ELSEVIER

Forensic Science International 138 (2003) 75–84

Forensic Science International

www.elsevier.com/locate/forsciint

Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their Vis spectra

Nicholas C. Thanasoulias*, Nikolaos A. Parisis, Nicholaos P. Evmiridis

Laboratory of Analytical Chemistry, Department of Chemistry, University of Ioannina, University Campus, 451 10 Ioannina, Greece

Received 16 April 2003; received in revised form 24 August 2003; accepted 25 August 2003

Abstract

Fifty blue ball-point pen inks of five different brands were examined on the basis of the Vis spectrum of their ethanolic solutions with a view to achieving good discrimination between them. Samples were dissolved in absolute ethanol and their absorbance values in the range of 400–750 nm, after appropriate transformations, were used as variables in the multivariate statistical techniques of cluster analysis (CA), principal component analysis (PCA) and discriminant analysis (DA). These techniques were used successively so that an effective and meaningful discriminant model was calculated in the final step. The initial 351 variables ($\log_{10}$ transformed ink absorption values at each wavelength) were subjected to a *K*-means CA over the objects (samples) and only 20 variables were retained. Principal component analysis was used to detect any outliers (four samples were removed) and the remaining samples were re-subjected to PCA to decide how many variables to enter into DA and whether original variables or components should be used. It was found that the first three principal components (in accordance with the Kaiser criterion) were good descriptors of the 20 original variables (96.97% of the data variance was explained) and their use as latent variables in DA lead to low average variable redundancy (33.6%) in the discriminant model. The calculated model had a Wilks' λ of 8.98×10^{-5} and was statistically significant at the $P = 0.05$ level. The post hoc classification of the training dataset was 100% correct. From the DA results and the component loadings it was found that discrimination was achieved on the basis of differences in the shape of the absorption bands as well as their relative intensities. The method was therefore deemed appropriate for supporting exclusionary forensic purposes.

Keywords: Forensic chemistry; Forensic document examination; Principal component analysis; Discriminant analysis; Vis spectrophotometry; Blue ball-point pen ink

1. Introduction

Ink analysis is an important forensic procedure that can reveal useful information about questioned documents. Most of its applications regard the detection and confirmation of alterations to documents with significant financial value such as insurance claims, wills, contracts and tax returns. These modifications can be confirmed by comparison of the inks used to produce the questioned document or estimation of the time at which the various sections of the document were written [1]. It is therefore evident that there is a great need for the development of instrumental methods that will allow an in-depth examination of the inks used to produce a document and at the same time rigid statistical protocols are necessary to be followed so that conclusions regarding ink similarity can be drawn on an objective basis at predefined confidence levels.

The aforementioned needs rise from the fact that documents are rather complex systems that consist of primary and secondary materials [2]. Primary materials comprise the document support (e.g. paper, cardboard, polymer, etc.) and the text (e.g. ink deposits, carbon copies, photocopier toner, pencil, etc.). Secondary materials usually appear on a document as a result of corrections and manipulations and include correcting materials, erasure residues, adhesives, stains, fingerprints, etc.

* Corresponding author. Tel.: +30-26510-98399; fax: +30-26510-44831.
E-mail address: nthanas@cc.uoi.gr (N.C. Thanasoulias).

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2003.08.014

Modern inks contain a plethora of substances that aim to improve the ink characteristics [1,3]. Obviously, the most important component is the coloring material. This comes in the form of dyes, pigments or a combination of both. Dyes can be acidic or basic and are soluble in the liquid body of the ink that is also known as the vehicle. On the other hand, pigments are finely ground multimolecular granules that are insoluble in the vehicle. The vehicle, whose composition affects the flowing and drying characteristics of the ink, can consist of oils, solvents and resins. Another class of substances is used to finely tune the characteristics of the ink. These substances include driers, plasticizers, waxes, greases, soaps and detergents.

The techniques regarding the analysis of inks can be divided into destructive and non-destructive ones with regard to the changes they bring about to the questioned document. In destructive methods, a portion of the ink has to be removed from the document prior to the analysis [4]. On the other hand, non-destructive methods involve the observation of ink on the document by means of a reflectance technique that allows the recording of the ink spectral characteristics without removing the sample from its support.

Chromatographic and electrophoretic methods have always been favored by scientists as far as the analysis of inks is concerned. This preference stems from the fact that inks are rather complex mixtures that require separation of their constituents if good discrimination is to be achieved. Paper chromatography has been among the oldest destructive methods employed in ink analysis and has been used especially for organic dye based inks [5–7] since the older iron-gallotanate inks are difficult to separate by chromatography. Thin layer chromatography (TLC) [2,8–10] and its variants, including disk chromatography [11] and high performance thin layer chromatography (HPTLC) [12–14], have gradually replaced paper chromatography and have proved to be a satisfactory equivalent for separating ink components. Observation of thin layer chromatograms under alternative light sources, the use of infrared luminescence and microspectrophotometry have also been employed in an attempt to achieve better characterization of the TLC bands [15–17]. Although TLC remains the preferred method for ink analysis due to its low cost and relative simplicity, high performance liquid chromatography (HPLC) has been used as an alternative destructive technique that has the advantage of higher resolution and is also capable of detecting colorless components in the ink matrix [18–20]. Characterization of inks has also been achieved by means of traditional electrophoretic methods, but with increased equipment cost and a demand for larger samples [21,22]. Better results in terms of resolution, quantitation and analysis time have been achieved with the use of capillary electrophoresis (CE) [1,23–25].

Non-destructive techniques, although being the most useful ones with regard to document integrity, have not been developed and exploited thoroughly. The observation of documents under alternative light sources with the naked eye is probably the most commonly used method [26]. Problems associated with the subjectivity of the human eye have been surpassed with the use of suitable detectors that can measure the reflected radiation from the samples at different wavelengths, thus offering further information regarding the nature and composition of the ink [2].

Although much research has been carried out for the development of efficient analytical methods regarding the composition of inks, chemometrics is an area that has not been extensively used to explore and support the analytical results. Multivariate chemometrics in particular is a powerful tool when dealing with multi-component systems and allows the extraction of maximum information from complicated datasets. Since forensic science is a discipline that must draw its conclusions on a purely objective basis whenever this is possible, it is mandatory for forensic scientists to follow rigid statistical protocols in reaching decisions regarding experimental data. Therefore, in our present work, we have tried to explore the usefulness of multivariate chemometrics in the discrimination of blue ball-point pen inks. This was achieved by extracting ink dyes in ethanol, recording the Vis spectra of the extracts and applying cluster analysis (CA), principal component analysis (PCA) and discriminant analysis (DA) to the spectral data. Although the use of multivariate chemometrics with UV-Vis spectra is usually avoided, as these spectra show broad bands that cause multicollinearity problems, we have followed a statistical protocol presented in previously published work [27] that allowed us to decide on the number and quality of the variables (original ones or principal components) to be used so that an effective discrimination of the inks could be achieved. Readers who are not familiar with the statistical techniques presented herein are referred to Appendix A at the end of the paper where the fundamentals of cluster analysis, principal component analysis and discriminant analysis are given.

2. Materials and methods

2.1. Samples, their preparation and measurements

Five commercially available brands of blue ball-point pen inks (coded as: BI, FC, PE, PI and ST) were used for the study. For each brand, 10 pens of the same batch were sampled once by means of a stainless steel needle that was used to penetrate the wall of the plastic ink reservoir and transfer a small portion of the sample (less than 1 mg) into the solvent. Each pen was sampled using a different needle and the ink stained tip of the needle was submerged in a test tube containing 10.0 ml absolute ethanol (MERCK, 99.8% v/v). After part of the ink had dissolved in the solvent, the solution was agitated and centrifuged at 2000 rpm for 10 min and the absorbance of the supernatant liquid was measured on a JENWAY 6405 UV-Vis spectrophotometer in

a 1.00 cm quartz cell against absolute ethanol as the blank. The scanning range was 400–750 nm at 1 nm intervals.

2.2. Statistical procedures

The absorbance values for each sample were divided by the total absorbance resulting from each spectrum and the results were multiplied by 100. The $\log_{10}(\%A)$ values were calculated to ensure normality of the data. All statistical analyses were performed on a personal computer running STATISTICA Version 4.3 for Windows by StatSoft Inc., 1993. The modules used were cluster analysis, factor analysis (FA) and discriminant analysis. Varimax rotation was used in PCA.

3. Results and discussion

The use of $\%A$ values instead of raw absorbance data was necessary since the resulting spectra would have to be normalized per unit mass of sample if they were to be comparable. However, ink samples of the size used in this study were extremely small (less than 1 mg) and very difficult to weigh and we therefore judged it necessary to use the formula:

$$\%A_k = \frac{A_k}{\sum_{i=400}^{750} A_i} \times 100$$

This transformation overcomes the problem of differences in the weights of the samples used to prepare the solutions and the spectra become comparable. The average spectrum for each pen brand (as $\%\bar{A}$ versus λ) is shown in Fig. 1. Each spectrum represents the average absorption of inks coming from the same batch. The wavelength range chosen in the study was that of the visible region in an attempt to imitate the response of the human eye to the coloring constituents of inks. From the spectra, it can be seen that the inks studied fell into two main categories: inks FC, PE, ST demonstrated a single-peaked spectrum, whereas inks BI and PI possessed double-peaked spectra. The first absorption band had its peak in the range of 570–600 nm and was common for all inks. Electromagnetic radiation of these wavelengths appears yellow–orange to the naked eye and is absorbed by blue–cyan substances. The band mentioned above was therefore attributed to the blue dyes and pigments contained in all samples. The second band, which was present only in the case of samples BI and PI, had its peak in the range of 660–670 nm. Radiation in this range is orange-red in color and is absorbed by blue–green substances. Indeed, when observed with the naked eye, the solutions of these inks had colors that differed slightly from those of inks FC, ST and PE. Although these average spectra allowed a rough discrimination of the samples, nothing could be said about the differences observed in the spectra of inks FC, ST and PE as these spectra represented only the between sample variance with nothing being known about the within sample variance. Therefore, advanced statistical tests were judged necessary for the establishment of a completely objective discriminating protocol.

As both PCA and DA are parametric statistical techniques, it was necessary for the data ($\%A$ values) to be $\log_{10}$ transformed to ensure normality. Since samples for each brand were considered to belong to different populations and only 10 samples from each population were available, no normality tests were carried out, as these tests are known to work well for large sample sizes only. From now on, the $\log_{10}(\%A)$ values will be referred to as the original variables.

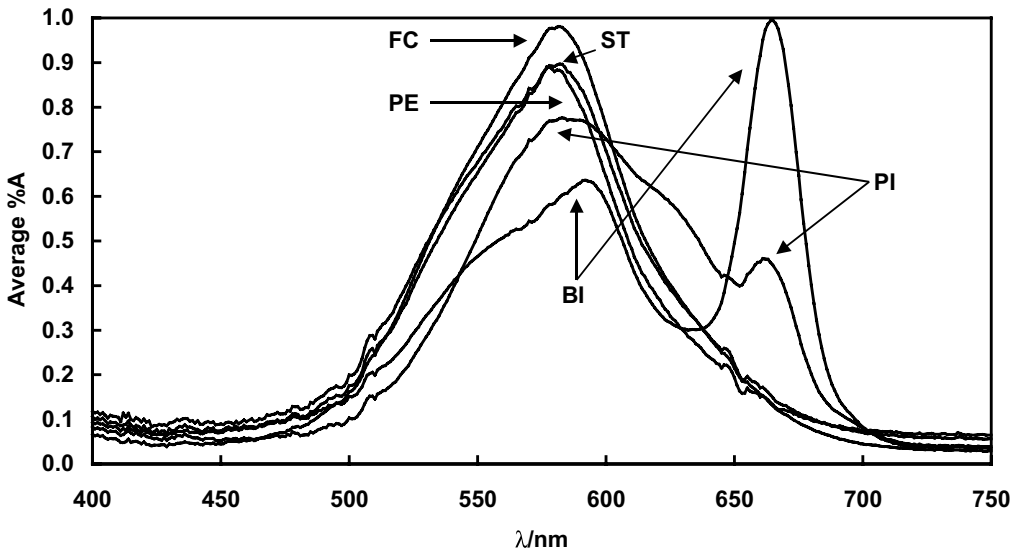

Fig. 1. Average $\%A$ vs. λ for all pen brands.

STATISTICA Version 4.3 for Windows does not allow more than 300 variables to be used in any of its analyses. However, the ink spectra were recorded in the range of 400–750 nm at 1 nm intervals and the $\log_{10}$ transformed $\%A$ values at every wavelength were to be taken as the original variables. This meant that 351 variables were available and some of them would have to be removed for the dataset to comply with the STATISTICA limitations. Feature reduction of this kind can be achieved by performing a cluster analysis (K-means) on the variables over the objects (ink samples) [28]. According to this technique, variables carrying similar information about the objects are expected to form clusters from which the most representative variables (the ones closest to the cluster centroids) can be chosen. Although the maximum number of clusters that can be formed by STATISTICA is 50, we were able to calculate only 20 clusters and use the relevant variables, since the formation of 50, 40 or 30 clusters yielded variables that proved to be highly correlated and all subsequent attempts to run a PCA resulted in a singular correlation matrix.

In the first place, PCA was necessary for the removal of any outliers in the data. The presence of outliers can affect the results of DA through an overestimation of the within sample variance, something that can reduce the effectiveness of the discriminant model. Although the deletion of some data may appear to introduce bias in the analysis, in this case, such a procedure was necessary since for each brand all 10 pens were of the same batch and their spectra were expected to come from the same population. In multivariate systems, PCA can aid the observation of outliers by projection of the data on a plane after Varimax rotation of the first two extracted components (PC_1 and PC_2). This plane is shown in Fig. 2 from which it can be seen that samples FC1, PI4, PI7 and ST8 should be considered outliers and their existence was attributed to measurement problems. These outliers were removed from the dataset.

The next step in the study regarded the quality of the discriminant model calculated in DA. Principal component analysis was successively applied to the original dataset and each time fewer and fewer components were extracted and rotated. For each component, the variable with the highest loading was retained and the resulting new subset of variables was used in DA. This method has the advantage of easier interpretation of the DA results since reference to the original wavelengths can be made directly, but multicollinearity is still present and high variable redundancy is observed in DA. The characteristics of the discriminant model were also checked when components were used instead of the original variables. When this method is followed, the interpretation of the results is more difficult as one has to look at the component loadings to decide which variables they represent. However, the technique handles the problem of multicollinearity effectively due to the orthogonal nature of the extracted components.

To avoid capitalizing on chance, one has to use as few variables as possible when running a DA. A useful rule of thumb [28] dictates that m variables should be used when n objects exist so that:

$$m < \frac{n}{3}$$

Taking the removed outliers into account, 46 objects remained and that meant that only up to 15 variables (or components) should be entered into the DA module. The two main criteria (Kaiser criterion [29] and the scree test [30]) used in deciding how many components to extract in PCA were also taken into account when interpreting the DA results. The scree plot (Fig. 3) for the given dataset showed

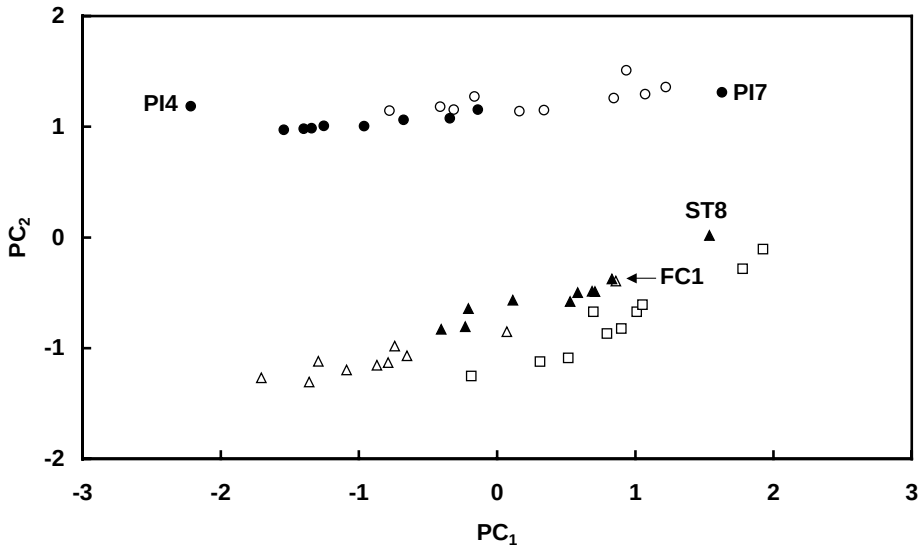

Fig. 2. Principal component graph for the detection of outliers ((○) BI, (△) FC, (□) PE, (●) PI, (▲) ST).

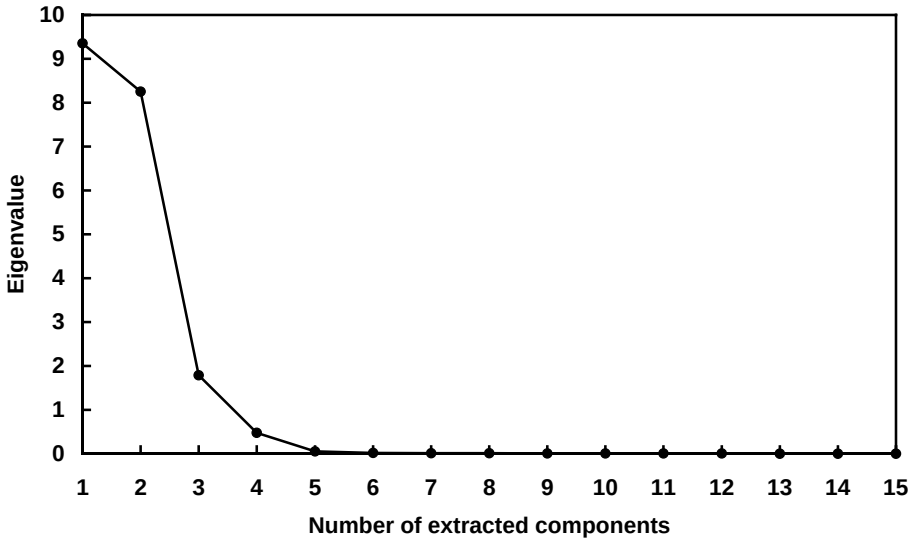

Fig. 3. Scree plot for deciding on the number of variables or components to be used in discriminant analysis ((●) eigenvalue of extracted component).

that the first three components had eigenvalues greater than unity (Kaiser criterion) while the first four or five components satisfied the scree test.

Changes in Wilks' λ values versus the number of original variables or components used are shown in Fig. 4. The use of a logarithmic scale for the Wilks' λ axis (showing values divided by 10^{-3}) was mandatory since a wide range of values had to be covered. As it was theoretically expected, the discriminatory power of the model increased as more and more variables (or components) were used. The change in the slope observed at the point that corresponded to three variables (or components) denoted that little improvement was achieved by further increasing the number of variables (or components). This point was in agreement with the Kaiser criterion result. Slightly better (smaller) Wilks' λ values were observed when components were used instead of original variables. At the same time, the percent average redundancy of the original variables in the discriminant

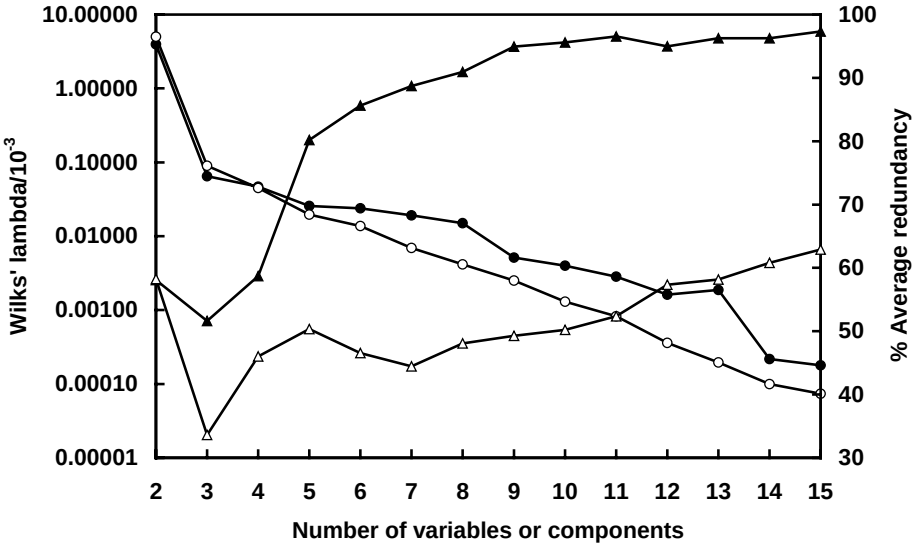

Fig. 4. Wilks' λ and percent average redundancy of variables or components in discriminant analysis ((●) Wilks' λ based on original variables, (○) Wilks' λ based on principal components, (▲) percent average redundancy of original variables, (△) percent average redundancy of principal components).

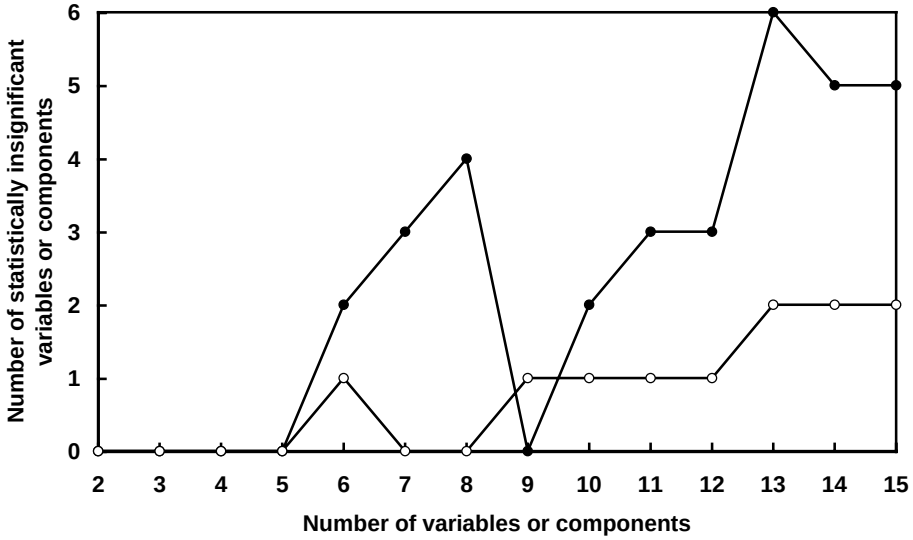

Fig. 5. Number of statistically insignificant variables or components in discriminant analysis ((●) number of statistically insignificant original variables, (○) number of statistically insignificant principal components).

model increased rapidly when more than four variables were used and eventually reached a plateau at ca. 96%. It is also worth mentioning that four variables is the predicted number of variables by the scree test. When principal components were used in the model, the same increase was observed at the point that corresponded to four components, but the final plateau did not exceed 63%. This was expected since principal components are orthogonal and uncorrelated by definition. The use of principal components as new latent variables in DA was therefore preferable.

Examination of the number of statistically insignificant (at $P = 0.05$) variables or components in the discriminant model (Fig. 5) showed that when five or less variables or components were entered into the DA module, they were all statistically significant at the $P = 0.05$ level. Again the use of principal components instead of original variables proved to be more adequate. In every case, the percent correct post hoc classification of the items in the training dataset was 100% except when only two variables or components were used, something that resulted in percent correct classification values of 91.3 and 84.8%, respectively.

Based on the findings mentioned above, we decided to use the scores of the first three principal components in the DA model. This allowed us to sacrifice as little discriminatory power as possible and also avoid entering too many new latent variables in the model. The three components used accounted for 96.97% of the total variance in the original data. The Wilks' λ of the calculated discriminant model was 8.98×10^{-5} and the model was found to be statistically significant at the $P = 0.05$ level with an average variable redundancy of 33.6%. The post hoc classification of the items in the training dataset was found to be 100% correct.

The three discriminant functions (canonical roots) calculated for the model were checked for their significance by means of a χ^2-test and successive removal of the roots until only one root remained. From the P-values in Table 1 it can be seen that all three roots were statistically significant at the $P = 0.05$ level.

The discriminant functions that were calculated were of the form:

$$L = b_0 + \sum_{i=1}^{3} b_i PC_i$$

The raw (b_i) and standardized (B_i) coefficients of the canonical roots are given in Table 2. Raw coefficients were

Table 1
Significance of discriminant functions

Removed roots	χ^2	d.f.	P-value
0	382.0371	12	<0.001
First	194.9673	6	<0.001
First, second	27.37222	2	<0.001

Table 2
Raw and standardized coefficients of discriminant functions

	Root 1	Root 2	Root 3
Raw coefficient			
b_0	7.06×10^{-15}	5.70×10^{-15}	-1.33×10^{-16}
b_1	−1.2365	0.2430	1.3203
b_2	3.8685	−6.6783	0.1139
b_3	−8.4158	−3.1056	−0.1416
Standardized coefficient			
B_1	−0.9193	0.1807	0.9816
B_2	0.5633	−0.9724	0.0166
B_3	−1.1575	−0.4271	−0.0195

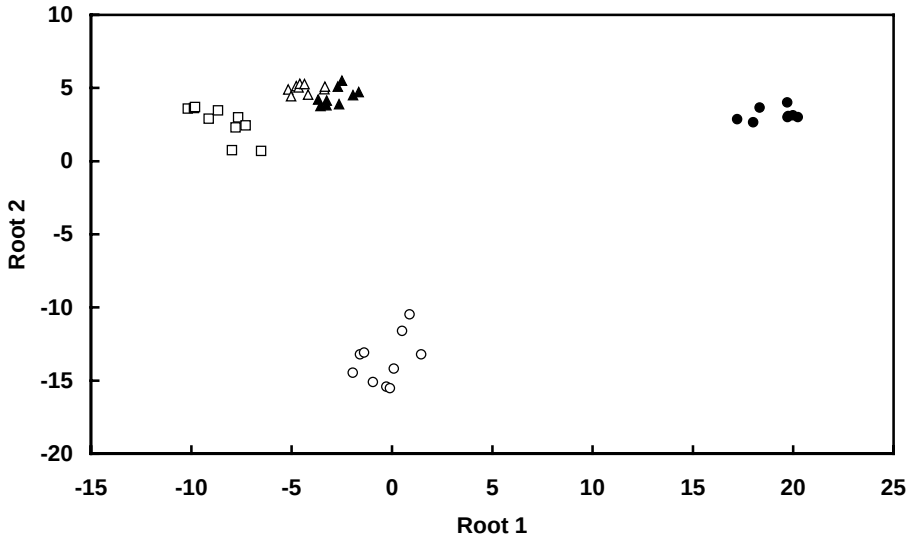

Fig. 6. Canonical score graph of the first two discriminant functions for the post hoc classification of the training dataset ((○) BI, (△) FC, (□) PE, (●) PI, (▲) ST).

Table 3
Statistically significant component loadings

PC_1			PC_2			PC_3		
Original variable	Loading	*P*-value	Original variable	Loading	*P*-value	Original variable	Loading	*P*-value
$\log_{10}(\%A_{418})$	0.9614	<0.001	$\log_{10}(\%A_{677})$	0.9980	<0.001	$\log_{10}(\%A_{495})$	0.7999	<0.001
$\log_{10}(\%A_{405})$	0.9551	<0.001	$\log_{10}(\%A_{660})$	0.9858	<0.001	$\log_{10}(\%A_{511})$	0.7374	<0.001
$\log_{10}(\%A_{443})$	0.9513	<0.001	$\log_{10}(\%A_{686})$	0.9748	<0.001	$\log_{10}(\%A_{482})$	0.7106	<0.001
$\log_{10}(\%A_{444})$	0.9483	<0.001	$\log_{10}(\%A_{696})$	0.7583	<0.001	$\log_{10}(\%A_{524})$	0.6330	<0.001
$\log_{10}(\%A_{707})$	0.9296	<0.001	$\log_{10}(\%A_{643})$	0.6645	<0.001	$\log_{10}(\%A_{537})$	0.4773	0.001
$\log_{10}(\%A_{728})$	0.9092	<0.001	$\log_{10}(\%A_{495})$	−0.3392	0.021	$\log_{10}(\%A_{464})$	0.4599	0.001
$\log_{10}(\%A_{464})$	0.8742	<0.001	$\log_{10}(\%A_{511})$	−0.5995	<0.001	$\log_{10}(\%A_{643})$	−0.7040	<0.001
$\log_{10}(\%A_{482})$	0.6910	<0.001	$\log_{10}(\%A_{524})$	−0.7661	<0.001	$\log_{10}(\%A_{622})$	−0.9469	<0.001
$\log_{10}(\%A_{696})$	0.5609	<0.001	$\log_{10}(\%A_{537})$	−0.8754	<0.001			
$\log_{10}(\%A_{495})$	0.4881	0.001	$\log_{10}(\%A_{570})$	−0.9760	<0.001			
$\log_{10}(\%A_{511})$	0.3059	0.039	$\log_{10}(\%A_{553})$	−0.9803	<0.001			

used for the calculation of the sample canonical scores and the drawing of the respective graph (Fig. 6). Only the first two roots are given since those were the ones that played the most important role in the discrimination. Root 1 was responsible for the separation ||PE||FC-ST-BI||PI||, whereas root 2 further aided the separation ||FC-ST||BI||. Root 3, which is not shown here, was responsible for the final separation ||FC||ST||.

Standardized coefficients were used for the estimation of the component contribution to the discrimination achieved by the respective discriminant function. From Table 2 it can be seen that PC_3 played the most important role in the discrimination achieved by root 1. From the statistically significant ($P < 0.05$) component loadings (Table 3) it can be seen that PC_3 mainly represented absorbance at 622 nm. At that wavelength, spectra belonging to the PI group demonstrated an absorbance shoulder (Fig. 1) that was absent in the other groups. The discriminatory power of root 1 is especially evident in Fig. 6 in which the PI group lies on the far right, completely separated from the other

Table 4
Contribution to discrimination and redundancy of components

Component	Partial Wilks' λ	*P*-value	Tolerance	Redundancy (%)
PC_1	0.2735	<0.001	0.5431	45.7
PC_2	0.0153	<0.001	0.7917	20.8
PC_3	0.0113	<0.001	0.6568	34.3

Table 5
Classification function coefficients

Coefficient	Ink group				
	BI	FC	PE	PI	ST
b_0'	−94.3730	−24.9465	−41.5852	−189.4998	−15.4220
b_1'	−3.1026	4.4046	12.4671	−22.5842	5.2652
b_2'	89.6619	−50.2556	−50.4302	52.7334	−40.2094
b_3'	45.1379	21.7563	63.0824	−170.7646	9.7896

groups. Table 2 also shows that root 2 was mainly affected by PC_2. From Table 3 it can be seen that this component represented absorbance in the range of 660–686 nm and at 553 and 570 nm. Fig. 1 shows that the first range corresponded to the second peak that characterized BI and PI inks with *%A* values being significantly higher in the case of BI inks. Absorption at 553 and 570 nm was also different for BI inks and corresponded to the ascending part of the band with a slope that was smaller than the slopes of the other curves at those wavelength points. These features were responsible for the separation of the BI group by means of root 2, something that is also evident in Fig. 6. Similarly, PC_1 scores affected the discrimination achieved by root 3 (Table 2). This component represented absorbance values in the range of 405–444 nm and at 707 and 728 nm (Table 3). All of these wavelengths corresponded to very low absorbance values at the spectra extremes (Fig. 1). Although unnoticeable to the naked eye, these subtle differences in absorption were the ones that achieved the separation of the somewhat overlapping FC and ST groups.

The overall contribution of the principal components to the discriminant model is shown in Table 4. The contribution followed the order $PC_3 > PC_2 > PC_1$ and all components were found to be statistically significant at the $P = 0.05$ level. This order is in perfect agreement with what was mentioned above about the role and nature of each discriminant function.

The DA was completed with the calculation of the so-called classification functions. These functions allow the post hoc classification of the items in the training dataset or the classification of new items. To classify new samples one would have to calculate the relevant component scores first and then enter the results into the classification functions. The function yielding the highest result would indicate the group to which the new sample belonged. Five classification functions were calculated (one for each group) and were of the form:

$$C = b_0' + \sum_{i=1}^{3} b_i' PC_i$$

Their coefficients are given in Table 5 and the post hoc classification of the training dataset by means of these functions was found to be 100% correct (no misclassifications occurred).

4. Conclusions

This study showed that excellent discrimination (100% correct classification of the training dataset) between inks of different brands could be achieved by examination of the Vis spectra of ethanolic ink solutions and application of a multivariate chemometrics protocol for the study of the analytical data. A discriminant model was calculated with a Wilks' λ of 8.98×10^{-5} and was found to be statistically significant at the $P = 0.05$ level. Problems with data multicollinearity were effectively dealt with by means of PCA, which was used for the calculation of three new latent variables (principal components) that were used in DA. The average redundancy of the latent variables in the discriminant model was 33.6%. The study showed that discrimination of the samples was based on differences that regarded (a) the shape of the first spectral band, which was present in all cases, (b) the intensity of the second band, which appeared only in the case of two inks and (c) the absorption of inks at the spectra extremes. The proposed method is similar to chromatography, in the sense that it detects the coloring materials in the sample, but avoids all the time consuming steps that characterize separation techniques. We believe that a more thorough study of the system could be based on the use of many more ink samples representative of the blue ballpoint pen ink population. Such an experiment should also regard the variability between inks of the same brand but of different batches and should involve either the extraction of inks from documents (destructive technique) or direct observation of the radiation reflected by the ink on the document (non-destructive technique). The calculated models should also be tested on new samples instead of items from the training dataset to ensure a better assessment of their usefulness. This would help establish an appropriate protocol that could allow forensic document examiners to draw objective conclusions regarding the similarity of inks used to write various sections of a document.

Appendix A. Statistical background

A.1. Cluster analysis

The term *cluster analysis* is used to describe a number of different classification algorithms. Generally, these

algorithms allow the organization of observed data into meaningful structures, thus promoting the development of taxonomies. In its simplest form, CA is based on a joining or tree clustering algorithm. Its purpose is to join objects into successively larger clusters (hierarchical tree) using some measure of similarity between the objects. The most commonly employed similarity measure in this technique is the Euclidian distance between the objects. If the objects have been measured on m variables, an m-dimensional space exists in which the Euclidian distance between objects k and l is given by the formula

$$D_{kl} = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2}$$

Another form of CA is the K-means algorithm that addresses a different problem, namely that of which objects belong to a certain predefined number of clusters. To answer this question one needs an iterative method according to which k random clusters are formed initially and then objects are moved between these clusters so that the within cluster variance is minimized, while the between cluster variance is maximized. Although it is more common to employ CA to classify objects on the basis of their variables, it is possible to run a K-means algorithm on the variables over the objects, thus achieving the grouping of variables that carry similar information about the objects.

A.2. Principal component analysis

Principal component analysis is a technique that belongs to the broader field of factor analysis. Generally, FA aims at (a) reducing the number of variables on which the objects of a dataset were measured and (b) detecting structure in the relationships between the variables. Principal component analysis achieves the aims mentioned above by using linear combinations of the original variables (manifest variables) to yield new variables called principal components or PCs (latent variables). The extraction of the PCs is successive with the first PC explaining most of the variance in the original data. The second PC can then be extracted to explain most of the remaining unexplained variance. This procedure can be repeated m times for m manifest variables until all the original variance has been explained. By definition, the extracted components are orthogonal and therefore uncorrelated. By extracting only the first two or three PCs one can project the objects of a dataset on a plane or in a three-dimensional space respectively and visualize an otherwise unperceivable m-dimensional space.

The correlations of the PC scores with the original scores on the m manifest variables are called component loadings and form the basis for the qualitative interpretation of the extracted components. A technique called Varimax rotation is used to ensure that the loading of a manifest variable is maximized on one component while it is minimized on all other components. This results in a much easier interpretation of the PCs. If some of the original variables are already correlated, they are expected to load highly on the same component.

Each extracted component is characterized by its eigenvalue which roughly corresponds to the number of manifest variables this component represents. For the decision concerning the number of PCs that should be extracted for a given dataset two criteria have been extensively used: the Kaiser criterion and the scree test. According to the Kaiser criterion, only components with eigenvalues greater than unity should be extracted, the rationale being that components representing less than one variable should not be taken into account. On the other hand, the scree test requires the plotting of eigenvalues against the number of extracted components and the determination of the point where the plot levels off. Beyond this point, no further improvement in variance explanation can be achieved and more components are not needed.

A.3. Discriminant analysis

Discriminant analysis resembles PCA in the sense that new latent variables are formed from the original manifest variables, but the requirement is that maximum separation of the objects is achieved. The new latent variables which are also called discriminant functions or canonical roots are a linear combination of the manifest variables and are orthogonal.

The discriminatory power of the model calculated this way is assessed by means of the Wilks' λ statistic, which is given by the formula

$$\lambda = \frac{\det(W)}{\det(T)}$$

where $\det(W)$ is the determinant of the within-groups variance–covariance matrix and $\det(T)$ is the determinant of the total variance–covariance matrix. The smaller the Wilks' λ value is, the more effective the model is. Another measure of the individual contribution of each variable in the discriminant model is the partial Wilks' λ statistic, which is the ratio of the Wilks' λ value after adding the respective variable over the Wilks' λ value before adding the variable. Smaller values of the partial Wilks' λ statistic denote a higher contribution of the respective variable.

The effectiveness of the discriminant model can also be checked by running a post hoc classification of the training dataset objects. According to this technique, the original objects are treated as new ones and are classified by means of the classification functions calculated for the respective model. A variant of this technique is based on the classification of entirely new objects and observation of any misclassifications.

References

[1] C. Vogt, J. Vogt, A. Becker, E. Rohde, Separation, comparison and identification of fountain pen inks by

capillary electrophoresis with UV-visible and fluorescence detection and by proton-induced X-ray emission, J. Chromatogr. A 781 (1997) 391–405.
[2] C. Roux, M. Novotny, I. Evans, C. Lennard, A study to investigate the evidential value of blue and black ball-point pen inks in Australia, Forensic Sci. Int. 101 (1999) 167–176.
[3] J.A. Zlotnick, F.P. Smith, Chromatographic and electrophoretic approaches in ink analysis, J. Chromatogr. B 733 (1999) 265–272.
[4] R. Saferstein, Criminalistics: An Introduction to Forensic Science, seventh ed., Prentice-Hall, Englewood Cliffs, NJ, 2001, p. 468.
[5] J.W. Brackett Jr., L.W. Bradford, Comparison of ink writing on documents by means of paper chromatography, J. Crim. Law Crim. Police Sci. 43 (1952) 530–539.
[6] D.A. Crown, J.V.P. Conway, P.L. Kirk, Differentiation of blue ball-point pen inks, J. Crim. Law Crim. Police Sci. 52 (1961) 338–343.
[7] M. Lederer, M. Schudel, Adsorption chromatography on cellulose V: a simple chromatographic system for the identification of inks, J. Chromatogr. 475 (1989) 451–456.
[8] G.R. Nakamura, S.C. Shimoda, Examination of microquantity of ball-point inks from documents by thin-layer chromatography, J. Crim. Law Crim. Police Sci. 56 (1965) 113–118.
[9] R.S. Verma, K.N. Prasad, G.J. Misra, Thin-layer chromatographic analysis of fiber-tip pen inks, Forensic Sci. Int. 13 (1979) 65–70.
[10] V.N. Aginsky, Forensic examination of 'slightly soluble' ink pigments using thin-layer chromatography, J. Forensic Sci. 38 (1993) 1131–1133.
[11] D. Doud, Chromatographic analysis of inks, J. Forensic Sci. 3 (1958) 486–492.
[12] J.A. Tappolet, High-performance thin-layer chromatography: its application to the examination of writing inks, Forensic Sci. Int. 22 (1983) 99–109.
[13] R.N. Totty, M.R. Ordidge, L.J. Onion, Comparison of the use of visible microspectrometry and high performance thin-layer chromatography for the discrimination of aqueous inks used in porous tip and roller ball pens, Forensic Sci. Int. 28 (1985) 137–144.
[14] K.M. Varshney, T. Jettappa, V.K. Mehrotra, T.R. Baggi, Ink analysis from typed script of electronic typewriters by high performance thin layer chromatography, Forensic Sci. Int. 72 (1995) 107–115.
[15] R.M. Kevern, Infrared luminescence from thin layer chromatograms of inks, J. Forensic Sci. Soc. 13 (1973) 25–28.
[16] R.D. Blackledge, M. Iwan, Differentiation between inks of the same brand by infrared luminescence photography of their thin-layer chromatography, Forensic Sci. Int. 21 (1983) 165–173.
[17] V.N. Aginsky, Comparative examination of inks by using instrumental thin-layer chromatography and microspectrophotometry, J. Forensic Sci. 38 (1993) 1111–1130.
[18] L.F. Colwell, B.L. Karger, Ball-point pen ink examination by high pressure liquid chromatography, J. Assoc. Off. Anal. Chem. 60 (1977) 613–618.
[19] P.C. White, B.B. Wheals, Use of a rotating disc multi-wavelength detector operating in the visible region of the spectrum for monitoring ball pen inks separated by high-performance liquid chromatography, J. Chromatogr. 303 (1984) 211–216.
[20] A. Löfgren, J. Andrasko, HPLC analysis of printing inks, J. Forensic Sci. 38 (1993) 1151–1160.
[21] J.W. Thompson, Identification of ink by electrophoresis, J. Forensic Sci. Soc. 7 (1967) 199–202.
[22] H.W. Moon, Electrophoretic identification of felt-tip pen inks, J. Forensic Sci. 25 (1980) 146–149.
[23] S. Fanali, M. Schudel, Some separations of black and red water-soluble fiber-tip pen inks by capillary zone electrophoresis and thin-layer chromatography, J. Forensic Sci. 36 (1991) 1192–1197.
[24] E. Rohde, A.C. McManus, C. Vogt, W.R. Heineman, Separation and comparison of fountain pen inks by capillary electrophoresis, J. Forensic Sci. 42 (1997) 1004–1011.
[25] J.A. Zlotnick, F.P. Smith, Separation of some black roller-ball pen inks by capillary electrophoresis: preliminary data, Forensic Sci. Int. 92 (1998) 269–280.
[26] A. Giles, The forensic examination of documents, in: P. White (Ed.), Crime Scene to Court: The Essentials of Forensic Science, The Royal Society of Chemistry, Cambridge, 1998, pp. 123–125.
[27] N.C. Thanasoulias, E.T. Piliouris, M.S.E. Kotti, N.P. Evmiridis, Application of multivariate chemometrics in forensic soil discrimination based on the UV-Vis spectrum of the acid fraction of humus, Forensic Sci. Int. 130 (2002) 73–82.
[28] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988, pp. 407–409.
[29] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1960) 141–151.
[30] R.B. Cattell, The scree test for the number of factors, Multivar. Behav. Res. 1 (1966) 245–276.