analysis (PLS-DA) [27,28]*

154 Building a classification model can be viewed as finding the best relationship between a
155 multivariate independent matrix **X**, whose i^{th} row contains the spectral fingerprint recorded on the

i^{th} sample, and a qualitative vector of responses. Accordingly, if a suitably designed dummy response matrix **Y** is introduced, traditional regression methods can be used also to tackle with classification problems. In particular, when dealing with a classification problem involving *m* classes, each training sample is associated with a dummy binary-coded *m*-dimensional **y** vector having all entries equal to zero except for the component corresponding to the class the sample belongs to, which is equal to 1. For instance, in a problem involving 6 classes, like the one considered in the present study, samples belonging to the first category will be described by the dependent vector [1 0 0 0 0 0], samples belonging to the second by the vector [0 1 0 0 0 0] and so on. Under these assumptions, it is possible to use traditional regression methods to operate classification, computing a calibration model relating the matrix of predictors and this dummy matrix of responses. As the name itself suggests, the core of the PLS-DA approach is the use of Partial Least Squares regression [31], which operates a bilinear decomposition of both the X- and Y-spaces, under the assumption that a relationship between the two internal spaces exists, to compute the model parameters. The result is a linear classifier that has proved to be statistically equivalent to Linear Discriminant Analysis (LDA, [32]), but that is also applicable to all the cases when LDA cannot be used (low number of samples with respect to variables and/or correlated indices) [27].

In order to interpret the results in terms of the most significant spectral regions, it is important to check which of the measured variables contribute the most to the definition of the model. In the case of PLS-based techniques, this kind of information can be summarized in an index called Variable Importance in Projection (VIP [33]), a value that expresses whether a predictor is significant in the definition of the *F* latent vectors model for the prediction of a particular response. Mathematically, it is defined according to the formula:

$$VIP_j = \sqrt{N_{vars} \frac{\sum_{k=1}^{F} \left(b_k^2 \mathbf{t}_k^T \mathbf{t}\right)\left(w_{jk} / \|\mathbf{w}_k\|\right)^2}{\sum_{k=1}^{F} \left(b_k^2 \mathbf{t}_k^T \mathbf{t}\right)}} \qquad (1)$$

where $\mathbf{t}_k$ is the vector of sample scores along the k^{th} latent variable, $\mathbf{b}_k$ is the coefficient of the k^{th} PLS inner relationship, N_{vars} is the number of experimental variables and w_{jk} and $\mathbf{w}_k$ are the weight of the j^{th} variable for the k^{th} LV and the weight vector for the k^{th} LV, respectively. Since the average of squared VIP scores equals 1, 'greater than one rule' is generally used as a criterion to identify the most significant variables. Interpretation of the results can be further improved by inspection of the regression coefficients of the PLS model which, if opportunely examined, can indicate whether the

values of the different variables measured for samples coming from a specified category are higher or lower than those recorded on samples from all the other classes.

2.3.2 Soft Independent Modeling of Class Analogies (SIMCA) [29,30]

As stated above, the core of the class modeling approach is that each category is modeled independently on the others. In particular, SIMCA describes each class on the basis of a principal component model of opportune dimensionality, according to the equation:

$$\mathbf{X}_i = \mathbf{T}_A \mathbf{P}_A^T + \mathbf{E} \qquad (2)$$

where $\mathbf{X}_i$ is the sub-matrix of the original data set obtained by selecting only the samples from the i^{th} class, $\mathbf{T}_A$ and $\mathbf{P}_A$ are the matrices containing the first A scores and loading vectors, respectively, and $\mathbf{E}$ is the matrix of the residuals. Once the principal component model is computed, the class space is defined according to some statistically defined criterion for outlier detection. In particular, two statistical variables are used to express the degree of outlyingness of a sample with respect to the computed principal component model: T^2 which accounts for the distance of the sample within the model space and Q which represents its distance from the model space. The values of these two statistics for the analyzed samples are estimated from the scores matrix $\mathbf{T}_A$ and the residual matrix $\mathbf{E}$, respectively. The distance between each sample and the model of a category is then computed as "reduced distance" according to the equation:

$$d_i = \sqrt{\left(T^2_{red,i}\right)^2 + \left(Q_{red,i}\right)^2} = \sqrt{\left(\frac{T^2_i}{T^2_{lim}}\right)^2 + \left(\frac{Q_i}{Q_{lim}}\right)^2} \qquad (3)$$

where T^2_{lim} and Q_{lim} are threshold values for the two statistics corresponding to a selected percentile of the distributions, usually 95%, under the null hypothesis. Commonly, if the reduced distance of a sample exceeds $\sqrt{2}$, the sample is considered as an outlier and rejected by the class model; otherwise, if the distance is lower than this value, it is accepted and recognized as being part of that class.

3. Results and discussion

As anticipated in the introductory section, aim of this study was to build and validate reliable classification models for the traceability of pistachio nuts, coming from 6 different countries,

coupling near-infrared spectroscopy and chemometrics. To this purpose, a set of 483 pistachio samples from the 6 investigated countries was collected and analyzed by NIR spectroscopy, as described in section 2.2; the corresponding spectra (after averaging the 4 signals measured for each sample) are reported in Figure 1.

To relate the spectral fingerprints to the origin of the samples, discriminant (PLS-DA) and modeling (SIMCA) classification approaches were used and compared. However, since several instrumental effects can hinder or worsen the performances of the classification models, different kinds of spectral pretreatment were tested. As mentioned before, the chosen pretreatments were Multiplicative Scatter Correction (MSC), detrending, first and second derivatives, and the combination of MSC with any of the latter three (with MSC being applied first as recommended by Rinnan *et al.* [34]). Models built after these preprocessings were also compared with the ones calculated from raw spectra.

As, when dealing with supervised methods, validation of the models on an independent test set is of paramount importance to unbiasedly assess their predictive ability and performances, the whole data set made of 483 samples was then divided into training and test sets (the former to build the models, the latter to validate them). In order to maintain the same diversity in both sets, a sample splitting scheme based on the Duplex algorithm [35] was adopted. Indeed, the Duplex algorithm starts selecting the two objects in the data matrix that are farthest away from each other according to their Euclidean distance and putting them into the training set. Then, among the remaining candidates, the two objects farthest from each other are put into the test subset. At the next step, consecutive objects are selected and put alternatively in the training and test sets, the object added being the one farthest away from all the objects of the data matrix already selected in the considered set. To determine which object is the farthest one, a so-called maximin criterion, which is the same as in the Kennard and Stone algorithm [36], is used: the Euclidean distance between each candidate object and its closest neighbor already in the considered subset is computed and the object for which this distance is maximal is added.

In order to ensure that each of the 6 classes was adequately represented, the selection was performed separately for each category. Moreover, to account for the fact that different pretreatments had to be tested and that as much as possible of the variation after scatter or baseline removal was covered in the selection, at the same time having a unique sample splitting scheme to be able to compare the outcomes after the different preprocessings, a procedure recently designed by our group for another study [19] was adopted. In detail, Duplex algorithm was applied separately on each of the 8 data matrices corresponding to the different pretreatments (7 preprocessings and the raw spectra). The selection was performed class-wise using a 2:1 training:test splitting ratio and

working on the principal component representation of the data matrices (considering 20 PCs per category). Accordingly, the frequency of selection of each sample as part of the test set was computed so that, eventually, all the individuals selected more than 50% of the times (i.e. at least 5 times out of 8) were included in the final test set (made, in total, of 163 samples: 19 from Bronte, 18 from India, 38 from Iran, 15 from Syria, 34 from Turkey, 39 from USA). All the remaining 320 samples constituted the training set (22 from Bronte, 23 from India, 83 from Iran, 25 from Syria, 86 from Turkey, 81 from USA). The effectiveness of the splitting procedure in keeping a comparable diversity among the two sets can be graphically evaluated in Figure 2, where the projection of the training and the test samples onto the space spanned by the first two principal components is shown for the different spectral preprocessing. It is evident that the proposed procedure allows to select test samples spanning the same space of the training objects, irrespectively of the pretreatment considered, so that the chosen splitting scheme can be used to properly compare the results obtained with the different approaches.

3.1 PLS-DA analysis

In a first stage of our study, classification models were built according to a discriminant approach using the PLS-DA algorithm. In particular, each of the 8 data sets corresponding to the different spectral pretreatments was processed individually, after mean centering. In each case, the optimal complexity of the classification models was chosen as the one which led to the minimum overall classification error in cross-validation (10 cancelation groups). The results are summarized in Table 1, where the correct classification rates for each of the 6 classes and the overall one in calibration and cross-validation are reported for the different spectral preprocessing. It can be observed from the Table that MSC followed by detrending is the pretreatment leading to the best results in cross-validation, the corresponding PLS-DA model (built using 18 LVs) resulting in a classification error of 4.48% in calibration and 5.77% in cross-validation. Moreover, investigation on the classification accuracy for the different classes shows that the model built on data pretreated by MSC and detrending results in non error rates in cross-validation higher than 93% with the only exception of Iran, for which it is anyway slightly less than 90%.

This optimal model was then validated on the external test set, and the results are reported in Figure 3, where the values of the 6 components of the predicted **y** vector are reported for each sample. Assignation of an unknown sample to one of the 6 investigated categories is made based on the values of this predicted response vector: the sample is predicted to belong to the category corresponding to the highest value of the component. For the sake of easier visualization, horizontal

lines were added to the Figure to indicate the y values above which a sample was assigned to the particular category. It can be seen that the optimal PLS-DA model built on the training set was able to correctly predict the country of origin of most of the validation samples, thus confirming the effectiveness of the approach. The validation results are also reported in Table 2 in terms of non error rate in prediction both on individual categories and on the whole test set. These results indicate that the class belonging of unknown samples can be predicted very accurately (prediction error is lower than 10% for all categories with the only exception of Iran, analogously to what already observed in cross-validation).

The goodness of what resulted to be the best PLS-DA classification model, i.e. the one built on spectral data pretreated with MSC and detrending, can also be graphically visualized in Figure 4, where the projection of the training (cross-validated scores) and test samples onto the space spanned by the first three PLS-DA latent variables is displayed. It is apparent from the Figure that the test samples lie well into the space spanned by the training ones. Moreover, it is also possible to observe relatively well the grouping of samples from the different categories even if, due to the complexity of the model (18 LVs), it is difficult to appreciate the separation between the classes on this three dimensional representation.

In order to identify the spectral frequencies which contribute the most to the discriminant model, VIP scores [33] were computed and examined: as already described in section 2.3.1, VIP is an index accounting for the contribution of individual experimental variables to the bilinear model and it is scaled in such a way that indices having VIP larger than 1 are considered to be significant. Information on the VIP score was integrated with the examination of the regression coefficient for the interpretation of the model. Indeed, even if the presence of nonorthogonal contributions to the signal can perturb the shape of the regression coefficient vector so that it no longer looks like the pure spectrum [37], inspection of its values, although not straightforward, still can provide useful information. Accordingly, for the sake of interpretation variable significance estimated by VIP scores and the values of the regression coefficients were graphically represented in Figure 5, superimposed to the average spectral profile recorded on the samples after the optimal preprocessing. It can be seen from the Figure that the features identified as relevant by the model (which are also summarized in Table 3) correspond to spectrally meaningful frequencies, and that while most of the spectral intervals are common to all the investigated categories, there are also some significant differences. In particular, the spectral regions that appear to be relevant for all the categories involve the peaks at around 4500–5000 cm^{-1}, which may be attributed to combination bands of C-C and C-H stretching vibration, the signals between 5650 and 6000 cm^{-1}, due to the combination bands and first overtone of C-H bonds, and those between 7074 and 7180 cm^{-1}, which

can be ascribed to C-H bonds combination band. On the other hand, portions of the band between 8000 and 9000 cm^{-1} (second overtone of methylenic stretching vibrations) are significant only in the definition of the categories Turkey and USA.

3.2 SIMCA analysis

In a second stage, class-modeling approach was also used to process the same data set. Indeed, even if very satisfactory results were obtained using PLS-DA, as described in the previous sub-section, the asymmetry in the number of available samples for each category together with the difference in heterogeneity of the classes due, for instance, to the uneven geographical distribution of the productive areas in the investigated countries (production in Bronte being limited to a very narrow area compared to e.g. Iran, where numerous cultivation sites scattered over the country exist), could be the cause of the non perfect classification rates obtained. Moreover, in problems like those concerning the traceability of foodstuff, where the question to be answered is "Is the product coming from country X as declared?", one is more interested in assessing whether the investigated sample is compatible with the model of a specific category, which is exactly what class-modeling does. In this investigation, class-modeling on the data set containing the spectral fingerprints of pistachio samples was performed by means of the SIMCA algorithm. Accordingly, independent class models were built for each of the 6 investigated categories, whose optimal complexity was chosen as the number of principal components corresponding to the highest geometrical average of sensitivity and specificity in 6-fold row-wise cross-validation. As in the case of PLS-DA, the effect of the spectral pretreatments on the classification ability was evaluated, by comparing the cross-validated results of SIMCA modeling on the 8 matrices corresponding to the different preprocessings considered. Consistently with what already observed for the discriminant approach, also with SIMCA the best spectral preprocessing resulted to be the coupling of MSC with detrending, which allowed to achieve the highest values of sensitivity and specificity for all the categories. Indeed, most of the class models result in a perfect sensitivity and a rather good specificity in calibration while, in the cross-validation phase, sensitivity decreases significantly for many of the considered pretreatments. Besides, irrespectively of the pretreatment, the models for the categories Iran and Turkey have a significantly lower specificity than the others. This could be explained by the fact that the samples coming from these two regions are produced in various areas scattered in the country, far away from one another and characterized by different climates, environmental conditions and latitude values: to take into account this heterogeneity, keeping reasonable values of sensitivity, the class models have to be wide at the expense of specificity. This hypothesis was also supported by the observation that, for each of the two countries (in particular,

for Iran), there is a marked difference among the spectral fingerprints of the producing regions represented in the data set. When the models built after the best spectral pretreatment (MSC+detrending) were applied to the external test set, very good results were obtained (Table 4). Indeed, sensitivity and specificity values for the validation samples were over 80% for most of the categories (the only exceptions being the sensitivities of Syria and Turkey models), being in many cases even higher than 90%. These results can be graphically evaluated in Figure 6, where the projections of the training (cross-validated predictions) and test samples onto the model space of each single category are reported. This representation allows to easily visualize which samples are accepted or rejected by the different class models: the dotted lines in the graphs correspond to the threshold values of reduced distance, below which the samples are accepted by the model of the considered category, as described in Section 2.3.2. The Figure shows clearly that all the models are very sensitive both in cross-validation and on the external test set but that some of them, particularly Turkey and Iran, have lower specificity.

When more than one category is modeled, it is possible to check whether the samples are accepted by one, more than one or none of them. This could be useful in order to turn SIMCA into a discriminant classifier by assigning each sample to the category it is closer to. This information can be easily visualized building a so-called Coomans plot [38], a graph where the two axes represent the distance of the samples to each of the two class models under study. As an example, **Error! Reference source not found.** shows the Coomans plot for the Bronte and USA class models built on the spectral data pretreated by MSC + detrending algorithms. The dotted black lines correspond to the threshold distances values (in our case $\sqrt{2}$) and cut the plot in four different regions: the uppermost left and the lowermost right will correspond to unambiguous acceptance by a single category model (respectively Bronte and USA), the lowermost left to acceptance by both classes while the uppermost right to rejection by both category models. Most of the samples coming from the two different geographical areas lye inside the space of the corresponding class model. Only few samples coming from the other four zones were accepted by the two different class models, while the remaining samples are found in the uppermost right region of the plot. Moreover, none of the samples are accepted by both models. The diagonal line bisecting the plot represents the discriminant classification boundary so that all the samples lying above it are classified as from Bronte, while all the samples lying below it are predicted as from USA. Based on these considerations, discriminant classification based on SIMCA models would result in 100% (Bronte) and 100% (USA) both in cross-validation and on the external test set.

3.3 A closer look on Bronte

As reported in the Introduction, one of the 6 classes investigated in this study, "Pistacchio verde di Bronte", is the only pistachio product with a Protected Designation of Origin. Therefore, from the standpoint of the analytical control of frauds on certified foodstuff it would be of outmost importance to have an accurate method for checking the authenticity of this product. In this respect, the results of this study appear quite promising, both if the discriminant or if the class-modeling approach are concerned. Indeed, by looking at the results of PLS-DA reported in Section 3.1 one could see that for the class Bronte a non-error rate in prediction higher than 95% was obtained. Additionally, by inspecting the plots in Figure 3 it can be observed that only a very low number of samples from other categories are erroneously predicted as belonging to Bronte.

As far as SIMCA is concerned, the results in Table 3 indicate that the model of the category Bronte has very high sensitivity and specificity in prediction, as requested for a reliable traceability model.

4. Conclusions

In this study, the potential of NIR spectroscopy coupled to chemometric discriminant and class-modeling pattern recognition techniques for the traceability of pistachio nuts samples was demonstrated. Classification models with high accuracy were built to recognize the geographical origin of the samples, as evaluated on an external test set. The outcomes for both the different classification approaches are very satisfying: the origin of over 95% of validation samples was correctly predicted using PLS-DA and these results were confirmed by SIMCA modeling of the same data, which allowed to build very sensitive and highly specific models for authenticating the provenance of pistachio nuts. In particular, the results obtained for the category Bronte, which is the only type of pistachio having a Protected Denomination of Origin, appear really promising in the light of the possibility of building a traceability model for this product.

Furthermore, a first attempt of interpreting the observed differences in terms of significant spectral bands by means of the inspection of VIP scores was made: discriminant information was found to be associated to meaningful signals corresponding to methylene overtones and C-H and C=C combination bands.

Finally, it can be concluded that NIR spectroscopy coupled to chemometric classification techniques is a powerful tool to trace pistachio nuts samples, allowing a fast, cheap and non-invasive/non-destructive analysis.

References

[1] E. Arena, S. Campisi, B. Fallico, E. Maccarone, Distribution of fatty acids and phytosterols as a criterion to discriminate geographic origin of pistachio seeds, Food Chem. 104 (2007) 403-408.

[2] S. Kafkas, H. Ozkan, B.E. Ak, I. Akar, H.S. Atli, S. Koyuncu, Detecting DNA polymorphism and genetic diversity in a wide pistachio germplasm: comparison of AFLP, ISSR and RAPD markers, J. Am. Soc. Hortic. Sci. 131 (2006) 522-529.

[3] E. Tsantili, C. Takidell, M. Christopoulos, E. Lambrinea, D. Rouskas, P. Roussos, Physical, compositional and sensory differences in nuts among pistachio (Pistachia vera L.) varieties, Sci. Hortic. 125 (2010) 562–568.

[4] M. Bellomi, B. Fallico, Anthocyanins, chlorophylls and xanthophylls in pistachio nuts (Pistacia vera) of different geographic origin, J. Food Compos. Anal. 20 (2007) 352-359.

[5] G. Ballistreri, E. Arena, B. Fallico, Characterization of triacylglycerols in Pistacia Vera L. oils from different geographic origins, Ital. J. Food Sci. 22 (2010) 69-75.

[6] Food and Agriculture Organization. Food and Agricultural commodities production, 2011. http://faostat.fao.org/site/339/default.aspx Accessed 22/04/12

[7] K. Zur, A. Heier, K.W. Blaas, C. Fauhl-Hassek, Authenticity control of pistachios based on 1H- and 13C-NMR spectroscopy and multivariate statistics, Eur. Food Res. Technol. 227 (2008) 969–977.

[8] A. Fabbri, C. Valenti, The sicilian pistachio industry: an overview, Acta Hort. 470 (1998) 43-49.

[9] L. Di Marco, Il pistacchio in Sicilia: situazione e prospettive, Agricoltura Ricerca 79 (1987) 9.

[10] K.A. Anderson, B.W. Smith, Effect of Season and variety on the differentiation of geographic growing origin of pistachios by stable isotope profiling, J. Agric. Food Chem. 54 (2006) 1747-1752.

[11] S.M. Dyszel, B.C. Pettit, Determination of the country of origin of pistachio nuts by DSC and HPLC, J. Am. Oil Chem. Soc. 67 (1990) 947–951.

[12] K.A. Anderson, B.W. Smith, Use of chemical profiling to differentiate geographic growing origin of raw pistachios, J. Agric. Food Chem. 53 (2005) 410-418.

[13] S. Armenta, S. Garrigues, M. De La Guardia, Green Analytical Chemisty, Trends Anal. Chem. 27 (2008) 497-511.

[14] D.H. Ma, X.C. Wang, L.P. Liu, Y. Liu, Current progress in food geographical origin traceability by near infrared spectroscopy technology, Guangpuxue Yu Guangpu Fenxi 31 (2011) 877-881.

[15] M. Casale, C. Casolino, P. Oliveri, M. Forina, The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil, Food Chem. 118 (2009) 163-170.

[16] T. Woodcock, G. Downey, C.P. O'Donnell, Confirmation of declared provenance of European extra virgin olive oil samples by NIR spectroscopy, J. Agric. Food Chem. 56 (2008) 11520-11525.
[17] N. Sinelli, E. Casiraghi, D. Tura, G. Downey, Characterization and classification of Italian virgin olive oils by near- and mid-infrared spectroscopy, J. Near Infrared Spectrosc. 16 (2008) 335-342.
[18] O. Galtier, N. Dupuy, Y. Le Dreau, D. Ollivier, C. Pinatel, J. Kister, J. Artaud, Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra, Anal. Chim. Acta 595 (2007) 136-144.
[19] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: A case study, Anal. Chim. Acta 717 (2012) 39-51.
[20] S.M. Sun, B.L. Guo, Y.M. Wei, M.T. Fan, Application of near infrared spectral fingerprint technique in lamb meat origin traceability, Guangpuxue Yu Guangpu Fenxi 31 (2011) 937-941.
[21] R. Karoui, A.M. Mouazen, E. Dufour, L. Pillonel, E. Schaller, J. De Baerdemaeker, J.O. Bosset, Chemical characterization of European Emmental cheeses by near infrared spectroscopy using chemometric tools, Int. Dairy J. 16 (2006) 1211-1217.
[22] T. Woodcock, G. Downey, J.D. Kelly, C. O'Donnell, Geographical classification of honey samples by near-infrared spectroscopy: a feasibility study, J. Agric. Food Chem. 55 (2007) 9128-9134.
[23] H. Martens, S.A. Jensen, P. Geladi, Multivariate linearity transformations for near infrared reflectance spectroscopy. In: Proc. Nordic Symposium on Applied Statistics, Stokkland Forlag, Stavanger, Norway, 1983, 205–234.
[24] P. Geladi, D. MacDougall, H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc. 39 (1985) 491-500.
[25] R.J. Barnes, M.S. Dhanoa, S. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772-777.
[26] A. Savitzky, M. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627-1639.
[27] S. Wold, C. Albano, W.J. Dunn III, K. Esbensen, S. Hellberg, E. Johansson, M. Sjöström, Pattern recognition: finding and using regularities in multivariate data, in: H. Martens, H. Russwurm Jr. (Eds.), Food research and data analysis, Elsevier Applied Science, Barking, UK, 2009, pp. 147-188.
[28] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166-173.

[29] S. Wold, Pattern recognition by means of disjoint principal components models, Pattern Recogn. 8 (1976) 127-139.
[30] S. Wold, M. Sjöström, SIMCA: a method for analysing chemical data in terms of similarity and analogy, in: B.R. Kowalski (Ed.), Chemometrics, theory and application, 52nd ed., American Chemical Society, Washington, DC, 1977, pp. 243-282.
[31] S. Wold, H. Martens, H. Wold, Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad. Ruhe, A., Kagstrom, B., Eds.; Springer-Verlag: Heidelberg, Germany, 1983; pp. 286-293.
[32] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Hum. Genet. 7 (1936) 179–188.
[33] S. Wold, E. Johansson, M. Cocchi, 3D QSAR, in: H. Kubinyi (Ed.), Drug Design: Theory, Methods and Applications, Escom Science Publishers, Leiden, The Netherlands, 1993, pp. 523–550.
[34] Å. Rinnan, F. van den Berg, S. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trends Anal. Chem. 28 (2009) 1201-1222.
[35] R. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415-428.
[36] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[37] M.B. Seasholtz, B. Kowalski, Qualitative information from multivariate calibration models. Appl. Spectrosc. 44 (1990) 1337-1348.
[38] D. Coomans, I. Broeckaert, M.P. Derde, A. Tassin, D.L. Massart, S. Wold, Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles, Comput. Biomed. Res. 17 (1984) 1-14.

Figure captions:

Figure 1 – Raw spectra of the 483 pistachio samples analyzed in this study (after averaging the 4 signals recorded on each nut).

Figure 2 - Representation of the data splitting between the training and test set as a function of the different spectral pretreatments. Data are projected on the space spanned by the first two principal components. (●) Bronte; (■) India; (♦) Iran; (▼) Syria; (▲) Turkey; (★) USA; empty symbols correspond to training samples and filled symbols correspond to the test samples.

Figure 3 – PLS-DA analysis after MSC+detrending: predicted values of the dummy vector components corresponding to the different categories for test set samples. Horizontal lines indicate the threshold above which a sample is assigned to that particular class. (●) Bronte; (■) India; (♦) Iran; (▼) Syria; (▲) Turkey; (★) USA.

Figure 4 – PLS-DA analysis after MSC+detrending: projection of the training (cross-validateed scores) and test samples onto the space spanned by the first three latent variables. (●) Bronte; (■) India; (♦) Iran; (▼) Syria; (▲) Turkey; (★) USA; empty symbols correspond to training samples and filled symbols correspond to the test samples.

Figure 5 – PLS-DA analysis after MSC+detrending: graphical representation of the regression vectors for the 6 categories superimposed to the average pretreated spectrum. The regression vector components are colored according to their VIP score: significant variables (VIP>1) are colored in red, while those estimated as not relevant in green).

Figure 6: SIMCA on near-infrared data after MSC+detrending: projection of the training (cross-validated estimates) and test samples onto the model spaces of the 6 investigated categories. (●) class Bronte; (■) class India; (♦) class Iran; (▼) class Syria; (▲) class Turkey; (★) class USA. Empty symbols correspond to training samples and filled symbols correspond to the test samples.

Figure 7: SIMCA on near-infrared data after MSC+detrending: Coomans plot comparing the models of classes Bronte and USA. (●) class Bronte; (■) class India; (♦) class Iran; (▼) class Syria; (▲) class Turkey; (★) class USA. Empty symbols correspond to training samples and filled symbols correspond to the test samples.

Table 1- PLS-DA analysis: comparison of the correct classification rates in calibration and cross-validation obtained after the different spectral pretreatments.

	LV	Bronte		India		Iran		Syria		Turkey		USA		Overall	
		Cal.	CV	Cal.	CV	Cal.	CV	Cal.	CV	Cal.	CV	Cal.	CV	Cal.	CV
No pretreatment	9	95.04%	95.04%	94.61%	89.59%	76.38%	76.59%	93.63%	93.46%	86.09%	85.3%	86.66%	86.25%	85.53%	84.90%
Detrending	7	94.37%	94.37%	74.12%	69.43%	82.95%	80.45%	91.63%	91.29%	90.8%	90.8%	90.22%	89.80%	87.73%	86.61%
1st derivative	7	96.14%	95.97%	86.73%	83.71%	80.84%	78.76%	90.78%	90.61%	87.74%	86.36%	87.72%	88.13%	86.69%	85.64%
2nd derivative	15	98.49%	97.82%	98.82%	94.31%	92.98%	89.91%	93.97%	91.46%	88.96%	84.99%	97.71%	96.67%	93.97%	91.28%
MSC	14	97.32%	94.87%	91.78%	90.94%	85.39%	83.97%	89.12%	86.95%	87.47%	85.45%	91.28%	88.58%	89.01%	87.02%
MSC+ detrending	18	97.48%	97.32%	96.82%	95.3%	90.54%	89.48%	96.47%	93.97%	95.17%	93.15%	99.79%	99.17%	95.52%	94.23%
MSC+ 1st derivative	15	97.82%	97.15%	94.14%	93.13%	94.46%	91.81%	95.63%	89.29%	86.95%	84.71%	97.93%	95.43%	93.62%	91.08%
MSC+ 2nd derivative	16	98.66%	93.61%	99.66%	96.82%	96.93%	93.64%	95.80%	91.63%	89.18%	87.22%	97.92%	96.05%	95.32%	92.59%

Table 2: PLS-DA analysis on NIR data after MSC + detrending pretreatment: modeling and validation results

Classes	Non error rate in calibration	Non error rate in cross-validation	Non error rate in prediction
Bronte	97.48%	97.32%	95.14%
India	96.82%	95.30%	90.29%
Iran	90.54%	89.48%	83.59%
Syria	96.47%	93.97%	93.63%
Turkey	95.17%	93.15%	91.71%
USA	99.79%	99.17%	99.19%

Table 3 – PLS-DA model: features identified as relevant for all the classes

Spectral Region	Vibrational modes
4500–5000 cm^{-1}	combination bands of C-C and C-H stretching vibration
5650–6000 cm^{-1}	combination bands and first overtone of C-H bonds
7074–7180 cm^{-1}	C-H bonds combination band
8000–9000 cm^{-1} *	second overtone of methylenic stretching vibrations

*Meaningful only for the categories Turkey and USA

Table 4: Results of SIMCA analysis on NIR data after MSC + detrending pretreatment for both training and test sets

Classes	LV	Calibration		Cross-validation		Prediction	
		Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Bronte	4	95.45%	95.64%	72.73%	95.30%	89.47%	96.53%
India	7	100.00%	90.57%	65.22%	93.60%	83.33%	98.62%
Iran	5	98.80%	68.35%	93.98%	67.93%	92.11%	76.80%
Syria	1	88.00%	88.81%	88.00%	87.46%	73.33%	82.43%
Turkey	13	95.35%	79.06%	84.88%	80.77%	73.53%	80.62%
USA	10	93.83%	100.00%	85.19%	100.00%	87.18%	99.19%

Figure 1

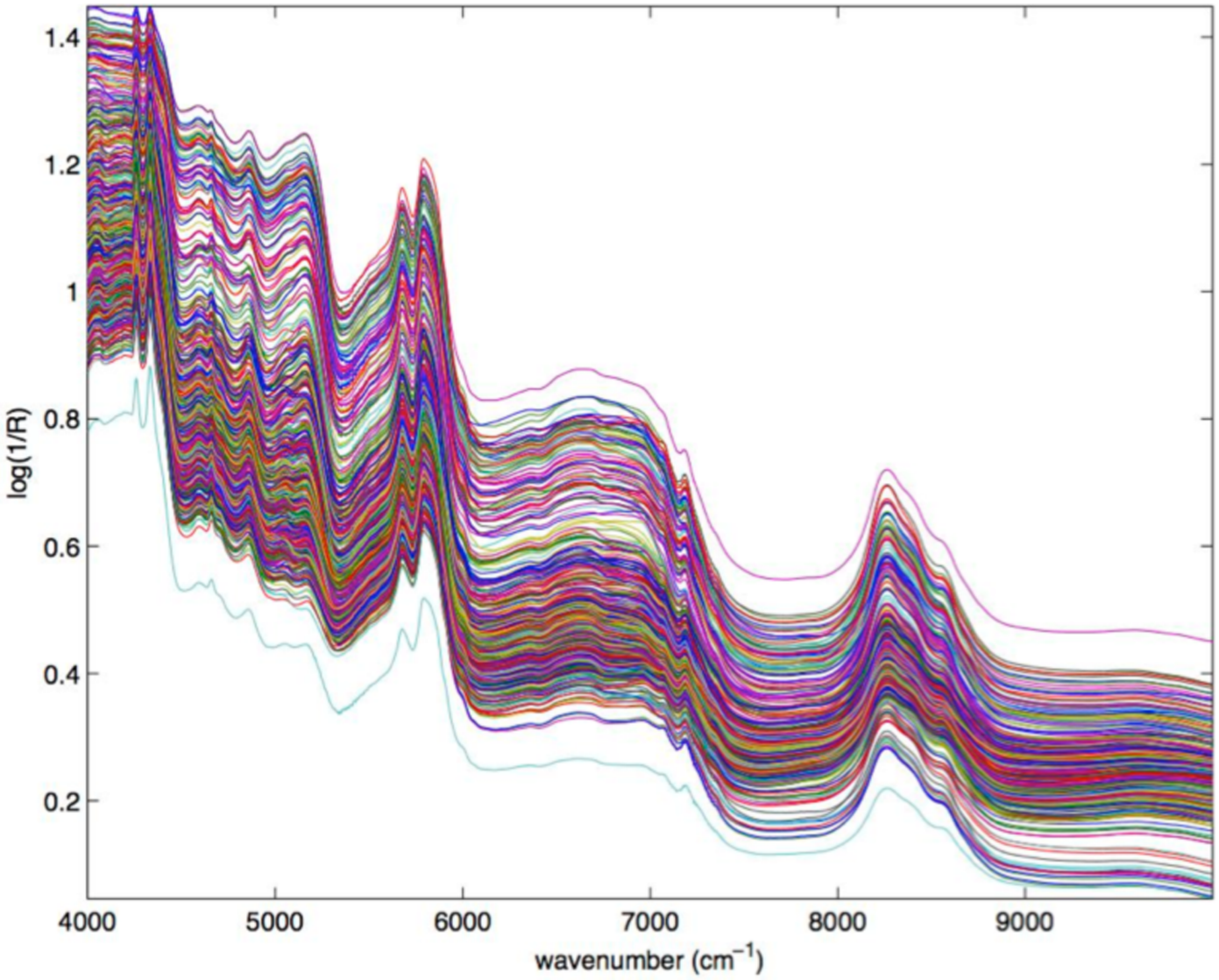

1.4
1.2
1
0.8
0.6
0.4
0.2
log(1/R)
4000
5000
6000
7000
8000
9000
wavenumber (cm^{-1})

Figure 2a

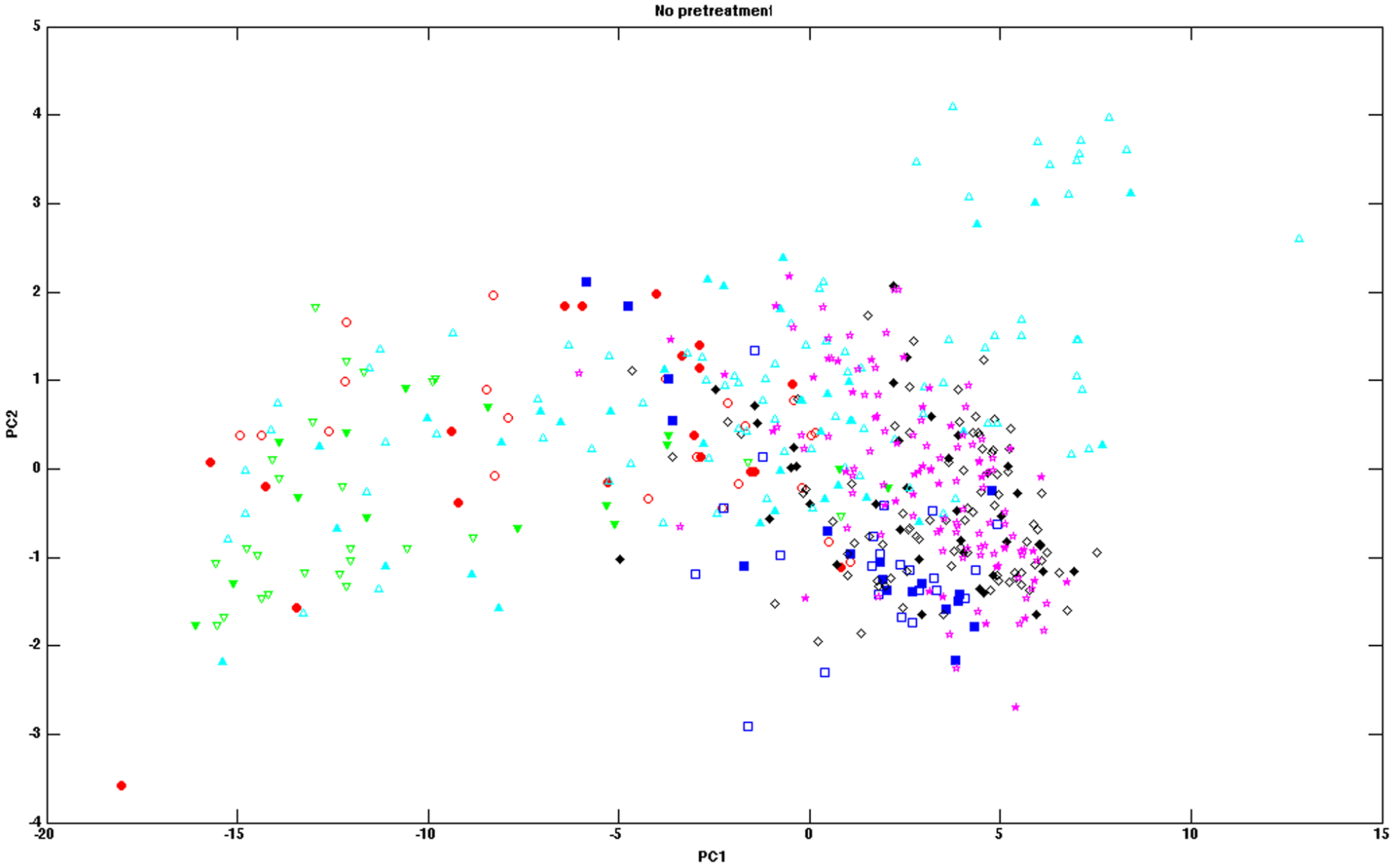


Figure 2b

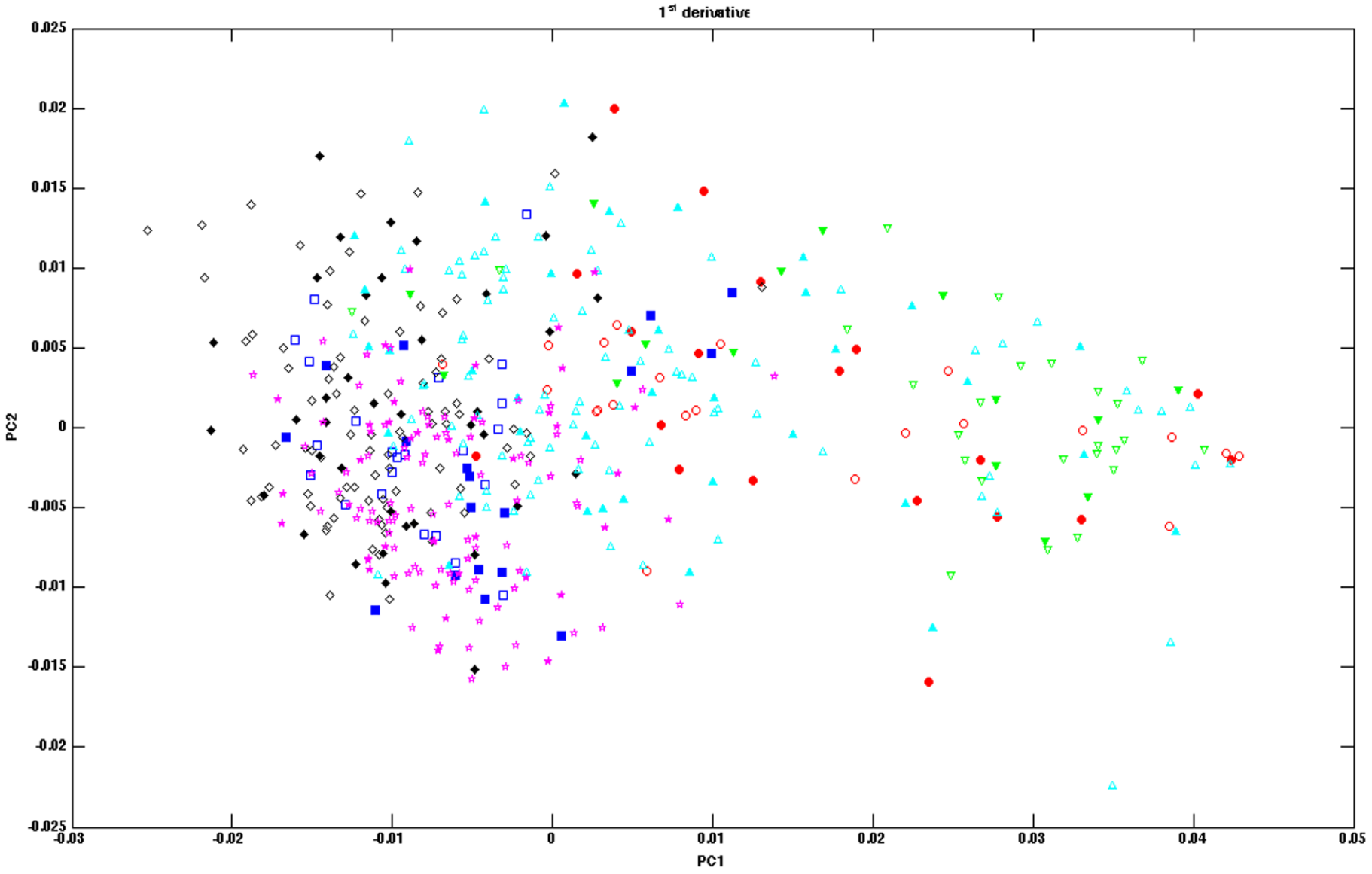

1st derivative
PC2
PC1
0.025
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
-0.025
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04
0.05

Figure 2c

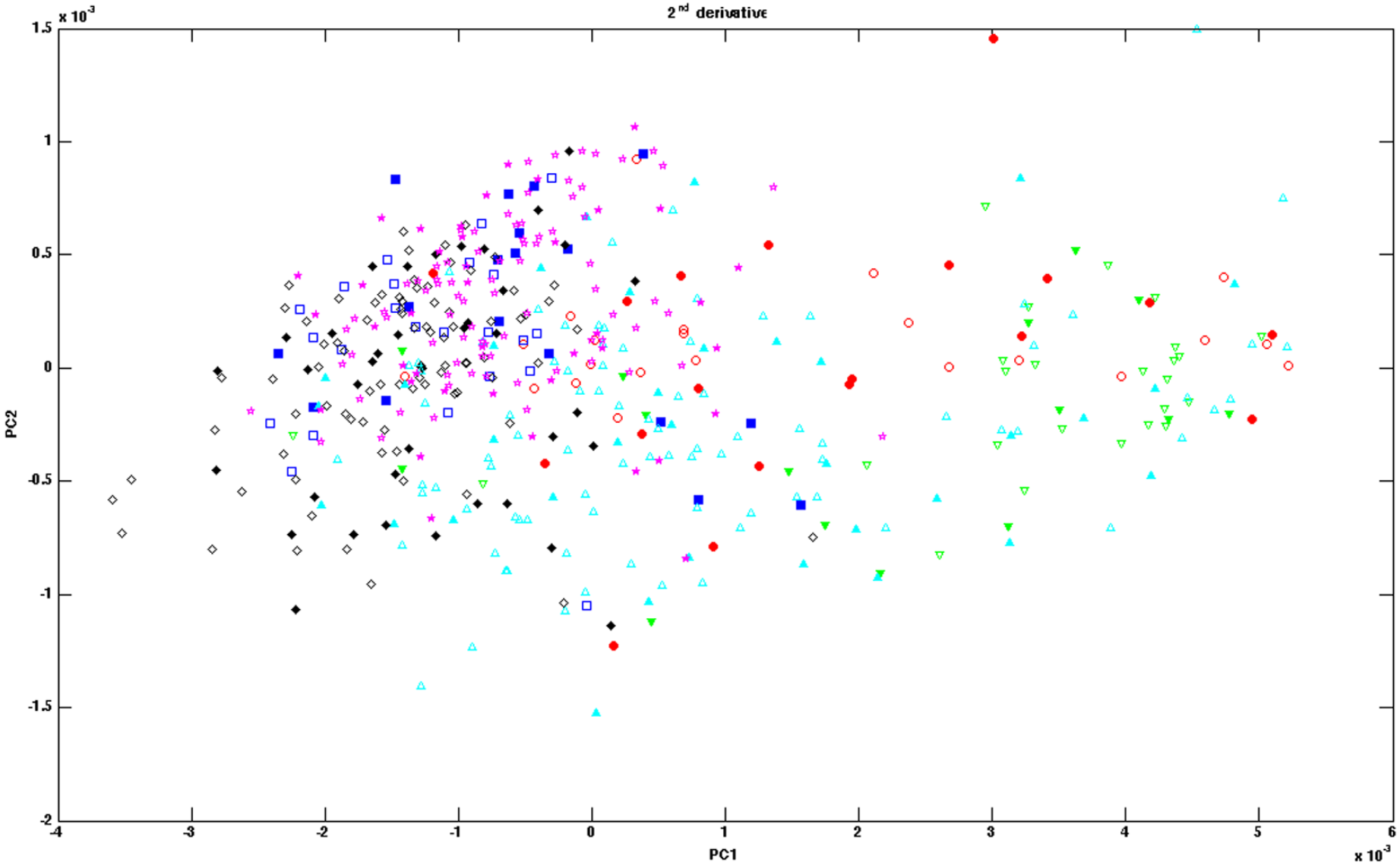

2nd derivative
x 10-3
1.5
1
0.5
0
-0.5
-1
-1.5
-2
PC2
-4
-3
-2
-1
0
1
2
3
4
5
6
PC1
x 10-3

Figure 2d

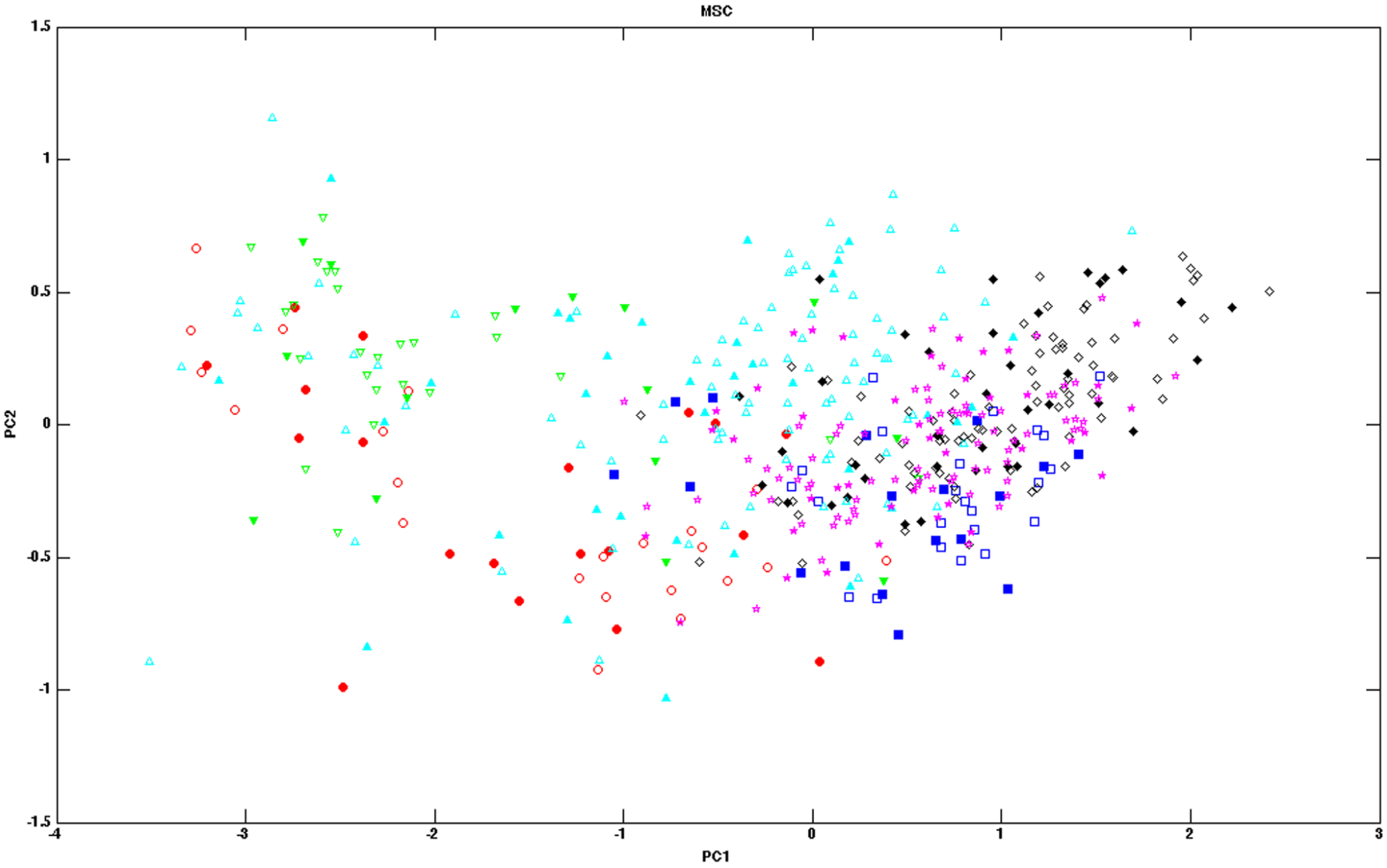


Figure 2e

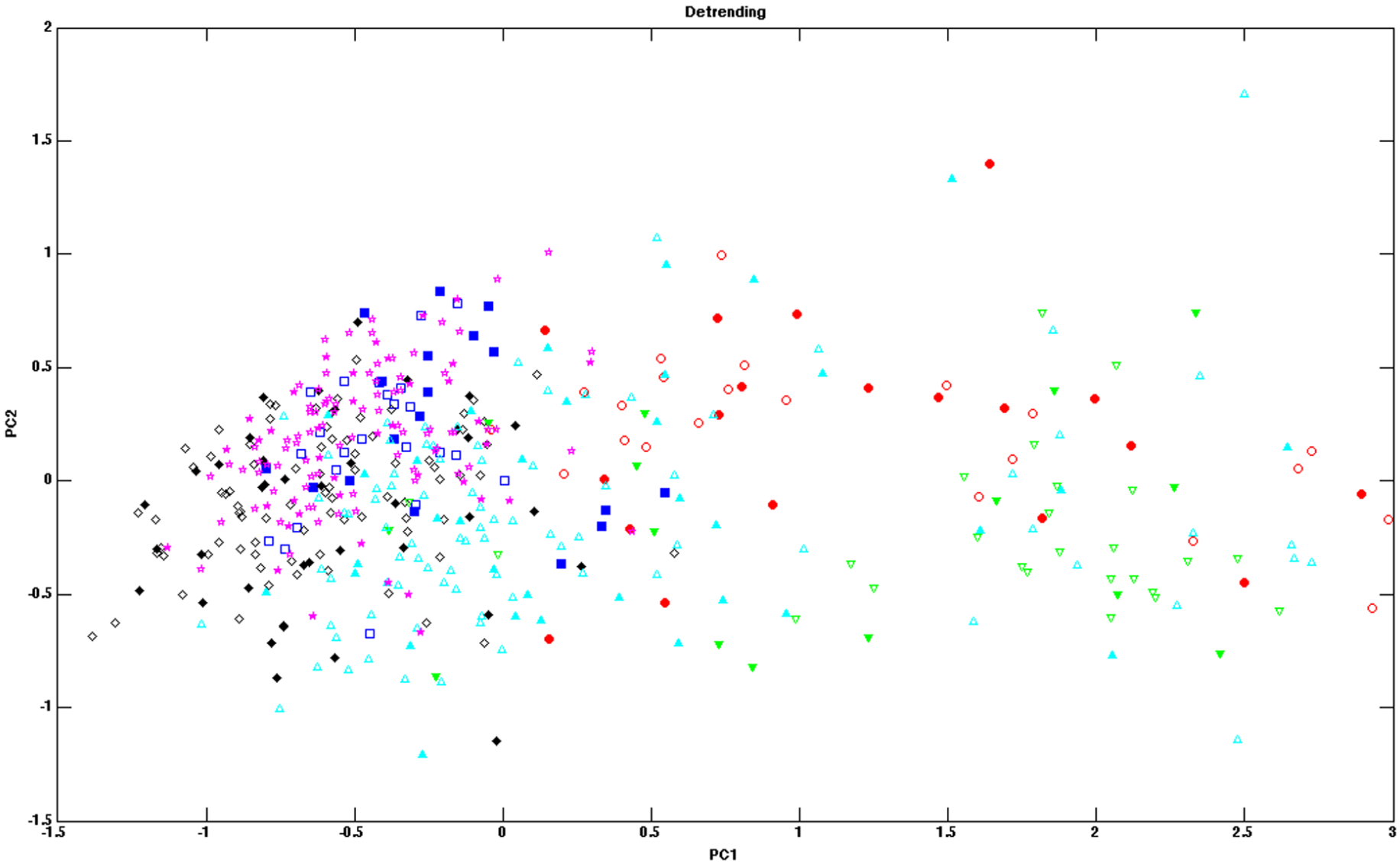


Figure 2f

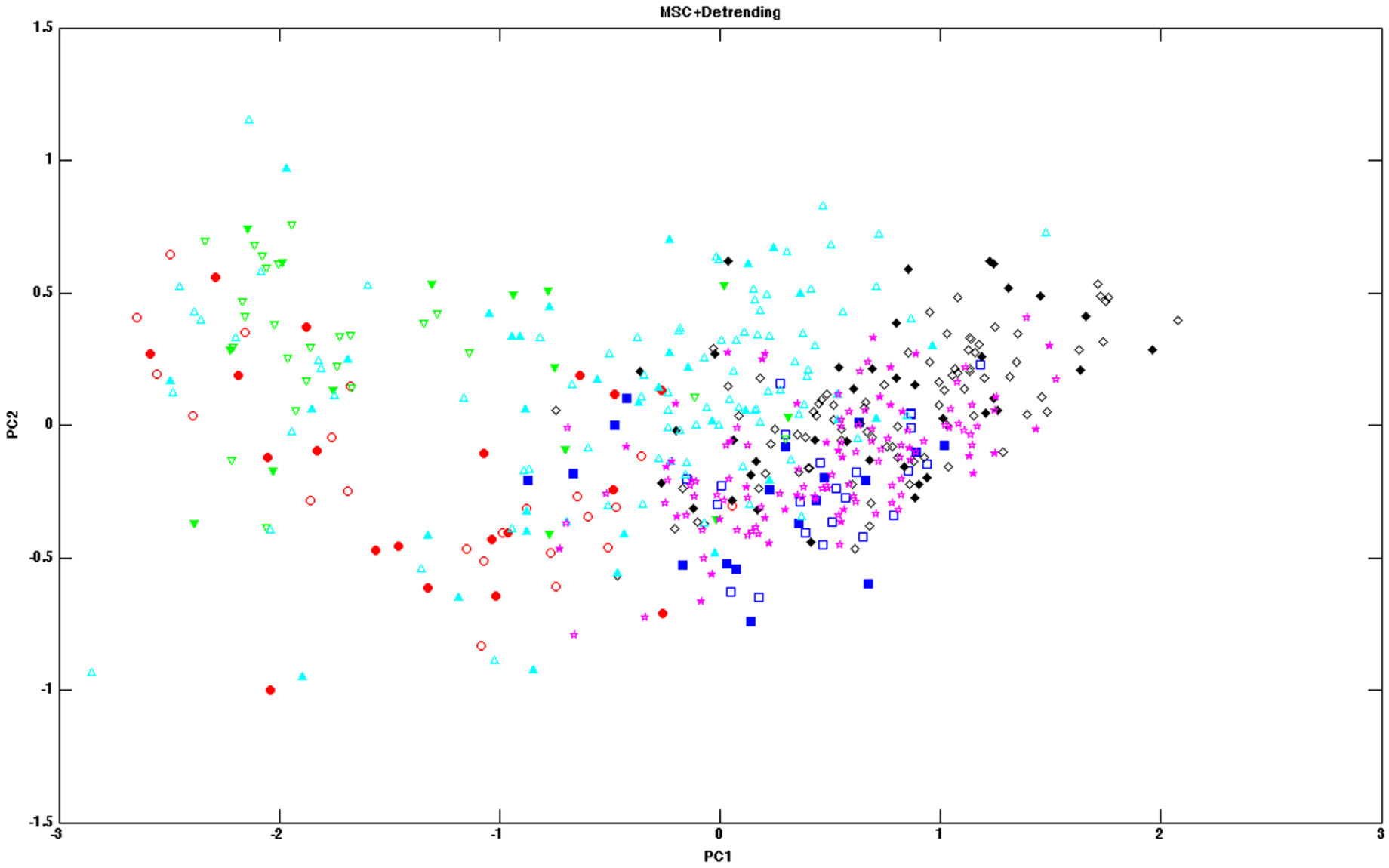


Figure 2g

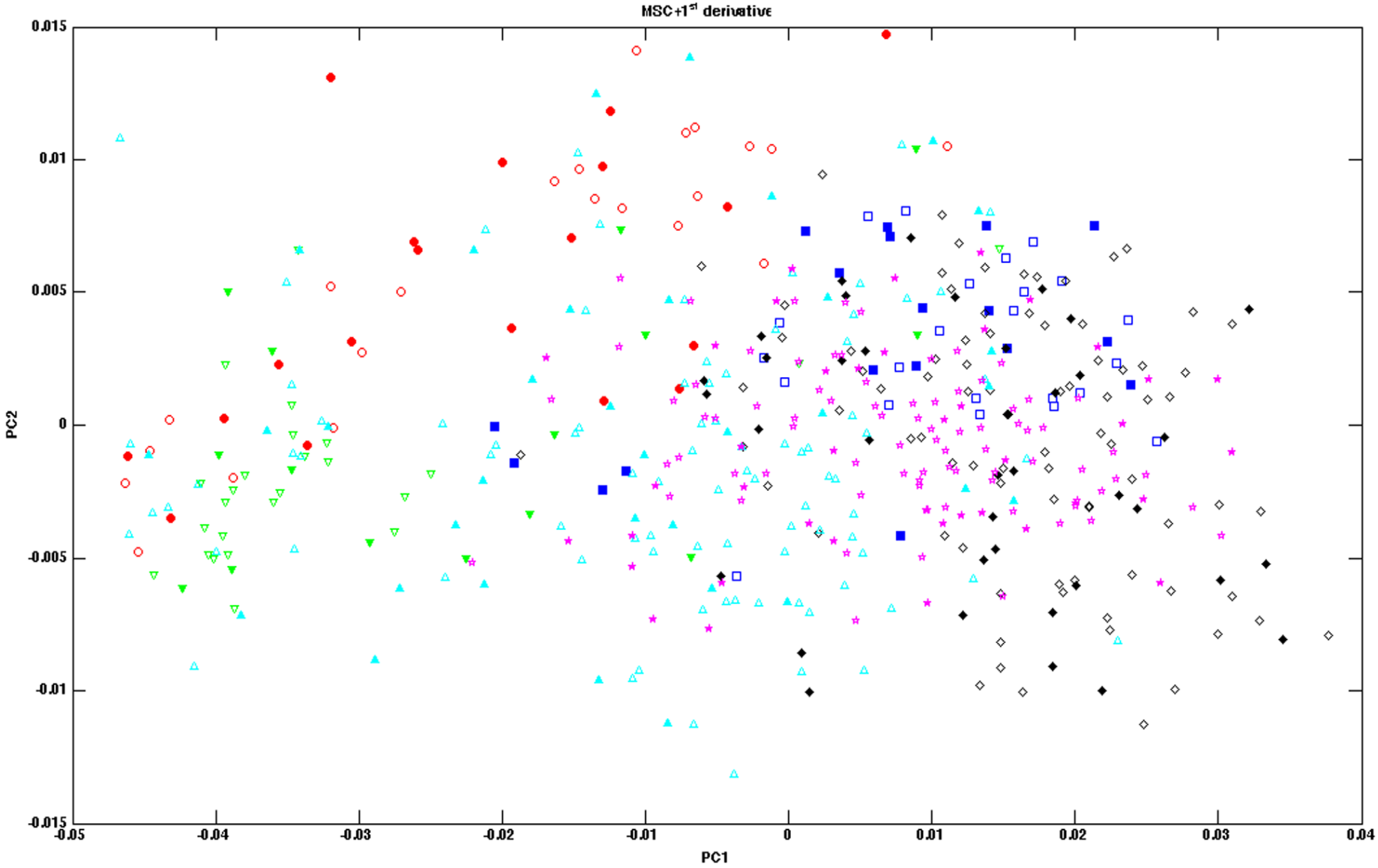

MSC+1st derivative
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
PC2
-0.05
-0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04
PC1

Figure 2h

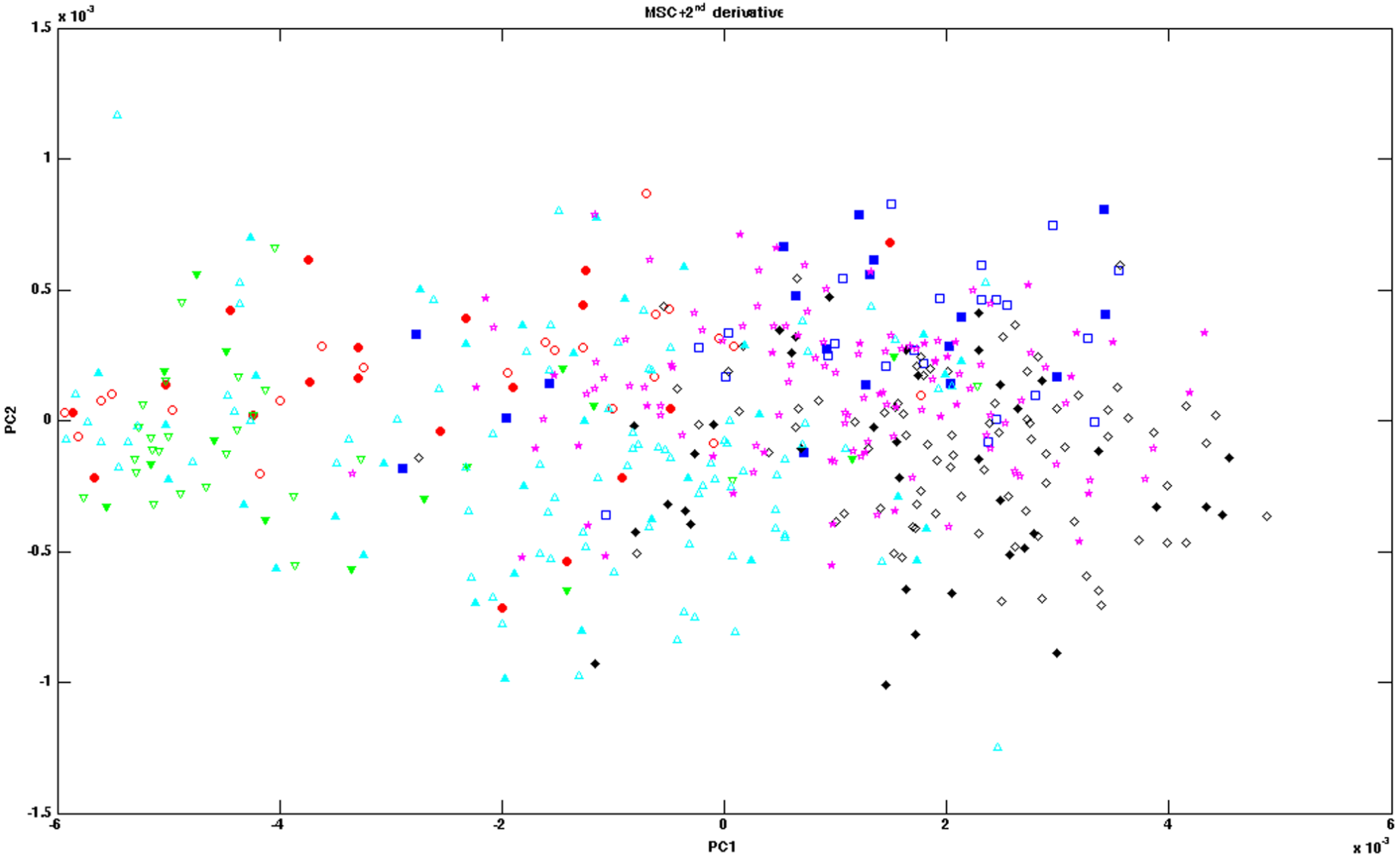
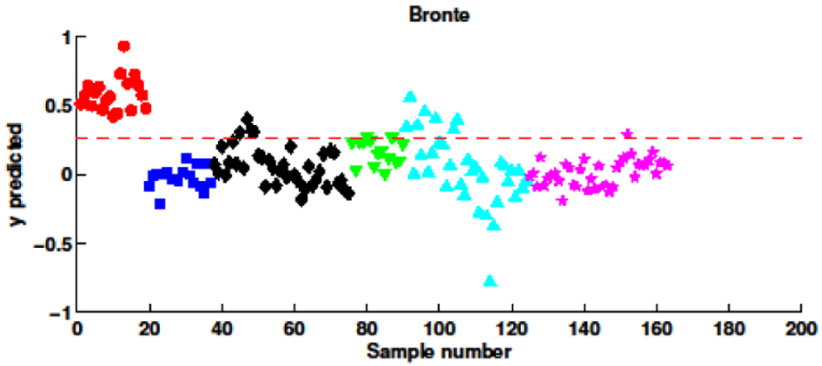
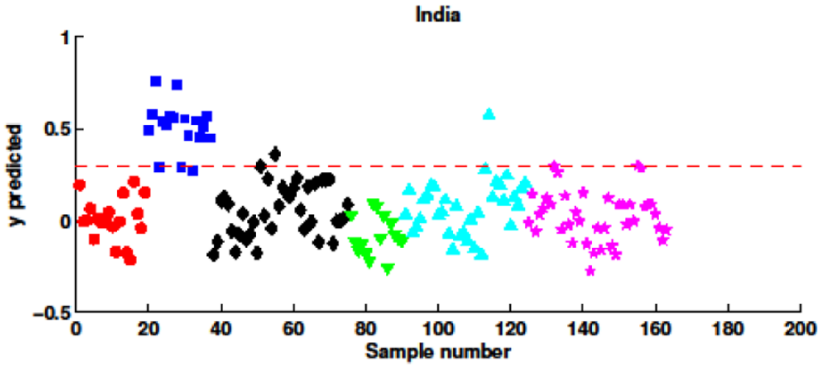
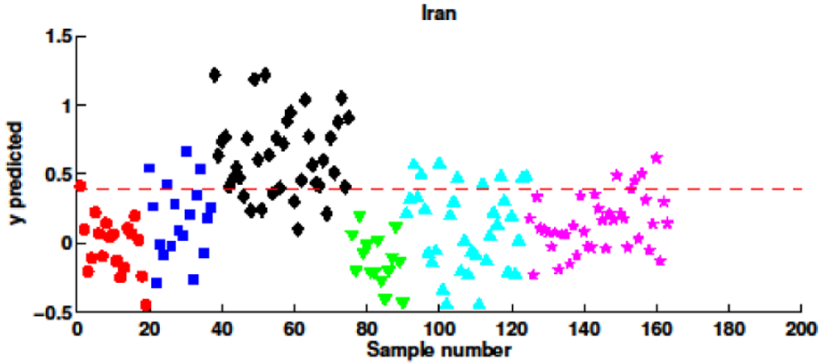
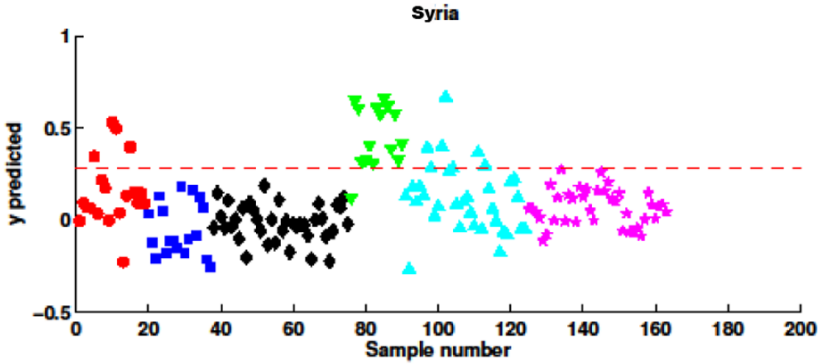


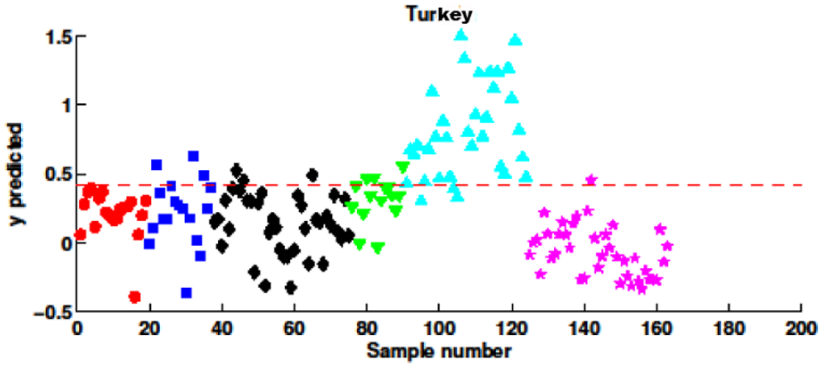
Figure 3


Bronte
y predicted
Sample number


India
y predicted
Sample number


Iran
y predicted
Sample number


Syria
y predicted
Sample number


Turkey
y predicted
Sample number

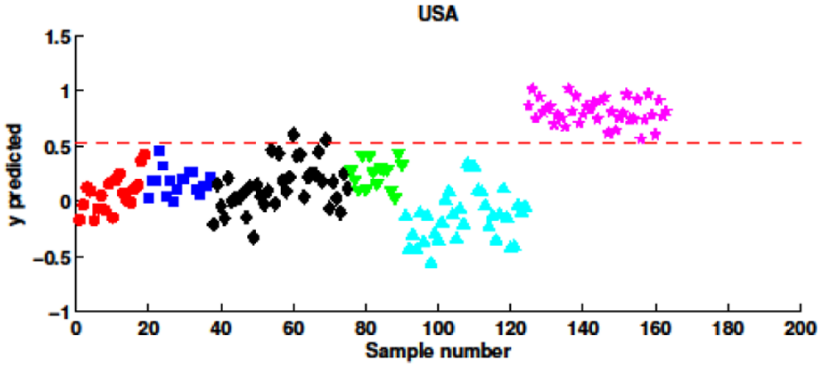

USA
y predicted
Sample number

Figure 4

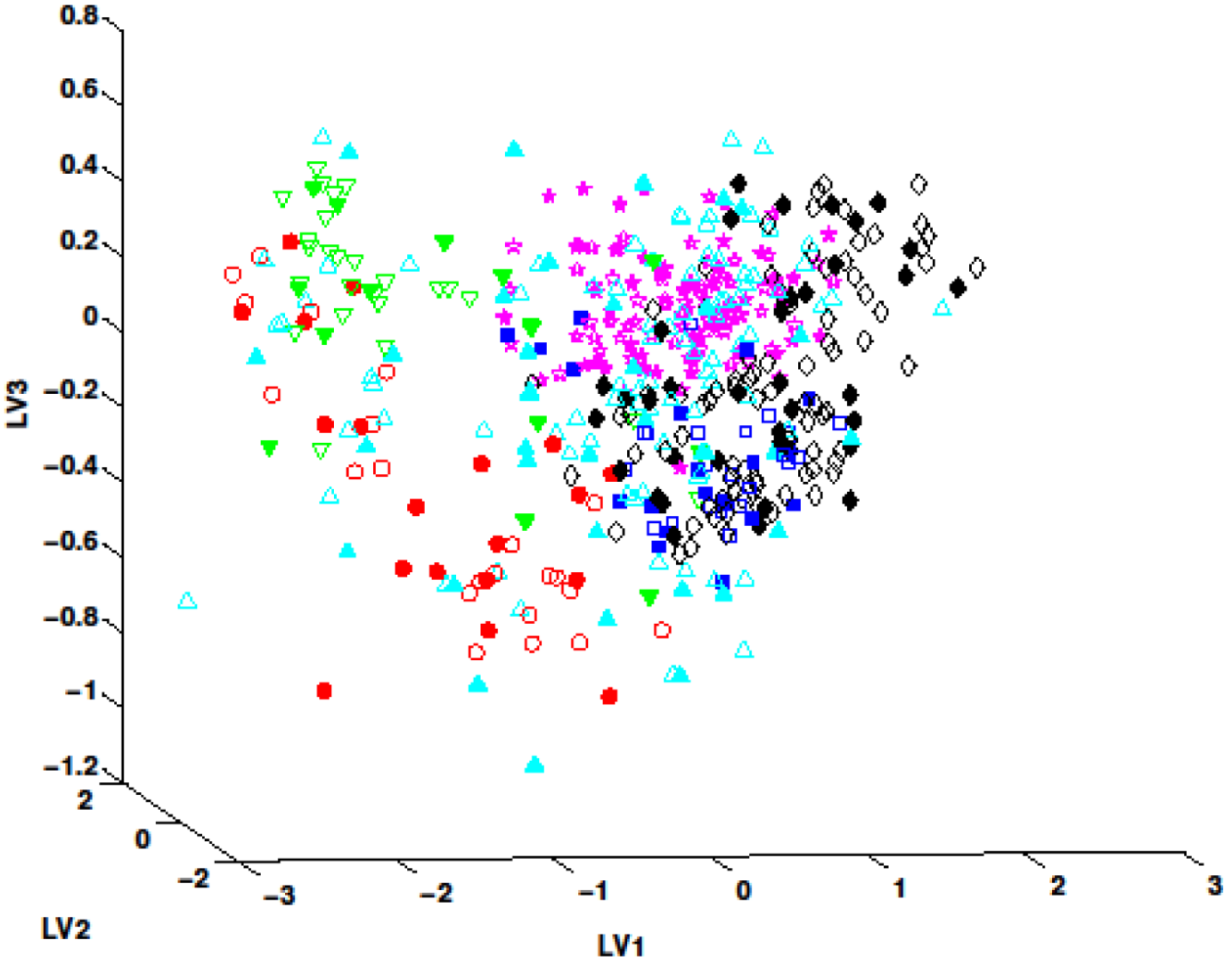


Figure 5

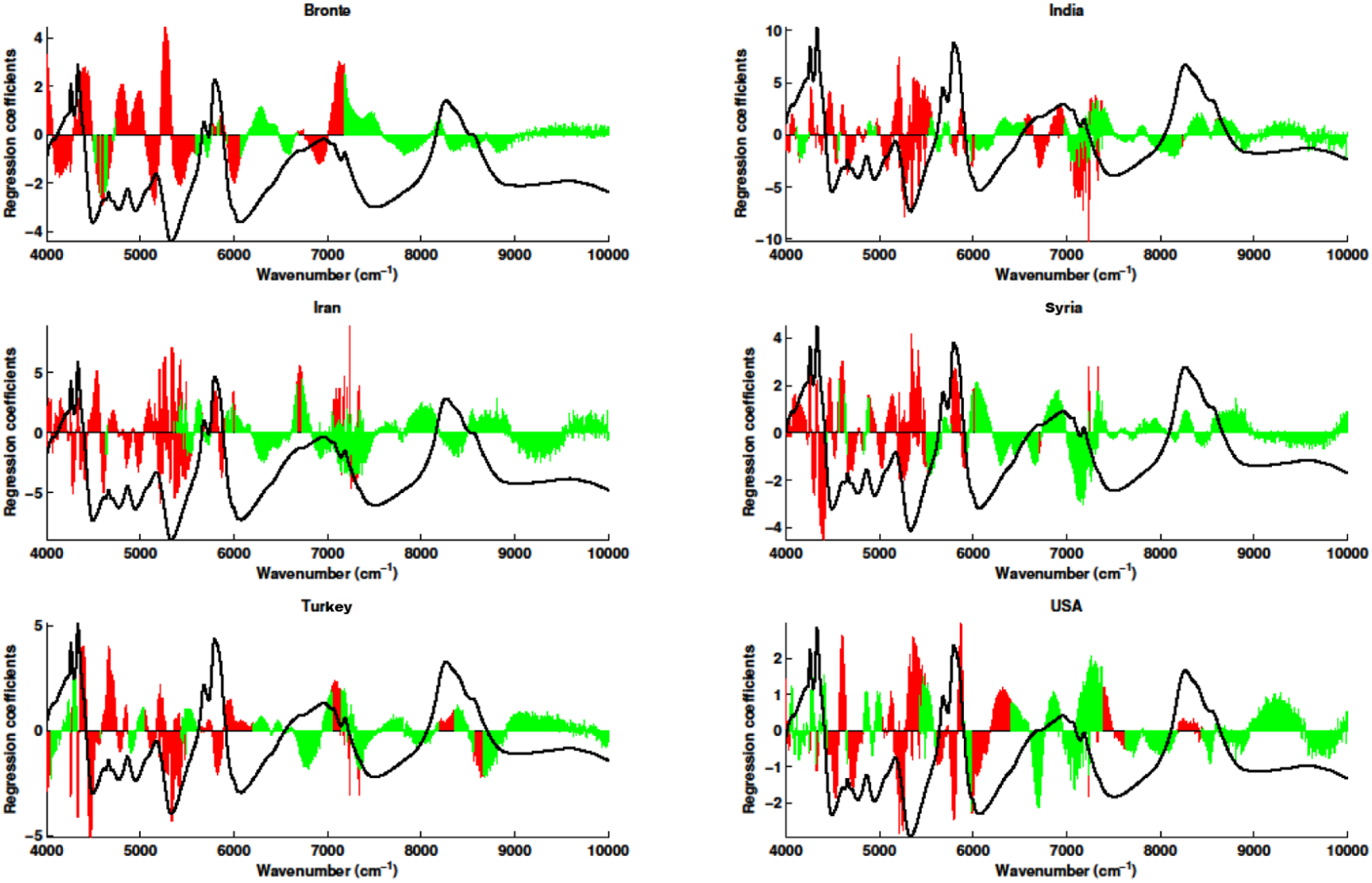

Bronte
India
Iran
Syria
Turkey
USA
Regression coefficients
Wavenumber (cm^{-1})
4000
5000
6000
7000
8000
9000
10000

Figure 6

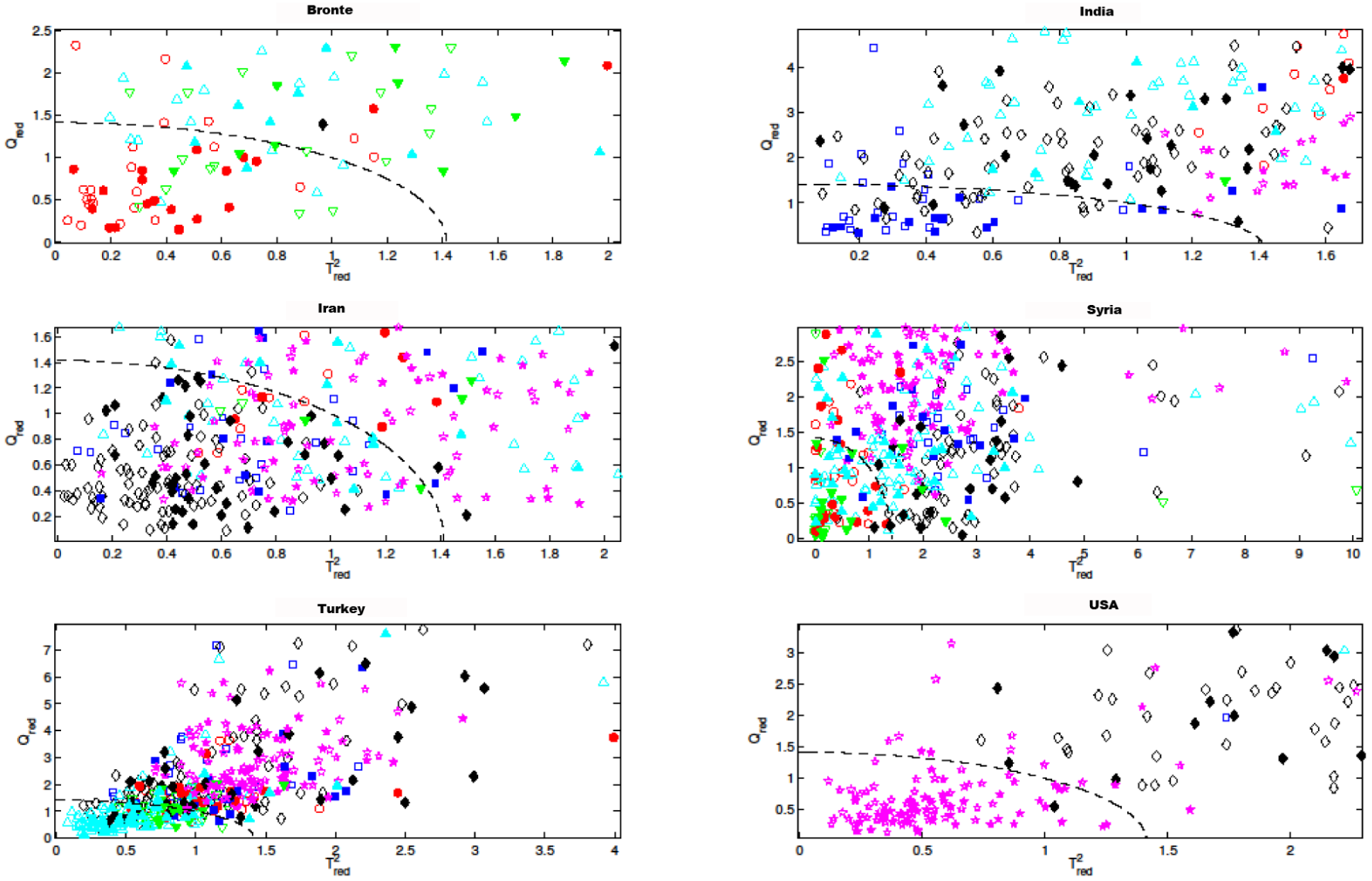
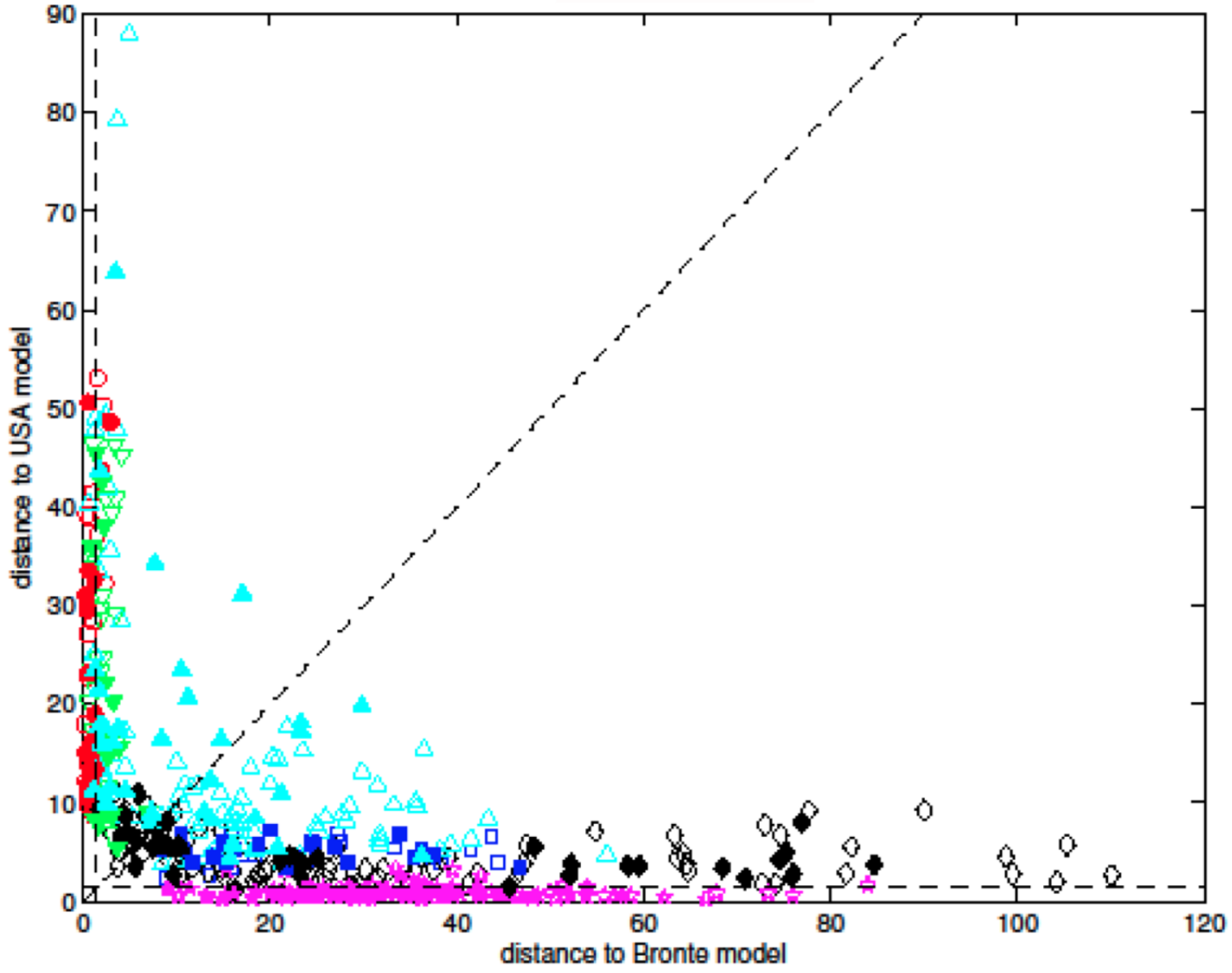


Figure 7


distance to USA model
distance to Bronte model
0
10
20
30
40
50
60
70
80
90
0
20
40
60
80
100
120