Research article

JOURNAL OF CHEMOMETRICS

Received: 15 June 2015, Revised: 15 May 2016, Accepted: 17 May 2016, Published online in Wiley Online Library: 1 July 2016

(wileyonlinelibrary.com) DOI: 10.1002/cem.2807

# A nonlinear principal component analysis to study archeometric data

**Alessandro Bitetto[a], Annarosa Mangone[b], Rosa Maria Mininni[a]* and Lorena C. Giannossa[b]**

**Statistical techniques, when applied to data obtained by chemical investigations on ancient artworks, are usually expected to recognize groups of objects to classify the archeological finds, to attribute the provenance of items compared with earlier investigated ones, or to determine whether an archaelogical attribution is possible or not. The statistical technique most frequently used in archeometry is the principal component analysis (PCA), because of its simplicity in theory and implementation. However, the application of PCA to archeometric data showed severe limitations because of its linear feature. Indeed, PCA is inadequate to classify data whose behavior describe a curve or a curved subspace of the original data space. As a consequence of it, an amount of information is lost because the multidimensional data space is compressed into a lower-dimensional subspace including principal components. The aim of this work is then to test a novel statistical technique for archeometry. We propose a nonlinear PCA method to extract maximum chemical information by plotting data on the smallest number of principal components and to answer archeological questions. The higher accuracy and effectiveness of nonlinear PCA approach with respect to standard PCA for the analysis of archeometric data are shown through the study of Apulian red figured pottery (fifth–fourth century BC) coming from some of the most relevant archeological sites of ancient Apulia (Monte Sannace (Gioia del Colle), Egnatia (Fasano), Canosa, Altamura, Conversano, and Arpi(Foggia)).** 

**Keywords:** archeometry; auto-associative neural network; nonlinear principal component analysis, Apulian red figured pottery

## 1. INTRODUCTION

The application of scientific techniques for data analysis to the study of archaelogical materials, initially considered secondary, discloses proficient and necessary features to review old assumptions in history (sometimes grounded on weak evidences), to optimize restoration and conservation actions and to reply new issues.

Only a systematic archeometric characterization of finds, obtained through an integrated approach of analytical techniques, provides compositional and structural information of bulk and surfaces that allow to answer questions in archeology (cf. [1–5]). However, even only the chemical analysis can provide a large number of information. As for ceramic material, the elemental composition of ceramic bodies, which is essential to trace the provenance of the clayey raw materials employed, provides data to establish an archeological classification (cf. [6–8]), to assign pottery of unknown provenance to previously studied groups of objects (cf. [9]), to decide whether or not an assignment can be made (cf. [10,11]), or to supply useful clues to highlight differences in the manufacturing process (cf. [12,13]).

However, a process of data mining is necessary to extract the "archeological" information contained in the set. Further, because the results of investigations predominantly aimed to scholars in non-scientific fields, the information obtained should be immediately "readable".

For this purpose, scatter plots related to selected pairs of chemical elements or their combinations were used in the first archeometric studies. The part of information relating to the other measured variables, not necessarily correlated, was thus lost.

Later, different statistical techniques have been employed. One of the most widely used statistical methods in archeometry is the classical (linear) principal component analysis (PCA) (e.g., [14–17]). It is a low-dimensional projection technique based on the simple idea of defining arrays or planes in a multivariate space so that they can describe the intrinsic behavior of data in the best way (cf. [18, Ch. 8]).

The "visualization" of the results is then obtained by plotting the score and loading vectors of the different parameters in the sub-plane of the first two principal components or in the subspace of the first three principal components.

A major problem when using PCA is to determine the significant number of principal component variables to interpret

* *Correspondence to: Rosa Maria Mininni, Dipartimento di Matematica, Università di Bari Aldo Moro,Via E. Orabona 4, Bari 70125, Italy*
*E-mail: rosamaria.mininni@uniba.it*

a *A. Bitetto, R. M. Mininni*
*Dipartimento di Matematica, Università di Bari Aldo Moro, Via E. Orabona 4, Bari 70125, Italy*

b *A. Mangone, L. C. Giannossa*
*Dipartmento di Chimica, Università di Bari Aldo Moro, Via E. Orabona 4, Bari 70125, Italy*

data structures. As often as not, in archeometric works, to make the reading easier, the data set is compressed on the plane defined by the first two principal components. But this compression can be very inefficient when the first two principal components describe only a low percentage of the total data information. That is because the "linear" feature of PCA (i.e., the principal components are linear combinations of the original variables) is inadequate to classify data whose behavior describe a curve or a curved subspace of the original data space. Roughly speaking, nonlinear effects can become partly visible when data sets are analyzed using PCA. This can trouble the determination of the optimal number of significant principal components.

It is worth noting that a preliminary exploratory analysis to all samples investigated (Section 3.1) by log-transformation of data (with and without subsequent standardization), commonly used pre-treatment technique with compositional data (e.g., [19, Ch.5]), was carried out. The results obtained did not highlighted a better visualization compared with results using standardized data.

Our main objective is hence to visualize and analyze the potential nonlinear structure of data sets by components that are generalized from straight lines to curves. The components are required to explain as much information as possible in a least square error sense. Therefore, a nonlinear generalization of standard PCA by replacing linear surfaces with curved ones is needed to improve classifications of data when nonlinear and unknown characteristics are present.

The purpose of this study is twofold:

- to test novel statistical tools in archeometry that allow to provide maximum chemical data information by plotting data on a minimum number of significant principal components;
- to use the obtained results to answer archeological questions.

A nonlinear approach for the classical PCA to group the archeological finds has been proposed. The nonlinear PCA (NLPCA) methodology let to replace original data matrix with a new matrix of data with nonlinearities and noise removed such that all classical linear methods of multivariate analysis can be used. Roughly speaking, NLPCA approach generates the desired significant principal components when the original data set is transformed in a new data set of adequate quality in the sense that there are no additional artifacts but only a Gaussian noise contribution.

The NLPCA approach, which can be used for all materials commonly studied in the archeometric field, has been applied in this work to the "Apulian red figured pottery".

This paper focuses on the use of the NLPCA approach based on *auto-associative neural network* models (cf. [20,21]) to produce nonlinear subspace PCA decompositions. In the literature, there are several applications of NLPCA methods based on neural networks to different disciplines as geophysics, oceanography, robotics, agriculture, molecular biology, chemistry, pattern recognition and image processing, and industrial quality control (cf. [22–24] and references therein). In particular, we consider *hierarchical NLPCA* proposed in [25], which to the best of our knowledge has not been used for analysis of archeometric data. The hierarchical NLPCA method achieves a hierarchical order of nonlinear components similar to standard linear PCA.

## 2. STATISTICAL METHOD DESCRIPTION

We denote by $\mathcal{X} \subset \mathbb{R}^m$ the original data space given by $N$ observed samples of data in $m$ variables, say $\mathbf{x}_j = (x_{j1}, \ldots, x_{jm})^T$ $(j = 1, \ldots, N)$, and by $\mathcal{Z} \subset \mathbb{R}^k$, $k < m$, the component space, a subspace of $\mathcal{X}$ where we will investigate nonlinear relations between data. The NLPCA approach, proposed by Kramer [26], is based on an auto-associative neural network. The key feature of such networks is to perform an input dimensionality reduction in a nonlinear way. The strategy is based on mapping $m$ input variables into $m$ output variables (for this reason, this network is called auto-associative).

It is a typical multi-layer network of nodes (neurons), consisting of an *input layer*, some middle layers (they are hidden from the outside), and an *output layer* (Figure 1). The number of nodes in the input and output layers must be equal. It corresponds with the number $m$ of variables describing classified objects. The middle layers usually consist of three layers: the first one is the *mapping layer* that constructs a nonlinear mapping $\Phi : \mathcal{X} \to \mathcal{Z}$, to perform a compression of the input vectors $\mathbf{x}_j$ into equivalent vector samples $\mathbf{z}_j = (z_{j1}, \ldots, z_{jk})^T \in \mathcal{Z}$, exploiting fewer variables (score variables) that can be regarded as the significant nonlinear principal components. These vectors lie in an internal layer, the *bottleneck layer*, consisting of a fewer number, say $k < m$, of nodes. The number $k$ depends on the complexity of the classification task and corresponds to the number of variables that should ideally represent the essential characteristics of the observed samples (the nonlinear principal components). Finally, the third layer, the so-called *de-mapping* layer, performs the inverse nonlinear transformation $\Psi : \mathcal{Z} \to \mathcal{X}^*$ that reconstructs the original sample vectors from their lower-dimensional component representations. The output vectors $\mathbf{x}_j^* \in \mathcal{X}^* \subset \mathbb{R}^m$ $(j = 1, \ldots, N)$, which approximates the original data samples, lie in the output layer. Except for the input nodes, each node is a neuron. We consider *feed-forward* auto-associative neural networks; that is, the information moves only in one direction, forward, from the input nodes, through the hidden nodes and to the output nodes.

Note that the nonlinear subspace $\mathcal{Z}$ could not be unique; NLPCA provides the optimal subspace $\mathcal{Z}$ by minimizing the mean square reconstruction error

$$E = \frac{1}{Nm} \sum_{j=1}^{N} \|\mathbf{x}_j - \mathbf{x}_j^*\|^2 \tag{1}$$

Each neuron in one layer has directed connections to the neurons of the subsequent layer by an associated numerical weight.

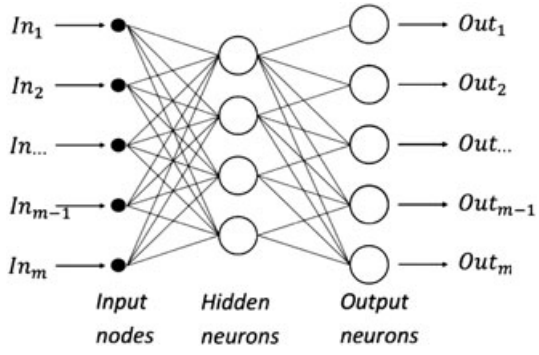


**Figure 1.** An example of feed-forward multi-layer network.



  

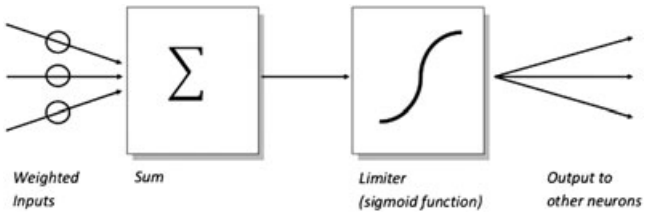


**Figure 2.** The structure of a neuron.

Indeed, the $i$th neuron of the $l$th layer receives a signal from the neurons belonging to the previous $(l-1)$th connected layer, and each of those signals, suitably weighted, is passed onto the next connected layer (Figure 2). The weight connecting two neurons serves to regulate the magnitude of signal that passes between them.

Afterwards, the weighted inputs are summed

$$u_i = \sum_r w_{ir} y_r^{(l-1)},$$

where $y_r^{(l-1)}$ are the output signals produced by neurons in the $(l-1)$th connected layer and transformed by a limiting function $f(\cdot)$, which scales the output to a fixed range of values

$$y_i^{(l)} = f(u_i) \tag{2}$$

The output $y_i^{(l)}$ is then passed onto all of nodes in the next layer. The function $f(\cdot)$ is called *activation function* and is generally chosen to be monotonic. There are different activation functions; the activation function used in the input, output, and bottleneck layers is a linear function, $f(u) = u$, while the activation function in the other layers is a differentiable nonlinear function. In this work, we applied a *logistic sigmoid* function,

$$f(u) = \frac{1}{1+\exp(-u)}.$$

The efficiency of a neural network strongly depends on the way the signals are propagated through the neurons under suitable weights tuning. Indeed, weights are unknown numbers. They are initially chosen at random, and their optimal estimates can be computed by adaptive algorithms. The process of adapting the weights is called *training*. A frequently used algorithm is the *backpropagation* method ([27, Ch. 7]). It consists on solving an optimization problem (by using the *gradient descent* algorithm) to minimize an error function. For a single layer $l$, the *error function* is defined as follows:

$$E(\mathbf{w}^{(l)}) = \frac{1}{2}\sum_i \left(\hat{y}_i^{(l)} - y_i^{(l)}\right)^2,$$

where the sum is evaluated on all neurons belonging to the same $l$th layer, $\mathbf{w}^{(l)}$ is the weight array related to $l$th layer, $\hat{y}_i^{(l)}$ is the output desired by the operator (the *target*), and $y_i^{(l)}$ is the output given by (2). The weights are the only parameters that can be modified to make the error function $E(\mathbf{w}^{(l)})$ as low as possible. We can thus minimize $E(\mathbf{w}^{(l)})$ by using an iterative process of gradient descent. Each weight in the $l$th layer is updated in $h$ steps,

$$\mathbf{w}^{(l)}(t+1) = \mathbf{w}^{(l)}(t) + \Delta\mathbf{w}^{(l)}(t), \quad t = 1, \ldots, h,$$

using the increment

$$\Delta\mathbf{w}^{(l)} = -\eta \frac{\partial E(\mathbf{w}^{(l)})}{\partial \mathbf{w}^{(l)}},$$

where $\eta$ is the *learning rate*, that is, a proportionality parameter, which defines the step length of each iteration in the negative gradient direction. In this work, we set $\eta = 0.1$ [28]. Training stops when the error reaches an acceptable level or when the error stops improving. In this way, we expect to find a minimum of the *total error function*: $E_{\text{tot}} = \sum_l E(\mathbf{w}^{(l)})$ (for more details, the reader can refer to [28, Ch.3 and 4]).

### 2.1. Hierarchical nonlinear principal component analysis

In the present paper, we apply an extension of NLPCA: the *hierarchical NLPCA*, proposed by Scholz and Vigario [25]. This methodology provides the optimal nonlinear subspace $\mathcal{Z}$ spanned by components, and at the same time, it preserves the order in which the nonlinear principal components appear, as well as the linear components in standard PCA. Further, because of its hierarchical nature, the method removes nonlinear correlations between components, and it is able to detect meaningful nonlinear features from real data, as will be shown in Section 3. In order to obtain ordered nonlinear components, Scholz and Vigario proposed an algorithm based on a neural network with a hierarchy of sub-networks that can extract ordered nonlinear principal components by controlling the corresponding reconstruction error. Roughly speaking, in order to perform the hierarchical NLPCA, the following *hierarchical error* function

$$E_H^{(k)} = E_1 + E_{1,2} + E_{1,2,3} + \cdots + E_{1,2,3,\ldots,k} \tag{3}$$

must be minimized. The single terms $E_1, E_{1,2}, \ldots, E_{1,2,\ldots,k}$ are the mean square reconstruction error defined in (1) when using only the first component, both the first and the second component, and so on, all the $k$ components, respectively. The error $E_H^{(k)}$ can be minimized by using an iterative process of gradient descent. The gradient $\nabla E_H^{(k)}$ is given by the following:

$$\nabla E_H^{(k)} = \sum_{i=1}^{k} \nabla E_{1,\ldots,i}.$$

The hierarchy constraints imposed to the feature space can then be interpreted as the research for a $k$-dimensional subspace $\mathcal{Z}$ of minimal error $E_H^{(k)}$ under the constraint that the $(k-1)$-dimensional subspace has also minimal error $E_H^{(k-1)}$. This is successively extended such that all $i$-dimensional subspaces $(i = 1, \ldots, k)$ have minimal error $E_H^{(i)}$. For this reason, the hierarchical NLPCA can be considered as a natural nonlinear extension of classical linear PCA.

In this work, we considered $k = 1, \ldots, 4$.

### 2.2. Confidence ellipsoids

A confidence ellipsoid is drawn assuming that data are sampled from an approximate multivariate Gaussian distribution with a nonzero covariance matrix $\Sigma_{xy}$. The magnitude of the ellipsoid axes depends on the variance of the data point. In the two-dimensional case, the orientation of the ellipse is determined by the covariance $\sigma_{xy}$. When data coordinates are uncorrelated (i.e., $\sigma_{xy} = 0$), the major axis of the ellipse is aligned with the x-axis.

In the experimental cases, we will study in Section 3, the sample data coordinates are correlated. Then we temporarily define a new coordinate system such that the ellipse becomes axis aligned and then rotate the resulting ellipse afterwards. The new coordinate system is defined along the orthogonal directions (the first two principal components, namely, PC1 and PC2) of greatest data variance. Thus, we may assume that the coordinates $(\xi, \eta)$ of the data points projected on the new axes will be approximately independent Gaussian distributed random variables with zero mean and variance–covariance matrix $\Sigma_{\xi\eta} = diag\left(\sigma_\xi^2, \sigma_\eta^2\right)$. As a consequence, $\left(\frac{\xi}{\sigma_\xi}\right)^2$ and $\left(\frac{\eta}{\sigma_\eta}\right)^2$ are independent chi-squared distributed random variables with one degrees of freedom. Then, the ellipse equation with respect to the new coordinate system is defined as follows:

$$\left(\frac{\xi}{\sigma_\xi}\right)^2 + \left(\frac{\eta}{\sigma_\eta}\right)^2 = s \tag{4}$$

The value of $s$ is chosen such that the resulting ellipse represents the selected $(1 - \alpha)$% confidence interval, that is, $s$ equals the quantile value $q_{1-\alpha}$ of the Chi-squared distribution with two degrees of freedom (e.g., a 95% confidence level corresponds to $s = q_{0.95} = 5.991$). Geometrically, PC1 and PC2 reflect the directions of the largest spread of the data. They are the eigenvectors of the variance-covariance (or correlation) matrix of the score matrix. Corresponding to each PC is an eigenvalue, namely, $\lambda_i$ $(i = 1, 2)$, which gives the amount of the spread in the data set explained by that PC. The major and minor axes of the ellipse are then oriented with respect to PC1 and PC2 axes. In the two-dimensional case, the orientation of the ellipse with respect to the original $(x, y)$ coordinate system is evaluated as follows:

$$a = \arctan\left(\frac{v_{1\eta}}{v_{1\xi}}\right),$$

where $\mathbf{v_1} = (v_{1\xi}, v_{1\eta})'$ is the eigenvector corresponding to $\lambda_1$. The length of the major and minor axes is evaluated by $2\sqrt{q_{1-\alpha}\,\lambda_i}$, $(i = 1, 2)$.

### 2.3. Calculation of statistics to compare principal component analysis and nonlinear principal component analysis fitting models

The underlying interrelationship between the original data matrix $\mathcal{X}$ and the reproduced data matrices, namely, $\mathcal{X}^*_{pca}$ and $\mathcal{X}^*_{nlpca}$, when using, respectively, the PCA and the hierarchical NLPCA methods is governed by the following:

$$\text{PCA:} \quad \mathcal{X} = \mathcal{X}^*_{pca} + \epsilon_{pca} = U_{pca}V^T + \epsilon_{pca} \tag{5}$$

$$\text{hierarchical NLPCA:} \quad \mathcal{X} = \mathcal{X}^*_{nlpca} + \epsilon_{nlpca} = f(U_{nlpca}) + \epsilon_{nlpca} \tag{6}$$

where $\epsilon_{pca}$ and $\epsilon_{nlpca}$ are the residuals or portions of data information unexplained by PCA and NLPCA method, because of experimental or systematic errors in the data. Both the errors represent statistically independent Gaussian noise with mean 0 and variance $\sigma^2$. $U_{pca}$ and $U_{nlpca}$ are the score matrices obtained, respectively, by PCA and hierarchical NLPCA. $V$ is the matrix of the loadings computed by PCA, and $f$ is a nonlinear function.

The efficiency of classical PCA and hierarchical NLPCA methods will depend on their ability to maximize the extracted data information by minimizing the mean square reconstruction error defined in (1). Hence, to evaluate how good the model (PCA or hierarchical NLPCA) fits the data, the following summary statistics will be computed in the statistical analysis described in Section 3

$$\text{Lack of fit:} \qquad Lof = \sqrt{\frac{\sum_{j=1}^{N}\sum_{i=1}^{m}\left(x_{ji} - x^*_{ji}\right)^2}{\sum_{j=1}^{N}\sum_{i=1}^{m}x_{ji}^2}} \tag{7}$$

$$R^2: \qquad R^2 = 1 - \frac{\sum_{j=1}^{N}\sum_{i=1}^{m}\left(x_{ji} - x^*_{ji}\right)^2}{\sum_{j=1}^{N}\sum_{i=1}^{m}x_{ji}^2} \tag{8}$$

In these equations, $x^*_{ji}$ are, respectively, the elements of the matrix $\mathcal{X}^*_{pca}$ or $\mathcal{X}^*_{nlpca}$. The numerator is the sum of squares of residuals, and the denominator is the total variance of the original dataset, assuming without any loss of generality that the dataset has zero mean. The statistics $R^2$ is the fraction of variance explained by PCA or hierarchical NLPCA.

The closer the *Lof* and $R^2$ are, respectively, to zero and one, the best model fitting of the experimental data has been achieved. When the fit is good, *Lof* provides more discrimination between models, as it is the case when noise is very low. When data fitting is not so good because of a larger noise contribution $R^2$ is then preferred.

## 3. RESULTS AND DISCUSSION

### 3.1. Case study: Apulian red figured pottery

"Apulian red figured pottery" is a local variant of the well-known Attic red figured production, based on painting on the vase a black glossy background developed during firing, saving figures from the ceramic body. It developed in Apulia between the third quarter of the fifth century BC and the end of the fourth century BC. This production, characterized by excellent drawing ability and remarkable quality, is one of the most important handicraft production group of figured pottery in Magna Graecia, largely commercialized both within and outside the region [29–32].

In this paper, the attention is focused on the analysis of findings coming from some of the most relevant archeological sites of ancient Apulia (Monte Sannace (Gioia del Colle), Egnatia (Fasano), Canosa, Altamura, Conversano, and Arpi (Foggia)).

The data subjected to statistical treatment are related to the elemental chemical composition of the ceramic body of the samples investigated, obtained by atomic spectroscopy analyses (atomic absorption spectroscopy and inductively coupled plasma mass spectrometry). Details of the applied analytical experimental procedure are reported in [12]. Furthermore, a multivariate statistical technique (Wilks' lambda test [33,34]) was applied as a criterion to discard those chemical elements that make more difficult a good archeological classification or are redundants.

Stylistic-typological and mineral-petrographic analyses have been performed on the same fragments with different and complementary techniques (polarized-light optical microscopy, scanning electron microscopy with energy dispersive spectrometry, and X-ray diffraction). The obtained results from optical microscopy, scanning electron microscopy, and X-ray diffraction allow to justify the grouping shown by the statistical analysis, highlighting morphological and minero-petrographical differences between objects in different clusters.



  

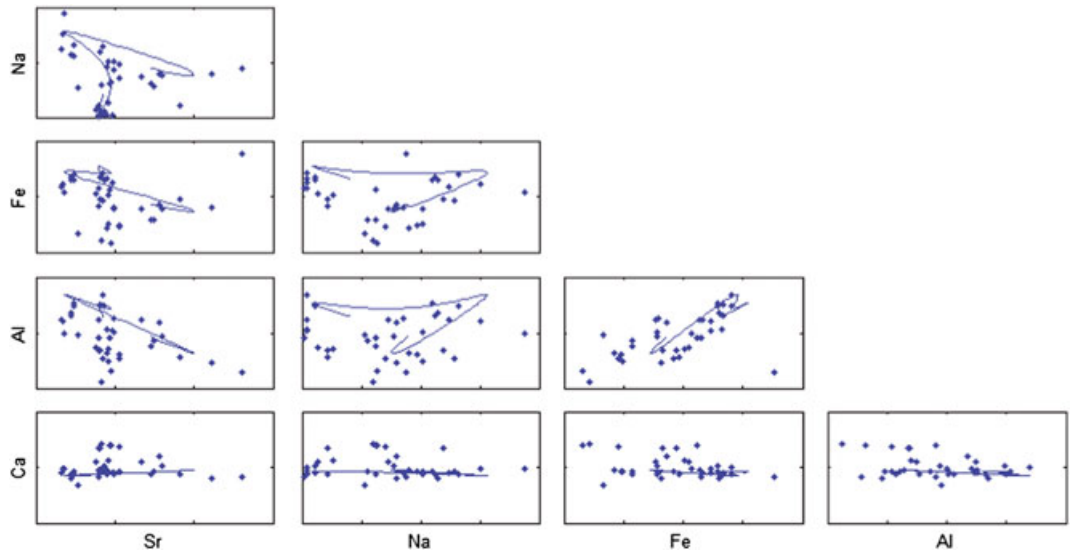


**Figure 3.** Scatter plots of pair-wise combinations of five variables classifying the elemental chemical composition of ceramic bodies from Monte Sannace archeological site. The extracted nonlinear first principal component, marked by a line, shows a strong nonlinear behavior.

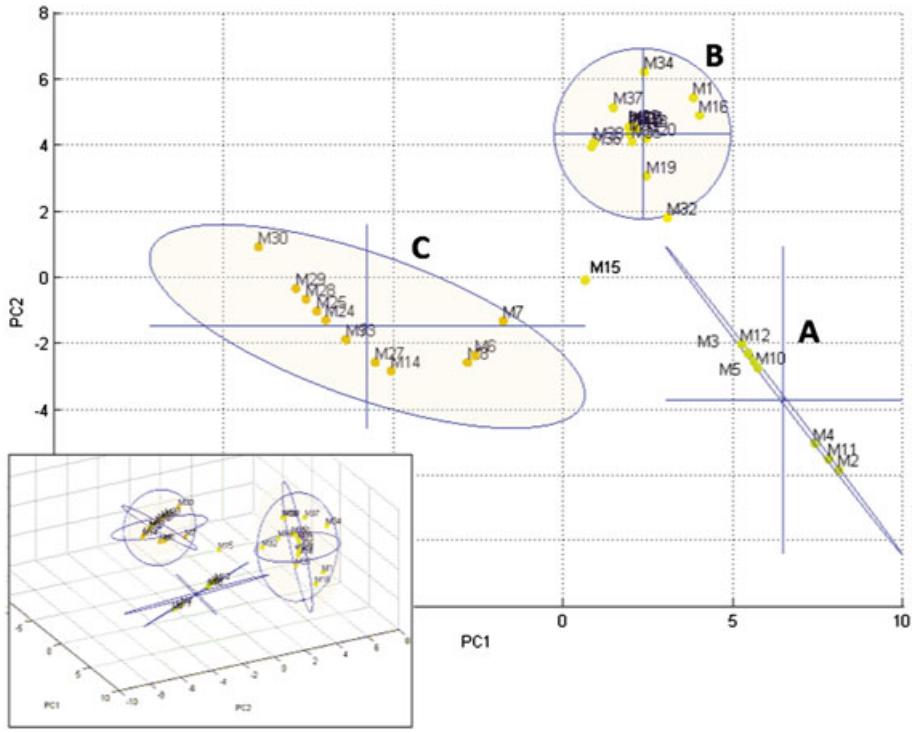


**Figure 4.** Scores plot related to the finds from Monte Sannace for the first $k = 2$ PCs. Ninety-five percent confidence ellipses are also reported. Inset: scores plot and confidence ellipsoids for the first $k = 3$ PCs.

The sample data coming from Monte Sannace and Egnatia archeological sites were previously analyzed by classical PCA (cf. [12,13]). The results provided very interesting clues from archeological/archeometric point of view, but some assignments were not perfectly clear, especially for Monte Sannace findings. The hierarchical NLPCA approach was then tested.

In the following subsections, statistical analysis was performed on data correlation matrix. Score plots and 95% confidence ellipsoids were systematically outlined for each data set.

### 3.1.1. *Monte Sannace archeological site*

Chemical data of ceramic bodies of 37 samples classified with respect to 10 selected variables were subjected to statistical analysis by hierarchical NLPCA.

The sampling involved fragmentary material and artifacts (amphoras, craters, plates, *skyphoi*, and bowls *oinochoi*). A comprehensive archeological survey was therefore conducted on the most part of the recovered material, that is, a stylistic and chronological classification that takes the belonging context into account and is supported by the comparison with other

**Table I.** Hierachical nonlinear principal component analysis applied to Monte Sannace data set: eigenvalues, hierarchical error, *Lof*, and $R^2$ when $k = 1, 2, 3, 4$ PCs are considered

| PCs | Eigenvalues | Hierarchical error $E_H^{(k)}$ | $E_{1,..,k}$ $\left(E_H^{(k)} - E_H^{(k-1)}\right)$ | *Lof* | $R^2$ |
|---|---|---|---|---|---|
| $k = 1$ | 4.42 | 0.2409 | — | 0.58 | 0.66 |
| $k = 2$ | 3.10 | 0.3408 | 0.0999 | 0.46 | 0.80 |
| $k = 3$ | 1.79 | 0.3810 | 0.0402 | 0.33 | 0.89 |
| $k = 4$ | 0.69 | 0.3897 | 0.0087 | 0.26 | 0.94 |

pottery [17]. The samples come from the archeological site of Monte Sannace except for sample M15 coming from a different site of the same area (the archaic-classical necropolis of Santo Mola (Gioia del Colle)). From archeological studies showing the presence of Greek letters and peculiar accessory decorations (palms and rolls), archeologists deduced that sample M5 comes from a Greek colony or even from Attica. Samples M1 and M15–M21 are dated from the fourth century BC, and samples M2–M5 are dated from the end of fifth to the beginning of fourth century BC, whereas samples M6–M14 and M23–M38 cannot be dated because they are fragments.

Indeed, the 37 chemical data samples were already analyzed by applying the classical PCA method on standardized data (zero mean and unit variance), and the statistical results were published in [12]. The statistical results obtained in [12] stressed the grouping of the objects into three clusters (A, B, and C). This splitting agrees with the different dating of samples. Moreover, the grouping in respect to less ancient objects (custers B and C) reflected two different production technologies: the "classic" Attic manufacturing (cluster B) and the "non classic" one, with an engobe layer on the ceramic body (cluster C). In detail, cluster A contained samples M2–M5 and M10–M12; cluster B contained samples M1, M15–M23, M26, and M31–M38, and cluster C samples M6–M9, M13, M14, M24, M25, and M27–M30.

The hierarchical NLPCA approach was then applied to data correlation matrix. An auto-associative neural network was used to extract the first $k = 1, .., 4$ nonlinear PCs in a hierarchical order.

Scatter plots in Figure 3 show nonlinear correlations among five of 10 variables that result to be the most important to the first nonlinear PC. This justifies the use of the NLPCA approach as already mentioned in Section 1. Figure 4 shows the scores profile plotted in the sub-plane of the first two and three PCs, respectively. A clear distribution of the finds in three clusters (A, B, and C) without any ambiguity is evident. Further, 95% confidence ellipsoids do not overlap contrary to PCA scores plot, where the location of some samples is not clear (cf. [12, Fig. 3]). The outlier position of the M15 sample is also highlighted. This results in the archeological evidence of a different provenance, as mentioned earlier. Therefore, the hierarchical NLPCA analysis allows to differentiate objects in close but distinct contexts.

In order to evaluate how good the hierarchical NLPCA method fits the data with respect to the standard PCA method, we excluded the sample outlier M15 from the data matrix and computed the summary statistics defined in (7) and (8) for both methods. It is worth noting that for PCA, the values of statistics $R^2$ can be equivalently computed as the corresponding eigenvalue divided by the total variance in the correlation matrix (in our case, the total variance equals 10, the number of variables). In (9), there are lists of the eigenvalues and the *Lof* and $R^2$ values corresponding to the number of principal components (PCs) extracted by PCA.

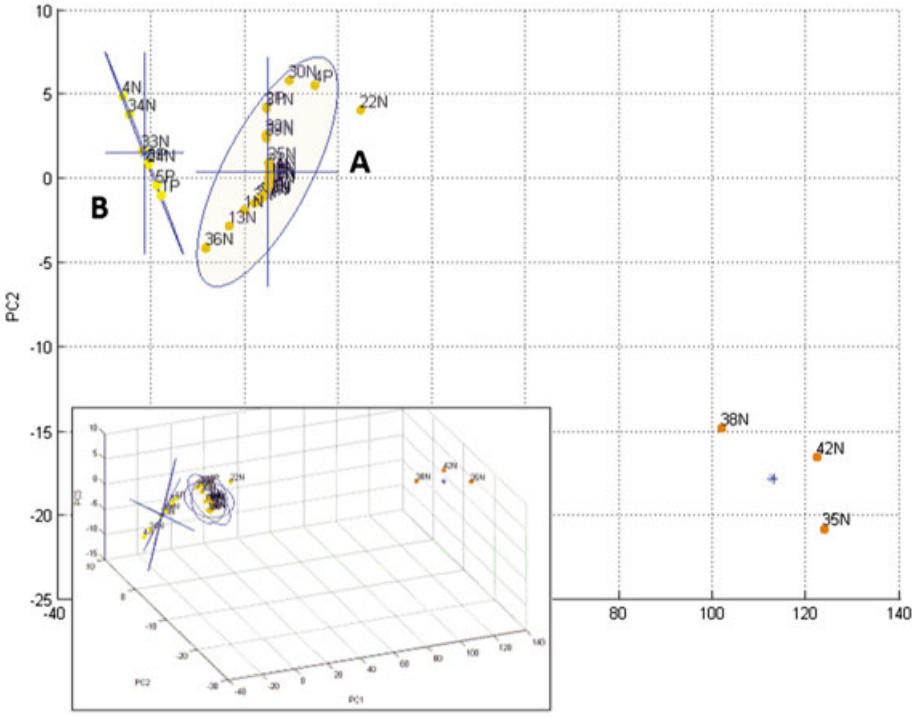


**Figure 5.** Scores plot related to the finds from Egnatia site for the first $k = 2$ PCs. Ninety-five percent confidence ellipses are also reported. Inset: scores plot and confidence ellipsoids for the first $k = 3$ PCs.



  

**Table II.** Hierachical nonlinear principal component analysis applied to data samples from Egnatia site: eigenvalues, hierarchical error, *Lof*, and $R^2$ when $k = 1, 2, 3, 4$ PCs are considered

| **PCs** | Eigenvalues | Hierarchical error $E_H^{(k)}$ | $E_{1,..,k}$ $\left(E_H^{(k)} - E_H^{(k-1)}\right)$ | *Lof* | $R^2$ |
|---|---|---|---|---|---|
| $k = 1$ | 6.14 | 0.2050 | – | 0.53 | 0.72 |
| $k = 2$ | 2.44 | 0.2842 | 0.0792 | 0.38 | 0.86 |
| $k = 3$ | 0.85 | 0.3193 | 0.0351 | 0.28 | 0.92 |
| $k = 4$ | 0.57 | 0.3415 | 0.0222 | 0.22 | 0.95 |

| *PCs* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Eigenvalues* : | 3.09 | 2.18 | 1.62 | 0.92 | 0.72 | 0.53 | 0.41 | 0.30 | 0.17 | 0.06 |
| *Lof* : | 0.83 | 0.69 | 0.56 | 0.47 | 0.38 | 0.31 | 0.23 | 0.15 | 0.08 | 0.00 |
| $R^2$ : | 0.31 | 0.53 | 0.69 | 0.78 | 0.85 | 0.91 | 0.94 | 0.98 | 0.99 | 1.00 |

(9)

When PCA method is applied to data samples, we obtain bad values of *Lof* and $R^2$ when the first $k = 3$ PCs are considered. The explained variance was just the 69% of the total one. This situation can be encountered for experimental data where noise contributions and/or nonlinear effects are important and the classical PCA algorithm does not account for them.

As for hierarchical NLPCA, Table I lists the eigenvalues corresponding to the first $k = 4$ PCs, the hierarchical error $E_H^{(k)}$, defined in (3) as sum of mean square reconstruction errors $E_{1,..,k}$, the *Lof* and $R^2$ statistics when, respectively, $k = 1, 2, 3, 4$ PCs are extracted. It is considered that the mean square reconstruction error is smaller when the number of significant PCs is larger. Note that when the number of PCs increased from $k = 2$ to $k = 3$ and from $k = 3$ to $k = 4$, the reconstruction error decreased by a small value. Indeed, Figure 4 shows that the first $k = 2$ principal components are sufficient to visualize the data structure. However, the fitting values given in terms of *Lof* and $R^2$ indicate that a better model fitting of the analyzed chemical data is achieved when at least the first $k = 3$ PCs are considered. It is worth pointing out that the parameter $R^2$ is more closer to 1 than *Lof* to zero, confirming a high noise contribution.

### 3.1.2. *Egnatia archeological site*

A statistical analysis similar to that performed in Section 3.1.1 was carried out on 39 artifacts dated back to the second half of the fourth century BC and coming from two different areas of Egnatia: the so called Penna Grande, in the farthest North-West zone of the town and the western necropolis (one of the most important extra-moenia funeral contexts of Egnatia during the Messapian and Roman ages). In this case, the results obtained by the classical PCA, reported in the literature [13], showed the presence of three outliers (35N, 38N, and 42N) and two clusters (A and B). All the samples were contained in cluster A except for samples 1P, 5P, 6P, 4N, 24N, 33N, and 34N, included instead in cluster B. From an archaeological point of view, the presence of outliers was explained by a different provenance of the pottery, whereas the grouping in clusters by a different production technology: the "classic" (cluster B) or "non-classic" (cluster A) Attic, with the use of the engobe, in accordance with Monte Sannace findings. The chemical data were classified with respect to the same group of selected variables considered in Section 3.1.1. The sample data from Egnatia site show also nonlinear correlations among variables. The application of hierarchical NLPCA to data correlation matrix yielded a good qualitative samples' separation when the obtained score matrix is plotted in the sub-plane of the first $k = 2$ PCs (Figure 5), similar to the visualization by PCA method with the first $k = 3$ PCs ([13, Fig.3]).

To test the efficiency of the hierarchical-NLPCA approach, three outliers, namely, the samples 35N, 38N, and 42N, were excluded from the data matrix. The statistical analysis was again performed for both PCA and NLPCA methods to compute and compare the corresponding eigenvalues and the statistics *Lof* and $R^2$. By the standard PCA, we obtain the following results.

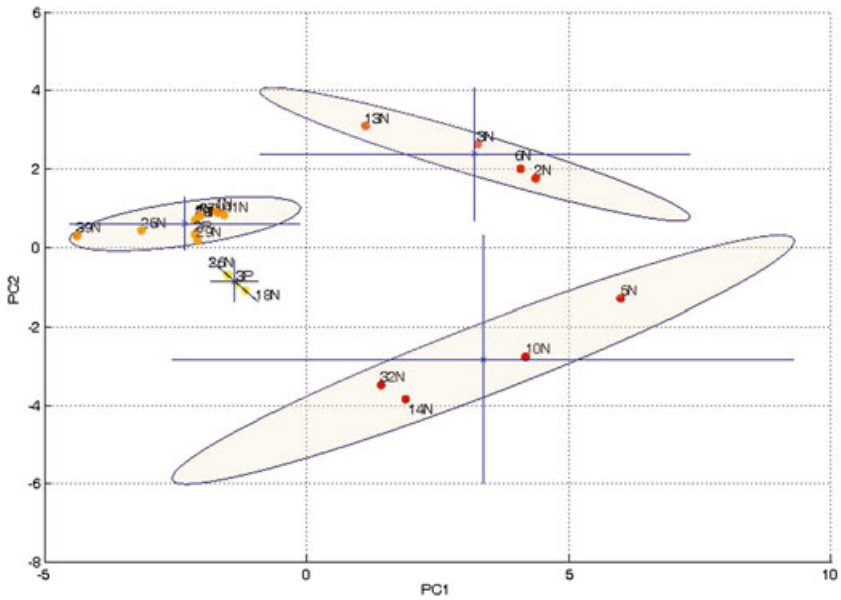


**Figure 6.** Hierarchical nonlinear principal component analysis: Scores plot and ninety-five percent confidence ellipses related to cluster A findings from Egnatia site for the first k = 2 PCS.

**Table III.** Elemental chemical composition of findings from Conversano, Altamura, Canosa, and Arpi archeological sites

| | Sample | (% w/w) | | | | | | (ppm) | |
|---|---|---|---|---|---|---|---|---|---|
| | | Ca | Mg | Na | Fe | Al | Ti | Cr | Ni |
| Conversano (BA) | C1 | 5.73 | 1.28 | 0.27 | 5.02 | 6.39 | 0.56 | 133 | 56 |
| | C2 | 6.03 | 1.28 | 0.30 | 4.81 | 5.97 | 0.55 | 135 | 49 |
| | C3 | 6.22 | 1.21 | 0.32 | 4.32 | 6.94 | 0.52 | 132 | 45 |
| | C4 | 4.55 | 1.10 | 0.37 | 4.71 | 6.29 | 0.55 | 135 | 57 |
| | C5 | 4.96 | 1.29 | 0.31 | 4.86 | 7.18 | 0.55 | 126 | 63 |
| | C6 | 5.70 | 1.36 | 0.33 | 5.15 | 7.04 | 0.47 | 126 | 53 |
| Altamura (BA) | A1 | 5.49 | 1.13 | 0.45 | 4.33 | 6.69 | 0.47 | 93 | 55 |
| | A2 | 6.16 | 1.30 | 0.51 | 3.88 | 6.89 | 0.43 | 92 | 63 |
| | A3 | 5.77 | 1.07 | 0.42 | 3.63 | 6.84 | 0.45 | 82 | 75 |
| | A4 | 5.86 | 1.12 | 0.48 | 4.55 | 6.55 | 0.52 | 96 | 56 |
| | A5 | 5.85 | 1.24 | 0.52 | 4.08 | 6.99 | 0.49 | 111 | 58 |
| | A6 | 6.01 | 1.43 | 0.50 | 4.51 | 6.83 | 0.53 | 116 | 55 |
| | A7 | 5.97 | 1.32 | 0.61 | 4.22 | 7.35 | 0.60 | 99 | 51 |
| | A8 | 6.13 | 1.10 | 0.45 | 4.69 | 7.33 | 0.52 | 112 | 56 |
| | A9 | 5.73 | 1.33 | 0.59 | 4.43 | 7.16 | 0.53 | 106 | 53 |
| | A10 | 5.05 | 1.09 | 0.31 | 4.29 | 6.02 | 0.52 | 113 | 65 |
| Canosa (BT) | Ca1 | 9.09 | 1.04 | 0.76 | 3.25 | 6.78 | 0.28 | 121 | 37 |
| | Ca2 | 5.68 | 0.73 | 0.88 | 2.80 | 4.87 | 0.25 | 137 | 25 |
| | Ca3 | 7.77 | 1.21 | 0.74 | 3.43 | 7.62 | 0.29 | 131 | 24 |
| | Ca5 | 8.69 | 1.28 | 0.73 | 4.15 | 7.76 | 0.34 | 126 | 30 |
| | CaA | 7.39 | 1.12 | 0.72 | 3.01 | 7.07 | 0.23 | 122 | 26 |
| | CaB | 7.47 | 0.94 | 0.77 | 2.96 | 5.89 | 0.26 | 118 | 23 |
| | CaC | 9.99 | 1.19 | 0.83 | 3.20 | 6.81 | 0.29 | 126 | 27 |
| | CaD | 8.80 | 1.40 | 0.83 | 3.43 | 7.57 | 0.29 | 126 | 26 |
| | CaE | 8.43 | 1.34 | 0.71 | 3.94 | 8.74 | 0.39 | 130 | 24 |
| Arpi (FG) | Ar2 | 5.01 | 1.14 | 0.96 | 4.43 | 8.30 | 0.42 | 117 | 78 |
| | Ar3 | 5.38 | 1.11 | 0.73 | 4.12 | 7.00 | 0.36 | 122 | 48 |
| | Ar4 | 6.07 | 1.02 | 0.79 | 4.06 | 7.59 | 0.39 | 122 | 39 |
| | Ar6 | 5.32 | 1.26 | 1.22 | 3.92 | 8.43 | 0.41 | 104 | 50 |
| | Ar7 | 5.02 | 1.13 | 1.25 | 4.06 | 8.51 | 0.42 | 98 | 49 |
| | Ar8 | 6.14 | 1.24 | 1.25 | 3.91 | 7.85 | 0.32 | 96 | 39 |
| | Ar9 | 6.00 | 1.27 | 1.08 | 3.85 | 7.98 | 0.34 | 88 | 37 |
| | ArA | 8.68 | 1.42 | 1.24 | 2.58 | 6.90 | 0.29 | 90 | 40 |
| | ArB | 9.35 | 1.30 | 1.16 | 3.06 | 8.30 | 0.42 | 71 | 55 |

| *PCs* : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Eigenvalues* : | 4.73 | 1.87 | 1.08 | 0.96 | 0.66 | 0.29 | 0.21 | 0.15 | 0.03 | 0.01 |
| *Lof* : | 0.73 | 0.58 | 0.48 | 0.37 | 0.26 | 0.20 | 0.14 | 0.06 | 0.03 | 0.00 |
| $R^2$ : | 0.47 | 0.66 | 0.77 | 0.86 | 0.93 | 0.96 | 0.98 | 0.99 | 0.99 | 1.00 |

(10)

Table II lists the statistical results when $k = 1, 2, 3, 4$ PCs are, respectively, extracted by the hierarchical NLPCA. In comparison with PCA results, the eigenvalues indicate that the amount of data spread is concentrated in the first $k = 2$ significant PCs. In addition, the fitting values $Lof = 0.28$ and $R^2 = 0.92$ are achieved when the first $k = 3$ PCs are considered. Further, when the number of PCs increased from $k = 2$ to $k = 3$ (and from $k = 3$ to $k = 4$), the reconstruction error $E_{1,..,k}$ decreased by a small value. Indeed, Figure 5 shows that the first $k = 2$ PCs are sufficient to describe the data structure. This confirms that the NLPCA method reduces the minimum number of significant principal components. Moreover, the parameter $R^2$ is more closer to 1 than *Lof* to zero, confirming a high noise contribution, in analogy with Monte Sannace findings.

To further test the potentiality of NLPCA approach, cluster A was analyzed separately (cluster B was not analyzed because of the low number of samples). Figure 6 shows that the first $k = 2$ PCs are sufficient to describe the data structure in group A and highlights a clear separation of the samples, based on the context of provenance (N "Western Necropolis", P "Punta Penna grossa") with the exception of samples 3P, 18N, and 26N. For these samples, however, a shift from one area to another cannot be excluded, because they come from tombs violated since the end of the eighteenth century.



  

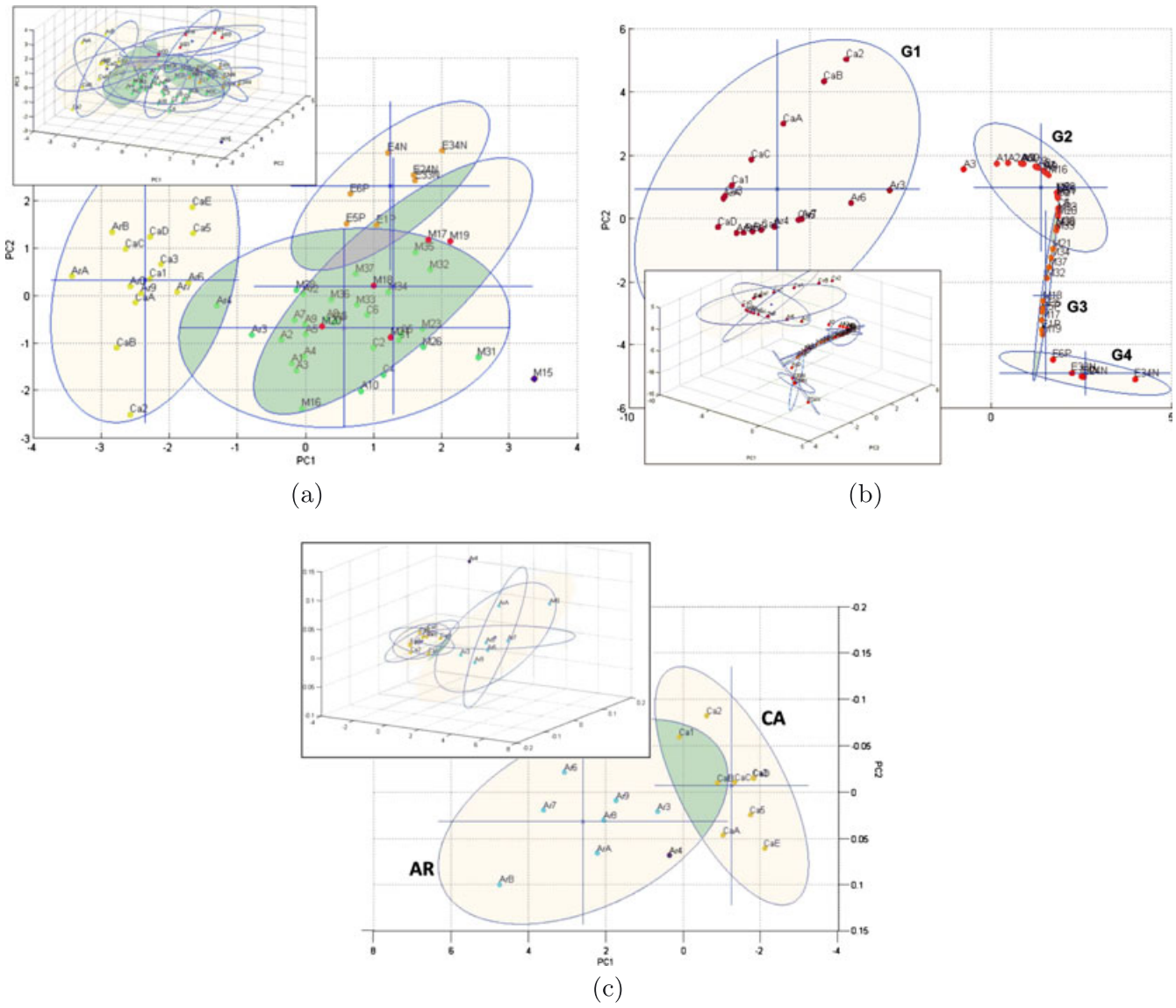


**Figure 7.** Scores plots and ninenty-five percent confidence ellipses related to the finds from different Apulian sites (A=Altamura, C=Conversano, M=Monte Sannace, E=Egnatia, Ar=Arpi, Ca=Canosa) for the first $k = 2$ PCs. Inset: scores diagram and confidence ellipsoids for the first $k = 3$ PCs. (a): PCA analysis. (b): hierarchical NLPCA analysis. (c): Scores plot and 95% confidence ellipses related to the finds of Canosa and Arpi for the first $k = 2$ PCs extracted by the hierarchical NLPCA method.

### *3.1.3. Monte Sannace, Egnatia, Arpi, Canosa, Altamura, and Conversano archeological sites*

The possibility of obtaining a separation of the objects according to different sites of discovery was checked. Although in the absence of certain indicators of production, for example, kiln waste, any differences found among the objects cannot be strictly linked to their provenance, a distinction according to the site of discovery can still provide an indication of a fragmented production.

Therefore, the chemical data matrix relative to the ceramic bodies of artifacts coming from Egnazia and Monte Sannace was extended to chemical data of finds, never analyzed before, and coming from other Apulian sites, in particular, finds coming from Niobid vase tomb of Arpi (except ArA and ArB), from Varrese hypogeum in Canosa, from tombs 1 and 3 of Via Ramunni (C3 and C1+C2, respectively), from tomb 1 of Via Verdi in Conversano, from tomb 1 of Via Reno (A1–A3), from tomb "a Grotticella" of via Ofanto (A6 and A8), and from tomb of "Scavo Agip" in Altamura. Chemical data (34 samples) were classified with respect to eight variables (cf. Table III). These samples were homogeneous for both period (fourth century BC) and technology (classic Attic). For this reason, we added items selected from cluster B (fourth century BC and classic Attic technology) from both Egnazia and Monte Sannace (24 samples).

A total of 58 samples were statistically analyzed by applying both classical PCA and hierarchical NLPCA. The obtained results were compared.

The PCA analysis extracted $k = 8$ PCs with the following statistical results.

| *PCs* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *Eigenvalues* : | 2.67 | 1.65 | 1.29 | 1.09 | 0.50 | 0.34 | 0.31 | 0.15 |
| *Lof* : | 0.82 | 0.68 | 0.55 | 0.40 | 0.32 | 0.24 | 0.14 | 0.00 |
| $R^2$ : | 0.33 | 0.54 | 0.70 | 0.84 | 0.90 | 0.94 | 0.98 | 1.00 |

(11)

**Table IV.** Hierachical nonlinear principal component analysis applied to data set from all the archeological sites and Canosa-Arpi sites: eigenvalues, hierarchical error, *Lof*, and $R^2$ when $k = 1, 2, 3, 4$ PCs are considered

| Archeological Site | PCs | Eigenvalues | Hierarchical Error $E_H^{(k)}$ | $E_{1,..,k}$ $\left(E_H^{(k)} - E_H^{(k-1)}\right)$ | *Lof* | $R^2$ |
|---|---|---|---|---|---|---|
| All | $k = 1$ | 4.36 | 0.2898 | — | 0.67 | 0.55 |
| | $k = 2$ | 2.33 | 0.4287 | 0.1389 | 0.47 | 0.78 |
| | $k = 3$ | 0.81 | 0.4742 | 0.0455 | 0.37 | 0.86 |
| | $k = 4$ | 0.50 | 0.5310 | 0.0568 | 0.26 | 0.93 |
| Canosa-Arpi | $k = 1$ | 4.56 | 0.0834 | — | 0.39 | 0.85 |
| | $k = 2$ | 1.95 | 0.1012 | 0.0178 | 0.26 | 0.93 |
| | $k = 3$ | 1.49 | 0.1194 | 0.0182 | 0.18 | 0.97 |

These results indicate that the value of $R^2$ closer to 1 is obtained when data are projected in a subspace including at least $k = 4$ PCs. The corresponding lack of fit is large, confirming a high noise contribution. Figure 7(a) shows the scores profile plotted in the sub-plane of the first $k = 2$ and $k = 3$ significant PCs. Several overlapping areas are pointed out.

On the other hand, the analysis performed by hierarchical NLPCA (Figure 7(b)) yielded a clear samples' separation in macrogroups, which reflect the ancient geographic division of Apulia: Messapia (cluster G4), Peucetia (clusters G2+G3), and Daunia (cluster G1).

A comparison between Table IV and the results obtained by standard PCA indicate that the first $k = 2$ PCs extracted by NLPCA are sufficient to describe the data structure better than PCA did. In fact, when the number of PCs increased from $k = 2$ to $k = 3$ (and from $k = 3$ to $k = 4$), the reconstruction error $E_{1,..,k}$ decreased by a small value.

A further application of hierarchical NLPCA on cluster G1 allowed to separate finds from Daunia (Canosa and Arpi) and, hence, all the analyzed finds with respect to the different sites of provenance (Egnazia cluster G4, Altamura and Conversano cluster G2, Monte Sannace cluster G3, Canosa cluster Ca, and Arpi cluster Ar). The statistical results are reported in Table IV. In this case, the fitting values $Lof = 0.26$ and $R^2 = 0.93$ are achieved when the first $k = 2$ principal components are considered. Figure 7(c) clearly highlights the total absence of overlap between finds from Canosa and Arpi.

## 4. CONCLUSIONS

In this paper, we analized ceramic body compositional data of archeological finds (Apulian red figured pottery). Our aim was to show that plotting this kind of nonlinearly correlated objects, represented in curved multi-dimensional spaces, by mapping them into a lower-dimensional subspace, provides the information sought better than do the PCA technique. In this paper, we considered the hierarchical NLPCA method as a natural nonlinear generalization of the standard PCA technique applied to the statistical classification of archeological findings. Indeed, this methodology based on Auto-Associative Neural Network models, provides the optimal (in the sense of a minimum square reconstruction error) subspace spanned by a minimum number, say $k = 2, 3$, of principal components, and, at the same time, the maximum chemical data information immediately "readable" by visualization. The hierarchical NLPCA approach, as it shown in Section 3, contributes to visualize differences between objects better than the PCA method, helping to distinguish the workshops and production manufacturing. In particular, the opportunity achieved by NLPCA to separate the objects according to the site of provenance is an important result in itself. Indeed, it makes possible to identify locally produced and imported finds by a statistical treatment of chemical compositional data of the ceramic bodies, which can be used to "reallocate" pottery whose information about the origin was lost (i.e., artifacts in museums and private collections or coming from clandestine markets), unfortunately representing a high amount of the entire class.

## REFERENCES


1. Mangone A, Giannossa LC, Laganara C, Laviano R, Traini A. Manufacturing expedients in medieval ceramics in Apulia. *J. Cult. Herit.* 2009; **10**: 134–143.
2. Mangone A, Giannossa LC, Laviano R, Fioriello CS, Traini A. Late Roman lamps from Egnatia: from imports to local production. Investigations by various analytical techniques to the correct classification of archaeological finds and delineation of technological feature. *Microchem. J.* 2009; **91**: 214–221.
3. Mangone A, Giannossa LC, Colafemmina G, Laviano R, Traini A. Use of various spectroscopy techniques to investigate raw materials and define processes in the overpainting of Apulian red figured pottery (4th century BC) from southern Italy. *Microchem. J.* 2009; **92**: 97–102.
4. Giannosa LC, Loperfido S, Caggese M, De Benedetto GE, Laviano R, Sabbatini L, Mangone A. A systematic characterization of fibulae from Italy: from chemical composition to microstructure and corrosion processes. *New J. Chem.* 2013; **37**(4): 1238–1251.
5. Giannossa LC, Acquaviva M, De Benedetto GE, Acquafredda P, Laviano R, Mangone A. Methodology of a combined approach: analytical techniques to identify technology and raw materials of thin walled pottery from Herculaneum and Pompeii. *Anal. Methods* 2014; **6**(10): 3490–3499.
6. Aruga R, Mirti P, Casoli A. Application of multivariate chemometric techniques to the study of Roman pottery (terra sigillata). *Anal. Chim. Acta* 1993; **276**: 197–204.
7. Mirti P, Gulmini M, Pace M, Elia D. The provenance of red figure vases from Locri Epizephiri (southern Italy): new evidence by chemical analyses. *Archaeometry* 2004; **46**: 183–200.
8. Mirti P, Perardi A, Gulmini M, Preacco Ancon MC. A scientific investigation in the provenance and technology of a black-figure amphora attributed to the Priam group. *Archaeometry* 2006; **48**(1): 31–43.
9. Mangone A, De Benedetto GE, Fico D, Giannossa LC, Laviano R, Sabbatini L, van der Werf I, Traini A. Multianalytical study of




  

archaeological faience from Vesuvian area as a valid tool to investigateprovenance and technological features. *New J. Chem.* 2011; **35**: 2860–2868.
10. Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L. *Principal Components and Factor Analysis Chemometrics: A Textbook*. Elsevier: Amsterdam, 1988.
11. Neff H. *Chemical Characterization of Ceramic Pastes in Archaeology*. Prehistory Press: Madison, Wisconsin, 1992.
12. Mangone A, Giannossa LC, Ciancio A, Laviano R, Traini A. Technological features of apulian red figured pottery. *J. Archaeol. Sci.* 2008; **35**(6): 1533–1541.
13. Mangone A, Caggiani MC, Giannossa LC, Eramo G, Redavid V, Laviano R. Diversified production of red figured pottery in Apulia (Southern Italy) in the late period. *J. Cult. Herit.* 2013; **14**: 82–88.
14. Aruga R. The problem of responses less than the reporting limit in unsupervised pattern recognition. *Talanta* 2004; **62**: 871–878.
15. Bellanti F, Tomassetti M, Visco G, Campanella L. A chemometric approach to the historical and geographical characterisation of different terracotta finds. *Microchem. J.* 2008; **88**: 113–120.
16. Pizarro C, Pérez-del-Notario N, Sáenz-Gonzalez C, Rodriguez-Tecedor S, Gonzalez-Saiz Depa JM. Matching past and presente ceramic production in the banda area (Ghana): improving the analytical performance of neutron activation analysis in archaeology using multivariate analysis techniques. *Archaeometry* 2012; **54** (1): 101–113.
17. Remolá JA, Lozano J, Ruisanchez I, Larrechi MS, Rius FX, Zupan J. New chemometric tools to study the origin of amphorae produced in the Roman Empire. *Trends Anal. Chem.* 1996; **15**(3): 137–150.
18. Miller JN, Miller JC. *Statistics and Chemometrics for Analytical Chemistry* (5th edn). Pearson Education Limited, Prentice Hall: Great Britain, 2005.
19. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. *Lecture Notes on Compositional Data Analysis*. University of Girona: Spain, 2007.
20. Sarcià SA, Cantone G, Basili VR. Auto-Associative Neural Networks to improve the accuracy of estimation models. In *Artificial Intelligence Applications for Improved Software Engineering Development: New Prospects*, Meziane F, Valeda S (eds). IGI Global: USA, 2009; 66–81.
21. Scholz M, Fraunholz M, Selbig J. Nonlinear Principal Component Analysis: Neural Network Models and Applications. In *Principal Manifolds for Data Visualization and Dimension Reduction*, Gorban AN, Kegl B, Wunsch DC, Zinovyev A (eds), LNCSE, vol. 58. Springer Berlin: Heidelberg, 2007; 44–67.
22. Linker R. Spectrum analysis by recursively pruned extended autoassociative neural network. *J. Chemometrics* 2005; **19**(9): 492–499.
23. de la Fuente RL-N, Garcia-Muñoz S, Biegler LT. An efficient nonlinear programming strategy for PCA models with incomplete data sets. *J. Chemometrics* 2010; **24**: 301–311.
24. Dimitrov I, Naneva L, Bangov I, Doytchinova I. Allergenicity prediction by artificial neural network. *J. Chemometrics* 2014; **28**: 282–286.
25. Scholz M, Vigario R. Nonlinear PCA: a new hierarchical approach. In *ESANN'2002 Proceedings - European Symposium on Artificial Neural Networks,* d-side publi, Verleysen M (ed). D-side Publications: Bruxelles, Belgium, 2002; 439–444.
26. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* 1991; **37**(2): 233–243.
27. Rojas R. *Neural Networks*. Springer-Verlag: Berlin, 1996.
28. Bishop CM. *Neural Network for Pattern Recognition*. Oxford University Press, 1995.
29. Trendall AD, Cambitoglou A. *The Red-figured Vases of Apulia; Late Apulian II*. Clarendon Press: Oxford, 1982.
30. Robinson EGD. Workshops of Apulian red-figure outside Taranto. In *EUMOUSIA. Ceramic and Iconographic Studies in Honour of Alexander Cambitoglou*, Descoeudres JP (ed), Meditarch, Supplement, vol. 1. University of Sidney: Sidney, 1990; 181–196.
31. Trendall AD. *Red Figure Vases of South Italy and Sicily. A Handbook*. Thames-Hudson: London, 1989.
32. Trendall AD, Cambitoglou A. *The Red-Figured Vases of Apulia; Early and Middle Apulian I*. Clarendon Press: Oxford, 1978.
33. Mardia K, Kent JT, Bibby J. *Multivariate Analysis*. Academic Press: London, 1979.
34. Pawlowsky-Glahn V, Buccianti A. *Compositional Data Analysis: Theory and Applications*. J. Wiley & Sons: New York, 2011.