# Combining excitation-emission matrix fluorescence spectroscopy, Parallel Factor Analysis, cyclodextrin-modified micellar electrokinetic chromatography and Partial Least Squares Class-Modelling for green tea characterization

Monica Casale[a,*], Benedetta Pasquini[b], Maryam Hooshyari[a], Serena Orlandini[b,*], Eleonora Mustorgi[a], Cristina Malegori[a], Federica Turrini[a], Maria Cruz Ortiz[c], Luis Antonio Sarabia[d], Sandra Furlanetto[b]

[a]*Department of Pharmacy, University of Genoa, Viale Cembrano 4, 16148 Genoa, Italy*

[b]*Department of Chemistry "U. Schiff", University of Florence, Via U. Schiff 6, 50019 Sesto Fiorentino, Florence, Italy*

[c]*Department of Chemistry, University of Burgos, Plaza Misael Bañuelos s/n, 09001 Burgos, Spain*

[d]*Department of Mathematics and Computation, University of Burgos, Plaza Misael Bañuelos s/n, 09001 Burgos, Spain*

*Corresponding authors. Tel.: +39 010 3532633 (M. Casale), +39 055 4573733 (S. Orlandini)

*E-mail addresses:* monica@difar.unige.it (M. Casale), serena.orlandini@unifi.it (S. Orlandini)

## Abstract

In this study, an alternative analytical approach for analyzing and characterizing green tea (GT) samples is proposed, based on the combination of excitation–emission matrix (EEM) fluorescence spectroscopy and multivariate chemometric techniques. The three-dimensional spectra of 63 GT samples were recorded using a Perkin–Elmer LS55 luminescence spectrometer; emission spectra were recorded between 295 and 800 nm at excitation wavelength ranging from 200 to 290 nm, with excitation and emission slits both set at 10 nm. The excitation and emission profiles of two factors were obtained using Parallel Factor Analysis (PARAFAC) as a 3-way decomposition method. In this way, for the first time, the spectra of two main fluorophores in green teas have been found. Moreover, a cyclodextrin-modified micellar electrokinetic chromatography method was employed to quantify the most represented catechins and methylxanthines in a subset of 24 GT samples in order to obtain complementary information on the geographical origin of tea. The discrimination ability between the two types of tea has been shown by a Partial Least Squares Class-Modelling performed on the electrokinetic chromatography data, being the sensitivity and specificity of the class model built for the Japanese GT samples 98.70% and 98.68%, respectively. This comprehensive work demonstrates the capability of the combination of EEM fluorescence spectroscopy and PARAFAC model for characterizing, differentiating and analyzing GT samples.


*Keywords:* Catechins; Cyclodextrin modified-micellar electrokinetic chromatography; Excitation-emission matrix fluorescence spectroscopy; Green tea; Methylxanthines; Parallel Factor Analysis

---

*Abbreviations*: CF, caffeine; (+)C, (+)-catechin; CyD-MEKC, cyclodextrin-modified micellar electrokinetic chromatography; EC, (-)-epicatechin; ECG, (-)-epicatechin gallate; EGC, (-)-epigallocatechin; EGCG, (-)-epigallocatechin gallate; EEM, excitation-emission matrix; GT, green tea; HPβCyD, (2-hydroxypropyl)-β-cyclodextrin; PARAFAC, Parallel Factor Analysis; PLS-CM, Partial Least Squares Class-Modelling; TB, theobromine.

## 1. Introduction

Tea is an aromatic beverage made from the leaves of *Camellia sinensis*, a plant native to Southeast Asia, cultivated and consumed by humans for thousands of years. Due to its attractive aroma and taste and its effect on reducing lifestyle-related diseases, tea is the most consumed beverage in the world. Green tea (GT) is made from unfermented leaves of *Camellia sinensis* and contains a high concentration of polyphenols, which are powerful antioxidants. The potential health benefits of GT, especially related to its antioxidant properties, have led to an increase of its consumption in the last decades. The principal compounds of GT having biological effects have been identified as catechins and xanthines [1]. Catechins show a strong antioxidant activity and exert antiinflammatory, antiarhtritic, antiangiogenic, neuroprotective, anticancer, antiobesity, antiatherosclerotic, anti-diabetic, antibacterial, antiviral and antidental caries effects. Xanthines are responsible for the stimulating effects; caffeine (CF) is a central nervous system and cardiac stimulant and has a diuretic effect, while theobromine (TB), which is present in lower amounts, has also a diuretic effect [1-7]. Among the most abundant catechins in GT there are (+)-catechin, ((+)C), (-)-epicatechin (EC), (-)-epigallocatechin (EGC), (-)-epicatechingallate (ECG), (-)-epigallocatechin gallate (EGCG) [8].

The composition of GT can be influenced by several parameters associated with growth conditions, such as genetic strain, season, climatic conditions, soil profile, growth altitude, horticultural practices, plucking season, shade growth, and with the region in which tea has been cultivated. The other factors that can influence the profile of bioactive compounds are manufacturing process (withering, steaming/pan-firing, rolling, oxidation/fermentation and drying) and storage [8,9]. Besides this huge variability, the price of tea greatly varies according to its geographical origin. Hence, the recognition of the origin of GT is crucial to protect the interests of both consumers and sellers [10,11]. Several analytical methods have been proposed together with chemometric techniques in order to characterize the geographical origins and/or varieties of teas [12-15]. However, most of these methods require expensive equipment and involve tedious sample preparation in order to discriminate GT samples from different geographical origins; as an example, Ye *et al*. [14] extracted the volatile organic components from the dried tea leaves by headspace solid-phase microextraction procedure, followed by GC–MS analysis.

In a previous paper coauthored by some of us [10], cyclodextrin-modified micellar electrokinetic chromatography (CyD-MEKC) was employed to simultaneously analyse the most represented catechins and methylxanthines in 92 GT samples of different geographical origin, and the comparison of the obtained data showed that Japanese commercial GT products contained a general lower level of catechins than Chinese GTs.

The contents of catechins and methylxanthines were thus used as chemical descriptors and potential indicators of the geographical origin. Considering this previous work as a starting point for further investigations, in the present study an alternative analytical approach was applied for identifying the differences in terms of active compounds content in

GT samples from different geographical origin. In order to reach this aim, 63 GT samples were analysed by fluorescence spectroscopy: 29 samples from Japan and 34 from China. The main reason of the choice of these two countries was the interest of the consumers in the comparison of Japanese and Chinese GTs in terms of active compounds content. As a matter of facts, Chinese GT tends to cost consumers much less than Japanese GT, for the massive prevalence of Chinese GT and thus the necessity of maintaining low prices by Chinese producers, and for the lack of space for the production of GT in Japan. Moreover, one of the main differences in GT processing between Chinese and Japanese producers is the way deactivation of enzymes is performed. Chinese GT is usually dry heated in order to deactivate oxidases, whereas in the case of Japanese GT steaming is employed. Besides, Japanese GT is usually shade grown [9]. Hence, we deemed it worthwhile to compare the GTs from these two countries in order to understand if the higher price of Japanese teas can be supported or not by the fact that it is a more prized tea for its higher antioxidant capacity.

In more detail, the innovative analytical approach presented is based on the combination of excitation–emission matrix (EEM) fluorescence spectroscopy and chemometric tools to extract useful information from a huge amount of data. The chemometric approach is a fundamental part of the interpretation of fluorescence spectral data of agro-food products due to the presence of many fluorophores, since the fluorescence of a sample consists of a number of overlapping signals not easily understandable without a proper data processing. Accordingly to these principles, three-dimensional fluorescence spectra were elaborated through PCA [16] after unfolding the data into matrices and through Parallel Factor Analysis (PARAFAC) [17] on three-way data as display methods. Moreover, SELECT [18] technique was applied for variable selection, in order to individuate the variables with the highest classification power, *i.e.* the most informative emission bands in discriminating between Japanese and Chinese GTs.

Finally, the content of catechins and methylxanthines was determined in a subset of 24 GT samples by the previously developed chiral CyD-MEKC method in order to obtain complementary information on the geographical origin of GT samples and to confirm what observed in our previous work [10], *i.e.* that the amount of all the considered compounds was higher for Chinese GTs, with the exception of ECG. A Partial Least Squares Class-Modelling (PLS-CM) was carried out on this subset of samples to develop a predictive model able to classify new GT samples according to the geographical origin using the CyD-MEKC data.

## 2. Materials and methods

### *2.1. Chemicals, solutions and samples*

The reference standards of (+)C, EC, EGC, ECG, EGCG, CF, TB, as well as boric acid, 86.1% phosphoric acid, sodium dodecyl sulphate (SDS), (2-hydroxypropyl)-β-cyclodextrin (HPβCyD, degree of substitution 0.6), were

purchased from Sigma-Aldrich (St. Louis, MO, USA). The standard stock solutions (1 mg $mL^{-1}$) of (+)C, EC, EGC, ECG, EGCG, CF, TB and of the internal standard syringic acid were prepared in a mixture of methanol/water in 15:85 ratio %v/v. Working standard solutions were obtained by dilution with water in a vial to 500 μL for achieving the desired final concentration values of the compounds.

A set of 63 GT samples of different varieties and from different geographical origins (29 from Japan and 34 from China) was selected for the study and analysis. In order to assure a good degree of representativity of the samples, the main sources of variability for GTs were considered, *i.e.* for Japanese GTs the different varieties, including Bancha, Gyokuro, Matcha, Sencha, Matcha Tsuru types, while for Chinese GTs the different zones (the ten provinces of Hunan, Fujian, Zhejiang, Anhui, Yunnan, Guandong, Jiangsu, Hubei, Shandong, Guanxi). Moreover, each geographical group included samples stored in different conditions and coming from different manufacturing processes. Supplementary Table S1 shows the description of the samples and the corresponding assigned code. The commercial GT samples were collected locally in specialized stores located in the cities of Florence and Genoa (Italy). A subset of 24 samples randomly selected including different types of Japanese GT and different zones of Chinese GT has been analyzed using the CyD-MEKC method for the quantitation of catechins and methylxanthines (Table 1).

*2.2. Preparation of GT samples*

In order to simulate the content of active compounds in a cup of tea, GT samples were prepared by infusion of tea leaves. The samples were prepared immersing 0.2 g of finely powdered tea leaves in 10 mL of water at 85 ºC for 5 min in a beaker. Then, the beaker containing tea leaves and water was transferred into an ice bath for 30 s to stop the infusion at the same moment for each sample. In order to remove the leaves before performing the analysis, the infusion was filtered using a filter paper (Albet® LabScience) with a porosity equal to 73 $g/m^2$.

*2.3. Instrumental*

*2.3.1. Capillary electrophoresis*

The CyD-MEKC method used for the determination of the compounds was derived from a previous study coauthored by one of us [15]. The analyses were carried out using a $^{3D}$CE instrument from Agilent Technologies (Waldbronn, Germany) controlled by the software $^{3D}$CE ChemStation (Agilent Technologies) for both acquisition and data management. Fused-silica capillaries (Unifibre, Settimo Milanese, Italy) of 33.0 total length, 8.5 cm effective length and 50 μm inner diameter were used. The detection was carried out by using the on-line DAD detector and the detection wavelength was 200 nm. Voltage and temperature were set at 15 kV and 25 °C, respectively. The background electrolyte was made by 25 mM borate-phosphate buffer pH 2.50 with the addition of 90 mM sodium dodecyl sulphate

and 25 mM HPβCyD. Total analysis time was about 8 minutes. Calibration was perfomed by the internal standard method, using syringic acid as internal standard. The method had been previously validated in terms of selectivity, linearity, repeatability, accuracy and sensitivity, showing adequate performances for the analysis of catechins and methylxanthines in GT, with LOQ values ranging from 0.05 to 0.7 μg $mL^{-1}$ [15]. Further information on the CE method and procedure may be found in mentioned Ref. [15].

*2.3.2. Fluorescence spectroscopy*

The EEM fluorescence measurements were performed directly on GT extracts at room temperature on a Perkin-Elmer LS55B luminescence spectrometer (Waltham, MA, USA). The excitation-emission matrices of the GT infusions were recorded using the standard cell holder and a 10 mm quartz SUPRASIL® cell with cell volume of 3.5 mL by PerkinElmer. The excitation spectra were recorded between 200 nm and 290 nm each 5 nm (19 recorded points), whereas the emission wavelengths ranged from 295 nm to 800 nm each 0.5 nm (1011 recorded points). The excitation and the emission monochromator slits were set to 10 nm. The FL WinLab software (PerkinElmer) was used to register the fluorescent signals.

*2.4. Multivariate data analysis*

*2.4.1. Data exploration*

PCA [16] is the most used tool in exploratory data analysis and it uses an orthogonal transformation to convert a set of correlated variables into a set of uncorrelated variables called principal components. This approach makes it possible to visualize in a comprehensive way the dataset starting from a two-dimensional data matrix. According to the specific nature of EEM data, organized in a three-dimensional data array, for performing PCA a step of unfolding of the matrix is requested, while with the PARAFAC algorithm it is possible to directly model n-way data. In the case of three-way data, like the EEM data, PARAFAC decomposes a data array $\underline{\mathbf{X}}$ with dimension I × J × K into three loading matrices **A**, **B** and **C,** being their columns $a_i$, $b_j$ and $c_k$ respectively. The trilinear PARAFAC model is expressed as follows:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad i = 1, 2,\ldots, I; \quad j = 1, 2,\ldots, J; \quad k = 1, 2, \ldots, K \qquad (1)$$

where $x_{ijk}$ is the element in the position *i*, *j*, *k* of the three-way array $\underline{\mathbf{X}}$; *F* is the number of factors; $a_{if}$, $b_{jf}$ and $c_{kf}$ are the elements of the matrices **A** (I × F), **B** (J × F) and **C** (K × F), respectively; $e_{ijk}$ represents the generic element of the residual array $\underline{\mathbf{E}}$ (I × J × K). The PARAFAC model is found by minimizing the sum of squares of the residuals.

The excitation-emission fluorescence matrices obtained for several samples can be arranged into a three-way array and the PARAFAC decomposition can be applied for the analysis of fluorescent data. In this case, $\underline{\mathbf{X}}$ contains the fluorescence intensity at the *k*-th excitation wavelength and *j*-th emission wavelength recorded for the *i*-th sample. Therefore, the vectors $a_i$, $b_j$ and $c_k$ are the sample, emission and excitation profiles of the *f*-th fluorophore, respectively. The similarity between the trilinear PARAFAC model and the physical model for fluorescence can be found in Ref. [19].

Data are trilinear when the experimental data array is compatible with the structure in Eq. (1). The core consistency diagnostic (CORCONDIA) developed by Bro and Kiers [20] is an index that measures the degree of trilinearity of the experimental data array. A trilinear model has a value of CORCONDIA index close to 100%.

If the fluorescence data are trilinear and the appropriate number of factors has been chosen to fit the model, the PARAFAC decomposition provides unique profile estimations, and the achievement of the true underlying excitation and emission spectra for every fluorophore is ensured [17]. PARAFAC has been widely used due to this highly attractive uniqueness property [21], which could be used for the unequivocal identification of compounds.

*2.4.2. Variable selection*

The selection of the informative variables was performed by means of SELECT [18], a feature selection technique based on the stepwise decorrelation of the variables, which is implemented in the V-Parvus software [22]. This technique generates a set of decorrelated variables ordered according to their Fisher weights. At each step, SELECT searches for the variable with the largest classification weight. This variable is selected and decorrelated from the other variables; then the algorithm is repeated until a fixed number of variables is selected or the Fisher weight is lower than a specific cut-off value. SELECT presents an interesting characteristic: the fraction of the residual variance of the predictors after the orthogonalization can be used to select intervals of predictors with better classification performance.

*2.4.3. Class modeling*

PLS-CM [23] is a supervised method of classification between two categories (or classes), in our case Japanese or Chinese GT. It is a version of Partial Least Squares (PLS) algorithm with a binary response that makes it possible to model the probability distribution of the samples for each class and then performs a hypothesis test evaluating the α probability of type I error and the β probability of type II error. Class-model sensitivity (proportion of the samples of the class that are correctly assigned) and specificity (proportion of samples correctly rejected) are (1-α)·100 and (1-β)·100, respectively. The risk curve is the plot of β error versus α error probabilities.

*2.4.4. Software*

Data analysis was performed in the MATLAB environment [24], thanks to tailor made algorithms developed and implemented by the Authors. For the data processing, PCA, PARAFAC and PLS-CM algorithms were applied, in order to extract the significant information embodied within data. For performing variable selection, the SELECT method was applied thanks to its implementation in the software V-Parvus [22].

## 3. Results and discussion

*3.1.Catechins and methylxanthines content*

The CyD-MEKC method previously described [15] was applied to the analysis of a subset of 24 GT samples in order to confirm our previous observations [10] and to lay the basis for the EEM data processing. By applying the CyD-MEKC method, the samples were characterized by means of $n$=7 variables, namely (+)C, EC, EGC, ECG, EGCG, CF and TB (mg $g^{-1}$, dry basis), obtaining a data matrix having 24 rows (samples) and 7 columns (variables), shown in Table 1. This data set was submitted to chemometric modeling starting from PCA as a display method and then applying the PLS-CM algorithm for class modeling purposes.

Firstly, PCA was performed on the data matrix to enhance the presence of structures inside the samples and to understand the correlation between the variables. Fig. 1 shows the loading (a) and the score (b) plots of the catechins ((+)C, EC, EGC, ECG, EGCG), CF and TB autoscaled data in the plane of the 2 first Principal Components, that explain the 86% of the total variance. From the loading plot it was possible to point out that the variable EGCG is the most important factor in PC1, followed by CF and EGC. All loadings are positive so that the samples with highest scores on PC1 have greater value in all the variables. On the contrary, loadings of PC2 have different sign: ECG has the highest positive loading and TB has the highest negative. Along PC1, the scores of the Japanese GT samples in relation to the scores of the Chinese GT samples are lower, indicating that in general Chinese GT samples were characterized by a higher content in the active compounds. This observation is in full agreement with what reported in our previous study [10].

In order to build the PLS-CM model, it is necessary to build a dummy vector containing the information about class membership; for this reason, a binary response was constructed considering the values 1 and 2 for the Japanese and Chinese GT, respectively (Table 1). The number of PLS latent variables that minimized the root mean square error in cross-validation (RMSECV) obtained by leave one out procedure was 3, and they explained the 81.68% of response with 90.05% of predictors variance. Fig. 2 shows the distribution of PLS fitted values for the Japanese and Chinese GT

samples. Both classes have normal distribution with mean values 1.09 and 1.91 and SD values 0.09 and 0.27, respectively.

In order to decide if an unknown sample belongs to one or another class, a threshold value, $t_v$, between 1 (GT from Japan) and 2 (GT from China) must be established. If the value estimated by PLS is higher than $t_v$ the sample is classified to belong to class 2 (China), while for estimated values lower than $t_v$ the sample is classified to belong to class 1 (Japan). A model for one class (e.g. "GT Japanese"), is in fact the acceptation region for the null hypothesis $H_0$: the sample belongs to "Japanese GT" class. Therefore, the evaluation of the quality of a class model is given by its sensitivity and specificity. Both parameters have been evaluated in cross-validation, being 98.70% and 98.68%, respectively. The risk curve, reported in Supplementary Fig. S1, is the plot of β versus α probabilities, where it is clear that both probabilities change in opposite directions, that is, α decreases when β increases and vice versa.

*3.2 Fluorescence spectra*

Fig. 3 shows two typical excitation-emission spectra of one Japanese (J1) and one Chinese GT sample (C1).

*3.2.1. Repeatability studies*

In order to assess the experimental variability and the repeatability in preparing the tea infusions, the analysis of two GT samples of different geographical origin (one from Japan and one from China) were replicated 3 times at a distance of time (one week). Supplementary Fig. S2 displays the score plot obtained by PCA of the spectral data after unfolding. PC1, which explains 97.8% of the total variance, clearly separates the 2 GT samples; on the contrary, the difference among the 3 replicates of the same sample is along PC2, which explains only 1.4% of the variance.

*3.2.2. PCA*

Two bands of the emission spectra were removed, namely from 295 to 350 nm and from 700 to 800 nm, due to the lack of information typical of these two areas (Fig. 3). The range between 350-700 nm was retained and used for data elaboration. A data matrix of dimension 63 × 13300 was built, where each row corresponded to the emission spectrum (700 wavelengths) obtained at each of the 19 excitation wavelengths for all the 63 GT samples measured. PCA was performed as unsupervised pattern recognition technique on this 'unfolded' matrix after the data had been mean-centered.

Fig. 4 shows the score plot on the plane PC1-PC4. It is possible to notice a discrimination between Japanese and Chinese GT samples along PC1, the direction explaining the 74.3% of the total variance, even if a certain overlap is present and the complete separation between the classes is not obtained. In the PC1-PC4 plot it can be also clearly

noticed that Matcha GT samples, considered one of the Japan's rarest and most precious GT variety, are grouped in a cluster in the orthogonal space at negative scores on PC1.

Looking at the loading profile on PC1 (Fig. 5), it is possible to notice the bands more informative along PC1 and thus useful for discriminating between Japanese and Chinese GTs, namely 410-450 nm and 500-600 nm. The first band (410-450 nm) shows positive loadings on PC1 and this suggests that it is related to active compounds content in GT from China; on the contrary the broad band (500-600 nm) has negative loadings, therefore it seems linked to chemical compounds characterizing the Japanese GTs.

*3.2.3. PARAFAC*

The EEM data recorded for the 63 samples analysed were arranged into a data array where the excitation wavelengths between 200 nm and 290 nm and the emission wavelengths between 295 nm and 800 nm were considered. Therefore, the dimension of this array was 63 × 1011 × 19 (where 63 are the samples, 1011 the emission wavelengths and 19 the excitation wavelengths). The PARAFAC decomposition of this array, without any constrain, required two factors (CORCONDIA of 100%, explained variance of 98.6%).

The plot of the loadings of the mode of the samples (first mode, Fig. 6a) is similar to the PCA score plot (Fig. 4) and it shows a rather clear discrimination between Chinese and Japanese GTs. The plot of the loadings of the mode of the emission (second mode, Fig. 6b) shows the emission spectra for two fluorophores, one with maximum around 420 nm and the other one with maxima at 500-550 nm. The plot of the loadings of the third mode (Fig. 6c) shows the excitation profiles. As can be seen in these plots, PARAFAC enabled to differentiate the infusions of GT according to the geographical origin (Chinese and Japanese). Moreover, due to the trilinearity of the data, it can be concluded that the two groups of fluorophores found with the PARAFAC model are the same in all the GT samples.

*3.2.4. Variable selection*

SELECT was applied as a variable selection technique in order to individuate the variables with the highest classification power, *i.e.* the most informative emission bands in discriminating between Japanese and Chinese GT samples. SELECT was applied on the unfolded data matrix of dimension 63 × 13300 where each row corresponded to the emission spectrum obtained for each excitation wavelength of each GT sample measured; the frequency histogram of the selections showed as the most selected variables the two bands 415-450 nm and 495-550 nm (Supplementary Fig. S3).

It is worthwhile to notice that the variables chosen by SELECT corresponded to the two bands highlighted by PARAFAC in the second mode, namely the emission spectra of two fluorophores. These outcomes are also in

293 agreement with the profile of the loading on PC1, that highlights the presence of two important bands, the first positive
294 at 410-450 nm and the second negative over 500 nm. Combining this information, it was possible to assume that the
295 first emission band (410-450 nm) is due to a fluorophore characterizing the Chinese GT samples and that the broad
296 band at 500-550 nm is related to the presence of compounds most abundant in the Japanese GT samples. The band at
297 410-450 nm probably corresponds to fluorescence emission of catechins, which are more abundant in Chinese samples.
298 The band at 500-550 nm is probably attributable to carotenoids, that are recognized to be in particularly high quantities
299 in Japanese tea, especially in Matcha, which contains 4 times more carotene than carrots and nine times more than
300 spinach [25]. The infuses of GT prepared for the analysis were noticed to be slight yellow-green color due to pigments
301 as chlorophylls and carotenoids; the quantities of pigment extracted in hot water are related to the concentrations of the
302 pigments in teas [26]. These observations were in agreement with the findings of Ref. [27], where the emission spectra
303 of various organic compounds which are known to be endogenous component of plant leaves were measured,
304 evidencing that catechins possess a fluorescence maximum near 440 nm and that β-carotene exhibits fluorescence
305 emission with a maximum near 530 nm.
306

307 ## 4. Conclusions

308 The aim of the present study was to evaluate the possibility of using EEM fluorescence spectroscopy as a rapid
309 analytical method for analyzing and characterizing GT samples, distinguishing between different geographical origins
310 (China or Japan). The experimental data, given their complex and multivariate nature, were elaborated with
311 chemometric techniques with the aim of extracting the useful information contained therein. PCA was applied, as a
312 display technique, on the "unfolded data" and PARAFAC was performed on three-dimensional arrays. The PCA results
313 were visualized by means of the score plot related to PC1 and PC4, which explained 76.8% of the total variance making
314 it possible to distinguish Chinese and Japanese samples. The separation between the two geographical origins was
315 mainly along PC1. Using PARAFAC, it was possible to perform the decomposition of the three-dimensional emission-
316 excitation matrix: the information on the first mode was similar to that observed by applying PCA to the matrix after
317 unfolding and it demonstrated that fluorescence spectroscopy is a promising and fast analytical method to characterize
318 GT samples on the basis of their geographical origin. PARAFAC on the second mode also highlighted the emission
319 spectra of two fluorophores, one with a maximum around 420 nm and the other with a maximum at 500-550 nm. These
320 bands correspond to the variables with the highest loadings on PC1 and also correspond to the variables selected by the
321 SELECT algorithm, that are those with the highest discriminating power between Japanese and Chinese GT samples.
322 The band around 420 nm was assumed to correspond to the fluorescence emission of catechins, which are more
323 abundant in the Chinese samples, and the band around 500-550 nm was attributed to carotenoids. Moreover, the CyD-

MEKC method was applied for the analysis of a subset of 24 GT samples confirming that catechins are more abundant in Chinese samples. In addition, the PLS-CM built with these data made it possible to distinguish Japanese from Chinese GT samples with a sensitivity and specificity of 98.70 and 98.68%, respectively.

## Acknowledgements

The Authors would like to acknowledge La Via del Tè (Florence, Italy) for the kind gift of tea samples.

## References


[1] Y. Suzuki, N. Miyoshi, M. Isemura, Health-promoting effects of green tea, Proc. Jpn. Acad. B-Phys. 88 (2012) 88–101.

[2] P. Bogdanski, J. Suliburska, M. Szulinska, M. Stepien, D. Pupek-Musialik, A. Jablecka, Green tea extract reduces blood pressure, inflammatory biomarkers, and oxidative stress and improves parameters associated with insulin resistance in obese, hypertensive patients, Nutr. Res. 32 (2012) 421–427.

[3] C. Cabrera, R. Artacho, R. Giménez, Beneficial effects of green tea – a review, J. Am. Coll. Nutr. 25 (2006) 79–99.

[4] R. Cooper, Green tea and theanine: health benefits, Int. J. Food Sci. Nutr. 63 (2012) 90–97.

[5] P. Velayutham, A. Babu, D. Liu, Green tea catechins and cardiovascular health: an update, Curr. Med. Chem. 18 (2008) 1840–1850.

[6] H. Wang, G.J. Provan, K. Helliwell, Tea flavonoids: their functions, utilization and analysis, Trends Food Sci. Technol. 11 (2000) 152–160.

[7] J.-M. Yuan, C. Sun, L.M. Butler, Tea and cancer prevention: epidemiological studies, Pharmacol. Res. 64 (2011) 123–135.

[8] M. Bonoli, P. Colabufalo, M. Pelillo, T. Gallina Toschi, G. Lercker, Fast determination of catechins and xanthines in tea beverages by micellar electrokinetic chromatography, J. Agric. Food Chem. 51 (2003) 1141–1147.

[9] A. Kosińska, W. Andlauer, Antioxidant Capacity of tea: effect of processing and storage, in: V.R. Preedy (Ed.), Processing and Impact on Antioxidants in Beverages, Academic Press, Elsevier, Waltham, 2014, pp.109-120.

[10] B. Pasquini, S. Orlandini, M. Goodarzi, C. Caprini, R. Gotti, S. Furlanetto, Chiral cyclodextrin-modified micellar electrokinetic chromatography and chemometric techniques for green tea samples origin discrimination, Talanta 150 (2016) 7-13.

[11] G. Ma, Y. Zhang, J. Zhang, G. Wang, L. Chen, M. Zhang, T. Liu, X. Liu, C. Lu, Determining the geographical origin of Chinese green tea by linear discriminant analysis of trace metals and rare earth elements: Taking Dongting Biluochun as an example, Food Control 59 (2016) 714-720.

[12] P.H. Gonçalves Dias Diniz, M. Ferreira Barbosa, K.D. Tavares de Melo Milanez, M.F. Pistonesi, M.C. Ugulino de Araújo, Using UV–Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup, Food Chem. 192 (2016) 374-379.

[13] N.S. Ye, A minireview of analytical methods for the geographical origin analysis of teas (Camellia sinensis), Crit. Rev. Food Sci. Nutr. 52 (2012) 775-780.

[14] N. Ye, L. Zhang, X. Gu, Discrimination of green teas from different geographical origins by using HS-SPME/GC–MS and pattern recognition methods, Food Anal. Methods 5 (2012) 856-860.

[15] R. Gotti, S. Furlanetto, S. Lanteri, S. Olmo, A. Ragaini, V. Cavrini, Differentiation of green tea samples by chiral CD-MEKC analysis of catechins content, Electrophoresis 30 (2009) 2922-2930.

[16] I.T. Joliffe, Principal Component Analysis, Springer-Verlag, New York, 2002.

[17] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149-171.

[18] M. Forina, S. Lanteri, M. Casale, M.C. Cerrato Oliveros, Stepwise orthogonalization of predictors in classification and regression techniques: An “old” technique revisited, Chemom. Intell. Lab. Syst. 87 (2007) 252–261.

[19] M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, D. Giménez, Identification and quantification of ciprofloxacin in urine through excitation-emission fluorescence and three-way PARAFAC calibration, Anal. Chim. Acta 642 (2009) 193-205.

[20] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. 17 (2003) 274-286.

[21] M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, A. Herrero, S. Sanllorente, C. Reguera, Usefulness of PARAFAC for the quantification, identification, and description of analytical data, in: A. Muñoz de la Peña, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (Eds.), Data Handling in Science and Technology: Fundamentals and Analytical Applications of Multiway Calibration, Elsevier, Amsterdam, 2015, pp. 37-81.

[22] M. Forina, S. Lanteri, C. Armanino, M.C. Casolino, M. Casale, P. Oliveri, V-PARVUS 2014, an extendable package of programs for explorative data analysis, classification and regression analysis, Dept. of Pharmacy, University of Genoa.

[23] M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, Tutorial on evaluation of type I and type II errors in chemical analyses: from the analytical detection to authentication of products and process control, Anal. Chim. Acta 674 (2010) 123–142.

[24] MATLAB version 8.4.0.150421 (R2014b), The Mathworks, Inc., Natick, MA, 2014.

[25] N. Hall, The Tea Industry, first ed., Woodhead Publishing, Cambridge, 2000, p. 21.

[26] Y. Suzuki, Y. Shioi, Identification of chlorophylls and carotenoids in major teas by high-performance liquid chromatography with photodiode array detection, J. Agric. Food Chem. 51 (2003) 5307−5314.

391 [27] M. Lang, F. Stober, H.K. Lichtenthaler, Fluorescence emission spectra of plant leaves and plant constituents,
392 Radiat. Environ. Bioph. 30 (1991) 333–347.
393

394 **Figure Captions**
395

396 **Fig. 1.** PCA (a) loading plot and (b) score plot of catechins and methylxanthines data.

397

398 **Fig. 2.** Normal distribution fitted for Japanese GT samples (in blue) and Chinese GT samples (in red).

399

400 **Fig. 3.** A typical excitation-emission spectra of (a) a Japanese (J1) and (b) a Chinese (C1) GT sample.

401

402 **Fig. 4.** PCA score plot on the PC1-PC4 plane for the fluorescence data. Matcha samples are indicated in green in the
403 plot.

404

405 **Fig. 5.** Loading profile on PC1.

406

407 **Fig. 6.** PARAFAC results: (a) loading plot of the mode of the samples (first mode); explained variance 98.6%
408 (F1=96.0% and F2=2.6%); (b) loading plot of the emission mode (second mode); (c) loading plot of the excitation mode
409 (third mode).

410