Joint ICTP-IAEA School on Novel Experimental Methodologies for Synchrotron Radiation Applications in Nano-science and Environmental Monitoring
17 November - 28 November 2014

Héctor Jorge Sánchez

Multivariate Methods

Principal Components Analysis

Summary

Introduction

- Aims
- Introduction to PCA
- Advantages of PCA

Basic Statistics

- Basic definitions
- Covariance matrix and correlation

Principal Component analysis

- What is it?
- PCA and linear algebra
- PCA and geometry

Applications

- Chinese porcelains classification
- Dog hair analysis

Aims

To describe a multivariate statistical technique, applicable to x-ray spectrometry.

To show some applications of Principal Components Analysis methodology.

Aims

Multivariate data set

Several samples (*n*) with several variables (*p*) por each simple

Multivariate Analysis

Principal Components Analysis

analysis of the covariance structure among

Introduction to PCA

Principal Components Analysis (PCA)

A mathematical tool of linear algebra that allows to:

- Describe the total variability of a set of multivariate observations, representing the cases in a reduced dimension space with respect to the dimension space of the original variables.
- Explore the covariance among variables.
- Identify the most important variables that explain the variability of the data set.

Advantages of PCA

Advantages of PCA

Reducing data set dimension

Analysis of variables

Gathering information for future samplings

Basic Statistics

Basic definitions

Aritmetic mean of the i-esima variable

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \qquad i = 1, 2, \cdots, p$$

Variance of the variable *i*

$$\sigma_i^2 = \sigma_{ii} = \frac{1}{n}\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 \qquad i = 1, 2, \cdots, p$$

Covariance between the variables *i* and *k*

$$\sigma_{ik} = \frac{1}{n}\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k) \qquad i = 1, 2, \cdots, p \quad k = 1, 2, \cdots, p$$

Correlation coefficient

$$Corr(x_i, x_k) = r_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii} . \sigma_{kk}}}$$

Héctor Jorge Sánchez

Basic Statistics

Covariance Matrix

$$\mathbf{\Sigma} = Cov(\mathbf{x}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Correlation Matrix

$$\mathbf{R} = Corr(\mathbf{x}) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

Basic Statistics

Calculating the covariance matrix

- Given $\vec{x}_i = (x_{i_1}, x_{i_2}, \cdots, x_{i_p})$

- we can define the matrix:

$$\mathbf{x}_{n\times p} = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_n \end{pmatrix} = \begin{pmatrix} x_{1_1} & x_{1_2} & \cdots & x_{1_p} \\ x_{2_1} & x_{2_2} & \cdots & x_{2_p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_1} & x_{n_2} & \cdots & x_{n_p} \end{pmatrix}$$

- and the matrix, centered to the coordinate origin defined by the mean values:

$$\tilde{\mathbf{x}}_{n\times p} = \begin{pmatrix} x_{1_1} - \bar{x}_1 & x_{1_2} - \bar{x}_2 & \cdots & x_{1_p} - \bar{x}_p \\ x_{2_1} - \bar{x}_1 & x_{2_2} - \bar{x}_2 & \cdots & x_{2_p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_1} - \bar{x}_1 & x_{n_2} - \bar{x}_2 & \cdots & x_{n_p} - \bar{x}_p \end{pmatrix}_{n\times p}$$

Basic Statistics

Calculating the covariance matrix

$$\tilde{\mathbf{x}}'_{p\times n}\tilde{\mathbf{x}}_{n\times p} =$$

$$= \begin{pmatrix} \sum_{i=1}^{n}(x_{i_1}-\bar{x}_1)^2 & \sum_{i=1}^{n}(x_{i_1}-\bar{x}_1)(x_{i_2}-\bar{x}_2) & \cdots & \sum_{i=1}^{n}(x_{i_1}-\bar{x}_1)(x_{i_p}-\bar{x}_p) \\ \sum_{i=1}^{n}(x_{i_2}-\bar{x}_2)(x_{i_1}-\bar{x}_1) & \sum_{i=1}^{n}(x_{i_2}-\bar{x}_2)^2 & \cdots & \sum_{i=1}^{n}(x_{i_2}-\bar{x}_2)(x_{i_p}-\bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}(x_{i_p}-\bar{x}_p)(x_{i_1}-\bar{x}_1) & \sum_{i=1}^{n}(x_{i_p}-\bar{x}_p)(x_{i_2}-\bar{x}_2) & \cdots & \sum_{i=1}^{n}(x_{i_p}-\bar{x}_p)^2 \end{pmatrix}$$

$$= n \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} = n\mathbf{\Sigma}_{p\times p}$$

$$\Rightarrow \frac{1}{n}\tilde{\mathbf{x}}'_{p\times n}\tilde{\mathbf{x}}_{n\times p} = \mathbf{\Sigma}_{p\times p}$$

Basic Statistics

The correlation matrix

- It is the standardized covariance matrix:

$$\mathbf{R} = \begin{pmatrix} 1 & \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}\sigma_{pp}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{22}\sigma_{11}}} & 1 & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}\sigma_{11}}} & \frac{\sigma_{p2}}{\sqrt{\sigma_{pp}\sigma_{22}}} & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

$$-1 \leq r_{ij} \leq 1$$

Principal Components Analysis

How do we understand it?

Geometrically

Stablishing a new coordinate system by centering and rotating the original system using $X_1, X_2, \ldots, X_p$ as new axes.

Algebraically

PCA operates on $\boldsymbol{\Sigma}$ or $\mathbf{R}$, looking for particular linear combinations among the p original variables $X_1, X_2, \ldots, X_p$.

Héctor Jorge Sánchez

Principal Components Analysis

PCA algebraically

- Let $\mathbf{x}' = (\vec{x_1}, \vec{x_2}, \cdots, \vec{x_p})$ with a covariance matrix Σ of eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$
- Considering the system $\vec{z_i} = \mathbf{a}_i \cdot \mathbf{x}$

$$\begin{aligned}
\vec{z_1} &= \mathbf{a}_1' \cdot \mathbf{x} = a_{11}\vec{x_1} + a_{21}\vec{x_2} + \cdots + a_{p1}\vec{x_p} \\
\vec{z_2} &= \mathbf{a}_2' \cdot \mathbf{x} = a_{12}\vec{x_1} + a_{22}\vec{x_2} + \cdots + a_{p2}\vec{x_p} \\
&\vdots \\
\vec{z_p} &= \mathbf{a}_p' \cdot \mathbf{x} = a_{1p}\vec{x_1} + a_{21}\vec{x_2} + \cdots + a_{pp}\vec{x_p}
\end{aligned}$$

- then

$$\begin{cases} \text{Var}(\vec{z_i}) = \mathbf{a}_i'\Sigma\mathbf{a}_i & i = 1, 2, \cdots, p \\ \text{Cov}(\vec{z_i}, \vec{z_k}) = \mathbf{a}_i'\Sigma\mathbf{a}_k & i, k = 1, 2, \cdots, p \end{cases}$$

- PRINCIPAL COMPONENTS $\Longrightarrow Z_1, Z_2, \ldots, Z_p$ linear combinations of null covariances, whose variances are maximal

Héctor Jorge Sánchez
Joint ICTP-IAEA School on Novel Experimental Methodologies for Synchrotron Radiation Applications in Nano-science and Environmental Monitoring - 17 November - 28 November 2014

11/10/2014

Principal Components Analysis

PCA algebraically

- We look for the máximum of $$\lambda = \frac{Var(z)}{||\mathbf{a}||} = \frac{\mathbf{a}'\Sigma\mathbf{a}}{\mathbf{a}'\mathbf{a}}$$
- The máximum λ is the máximum eigenvalue of $(\mathbf{\Sigma} - \lambda\mathbf{I})\mathbf{a} = 0$
- The normalized eigenvector $\mathbf{a_1}$ corresponding to the highest eigenvalue λ_i is the coefficient vector in $\vec{z_1} = \mathbf{a}_1\mathbf{x}$
- The normalized eigenvector $\mathbf{a_2}$ corresponding to the second highest eigenvalue λ_2 is the coefficient vector in $\vec{z_2} = \mathbf{a}_2\mathbf{x}$
- The normalized eigenvector $\mathbf{a_p}$ corresponding to the lowest eigenvalue λ_p is the coefficient vector in $\vec{z_p} = \mathbf{a}_p\mathbf{x}$

Principal Components Analysis

PCA algebraically

- The total variance of the system is, therefore, the sum of the eigenvalue

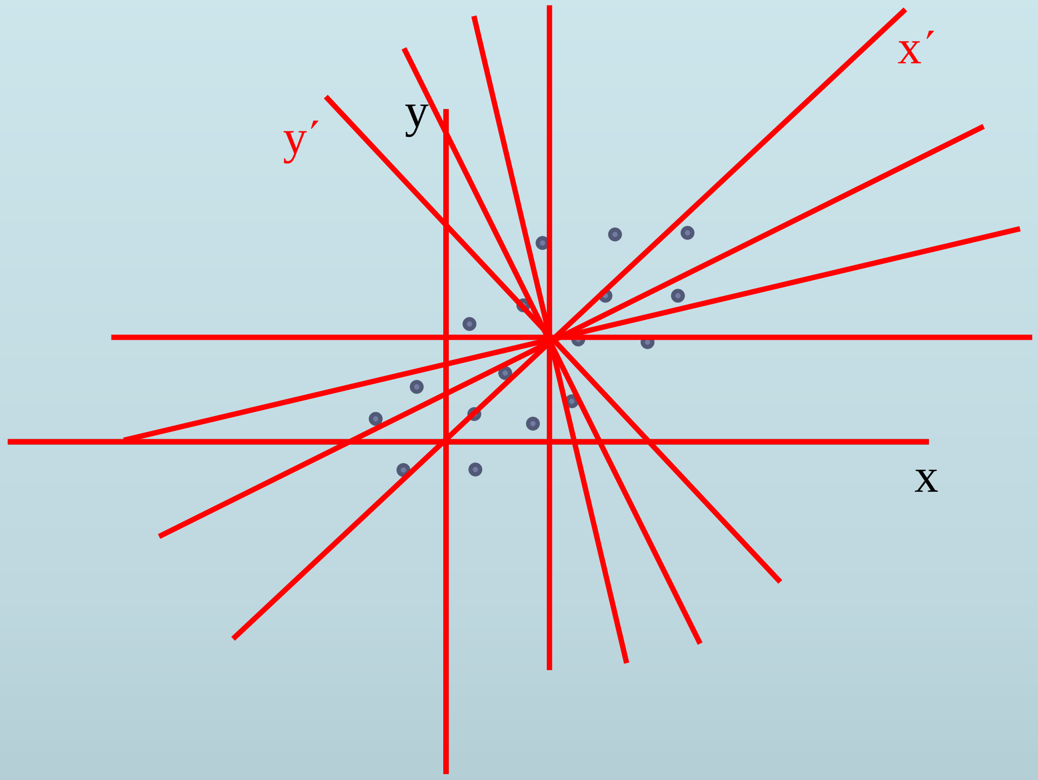
$$Total\ Variance = \sigma_{11} + \sigma_{22} + .. + \sigma_{pp} = \lambda_{11} + \lambda_{22} + .. + \lambda_{pp}$$

- Hence, the proportion of the variance explained by the kth component is:

$$Proportion\ of\ the\ k\text{th}\ variability = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

Principal Components Analysis

PCA Geometrically



Applications

EDXRF studies of porcelains (800–1600 A.D.) from Fujian, China with chemical proxies and principal component analysis

J. Wu *et al., X-Ray Spectrom.* **29**, 239–244 (2000)

41 Dehua porcelain samples from three different regions of China.

- Xunzhong: Qudou-Gong (DQ) wares (960 – 1368 a.d., Song-Yuan dynasty)
- Gaide: Wanping-Lun (DWP) wares (960 - 1368 a.d., Song-Yuan dynasty)
- Meihu: Mulin (DM) wares (618 - 960 a.d., Tang dynasty)

Applications

Major and minor elements present in the samples.

(9 variables: *Si; Al; Fe; Ti; Ca; Mg; K; Na2O and Mn*).

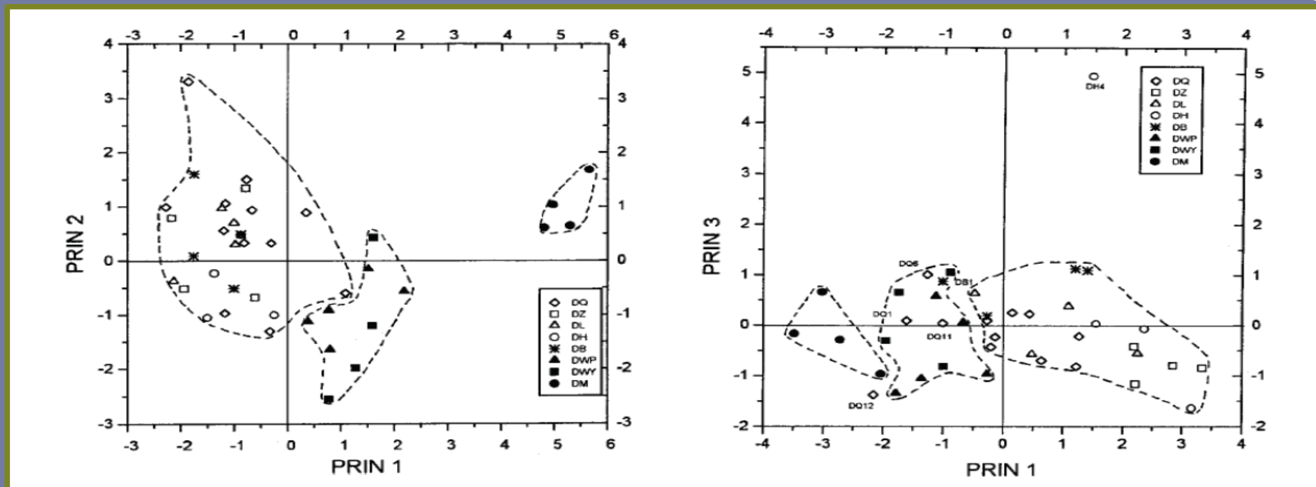
Trace elements present in the samples.

(9 variables: *Cr2; Ni; Cu; Zn; Rb; Sr; Y; Zr and Ba*).

Major, minor		**Traces**	
Eigenvalue	Acc. %	Eigenvalue	Acc %
λ_1	49	λ_1	45
λ_2	63	λ_2	63
λ_3	75	λ_3	83

Accumulative percentage of the total variability explained by the first three principal components data matrix for major and minor elements, and trace elements.

Applications



Plot of first two p
compon
concentratio

Plot o
co
concentrati
–Ba

qu-Xu,
VP=Wanping-
=Mulin (Meihu)

The chemical compositions were used for recognizing the provenience of Dehua porcelain. The 41 samples from eight kiln sites are distributed in three areas, corresponding to their original places of production, Xunzhong, Gaide and Meihu towns, respectively. Principal component analysis (PRIN 1, PRIN 2 and PRIN 3) reveals well defined regions for the samples. However, some the data points are very scattered because some concentration of the trace elements appears in abnormal values.

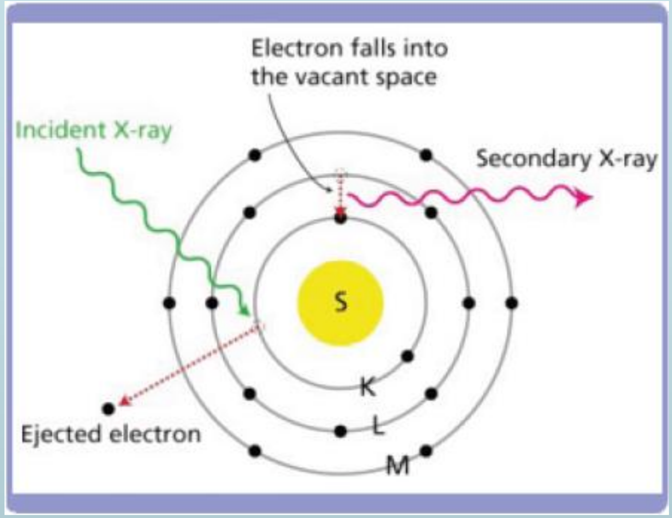
Applications

X-ray scattering processes and chemometrics for differentiating complex samples using conventional EDXRF equipment

M. I. Bueno *et al.*, *Chemometrics* **78**, 96-102 (2005)

Thirty-four hair samples of poodle dogs (of known age, hair color, gender, health status, and living environment).

- Samples were irradiated with a rhodium x-ray tube (50 kV, 100 s)
- The scattering and fluorescent spectra coming from the sample were recorded



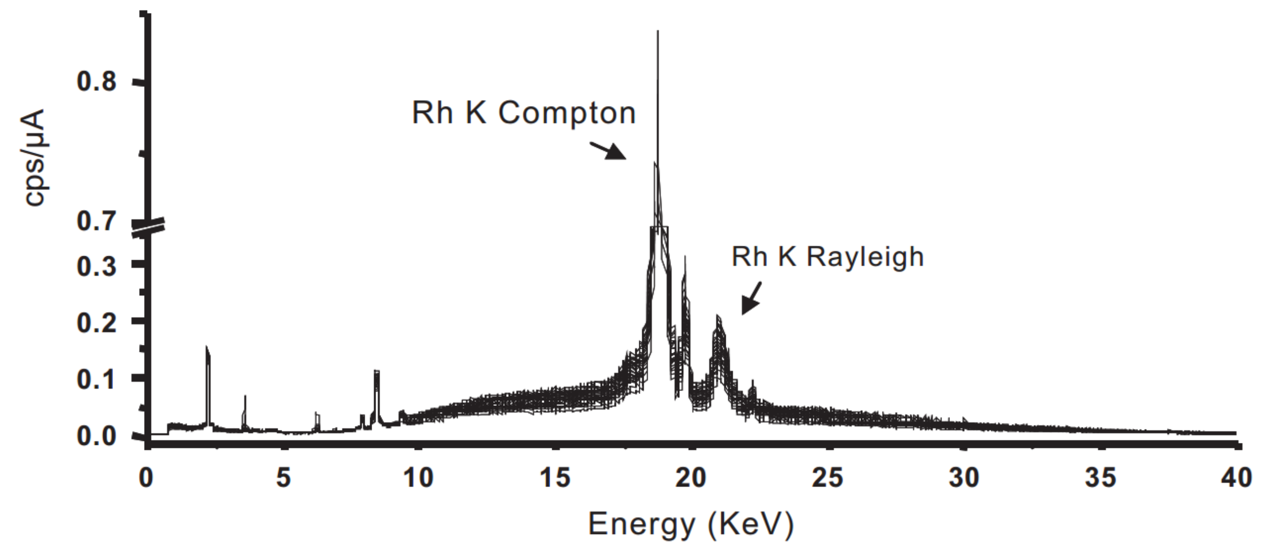
Applications

The spectra consist of counting channels discriminated by energy.

Thirty four spectra were recorded. 2014 channels for spectrum.

M. I. Bueno *et al.*, *Chemometrics* **78**, 96-102 (2005)

Sample



. Superimposed spectra of 34 hair samples of poodle dogs.

Applications

Spectrum processing

One data matrix was constructed in such a way that each row corresponded to the spectrum of a sample and each column to their respective energy values.

PCA was applied to this matrix generating new variables, the "Principal Components". The number of PC was the same as the number of columns in the matrix.

Applications

The proportion of the variance explained by the first six principal components

M.I. Bueno *et al.*, *Chemometrics* **78**, 96-102 (2005)

Principal Component	Explained Variance [%]
1	98.39
2	0.90
3	0.60
4	0.01
5	0.01
6	0.01

Applications

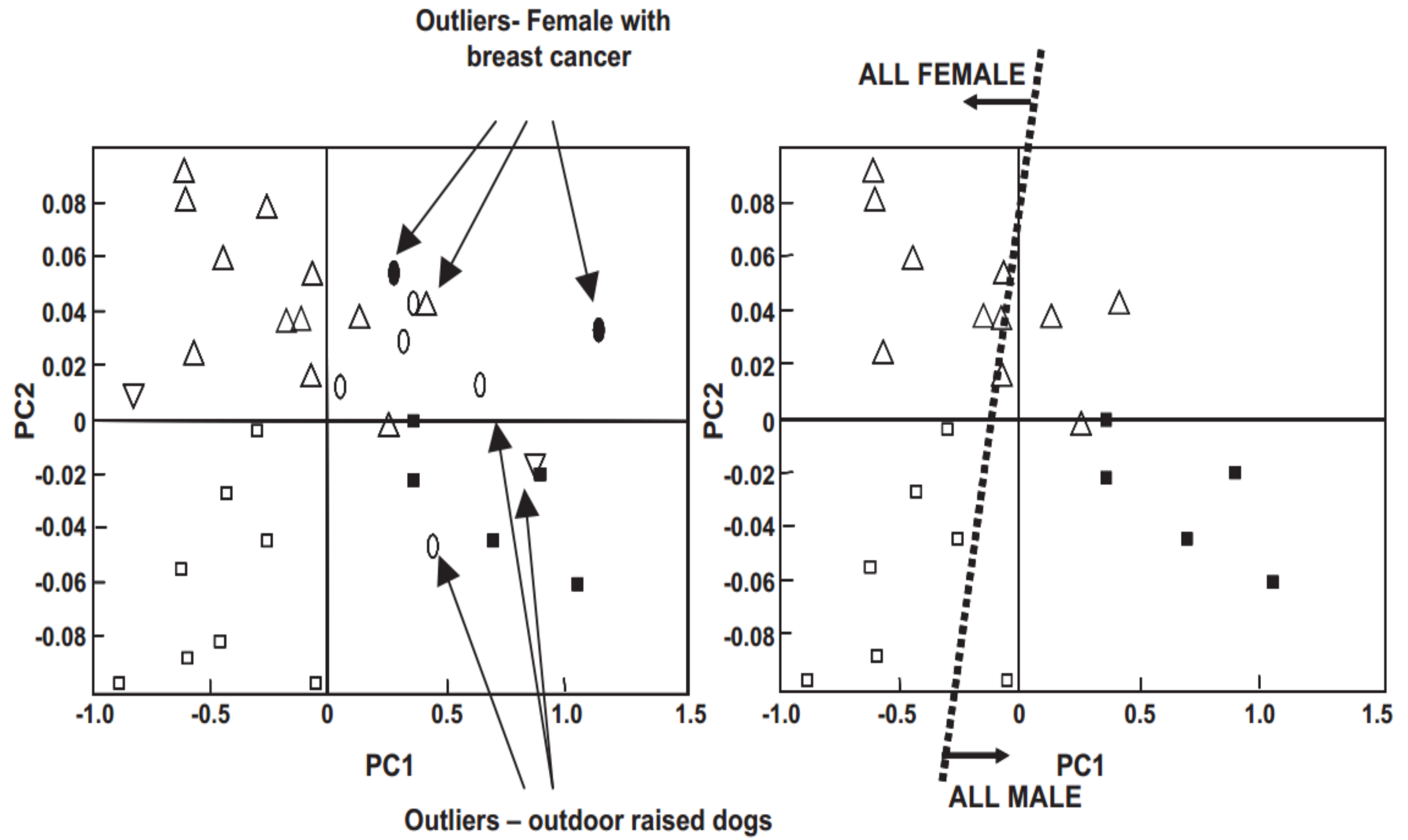


Fig. 2. Scores plots of dog hair samples after spectra processing by PCA. Left: PC2 versus PC1 for all dogs [(△) light brown hair; (■) black hair; (□) white hair]. Right: PC2 versus PC1 without considering the outliers [(○) white hair from sick dogs; (●) black hair from sick dogs; (▽) light brown hair from sick dogs].

Conclusions

What do you thing?

What can you conclude?

Héctor Jorge Sánchez