Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Application of bivariate statistics to full wine bottle diamagnetic screening data

S.J. Harley, V. Lim, M.P. Augustine*

Department of Chemistry, One Shields Avenue, University of California, Davis, CA 95616, United States

ARTICLE INFO

Article history:
Received 23 November 2011
Accepted 20 December 2011
Available online 29 December 2011

Keywords:
Principal component analysis
Bivariate correlated Student distribution
Statistics
Diamagnetic
Wine

ABSTRACT

A bivariate correlated Student distribution is applied to full wine bottle diamagnetic screening measurements. Previous work involving a limited number of rare wines indicated that like wines cluster in a plot of the first two principal component scores derived from a covariance matrix of the diamagnetic screening measurements. This study extends the approach to a much larger, statistically meaningful sixty bottle wine library where bivariate statistics are used to comment on the measured data. The full bottle diamagnetic screening of thirty-six identically labeled, sealed bottles of wine obtained from four different sources combined with principal component analysis data reduction followed by treatment with a bivariate distribution permit the effect of wine transport and storage to be observed. The usefulness and future success of the method towards the identification of counterfeit wines is mentioned.

1. Introduction

The wine industry has long sought a way to authenticate their products. The ease associated with counterfeiting wine, often involving minimal bottle, wine, and cork expense, coupled with the hundreds of percent investment return has resulted in an explosion of counterfeit bottles entering the market [1–3]. As auction houses have been held legally liable for the distribution of counterfeit wine [4], the need for authentication is a priority for the wine industry. The use of modern analytical instrumentation in wine fingerprinting is an attractive way to authenticate wine [5–15]. This can be accomplished by comparing chemical concentration results for a potential counterfeit wine to known authentic standards. Although in practice these approaches can separate different wine vintages and thus presumably identify counterfeit wine, all of the methods are destructive and require that the wine bottle be violated to obtain a sample for analysis. As opening or piercing the collectible wine bottle immediately ruins the investment, the wine industry has introduced external anti-counterfeiting measures to sealed bottles such as proprietary invisible inks [16], holographic images [17], and laser etching [18]. Although these efforts are effective deterrents for the counterfeiting of newly bottled wine, they do not address either older wine bottled before these external container modifications were introduced or refill counterfeiting [19].

In response to the need for a way to noninvasively and nondestructively screen full intact bottles of wine for counterfeits, devices that probe the wine dependent electric and magnetic susceptibility were constructed [20,21]. At low frequencies these susceptibilities are sensitive to the specific concentrations of organic and inorganic molecules and ions in bottled wine. These amounts depend on the viticultural and enological wine history – properties that afford region and vintage specific information. Both of these approaches work by measuring the amplitude and phase difference between the applied and measured time dependent electric or magnetic fields induced in the wine bottle by the generated electric polarization or magnetization. Up to several thousand amplitude attenuation and phase retardation values in the $500\,\text{kHz} < \nu < 30\,\text{MHz}$ frequency range are obtained from a homebuilt magnetic screening device [21]. Principal component analysis (PCA) is applied to this large data set to reduce dimensionality. To date, truncation of the PCA treated data to just the first two principal components $\overrightarrow{PC1}$ and $\overrightarrow{PC2}$ recovers 94% of the entire data set variance in the worst case. In other words, even though thousands of data points describing the behavior of either the low frequency electric or magnetic susceptibility of bottled wines in a massive collection or library are measured, PCA reduces the amount of data to just two points for each bottle studied. This information corresponds to projections of the measured data for a given bottle onto the first two principal components $\overrightarrow{PC1}$ and $\overrightarrow{PC2}$ for the entire library. A two dimensional plot of these principal component scores, PC_1 and PC_2, leads to clustering of the data for identical full bottles of wine. Specifically, PCA transformed and reduced data for identical wines, those from the same vineyard, type, and vintage, cluster in a PC_1 and PC_2 scores plot while different wines cluster in other regions of the PC_1 and PC_2 scores space [20,21]. Comparison of a potential counterfeit wine bottle to library results containing a collection of known authentic bottles will lead to a point on the

* Corresponding author. Tel.: +1 530 754 7550; fax: +1 530 752 8995.
E-mail addresses: augustin@chem.ucdavis.edu, maugust@ucdavis.edu (M.P. Augustine).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.12.068

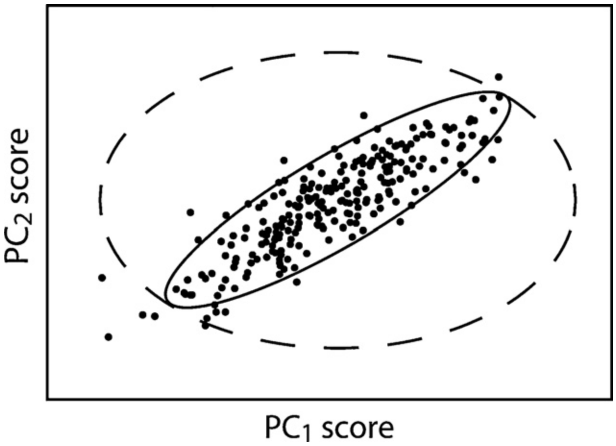

Fig. 1. Example PC$_1$ and PC$_2$ score subset with equiprobability contours corresponding to correlated and uncorrelated distributions. The solid and dashed ellipses, respectively, correspond to the same probability contours for a correlated and an uncorrelated Gaussian distribution.

PC$_1$ and PC$_2$ scores plot. The bottle *is* a member of the appropriate data cluster in the case that the suspect bottle is authentic and *is not* a member if the suspect bottle is a counterfeit.

To date fairly small collectible wine libraries have been studied with these noninvasive wine screening approaches as the emphasis of previous work involved exploring the potential of low frequency dielectric and diamagnetic studies in rare wine analysis [20,21]. The limited library size permitted visual arguments regarding suspect counterfeit bottle membership to known authentic wine data clusters. Unfortunately, these empirical arguments reveal nothing about the statistical certainty of the membership of a potential counterfeit wine data point to an authentic wine data cluster in the PC$_1$ and PC$_2$ scores plot. Both *K*-means clustering [22] and hierarchal analysis [23] are popular statistical methods often applied to data cluster problems but neither of these two approaches is useful here as they screen data with the purpose of identifying clusters. This information is already known as the electric and magnetic data for wines of like vineyard source, type, and year defines clusters in the PC$_1$ and PC$_2$ scores plot. What is required is a way to mathematically and statistically define data cluster attribution or membership. The approach discussed below uses the PC$_1$ and PC$_2$ scores corresponding to a collection of known identical authentic wines to establish a cluster membership probability. Comparison of the PC$_1$ and PC$_2$ scores for a suspect counterfeit bottle to this authentic wine probability map yields a statistically sound estimate of the suspect bottle authenticity.

Common approaches toward the application of statistical cluster attribution employ uncorrelated statistics. These are often justified with the claim that PCA uncorrelates the principal components [24]; the correlation coefficient of the transformed data goes to zero for the entire dataset. However, there is nothing that precludes the formation of correlated clusters within a principal component scores plot for portions of the data set. Since this application demands the definition of a statistical certainty for individual clusters within the entire PC$_1$ and PC$_2$ score space instead of a statistical description of the entire PC$_1$ and PC$_2$ dataset, correlated statistics are required. Fig. 1 visually demonstrates the statistical error introduced by the application of uncorrelated statistics to a subset of a synthetic PC$_1$ and PC$_2$ score distribution. The numerical and statistical details of the contours or ellipses overlaid on the synthetic PC$_1$ and PC$_2$ score distribution in Fig. 1 will be discussed in detail later in connection with real data. Both ellipses refer to the same probability. Specifically, any given point along these curves has an equal probability of belonging to the cluster. The solid and dashed ellipses correspond to the application of correlated and uncorrelated statistics to the PC$_1$ and PC$_2$ score cluster, respectively. Any point along the solid ellipse is the same distance from the correlated cluster and thus has the same probability of belonging to the cluster. The same empirical argument applied to the dashed ellipse reveals that some parts of the uncorrelated equal probability dashed line actually have a lower or higher position dependent probability of belonging to the cluster. This graphical comparison suggests that the application of uncorrelated statistics falsely presumes that a PC$_1$ and PC$_2$ score close to the cluster will have the same statistical membership probability as a score displaced from the cluster. This is an error too significant to retain statistical integrity.

The next section describes the application of bivariate statistics appropriate for statistically limited amounts of data to this wine screening approach. The statistical protocol is then applied to PC$_1$ and PC$_2$ score clusters determined from full bottle diamagnetic screening data for a thirty-six bottle wine library. These bottles are comprised of three, twelve bottle sets of different vintage wine with known viticultural and enological history. The combination of noninvasive full bottle wine analysis [21], PCA data reduction, and the statistical method described in the next section were collectively used to explore the effects of transportation and storage on one of these twelve bottle sets. Twenty-four additional bottles of the same vintage and type of wine were obtained from three different sources and diamagnetically screened. The resulting amplitude and phase data were treated with PCA and statistically examined to glean insight into the relationship between measured data, wine history, transportation, storage, and the ability of the full bottle hardware to identify counterfeit wine.

2. Statistical analysis

As mentioned above, PCA treatment of measured full bottle dielectric and diamagnetic data leads to like wine clustering in the PC$_1$ and PC$_2$ scores two-dimensional space. Combining this fact with the limited number of available authentic wines $n \approx 2$–5, a number much less than the $n > 29$ samples necessary to permit application of Gaussian statistics, prompted the use of the bivariate correlated Student distribution

$$P(\mathrm{PC}_1,\mathrm{PC}_2)=\frac{\left|\Sigma^{-1}\right|^{1/2}}{2\pi}\left(1+\sum_{i,j=1}^{2,2}\frac{(\Sigma^{-1})_{i,j}\mathrm{PC}_i\mathrm{PC}_j}{n}\right)^{-(n+2)/2} \tag{1}$$

to measured data clusters in the PC$_1$ and PC$_2$ score plot. The PC$_i$ and PC$_j$ principal component scores in Eq. (1) correspond to the projection of the measured data onto the $\overrightarrow{\mathrm{PC}1}$ and $\overrightarrow{\mathrm{PC}2}$ principal components, n is the number of wines in the library, and the scaling matrix is

$$\Sigma=\begin{bmatrix}\sigma_{\mathrm{PC}_1\mathrm{PC}_1} & \rho\sigma_{\mathrm{PC}_1}\sigma_{\mathrm{PC}_2}\\ \rho\sigma_{\mathrm{PC}_1}\sigma_{\mathrm{PC}_2} & \sigma_{\mathrm{PC}_2\mathrm{PC}_2}\end{bmatrix}. \tag{2}$$

Since only two principal components are used in this analysis, the definition of the PC$_1$ and PC$_2$ score labels on the variance and standard deviations in Eq. (2) as x and y, respectively allows the PC$_1$ standard deviation to be written as

$$\sigma_x=\frac{1}{\sqrt{n-1}}\sum_{i=1}^{n}\left|x_i-\bar{x}\right| \tag{3}$$

with a similar definition for σ_y, the PC$_2$ score standard deviation. Likewise the PC$_1$ score variance is

$$\sigma_{xx}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) \tag{4}$$

with a similar expression for the PC$_2$ variance σ_{yy}. Finally, the PC$_1$ and PC$_2$ score covariance is

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}), \tag{5}$$

and the correlation is defined as $\rho=\sigma_{xy}/\sigma_x\sigma_y$. Application of these definitions to the scaling matrix in Eq. (2) and the bivariate correlated Student distribution in Eq. (1) followed by expansion of the sum in Eq. (1) yields

$$P(x,y) = \frac{1}{2\pi\sqrt{-(\rho^2-1)\sigma_{xx}\sigma_{yy}}} \times \left[\frac{n(\rho^2-1)\sigma_{xx}\sigma_{yy} - x^2\sigma_{yy} + 2xy\rho\sigma_x\sigma_y - y^2\sigma_{xx}}{n(\rho^2-1)\sigma_{xx}\sigma_{yy}}\right]^{-(n+2)/2} \tag{6}$$

in the original two dimensional PC$_1$ and PC$_2$ scores space. The volume under the $P(x,y)$ distribution shown in Eq. (6) bounded by a constant $P(x_0,y_0)$ contour can be used to attach statistical significance to measured data. Unfortunately, the $P(x,y)$ distribution as written in Eq. (6) is not immediately amenable to the calculation of this volume or membership likelihood as the equal probability contour defined by $P(x_0,y_0)$ surrounding the peak of the $P(x,y)$ distribution is not single valued. To accomplish this integration and to avoid handling multi-valued, ill defined functions, the distribution in Eq. (6) is re-expressed in terms of the Eigen basis of the scaling matrix Σ as

$$P(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y}\left\{1+\frac{1}{n}\left[\left(\frac{X}{\sigma_X}\right)^2+\left(\frac{Y}{\sigma_Y}\right)^2\right]\right\}^{-(n+2)/2} \tag{7}$$

where the capital X and Y letters label the independent variables and σ_X and σ_Y are the uncorrelated standard deviations along these two coordinates. They are given by the eigenvalues of the scaling matrix as

$$\sigma_X^2 = \sigma_{XX} = \frac{1}{2}\left((\sigma_{xx}+\sigma_{yy}) - \sqrt{4\sigma_{xy}^2+(\sigma_{xx}-\sigma_{yy})^2}\right)$$
and
$$\sigma_Y{}^2 = \sigma_{YY} = \frac{1}{2}\left((\sigma_{xx}+\sigma_{yy}) + \sqrt{4\sigma_{xy}^2+(\sigma_{xx}-\sigma_{yy})^2}\right). \tag{8}$$

In the $n\to\infty$ limit the $P(X,Y)$ distribution shown in Eq. (7) reduces to a two dimensional Gaussian peak with standard deviations σ_X and σ_Y along the X and Y coordinates, respectively. Rearrangement of Eq. (7) in terms of the constant $P(X_0,Y_0)$ gives the ellipse defined by

$$\left(\frac{X}{\sigma_X}\right)^2+\left(\frac{Y}{\sigma_Y}\right)^2 = n[(4\pi^2P(X_0,Y_0)^2\sigma_{XX}\sigma_{YY})^{-1/(n+2)}-1] \tag{9}$$

When written in this way a particular choice of $P(X_0,Y_0)$ generates an elliptical contour surrounding the $P(X,Y)$ distribution function peak. The volume of this peak within the elliptical contour, referred to here as η and defined by the constant $P(X_0,Y_0)$, can be used to make statistically significant comments regarding membership of experimental data for suspect counterfeit wines to principal component data clusters for known authentic wine. The volume of $P(X,Y)$ within the region bounded by the $P(X_0,Y_0)$ elliptical contour or equivalently the value of η can be determined by using Eq. (9) to solve for the limits of integration over Y in terms of X as

$$Y_\pm = \pm\sigma_Y\sqrt{n[(4\pi^2P(X_0,Y_0)^2\sigma_{XX}\sigma_{YY})^{-1/(n+2)}-1]-\left(\frac{X}{\sigma_X}\right)^2} \tag{10}$$

The integration limits for X are similarly determined as

$$X_\pm = \pm\sigma_X\sqrt{n[(4\pi^2P(X_0,Y_0)^2\sigma_{XX}\sigma_{YY})^{-1/(n+2)}-1]} \tag{11}$$

from Eq. (9) by solving for X with $Y=0$. The value for η is then given by

$$\eta = \frac{1}{\pi\sigma_X\sigma_Y}\int_0^{X_+}\int_{Y_-}^{Y_+}\left\{1+\frac{1}{n}\left[\left(\frac{X}{\sigma_X}\right)^2+\left(\frac{Y}{\sigma_Y}\right)^2\right]\right\}^{-(n+2)/2} \times dYdX\times 100 \tag{12}$$

When X_+ and Y_+ are extended to $+\infty$ and Y_- is extended to $-\infty$, $\eta=100\%$ meaning that the elliptical contour contains the entire distribution. Finite values of these integration limits yield $\eta<100\%$.

3. Experimental

A wine library containing sixty sealed, full, intact bottles of Pianetta wine were prepared. Thirty-six of these bottles correspond to three sets of twelve bottles of the 2005 Cabernet Sauvignon, the 2006 Cabernet Sauvignon, and the 2006 Petite Syrah. All of the bottles in these sets were obtained directly from the winery. The grapes used to prepare each of these separate wines were from the same separate areas of the vineyard. The casks used for each of the separate wines were identical and, following bottling, were stored in a cellar at 14 °C and 61% humidity. Twenty-four additional bottles of the 2005 Cabernet Sauvignon divided into one set of twelve bottles and two sets of six bottles were obtained from three additional sources. The set of twelve bottles was obtained from a collector that transported the bottles from the winery prior to storage at atmospheric conditions in Fremont, CA. The two additional sets of six bottles were obtained from wine merchants in Salinas, CA and Monterey, CA. Nothing is known about the transportation and storage history of the two six bottle sets.

The diamagnetic absorption characteristics for all sixty bottles were obtained using a noninvasive, nondestructive full bottle magnetic susceptibility based wine screening device described in detail elsewhere [21]. Here the amplitude attenuation and phase retardation of an applied oscillating magnetic field at fifteen frequencies in the 500 kHz < ν < 30 MHz range yielded thirty experimental data points for each bottle and served as input for PCA.

All PCA calculations and data handling were accomplished with MATLAB while the statistical methods used to establish the data confidence limits described in Section 2 above were implemented with Mathematica.

4. Results

A summary of the PC$_1$ and PC$_2$ scores for the three sets of twelve bottles obtained directly from the Pianetta winery is shown in Fig. 2. The crosses, open inverted triangles, and open lozenges in Fig. 2, respectively correspond to the PC$_1$ and PC$_2$ scores for the 2005 Cabernet Sauvignon, the 2006 Cabernet Sauvignon, and the 2006 Petite Syrah. A comparison of the PC$_1$ and PC$_2$ scores for the additional three sets of the 2005 Cabernet Sauvignon obtained from other sources to the PC$_1$ and PC$_2$ scores for the winery acquired wine is provided in Fig. 3. The crosses shown in Fig. 3(a) again pertain to the twelve bottles of the 2005 Cabernet Sauvignon obtained directly from the winery while the "*X*" symbols correspond to experimentally determined PC$_1$ and PC$_2$ scores for the twelve bottles of 2005 Cabernet Sauvignon acquired from the private wine collector. The inset in Fig. 3(a) is a uniform expansion of the PC$_1$ and PC$_2$ score axes to aid in data cluster visualization. The open circles and squares shown in Fig. 3(b) correspond to the PC$_1$ and PC$_2$ scores for the two sets of six bottles of the 2005 Cabernet Sauvignon obtained from the Salinas, CA and Monterey, CA based wine merchants. A summary of the PC$_1$ and PC$_2$ scores for the entire set of 2005 Cabernet Sauvignon measurements is provided in Fig. 3(c). The ellipses shown in Figs. 2 and 3 were generated using a

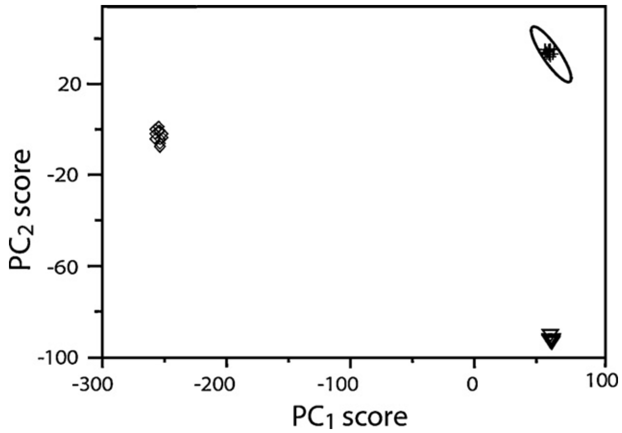

Fig. 2. Principal component scores plot corresponding to three sets of twelve bottles of Pianetta wine obtained directly from the winery. The crosses, open inverted triangles, and open lozenges, respectively, correspond to the first two principal component scores for PCA reduced full, sealed bottle diamagnetic screening data for the 2005 Cabernet Sauvignon, the 2006 Cabernet Sauvignon, and the 2006 Petite Syrah. The 0.5% ellipse shown as the solid line was generated from the data shown in Fig. 3.

Mathematica algorithm that implements the bivariate correlated Student distribution statistics described in Section 2. The value $\eta = 95\%$ was used to calculate the ellipses shown in Fig. 3(a) and (b) while $\eta = 99.5\%$ for the ellipse shown in Fig. 3(c) and reproduced in Fig. 2.

5. Discussion

Application of PCA to a set of identical sealed wine bottles instead of a library of different wines may seem attractive as significantly fewer data points are required and the resultant PCA transformed data would be inherently uncorrelated. This approach would eliminate the need for the statistical analysis used here. However, proceeding in this way is unattractive as a fingerprinting method. Restriction of the library to just one member limits the characterization of the cluster to just two values, σ_x and σ_y. On the other hand, consideration of the full library in the analysis expands the number of characteristic values to five corresponding to σ_x, σ_y, ρ, and the centroid x and y values. As the robustness of any identification method is proportional to the number of measured variables, full library analysis is the more informative method.

The combination of full bottle diamagnetic screening with PCA to separate different vintage and type of wine for twelve bottles each of three different vintages of Pianetta wine obtained directly from the vineyard is shown in Fig. 2. Despite the fact that the 2005 and 2006 Cabernet Sauvignon have the same color, texture, and taste established by sampling the bottle contents, the approach easily separates these wines based on cluster positions in the PC_1 and PC_2 scores plot provided in Fig. 2. The 0.5% probability ellipse shown in Fig. 2 will be discussed below. Fig. 2 demonstrates that the method is capable of easily separating wines obtained from the same vineyard. It is interesting to note that the PC_1 scores for the two Cabernet Sauvignons group together. This observation is reasonable as the wines are from the same vineyard. Overall, the constituents in the wine are the same; the type of grape and the location of the patch of grapes did not change. The noticeable changes in Cabernet Sauvignon PC_2 score could be related to annual variations in weather, fertilizer used during growth, fruit treatment during harvesting, etc. An understanding of this effect is the subject of ongoing work.

The wines in this estate stored library were prepared with grapes from the same vineyard area, fermented in the same cask, and stored in the same proper way. If the full bottle diamagnetic screening approach can successfully identify and essentially fingerprint collectible wine then the PC_1 and PC_2 score cluster for a given wine must be resolved from the PC_1 and PC_2 score clusters for other wines. Mechanisms not present in this estate wine library such as the effects of transport and storage, parameters that are often unknown prior to purchase of a collectible wine, can broaden the PC_1 and PC_2 score cluster distribution for a given wine. This effect can potentially interfere with wine comparisons using the susceptibility approach as PC_1 and PC_2 score clusters for different wines may no longer be resolved. The comparison and statistical analysis of estate stored wine PC_1 and PC_2 score clusters to those PC_1 and PC_2 score clusters for the same wine obtained from other sources with different transport and storage history is used to address this issue.

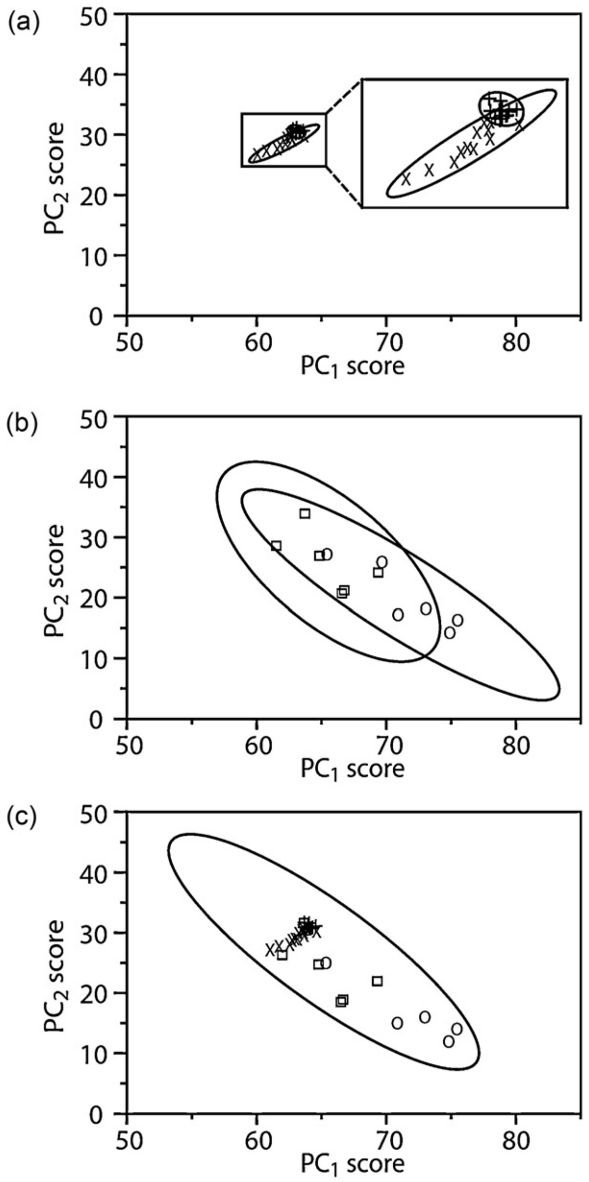

Fig. 3. Principal component scores plots comparing PCA reduced full, sealed bottle diamagnetic screening data for the Pianetta 2005 Cabernet Sauvignon obtained from four different sources. The cross and "X" symbols in Fig. 3(a) and (c), respectively, correspond to bottles acquired directly from the winery and from a private wine collector. The open circles and open squares in Fig. 3(b) and (c) pertain to bottles obtained from wine merchants in Salinas, CA and Monterey, CA, respectively. The 5% ellipses shown in (a) and (b) and the 0.5% ellipse shown in (c) were generated with a Mathematica algorithm that implements the statistics described in Section 2.

Twenty-four additional Pianetta 2005 Cabernet Sauvignon bottles from three different sources were used to increase the size of the estate wine library and to evaluate the effects of wine transport and storage on the measured results. It is likely that the three alternative sources for the Pianetta 2005 Cabernet Sauvignon adequately mimic anticipated conditions of rare collectible wine where the history of the wine prior to purchase is often not available or even known. One source, the private wine collector, probes the effect of PC_1 and PC_2 cluster distribution on known transport and storage history. Here wine was transported by automobile at room temperature from the Pianetta winery in Paso Robles, CA to Fremont, CA in 2006 and stored on a shelf subject to atmospheric temperature and humidity variations appropriate for Fremont, CA. Nothing is known about the transport and storage history of the Pianetta 2005 Cabernet Sauvignon bottles obtained from the wine merchants in Salinas, CA and Monterey, CA other than the fact that the wines were shipped from the merchant to the corresponding author's laboratory over the course of one week. Presumably these wines were either shipped to or transported by automobile from the winery to the merchant locations and upon reaching the merchant were appropriately stored in a temperature and humidity controlled cellar.

The PC_1 and PC_2 score comparisons show that the narrowly distributed PC_1 and PC_2 score cluster for the estate stored wine broadens when different transport and storage scenarios are included. It is surprising that the smallest increase in PC_1 and PC_2 score cluster distribution, seen in Fig. 3(a), corresponds to the wine improperly stored at atmospheric conditions while the larger PC_1 and PC_2 score cluster distribution increases, seen in Fig. 3(b), are observed for wines acquired from professional wine merchants. The storage argument for the increase in PC_1 and PC_2 score cluster distribution of the merchant wine may also be valid. However, the increase in PC_1 and PC_2 score cluster distribution size for the merchant wines in comparison to the smaller cluster size for the private collector wine in Fig. 3(a) suggests that wine storage is not the cause of the increased cluster size shown in Fig. 3(b). Rather, it is likely that the increase PC_1 and PC_2 score cluster distribution size for the merchant wines shown in Fig. 3(b) are transport dependent. Specifically, the merchant bottles were shipped at least once and the details regarding temperature, humidity, and bottle handling are unknown.

Although transportation seems to broaden the cluster size for the merchant wines, it is interesting to note it appears to have little impact on the orientation of the ellipse. Comparison of the merchant ellipses in Fig. 3(b) to the estate stored wine ellipse in Fig. 3(a) illustrates that the orientation of the ellipse has not changed. On the other hand, the orientation of the private collector wine ellipse in relation to the estate wine has changed. As seen in Eq. (6), ρ, more specifically the covariance σ_{xy}, controls the ellipse orientation. This suggests the correlation between PC_1 and PC_2 is consistent with wine transport.

The data cluster ellipses can be used to assign statistical membership probability for a suspect wine. If the PC_1 and PC_2 score for a suspect wine bottle, not in the original library or cluster distribution, lies within the ellipse it is a member of the cluster and likely an authentic wine. If, in the case of Fig. 3(a) and (b), the suspect wine bottle PC_1 and PC_2 score lies outside of the ellipse there is only a 5% probability that the bottle is a member of the cluster or equivalently that the suspect bottle is a different wine or a potential counterfeit. This statistical analysis suggests that the use of estate stored wines to generate authentication fingerprints to identify counterfeit wines could lead to misleading results as all of the alternative source wines yield PC_1 and PC_2 scores outside of the estate stored wine ellipse as shown in Fig. 3(a) and (b). It is the diamagnetic screening of known authentic collectible wine with different transport and storage history, factors automatically embedded in rare wine collections, that appropriately broadens the PC_1 and PC_2 score cluster distribution to provide an adequate comparison cluster for wine fingerprinting. This procedure is accomplished in Fig. 3(c) where a statistically generated ellipse with $\eta = 99.5\%$ is included for all thirty-six bottles of the Pianetta 2005 Cabernet Sauvignon. In this case if the PC_1 and PC_2 score for an additional bottle lies within the ellipse it is a member of the cluster or authentic whereas if the PC_1 and PC_2 score lies outside the ellipse there is only a 0.5% probability that the suspect wine belongs to the cluster or that the wine is different or a counterfeit.

The roughly order of magnitude increase in probability ellipse size recognized when including the effects of alternative wine source on the estate wine PC_1 and PC_2 score cluster illustrated in the comparison of Fig. 3(a) and (c) does not in any way preclude the separation and identification of different Pianetta wines. Although Fig. 3 suggests that transport and storage can dramatically effect a given wine PC_1 and PC_2 score cluster size, the PC_1 and PC_2 score scales in Fig. 2 in comparison to those in Fig. 3 make the effects of transport and storage insignificant when comparing different vintage wine from the same vineyard. To illustrate this comment, the 0.5% ellipse shown in Fig. 3(c) is appropriately scaled and overlaid on the PC_1 and PC_2 scores plot in Fig. 2.

Finally, it is worth commenting on the applicability of the bivariate correlated Student distribution to PC_1 and PC_2 scores determined from the full bottle diamagnetic screening device. This distribution function is appropriate for Gaussian like distributed data. In principal, the data obtained with this full bottle approach should be distributed in a Gaussian like fashion but in practice the PC_1 and PC_2 scores appear to be randomly distributed about some average PC_1 and PC_2 score value. In the case where abbreviated $n < 10$ libraries are used one could argue that an inadequate sampling is the source of an apparent random distribution. This effect should be absent in the $n = 36$ bottle library data reported in Fig. 3(c), however, along with clustering there is a non-Gaussian component to the data distribution. The data clustering effect is what was used to compare the three wine vintages in Fig. 2. Each vintage is statistically separate because the data for the vintage is a subset of the entire library. This is an acceptable consequence because each vintage has unique qualities that separate it from the other vintages. The same logic can be applied to compare bottles within a given vintage. The estate stored, merchant and private collector wines were all handled differently leading to small variations in their PC_1 and PC_2 scores. Thus although $n = 36$ different bottles of the same wine were examined, at most only 12 have a common history. Since this number is less than the $n > 29$ necessary to display Gaussian clustering behavior it is likely that inadequate sampling is the origin of the apparent random distribution.

6. Conclusion

The purpose of this study was to apply bivariate correlated Student distribution statistics to PCA reduced, full wine bottle, diamagnetic screening data. Here different Pianetta wine varieties obtained directly from the winery and three additional sources were used along with a recently developed wine screening device and PCA data reduction to explore the effects of transport and storage on sealed bottles of wine. Although the well separated PC_1 and PC_2 principal component score cluster distributions for different vintage and type of wine broaden when transport and storage effects are present, the mean cluster separation in the PC_1 and PC_2 score space remains significant enough to enable the determination of different wines and hence the identification of potential counterfeit wines.

Acknowledgment

Constant encouragement to pursue this work by Gene Mulvihill is gratefully acknowledged. A special thanks is extended to John Pianetta of Pianetta Winery for supplying estate stored wine.

References

[1] J.R. Wilke, The Wall Street Journal (2007).
[2] M. Frank, Wine Spectator (2007).
[3] E. McCoy, Bloomberg (2007).
[4] P. Bansal, Reuters (2007).
[5] D. Cozzolino, Non-Destructive Analysis by VIS-NIR Spectroscopy of Fluid(s) in its original container, Patent Appl. No. AU 2005100565 A4 (2005).
[6] C. Eliasson, N.A. Macleod, P. Matousek, Analytica Chimica Acta 607 (2008) 50–53.
[7] A.J. Weekley, P. Briuns, M.P. Augustine, Journal of Enology and Viticulture 53 (2003) 18–3214.
[8] A.J. Weekley, P. Briuns, M. Sisto, M.P. Augustine, Journal of Magnetic Resonance 161 (2003) 91–98.
[9] A. Carpentieri, M. Gennaro, A. Amoresano, Analytical and Bioanalytical Chemistry 389 (2007) 969–982.
[10] G.J. Martin, C. Guillou, M.L. Martin, M.T. Cabanis, Y. Tep, J. Aerny, Journal of Agricultural and Food Chemistry 36 (1988) 316–322.
[11] O. Lutz, Naturwissenschaften 78 (1991) 67–69.
[12] J.D. Greenough, H.P. Jackson, H.P. Longerich, Australian Journal of Grape and Wine Research (1997) 75–83.
[13] M.P. Day, B. Zhang, G.J. Martin, Journal of the Science of Food and Agriculture 67 (1995) 113–123.
[14] I.J. Košir, J. Kidri, Analytica Chimica Acta 458 (2002) 77–84.
[15] V.F. Taylor, H.P. Longerich, J.D. Greenough, Journal of Agricultural and Food Chemistry 51 (2003) 856–860.
[16] B.H. Scott, Business Wire (2007).
[17] De La Rue International, Case Study Brand Authentication for Romanian Wine, (2007).
[18] A. Afzali-Ardakani et al., Patent No. 6,817,538 B2 (2004).
[19] B. Wallace, The Billionaire's Vinegar, the Mystery of the World's Most Expensive Bottle of Wine, Crown, New York City, 2008.
[20] S. Harley, V. Lim, M. Augustine, Analytica Chimica Acta 702 (2011) 188–194.
[21] S. Harley, V. Lim, M. Augustine, Talanta 85 (2011) 2437–2444.
[22] J. MacQueen, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.
[23] A.K. Jain, Pattern Recognition Letters 31 (2010) 651–666.
[24] P. Serapinas, P.R. Venskutonis, V. Aninkevicius, Z. Ezerinkskis, A. Galdikas, V. Juzikiene, Analytical Methods 107 (2008) 1652–1660.