Statistics in Archaeology: New directions

Pedro Delicado*
Departament d'Economia i Empresa, Universitat Pompeu Fabra
Ramon Trias Fargas 25-27, 08005 Barcelona, SPAIN
delicado@upf.es, http://www.econ.upf.es/%7Edelicado

July 28, 1998

*This paper corresponds to the invited talk given at Barcelona, March 25, 1998, in the opening session of the *Computer Applications in Archaeology* meeting (CAA'98). Special thanks are due to J. A. Barceló, local organizer of the meeting. C. C. Beardah is thanked for allowing me to use data sets from Baxter and Beardah (1997). This work was partially supported by the Spanish DGES grant PB96-0300.

Abstract

Connections between Statistics and Archaeology have always appeared very fruitful. The objective of this paper is to offer an outlook of some statistical techniques that are being developed in the most recent years and that can be of interest for archaeologists in the short run.

Key Words: Artificial Neural Networks. Bayesian Statistics. Bootstrap. Multivariate Analysis. Nonparametric Statistics. Statistical Software.
JEL: C10

1 Introduction

Scientific research is not outside the world-wide extension process that many aspects of the live are experimenting at the end of the century. Connections between different areas are easier and easier. On the other hand, specialization is almost a requirement to be able to contribute significantly in any field of the scientific spectrum. So it is frequent to find researchers coming from very different areas who work on the same kind of problems.

This atmosphere favors that links between Statistics and Archaeology become even stronger than they traditionally were. This paper attempts to throw some light on the topics that statisticians are dealing with, in the hope that archaeologists are incorporating them to their usual research activities.

There are many stages in an archaeological research process where statistical problems are present. The problem of data collection appears first. Sampling techniques and experimental design offer good classical solutions and no new contributions are being referred here. We just turn to remark once again that an appropriate random selection of the data is crucial to validate posterior inference procedures.

Once data have been collected, the archeologist is concerned about *see* her data. Exploratory Data Analysis is then to her service. Here we refer to that matter when we explain some nonparametric methods (subsection 2.1) and multivariate methods (section 4).

Statistical inference is broadly present in Archaeology. Dating methods, typology (cluster analysis) and discriminant analysis are maybe the most popular of these procedures. Recent advances are compiled in sections 2, 3 and 6. Bootstrap and other resampling methods are the content of section 5. They turn out to be useful general tools for validating and calibrating inference methods.

A common problem in many practical studies is the simultaneous presence of qualitative and quantitative information. Some of the techniques developed here are specially adequate to deal with this problem. It is proper to emphasize multivariate methods based on distances (subsection 4.3) and Bayesian methods (section 6).

Section 7 quickly reviews some statistical packages. It also includes a list of Internet resources where statistical software related with Archaeology is accessible.

We are not reviewing the field known as Spatial Statistics notwithstanding that it is an important connection area between Statistics and Archaeology. Only some recent references are listed: Griffith (1997) is dedicated to Spatial Statistics, and Buck, Cavanagh, and Litton (1996) has a chapter about the Bayesian analysis of spatial data.

The rest of the paper is organized by statistical topics. It is hoped that at the end of the paper the correspondence between reviewed statistical methods and archaeological problems looks clear.

2 Nonparametric methods

We call *nonparametric methods* to the statistical techniques dealing with the estimation of functionals of the density or regression function. There are no parametric assumptions involved in the estimation (for instance, no normality

assumptions are made) or we can think that the parameter space has infinite dimension (for instance, each possible density function could be a parameter; then the parameter space would be the set of all possible functions, who has infinite dimension). Strictly speaking, nonparametric methods are not a novelty neither in Statistics nor in Archaeology (a simple histogram is a nonparametric estimator of a density function), but in our opinion all their potential has not been fully explored. We present here density estimation by kernel methods, and regression function estimation performed by three different approaches.

2.1 Kernel density and regression estimation

The objective is to estimate the density function (i.e. the value $f(x)$ of the density function of a random variable X at a point x, given a random sample of X: $X_1, \ldots, X_n$) or the regression function (i.e., the conditional expected value $E(Y|X = x)$ given a random sample of the variable (Y, X): $(X_1, Y_1), \ldots, (X_n, Y_n)$).

Kernel techniques are characterized by the use of a weight function (the *kernel* function) that permits give more mass to observed data X_i (or (X_i, Y_i)) near the point x when $f(x)$ (or $E(Y|X = x)$) is estimated.

Kernel density estimation can be considered as a way of smoothing the histogram. A specific reference in Archaeology is Baxter and Beardah (1997). More generic references are Silverman (1986) and Simonoff (1996)

Example 1

This example is based on data and ideas from Baxter and Beardah (1997). From 105 specimens of Romano-British waste glass, 11 variables were obtained measuring its chemical composition. Figure 1 represents the estimated density of the scores on the first principal component for each individual. The estimation is done with Beardah's routines KDE (see section 7). Clearly, there are two groups of glasses along the first principal component. □

Nonparametric regression (also know as smoothing techniques) is motivated as follows. Assume we have a dependent variable Y that can be explained by the independent variable X. A way to make more flexible linear regression is passing from

$$E(Y|X = x) = b_0 + b_1 x \quad \text{to} \quad E(Y|X = x) = m(x),$$

where m is an unknown function. The nonparametric estimation of $m(x)$ is done by computing local mean values:

$$\hat{m}(x) = \{\text{average of values } Y_i \text{ corresponding to values } x_i \text{ that are near the value } x\} = \text{Average}\{Y_i | x_i \text{ are near } x\}.$$

This is a sample version of $E(Y|X = x)$. A general reference is Simonoff (1996).

Example 2

We use again data from Baxter and Beardah (1997). In the figure 2 of that paper, we can see the scatter plot of the scores on the second principal component versus the scores on the first one. As we know, there is no linear relation between these two variables, but we can find nonlinear relation by using nonparametric

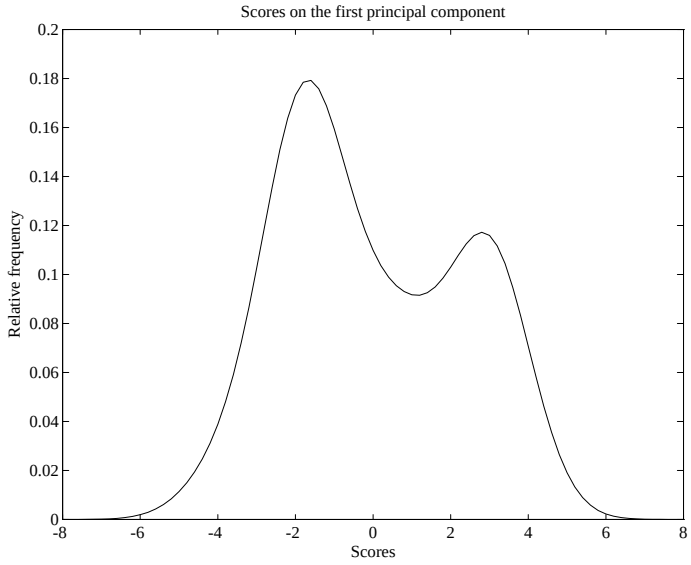


Figure 1: Example 1. Romano-British waste glass.

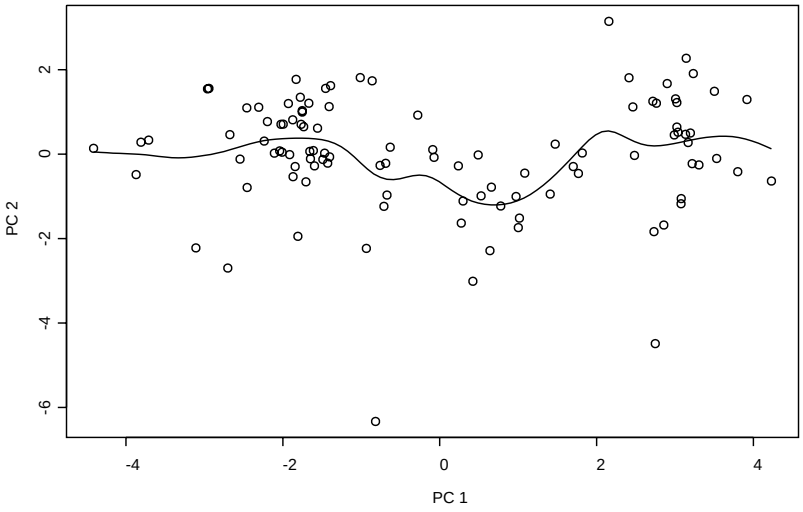


Figure 2: Example 2. Nonparametric regression for Romano-British waste glass.

regression: the second principal component can be nonlinearly explained by the first one. Figure 2 shows the result. □

Applications of nonparametric density and regression estimation include exploratory data analysis, cluster analysis (Baxter and Beardah 1997) and generalized additive models (GAM).

2.2 Generalized additive models (GAM)

We consider now the multiple linear regression model

$$E(Y|X) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p.$$

A nonparametric extension of it is the *additive model*:

$$E(Y|X) = b_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p),$$

where f_i are unknown functions that can be estimated by smoothing techniques. A complementary extension is the *Generalized Linear Model (GLM)*. For a known function g (the *link* function),

$$g(E(Y|X)) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p.$$

For instance, in the *logit* model Y is a 0-1 variable,

$$E(Y|X) = Prob(Y = 1|X) = \frac{e^{X'\beta}}{1 + e^{X'\beta}}.$$

The logit model is a GLM: if we define the *logit* transformation as $l(p) = \log(p/(1-p))$, for $p \in [0,1]$, then $l(E(Y/X)) = X'\beta$. The Generalized Additive Model (GAM) extends the original linear model in both directions:

$$g(E(Y|X)) = b_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p).$$

Venables and Ripley (1994) has some chapters dedicated to GLM and GAM. Comprehensive references are cited there.

2.3 Classification and regression trees (CART)

Classification and regression trees (CART) are nonparametric techniques for discriminant analysis and multiple regression. The key reference on this field is Breiman, Friedman, Olshen, and Stone (1984). Let us look at the simple case of discriminant analysis for two population. We observe

$$(Y_i; X_{1i}, \ldots, X_{pi}), i = 1, \ldots, n$$

where variable Y_i is 1 or 2, according to what population the case i is coming from. The objective is to predict the value of Y_i, given the information brought by $X_{1i}, \ldots, X_{pi}$. CART selects one of the p explanatory variables (that one with the biggest discrimination power, for instance, X_1) and divide the original sample into two parts: cases with $X_{1i} \geq C1$ (say, `Subsample 1`) and cases with $X_{1i} < C1$ (say, `Subsample 2`), as Figure 3 indicates. The choice of $C1$ is done

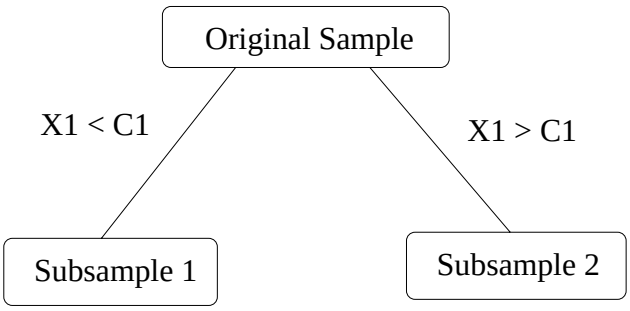


Figure 3: CART: first step.

in a way that these two new subsamples are as similar as possible to the original classes identified by Y_i.

Now, the same procedure is done in `Subsample 1` and `Subsample 2`. A binary tree is the output of the procedure. Each node is divided into two branches according to an observed variable. The final nodes have associated one of the values of Y: 1 or 2.

To classify a new observation $(X_1, \ldots, X_n)$, we let this new case running the tree from the first node to a final node (according to its values of X_i and to the splitting rules defining the successive intermediates nodes of the tree) and it is finally classified into the group indicated by the corresponding final node.

Example 3

Data consisting on measurements of 150 male Egyptian skulls from 5 different time periods (-4000, -3300, -1850, -200, 150) are considered. Data and original source can be found in Manly (1994). The objective is to discriminate between time periods based on the measures. Thirty skulls are measured from each period. Four measures are taken from each skull (see Figure 1.1 at Manly 1994):

$V1$	Maximal Breadth of Skull
$V2$	Basibregmatic Height of Skull
$V3$	Basialveolar Length of Skull
$V4$	Nasal Height of Skull

Figure 4 shows the final *classification tree* as the commercial package S-plus (see section 7) produces. Each final node contains a label indicating to which one of the five periods would be classified a skull with measures according to the path going from the original node to the final one. □

3 Artificial neural networks (ANN).

Artificial Neural Networks are very popular tools in Artificial Intelligence and Engineering. They are based on the connection of many very simple mathematical models, the *artificial neurons*, that imitates the work of a real neuron. We

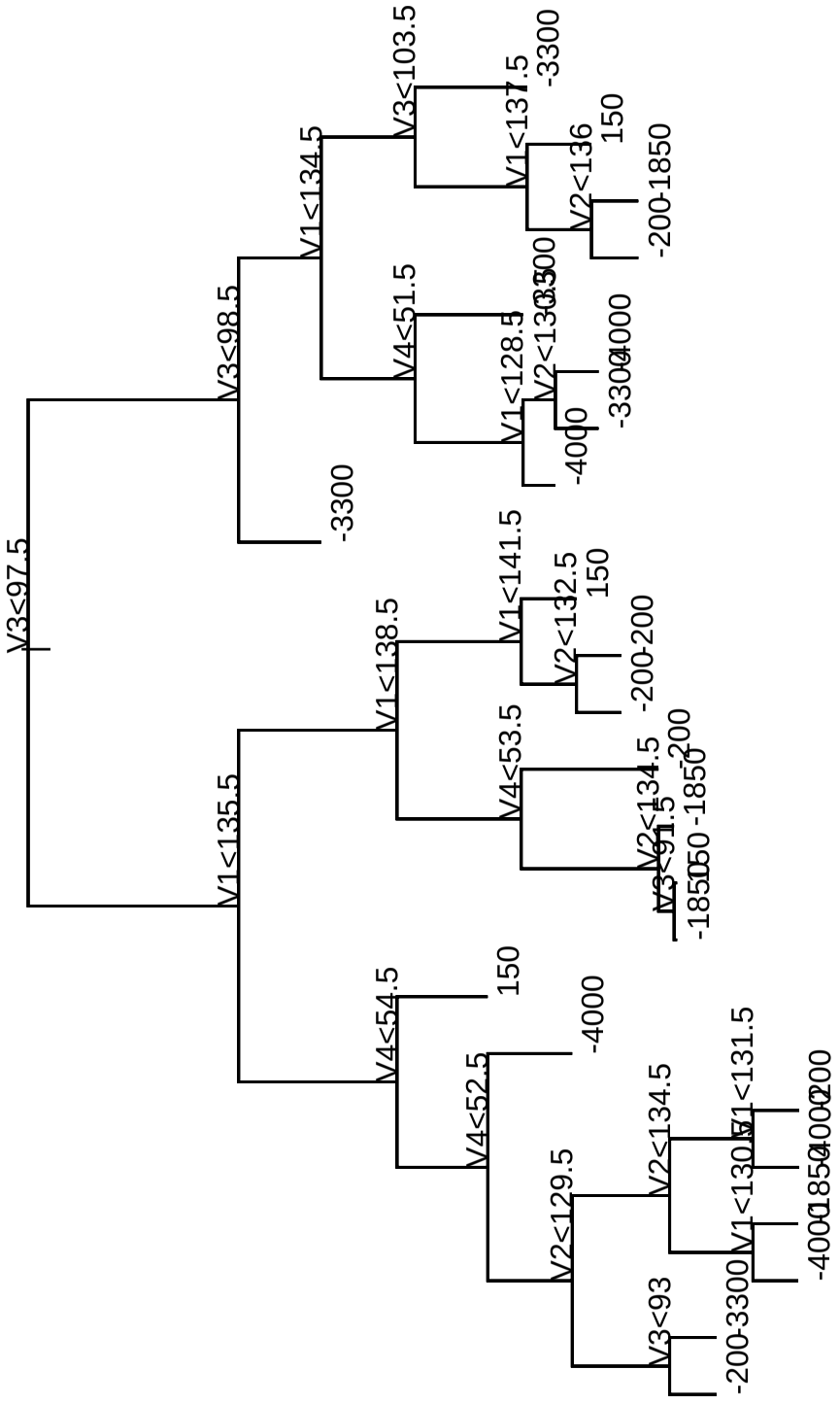
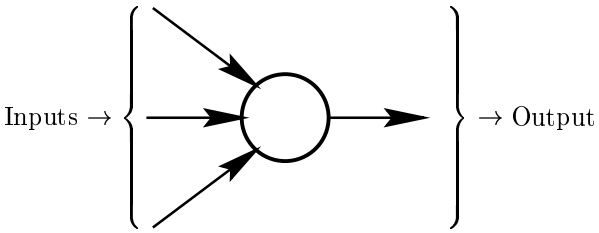


Figure 4: Example 3. CART for Egyptian skulls.

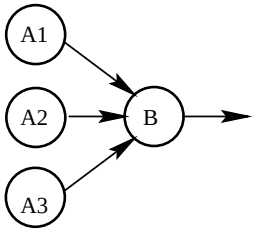
can think about an artificial neuron as a mechanism that transforms numerical input information into numerical output information:



$$\text{Output} = g(\text{Input 1}, \ldots, \text{Input p}),$$

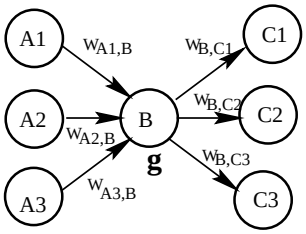
The function g is an activation function, that takes values near 1 for great inputs and values near 0 for low inputs.

Inputs for neuron B are the (weighted) outputs of other neurons $A1, \ldots, Ap$.



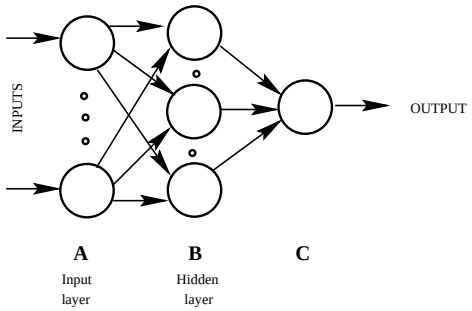
$$\text{Output}_B = g(w_1 \text{Output}_{A1} + \ldots + w_p \text{Output}_{Ap}),$$

Moreover, the output of B (modulated by some weights) is one of the inputs of (many) other neurons $C1, \ldots, Cq$.



There can be a lot of these B neurons. In fact, neurons Ai and Ci are of the same type as B. When many neurons and many interactions are considered, we obtain an Artificial Neural Network.

As an example, we present the *one hidden layer feed forward propagation ANN with a unique output*. There are three neuron layers (A, B and C) and the information goes from the first layer to the second layer and then to the third one.



The structure of the net is as follows.

- separate external information is given to each neuron in layer A,
- g_A is the identity function,
- $g_B = g_C = g$,
- there is only a final neuron in layer C, and
- the information produced by layer C is return outside.

Let x_i be the numerical information given to the i-th neuron in layer A and let y be the output obtained from neuron C. Then,

$$\begin{aligned} y = \text{Output}_C &= g(w_1^{BC}\text{Output}_{B1} + \ldots + w_p^{BC}\text{Output}_{Br}) = \\ &= g\left(\sum_{j=1}^{r} w_j^{BC}\text{Output}_{Bj}\right) = \\ &= g\left(\sum_{j=1}^{r} w_j^{BC} g\left(\sum_{i=1}^{p} w_i^{AB}\text{Output}_{Ai}\right)\right) = \\ &= g\left(\sum_{j=1}^{r} w_j^{BC} g\left(\sum_{i=1}^{p} w_i^{AB} x_i\right)\right). \end{aligned}$$

This ANN is a parametric family of nonlinear functions from $\mathbb{R}^p$ to $\mathbb{R}$ that, given inputs $(x_1, \ldots, x_n)$ returns the output value y. Each set of parameters $\{r, w_i^{AB}, w_j^{BC}\}$ determines a different function.

The fundamental property of the ANN is known as *Universal approximation property* and tells that every function from $\mathbb{R}^p$ to $\mathbb{R}$ can be approximated by one of these *one hidden layer ANNs*.

This property originates the statistical interest for ANN. Given observations $(y_i, x_i), i = 1, \ldots, n$, generated by the model

$$y_i = \Phi(x_i) + \varepsilon_i,$$

we can estimate Φ by means of a *one hidden layer ANN*.

Strictly speaking, this is a particular case of a nonlinear regression analysis. ANN literature has been developed (quite) independently and it has provided

important contributions. For instance, the estimations process (or *net training* process) is implemented in a very different way in ANN and in nonlinear regression analysis. More connections between both topics are needed.

Statistical applications of the ANN include discriminant analysis, regression, cluster analysis and nonlinear multivariate analysis.

4 Nonlinear multivariate analysis (NLMVA)

We present here some alternative tools to the well known Principal Component Analysis and Correspondence Analysis. Some of these techniques are quite recent, but others have been introduce in the statistical literature in the 80's.

4.1 NLMVA for discrete data: The Gifi system

Albert Gifi is the *nom de plume* for a group of authors related with the Department of Data Theory at the University of Leiden, The Netherlands. These authors compiled their work of more than 10 years in the book of Gifi (1990).

The book presents the particular idea of MVA that the Gifi team has. For instance, they affirm that all we can observe is discrete (but not all the *discrete* are equally rich). Moreover, no random variables are needed to be assumed the origin of data: data are enough to make Statistics.

The basic principle of the proposed techniques is the concept of *homogeneity*. The observed data and the observed variables are jointly transformed (in a nonlinear way) in order to obtain transformed objects as much *homogeneous* as possible. Not all transformation is always allowed: it depends on the data richness.

One special case of the proposed methodology is equivalent to *Multiple Correspondence Analysis* (MCA), but the scope of the book is wider than MCA. Continuous data can also be analyzed after a preliminary codification.

Some of the procedures developed in this book are included in the commercial package SPSS:

- HOMALS, similar to multiple correspondence analysis,
- PRINCALS, a nonlinear version of principal components also available for ordinal data,
- OVERALS, a nonlinear version of canonical correlation analysis.

4.2 NLMVA for continuous data: Principal curves

Principal curves are parameterized one dimensional curves that passe through the middle of a p-dimensional cloud of data. They are nonlinear generalizations of the first principal component. They were introduced in the work of Hastie and Stuetzle (1989). Some work has been done since then, but principal curves have not been very used, mainly because of the difficulties in the definition and implementation for the second (and posterior) principal curves, and also because the existing associated software has not been widely diffused. Other references are LeBlanc and Tibshirani (1994), Kégl, Krzyżak, Linder, and Zeger (1997), and Delicado (1998).

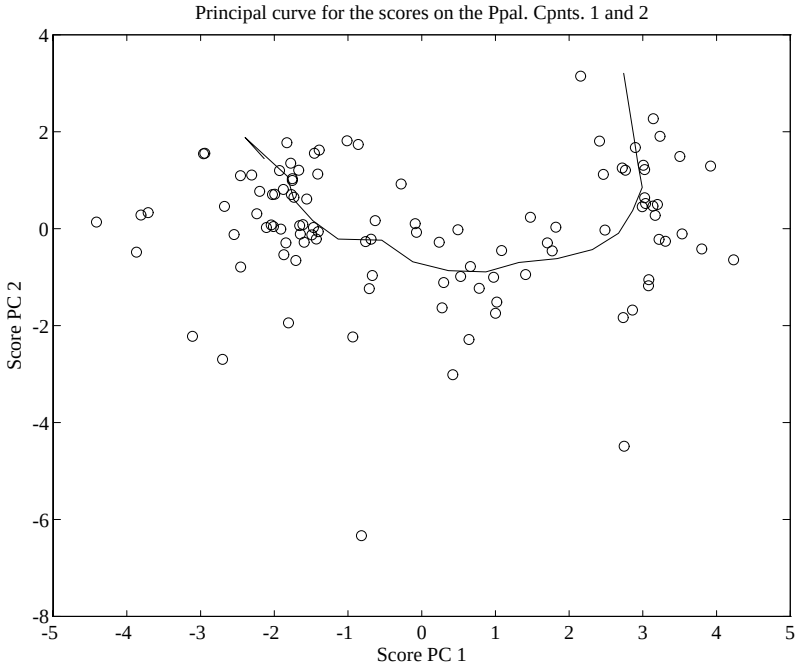


Figure 5: Example 4. Principal curve for Romano-British waste glass.

There exists a parallel neural network approach: Self-Organizing Maps and Generative Topographic Mapping. See, for instance, Bishop, Svensén, and Williams (1997).

Example 4

Figure 5 shows the first principal curve in the data set of the 105 glass scores on the two first principal components. □

4.3 Distance based methods

We have observed p characteristics of n objects, for instance,

	Characteristic 1	...	Characteristic p
Object 1	1	...	22.5
Object 2	0	...	29.0
⋮	⋮	⋮	⋮
Object n	1	...	17.3

Sometimes it is easier to give a distance between objects matrix $D = (d_{ij})$,

$$d_{ij} = \text{Distance}(\text{Object}_i, \text{Object}_j),$$

based on the p observed attributes, than proposing a joint model for these attributes. It is possible to mix qualitative and quantitative information (Gower's distance, for instance, but there exist alternative methods).

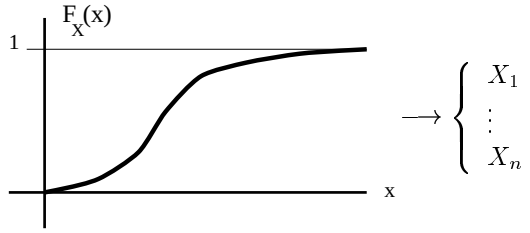
In the last years many work has been done in order to develop usual multivariate statistical analysis (principal components, discriminant analysis, regression analysis) from a distance matrix. In the Universidad de Barcelona there is an active group of researchers in this field (see Cuadras, Fortiana, and Oliva 1997). Multidimensional scaling is a precedent to this line of work.

Talking about distances, we have to refer to cluster analysis. Alternatives to the usual hierarchical methods are the pyramidal clusters, where the resulting clusters are allowed to be overlapped. More flexibility is obtained at the prize of more difficult interpretation of the results. As a reference, see Diday (1986).

5 Bootstrap and other resampling methods

Let us assume that we are interested in a particular characteristic X of a population and that we observe the value of this variable in a sample of similar objects. For instance, X can be the volume of some cups. Let $X_1, \ldots, X_n$ be our observations.

We can think that each X_i is the realization of the random variable X, which has unknown distribution function F_X.



Let $\mu = E(X)$ be the theoretical mean volume. Inference about μ is based on the sample mean

$$\hat{\mu} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Confidence intervals for μ can be constructed if we assume that F_X belongs to a particular parametric family and/or by means of asymptotic results. For instance, if $Var(X) < \infty$, we know that

$$\left(\bar{X}_n - 1.96\frac{S_X}{\sqrt{n}}, \bar{X}_n + 1.96\frac{S_X}{\sqrt{n}}\right)$$

is an asymptotic 95% confidence interval for μ. Let us observe that this interval has the form

$$(L_A, U_A)\,,$$

where L_A is the 2.5 percentile of the asymptotic distribution of $\bar{X}_n$, and U_A is its 97.5 percentile.

An alternative way to provide a confidence interval for μ could be as follows. Assume that we know F_X, so we can repeat as many times as we want the sampling process:

$$\begin{array}{lcccc} \text{Sample 1} & X_1^{(1)}, \ldots, X_n^{(1)} & \longrightarrow & \bar{X}_n^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ \text{Sample N} & X_1^{(N)}, \ldots, X_n^{(N)} & \longrightarrow & \bar{X}_n^{(N)} \end{array}$$

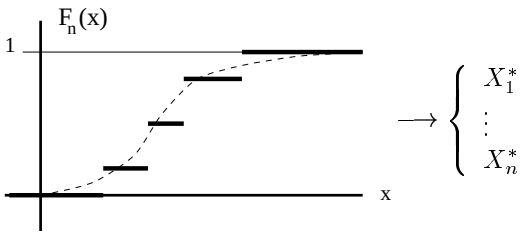
We have a size N sample of $\bar{X}_n$ from which we can build a confidence interval for μ as follows,

$$(L_N, U_N)\,,$$

where L_N is the 2.5 percentile of the sample $\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(N)}$, and U_N is its 97.5 percentile.

The problem of this approach arises when we realize that we do not know F_X. When a statistician does not know a population characteristic, usually he or she estimates it from the data. So we estimate the theoretical distribution function F_X by the empirical distribution function F_n of our initial sample $X_1, \ldots, X_n$:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i).$$



Efron (1979) introduces the term bootstrap to designate this resampling procedure (there exist other resampling methods, as the *jacknife*, also covered by this book). He proposes instead of sampling from F_X, taking samples from F_n, or equivalently, drawing values from the set $\{X_1, \ldots, X_n\}$ with replacement:

$$\begin{array}{lcccc} \text{Bootstrap sample 1} & X_1^{*(1)}, \ldots, X_n^{*(1)} & \longrightarrow & \bar{X}_n^{*(1)} \\ \vdots & \vdots & \vdots & \vdots \\ \text{Bootstrap sample N} & X_1^{*(N)}, \ldots, X_n^{*(N)} & \longrightarrow & \bar{X}_n^{*(N)} \end{array}$$

We *resample* our original sample. It is a valid procedure in many cases, but bootstrap does not always gives appropriate answers. Two are the direct advantages of bootstrap: first, in many cases bootstrap gives better results that asymptotic arguments, and second, sometimes the only possibility to make inference is by a resampling procedure.

The scope of bootstrap methods includes, among others, confidence intervals, hypothesis tests (as an example, see Delicado and del Río 1994) and times series. Two references are advisable: Efron and Tibshirani (1993) is a very well written book, recommended also as a very good book in Statistics; Davidson and Hinkley (1997) presents an updated review of this broad topic.

6 Bayesian methods

The Bayesian approach to Statistics is not at all new. Nevertheless we have considered appropriate to include it in this paper for several reasons. First, in

the last years the Bayesian contribution to the Statistics research is dramatically increasing. Moreover, the use of powerful computers permits to give a Bayesian answer to many problems that where not accessible to Bayesian statisticians few years ago. And finally, the Bayesian methodology is not widely used in Archaeology.

Recently has appeared a book that introduces Bayesian methods to archaeological community: Buck, Cavanagh, and Litton (1996). We can read in its preface:

> *The major advantage of the Bayesian approach is that it allows the incorporation of relevant prior knowledge or beliefs into the analysis.*

This sentence words an essential point of Bayesian Statistics.

6.1 An introduction to Bayesian methodology

We analyze a simple problem. We want to date a red pigmented ceramic found in a excavation. There are three possible periods for that kind of objects, *Period 1*, *Period 2* or *Period 3*, and we know that

$$\begin{aligned}\text{Prob}(red|Period\ 1) &= 0.2,\\ \text{Prob}(red|Period\ 2) &= 0.5,\\ \text{Prob}(red|Period\ 3) &= 0.8.\end{aligned}$$

The classical (or *frequentist*) answer is the following:

> *I should date my ceramic in the more likely period.*

That is, we take the maximum likelihood estimator, and the result is:

$$estimated\ period\ =\ 3.$$

The Bayesian answer is as follows. *A priori*, before observing the ceramic, we can assume that the three periods are equally probable:

$$\text{Prob}(Period\ 1) = \text{Prob}(Period\ 2) = \text{Prob}(Period\ 1) = \frac{1}{3}.$$

The Bayes' Theorem permits to calculate the probability of the event

$$A_i = \text{“}The\ ceramic\ corresponds\ to\ Period\ i\text{”}$$

conditioned to the observed fact

$$B = \text{“}The\ ceramic\ is\ red\text{”}.$$

$$P(A_i|B) = \frac{P(A_i \cap B)}{B} \Rightarrow P(A_i \cap B) = P(A_i|B)P(B).$$

But also,

$$P(A_i \cap B) = P(B|A_i)P(A_i).$$

So

$$P(A_i|B)P(B) = P(B|A_i)P(A_i) \Rightarrow$$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \text{ (Bayes' Theorem)}$$

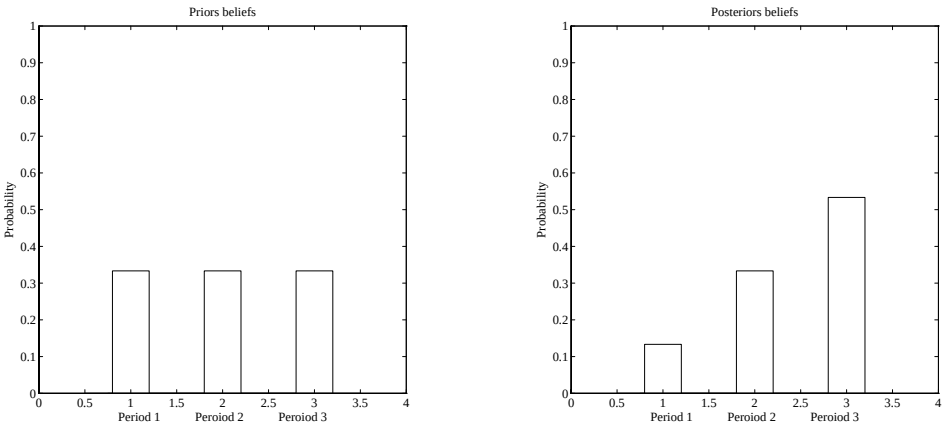


Figure 6:

We can compute Prob(*Period i*|*red*) and also Prob(*red*):

$$P(red) = P(red|Period\ 1)P(Period\ 1)+$$

$$+P(red|Period\ 2)P(Period\ 2) + P(red|Period\ 3)P(Period\ 3) = 0.5$$

and

$$P(Period\ 1|red) = \frac{0.2 \times \frac{1}{3}}{.5} = \frac{2}{15},$$

$$P(Period\ 2|red) = \frac{0.5 \times \frac{1}{3}}{.5} = \frac{5}{15},$$

$$P(Period\ 3|red) = \frac{0.8 \times \frac{1}{3}}{.5} = \frac{8}{15}.$$

So our prior beliefs have been modified in the experimental stage (the observation of the color of our ceramic) and now we have "*a posteriori*" beliefs.

A summary of Bayesian methodology could be as follows:.

1. Prior information is expressed as a probability distribution over the parameter space.

2. Likelihood function is, in fact, the conditional distribution of the observations given the parameter values.

3. Bayes' Theorem is used to combine prior information with experimental information and transform them into posterior information: another probability distribution over the parameter space.

Some of the positive points of Bayesian Statistics are the following. Bayesian approach is conceptually appealing (and simple). For instance, the probability that a parameter belongs to a Bayesian 95% confidence interval is really 0.95. Moreover, it is possible to include prior qualitative information into the inference process. It is also possible to progressively update the beliefs: the "posterior" information of today is the "prior" information of tomorrow.

On the other hand, some difficulties are inherent in Bayesian methodology. A prior distribution is always needed, even if you do not have such "a priori" information, and some results will strongly depend on that prior. Moreover, in medium and large size problems, the computation of the posterior distribution is extremely difficult. Many times only approximate solutions are available, as the produced by the Gibbs sampling method (see subsection 6.2).

We refer to Buck, Cavanagh, and Litton (1996), pp. 208-, to follow an application example of Bayesian methods to radiocarbon dating. There, it looks clear the superiority of these techniques when qualitative information has to be included in the analysis.

6.2 Computer intensive methods

The Bayes' Theorem version for absolutely continuous random variable is:

$$f_\theta(\theta|X=x) = \frac{f_X(x|\theta)f_\theta(\theta)}{\int f_X(x|\theta)f_\theta(\theta)d\theta},$$

where (X,θ) are random variables, θ is the unobserved parameter (that can be a vector of parameters), X is observed, $f_X(x|\theta)$ is the likelihood function and $f_\theta(\theta)$ is the prior distribution of θ. The result of the experimentation was the observation of the value x for X.

In many cases, we need to solve the denominator integral (and usually this is not an easy task). In other cases, we do not want the posterior distribution of all the parameter vector, but only the posterior distribution of some parameters. For instance, $\theta = (\mu, \sigma)$ and we want to know the posterior distribution of μ given the sample, with no attention of σ. Then,

$$f_\theta(\mu|X=x) = \int f_\theta(\mu,\sigma|X=x)d\sigma.$$

Again, we need to integrate.

The integration operation is usually not feasible. Then, Monte Carlo methods can help us to approximate the integrals. One of the most used Monte Carlo methods in this area is the *Gibbs sampling* (see again Buck, Cavanagh, and Litton (1996)). This technique has an additional advantage: only marginal conditional distribution must be specified

$$f(\mu|\sigma,x) \text{ and } f(\sigma|\mu,x)$$

instead of

$$f(\mu,\sigma|x).$$

7 Recommended Statistical Software

At our knowledge, no program exists implementing all the procedures presented in this paper. Nevertheless we could list some packages.

- S-plus. Many statistical developments are done in S-plus, and object oriented statistical commercial program that incorporates many of the most recent statistical techniques. A good book for S-plus is Venables and

Ripley (1994). S-plus is possibly the most updated commercial package. It also includes a module on Spatial Statistics. Some related web sites are `http://www.mathsoft.com` (the official web page of S-plus) and `http://www.stats.ox.ac.uk/pub/MASS`.

- OxCal. Bayesian inference (including Gibbs sampling) is possible by using this package from the Oxford Radiocarbon Laboratory. It is accessible at `http://units.ox.ac.uk/departments/archaeology`.

- KDE: Kernel Density Estimation MATLAB toolbox, by C.C. Beardah. See also Beardah and Baxter (1995). The routines can be downloaded at `ftp://ftp.maths.ntu.uk.ac/pub/ccb`.

- Some other Internet resources:

  - STATLIB (`http://www.stat.cmu.edu/statlib`) for Statistics and S-plus.
  - NEURAL CLASSIFICATION AND REGRESSION TREES. **Ntree** is a C library for the estimation of neural network smoothed versions of CART. (`http://www.informatik.uni-freiburg.de/pub/neural/`).
  - NEURAL NETWORKS IN MATLAB. See the web site `http://neural-server.aston.ac.uk/GTM/`.
  - GIFI SYSTEM. Source code for different compilers are available at `http://www.ucla.edu/gifi/`.
  - PRINCIPAL CURVES. The original S-plus program from Hastie and Stuetzle (1989) can be download at `http://www.stat.cmu.edu/S/principal.curve`. Many complete information about principal curves and software related with Kégl, Krzyżak, Linder, and Zeger (1997) is available at `http://www.cs.concordia.ca/%7Egrad/kegl/research/pcurves`. At `http://www.econ.upf.es/%7Edelicado/prcu`, you can find MATLAB routines implementing the principal curves approach described in Delicado (1998).

8 Conclusions

The possibilities of interaction between Archaeology and Statistics are extremely appealing. Joint projects of archaeologists and statisticians are certainly very promising. We borrow some words (here in italics) that Clive Orton (Orton 1997) write in the review of Buck, Cavanagh, and Litton (1996) for the *Journal of Archaeological Science*, and we finish the paper with a final advice:

> *My advice to archaeologists is* bear in mind Statistics, *and then* be sure that you know *a sympathetic statistician.*

References

Baxter, M.J. and C.C. Beardah (1997). Some archeological applications of kernel density estimates. *Journal of Archaeological Science*, **24**, 347–354.

Beardah, C.C. and M.J. Baxter (1995). Matlab routines for kernel density estimates and the graphical representation of archaeological data. Research Report 2/95, Nottingan Trent University, Dept. of Mathematics, Statistics and O. R.

Bishop, C. M., M. Svensén, and C. K. I. Williams (1997). GTM: A principled alternative to the self-organizing map. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche (Eds.), *Advances in Neural Information Processing Systems 9*, pp. 354–360. The MIT Press.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1984). *Classification and regression trees*. Chapman and Hall.

Buck, C.E., W.G. Cavanagh, and C.D. Litton (1996). *Bayesian Approach to Interpreting Archaeological Data*. Statistics in Practice. Wiley.

Cuadras, C.M., J. Fortiana, and F. Oliva (1997). The proximity of an individual to a population with applications to discriminant analysis. *Journal of Classification*, **14**, 117–136.

Davidson, A.C. and D.V. Hinkley (1997). *Bootstrap Methods and their Appication*. Cambridge University Press.

Delicado, P. (1998). Another look at principal curves and surfaces. Unpublished (`http://www.econ.upf.es/ delicado`).

Delicado, P. and M. del Río (1994). Bootstrapping the general linear hypothesis test. *Computational Statistics and Data Analysis*, **18**, 305–316.

Diday, E. (1986). Orders and overlapping clusters by pyramids. In J. de Leeuw, W. Heiser, J. Meulman, and F. Critchley (Eds.), *Multidimensional Data Analysis*, pp. 201–234. Leiden: DSWO.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *The Annals of Statistics*, **7**, 1–26.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley.

Griffith, D. A. (1997). *A casebook for spatial statistical data analysis*. Oxford.

Hastie, T. and W. Stuetzle (1989). Principal curves. *Journal of the American Statistical Association*, **84**, 502–516.

Kégl, B., A. Krzyżak, T. Linder, and K. Zeger (1997). Learning and design of principal curves. Technical Report preprint, Concordia Univ. and UC San Diego.

LeBlanc, M. and R. J. Tibshirani (1994). Adaptive principal surfaces. *Journal of the American Statistical Association*, **89**, 53–64.

Manly, B.F.J. (1994). *Multivariate Statiscal Methods. A primer. (2nd. ed.)*. Chapman and Hall.

Orton, C. (1997). Review of Bayesian Approach to Interpreting Archaeological Data (Buck et al., 1996). *Journal of Archaeological Science*, **24**, 575.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Simonoff, J.B. (1996). *Smoothing methods in statistics*. New York: Springer Verlag.

Venables, W.N. and B.D. Ripley (1994). *Modern Applied Statistics with S-plus*. Springer.