Principal Curves of Oriented Points: theoretical and computational improvements [1]

Pedro Delicado[2] and Mario Huerta[3]

[2]Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya
[3]Departament de Llenguatjes i Sistemes Informàtics, Universitat Politècnica de Catalunya

Summary

Principal curves where introduced by Hastie & Stuetzle (1989) as smooth parametric curves passing through the middle of a multidimensional data set. Delicado (2001) defines Principal Curves of Oriented Points, based on the fixed points of a function from $\mathbb{R}^p$ into itself. This definition is nonparametric and smoothing methods are used to find principal curves of a data set. Here we extend this work in two directions. First, we propose a bandwidth choice method based on the Minimum Spanning Tree of the data set. Second, we present an object oriented application that implements the principal curves computation for any dimension in a flexible recursive way. Examples on synthetic and real data are included.

Keywords: Bandwidth choice, clustering, Minimum Spanning Tree, object oriented programming, population pyramids.

[1] *Contact address:* Pedro Delicado, Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, C/ Pau Gargallo 5, 08028 Barcelona, SPAIN. This work was partially supported by the Spanish DGES grants PB98-0919 and BFM 2001-2327, and by the European Commission project HPCF CT-2000-00041. We are grateful to Valentín Navarro who helps us to pre-process population pyramids data set. The authors would like to thank two anonymous referees for their helpful comments and suggestions.
Running head: Principal Curves of Oriented Points.

[2]Pedro Delicado is Associate Professor, `delicado@upf.es`.

[3]Mario Huerta is PhD student in Computer Science, `mhuerta@lsi.upc.es`.

1 Introduction

Hastie & Stuetzle (1989) define *principal curves* (HSPC) as smooth curves passing through the middle of a multidimensional data set with a non elliptical configuration. Principal curves are nonlinear generalizations of the first principal component. In the same way, *principal surfaces* (resp., *principal manifolds*) are the generalization of the linear subspaces defined by the first two (resp., k) principal components. Hastie & Stuetzle (1989) give a formal definition of principal curves based on the concept of *self-consistency* (see (Tarpey & Flury 1996)). Principal surfaces are broadly covered in Hastie (1984).

Since then, several works on principal curves have appeared. Some of them are directly related with Hastie and Stuetzle's work: Banfield & Raftery (1992) reduce the estimation bias; under the definition of Tibshirani (1992) a given curve is always the principal curve of a random variable generated as that curve plus noise; LeBlanc & Tibshirani (1994) uses multivariate adaptive regression splines (see (Friedman 1991)) to develop estimation procedures of principal manifolds. Other papers follow different lines: Kégl, Krzyżak, Linder & Zeger (2000) introduce the concept of principal curves with a fixed length, and Sandilya & Kulkarni (2000) present curves with fixed turn; Smola, Mika, Schölkopf & Williamson (2001) find principal manifolds by minimizing the expected quantization error subject to some regularity restrictions; Delicado (2001) based its definition on *principal oriented points*; Verbeek, Vlassis & Krse (2002) fit polygonal lines with fixed number of segments to the data; Tan & Mavarovouniotis (1995) and Dong & McAvoy (1996) base their proposal on neural networks; Mulier & Cherkassky (1995) and Bishop, Svensén & Williams (1998) are supported on self organized maps and generative topographic mapping; Bishop & Tipping (1998), Tipping & Bishop (1999) and Chang & Ghosh (2001) use a normal mixture model based on the generative topographic mapping.

The paper of Chang & Ghosh (2001) does a good review of the field. They distinguish proposals valid only for principal curves from those that also permit manage with principal manifolds. They list the pros and the cons of the reviewed methods. Balázs Kégl dedicates a well documented web page to principal curves (`http://www.iro.umontreal.ca/~kegl/research/pcurves/`) including links to papers on principal curves, algorithms to find them, and some practical applications.

The present paper is based on Delicado (2001), where *Principal Curves of Oriented Points* (PCOP) were defined. Systems of principal curves of higher order were also introduced there, as a natural way to look for principal manifolds in the data. A brief introduction to PCOP is given in Section 2. Here we extend Delicado (2001) in several directions. In Section 3 we propose a bandwidth choice method for the smoother used in the PCOP construction.

This choice is based on a Minimum Spanning Tree previously computed, that also helps to detect clusters in PCOP intermediate steps (Section 4). Section 5 presents a flexible object oriented application of the PCOP method, allowing to manage data of any dimension and to extract systems of curves of any order. The design is recursive, as well as the general definition of PCOP in Delicado (2001). The paper finish with simulated and real data examples, developed at Section 6.

2 Principal curves of oriented points

The definition of a principal curve given in Delicado (2001) is based on a generalization of a local property of principal components for a multivariate normal distribution X: if v_1 is the direction of the first principal component then, for all $x \in \mathbb{R}^p$,

$$TV(X|X \in H(x, v_1)) \leq TV(X|X \in H(x, b)), \text{ for all } b \in S^{p-1},$$

where $H(x, b)$ is the hyperplane $\{y \in \mathbb{R}^p : (y - x)^t b = 0\}$, $TV(Y) = \text{Trace}(\text{Var}(Y))$ is the total variance of the random variable Y, and S^{p-1} is the unit ball in $\mathbb{R}^p$. Moreover $E(X|X \in H(x, v_1))$ belongs to the first principal component line for all x (see Proposition 1 in (Delicado 2001)).

Let now X be a p-dimensional absolutely continuous random variable with density f. The following functions are basic in the definition of principal curves:

$$\mu(x, b) = E(X|X \in H(x, b)), \ \phi(x, b) = TV(X|X \in H(x, b)),$$

$$b^*(x) = \arg \min_{b \in S^{p-1}} \phi(x, b), \ \mu^*(x) = \mu(x, b^*(x)), \ \phi^*(x) = \phi(x, b^*(x)).$$

We say that $b^*(x)$ is the *principal direction* of x. Smoothness of μ, ϕ, b^*, μ^* and ϕ^* depends on the smoothness of f. Figure 1 illustrates the definitions of b^* and μ^*.

Observe that, under normality, the iteration of μ^* leads (in one step) from any arbitrary point x_0 to a point x_1 in the first principal component line. Generalizing this property, Delicado (2001) defines the set $\Gamma(X)$ of *principal oriented points* (POP) of X as the set of fixed points of μ^*:

$$\Gamma(X) = \{x \in \mathbb{R}^p : x = \mu^*(x)\}.$$

Then, a curve α from I to $\mathbb{R}^p$ (where I is an interval in $\mathbb{R}$, possibly unbounded, and α is continuous and parameterized by the arc length) is a *principal curve of oriented points* (PCOP) of X if $\{\alpha(s) : s \in I\} \subseteq \Gamma(X)$. According to this definition, the first principal component line is a PCOP for the multivariate normal distribution.

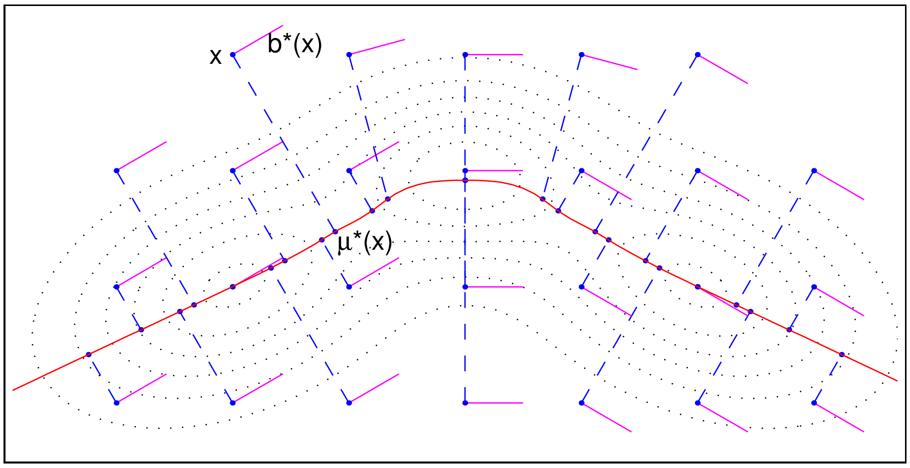

Figure 1: Some values of b^* and μ^* for the mixture of 3 bivariate normal distributions. The solid curve is the first *Principal Curve of Oriented Points* (PCOP).

Delicado (2001) proves that there exist POPs and that there exists a PCOP around any POP. Moreover, it is shown that only the first principal curve is a PCOP for the normal distribution, in contrast to HSPC definition (every principal component is a HSPC). In fact, the PCOP generalizes the variance-maximization property of the first principal component because it maximizes the total variance of the random variable projected along the curve.

Sampling counterparts of POPs and PCOPs are defined for data sets in Delicado (2001). Let $X_1, \ldots, X_n$ be a random sample from a multivariate random variable X. Smoothing techniques are used to compute sample expectations and variances, conditional to a hyperplane $H = H(x, b)$. Let X_i^H be the orthogonal projection of observation X_i onto H. A weight is associated to each projected observation, $w_i = K(\|X_i - X_i^H\|/h)$, where K is a kernel function and h is the smoothing parameter known as *bandwidth*. The smoothed expectation $\widetilde{\mu}(x, b)$ of the sample corresponding to H is defined as the weighted expectation of $\{X_i^H\}$ with weights $\{w_i\}$ (by definition, it belongs to $H(x, b)$). The smoothed variance corresponding to a hyperplane $H(x, b)$ is $\widetilde{\mathrm{Var}}(x, b)$, the weighted variance of the projected sample with weights $\{w_i\}$. The smoothed total variance is $\widetilde{\phi}(x, b) = \mathrm{Trace}(\widetilde{\mathrm{Var}}(x, b))$. From $\widetilde{\mu}$ and $\widetilde{\phi}$ there is a natural way to define $\widetilde{b}^*$, $\widetilde{\mu}^*$ and $\widetilde{\phi}^*$. Then the set of *sampling POPs* is $\tilde{\Gamma} = \{x \in I\!R^p : x \in \widetilde{\mu}^*(x)\}$, and a *sampling PCOP* is any curve $\hat{\alpha}$ inside $\tilde{\Gamma}$.

Any implementation of the sampling PCOP method should provide the following elements (the algorithm proposed in (Delicado 2001) and the application we present in Section 5 do it): the interval $I \subset I\!R$ where $\hat{\alpha}$ is defined ($\hat{\alpha}: I \to I\!R^p$); J points $s_1, \dots, s_J$ in I (without loss of generality, it is assumed that points s_j have zero mean); the values of $\hat{\alpha}$ in the points s_j; the principal directions $b^*(\hat{\alpha}(s_j))$; the density estimate of the random variable induced in I by the curve $\hat{\alpha}$:

$$\hat{f}_S(s_j) = C_1 \frac{1}{nh} \sum_{i=1}^{n} K_h \left(|(X_i - x_j)^t b_j| \right);$$

the central point of the data set along the curve: $\mathrm{E}_{PC}(X) = \hat{\alpha}(s = 0)$; the total variance in the normal hyperplane to $\hat{\alpha}$ at s_j: $\widetilde{\phi}(x_j, b_j)$; the total variability of the data along the curve $\hat{\alpha}$:

$$\widehat{TV}_{\hat{\alpha}} = \widehat{\mathrm{Var}}(S) + \int_I \widetilde{\phi}^*(\alpha(s)) \hat{f}_S(s) ds \simeq \widehat{\mathrm{Var}}(S) + \sum_j \widetilde{\phi}(x_j, b_j) \hat{p}_S(s_j);$$

the proportion of total variability explained by the estimated curve: $p_1 = \widehat{\mathrm{Var}}(S)/\widehat{TV}_{\hat{\alpha}}$.

In the framework of principal curves for a random variable X, the total variability of X along the curve α is defined as

$$TV_\alpha(X) = \mathrm{Var}(S) + \int_I TV[X|X \in H(\alpha(s), b^*(\alpha(s)))] f_S(s) ds. \quad (1)$$

If $TV[X|X \in H(\alpha(s), b^*(\alpha(s)))]$ is an appropriate measure for the dispersion of the $(p-1)$-dimensional conditional random vector $(X|X \in H(\alpha(s), b^*(\alpha(s))))$, then $TV_\alpha(X)$ is a good measure of the global variability of X. This is the case when these $(p-1)$-dimensional distributions are elliptical. Nevertheless, nonlinearities could also appear in the conditional distributions $(X|X \in H(\alpha(s), b^*(\alpha(s))))$. Therefore, the total variance should be replaced by a measure of the variability along a nonlinear principal curve.

The set of principal curves recovering nonlinear structures in the $(p-1)$-dimensional spaces, is called the *system of principal curves of second order*. There is not a common second principal curve for the whole distribution of X, but there is a different second principal curve for each point in the first one. Delicado (2001) defines recursively, by induction on the dimension p, systems of principal curves of any order. The basic element for these definitions is the *Generalized Total Variance* (GTV) along a curve in dimension p, that is obtained changing in (1) the total variance by the generalized total variance along principal curves of second order:

$$GTV(X) = \mathrm{Var}(S) + \int_I GTV[X|X \in H(\alpha(s), b^*(\alpha(s)))] f_S(s) ds. \quad (2)$$

The definition of GTV is also recursive. It is possible to determine what proportion of GTV is explained by the first principal curve (p_1), by the system of second principal curves (p_2) and so on. Principal components users are familiar with such a decomposition of total variability.

Figures and tables included in Delicado (2001) were made by a Matlab implementation of the algorithm presented there. That program was able to analyze data only in two or three dimensions. A much more flexible implementation is needed in order to be able to analyze any data set, independently of the data dimension and the number of required principal curves. It seems sensible to write a recursive code, because the recursive nature of the definitions of GTV and higher order PCOPs. Section 5 presents an object oriented application satisfying these requirements.

3 Bandwidth choice

A common characteristic of any nonparametric functional estimation technique is the presence of a *smoothing parameter* that controls the amount of smoothing in the estimated function. For example, the same nonparametric regression estimator can interpolate the observed points for some values of the smoothing parameter, and can produce a linear function for others. The right choice of this parameter is crucial for the estimator performance (see, for instance, Chapter 3 of (Wand & Jones 1995), and Chapter 5 of (Simonoff 1995)).

The smoothing parameter in the PCOP algorithm is the *bandwidth* h, which controls how near to a hyperplane $H(x,b)$ an observed point X_i should be, in order to be attracted by it when $\widetilde{\mu}(x,b)$ and $\widetilde{\phi}(x,b)$ are computed. When a PCOP is estimated, many local nonparametric regressions are done: every time a value of $\widetilde{\mu}(x,b)$ is calculated, we are doing $(p-1)$ simultaneous one-dimensional kernel regression evaluations, one for every dimension of the hyperplane $H(x,b)$. The same happens when $\widetilde{\mathrm{Var}}(x,b)$ is computed for any pair (x,b). In addition to that, a nonparametric density estimator is involved in the estimation process: the density of the random variable S, induced on I by the principal curve, is estimated jointly with the principal curve. Observe that the same bandwidth parameter h is used in regression estimation and in density estimation. So it is not trivial to decide how this parameter should be chosen.

A cross-validation procedure would be adequate, as a bandwidth choice setting. Given a value of h, its fitting would be measured by addition of the cross-validated errors. For an observed point, its cross-validated error is the prediction error of the principal curve fitted with the other observed data. Therefore, as many principal curves as observations might be calculated in order to evaluate a candidate bandwidth value h. The computational cost may be prohibitive.

It seems more practical to base the bandwidth choice in its side related with the density of S estimation. A fast choice rule, based on the normal reference rule, can be implemented. When a kernel density estimator is used (as it is our case), the expression of the value h minimizing the asymptotic expression of the estimator MISE, involves a constant depending on the second derivative of the unknown density function. A parametric model (the normal model) is assumed in order to estimate this constant and to obtain the following expression for the optimal h (see, e.g., (Scott 1992), p. 131, for more details):

$$h = c_K (3/4)^{1/5} \sigma n^{-1/5},$$

where σ is the standard deviation of the random variable of which the density is estimated, and c_K is a constant depending on the specific kernel function used in the estimation (for the normal kernel $c_K = 1$, and for the Epanechnikov kernel is $c_K = 2.214$; we are using the Epanechnikov kernel). In the usual density estimation problem, the available data are observations of the unknown random variable and its sampling standard deviation provides an estimate of σ.

Things are different in the PCOP framework: the estimated density corresponds to the random variable S, that is not directly observable. An estimation of the variance of S is available only after the estimation of the principal curve. Therefore, the standard deviation of S, σ_S, must be estimated in a different way. In Delicado (2001) this estimation was made by an *ad hoc* procedure, based on the ranges of the p observed variables. From this ranges, a measure of the *size* of the data set was derived. The proposed estimation of σ_S was this size multiplied by a constant, the value of which was calibrated by a rule of thumb.

Here we propose an alternative way of estimating σ_S. It is based on the computation of the Minimum Spanning Tree (MST) for the observed data, and on some semi-parametric assumptions for the underlying probability distribution. Our point is that the MST does not change very much when the data set is deformed by the joint action of a torsion and a curvature. Figure 2 (explained in detail below) is an example of that small distortion.

3.1 The Minimum Spanning Tree

Let E be a set of n points in $\mathbb{R}^p$. Each subset of E having two different elements is an *edge*. Let F be a subset of the $n(n-1)/2$ possible edges, and let (E, F) be a *graph*. Let W a set of Card(F) *costs*, each associated with an edge in F. Given a connected graph (E, F), a *Minimum Spanning Tree* (MST) is a subset T of F verifying two properties. First, every two points in E are connected by a path entirely contained in T. Second, the cost of T, defined as the sum of the elements in W corresponding to the elements in

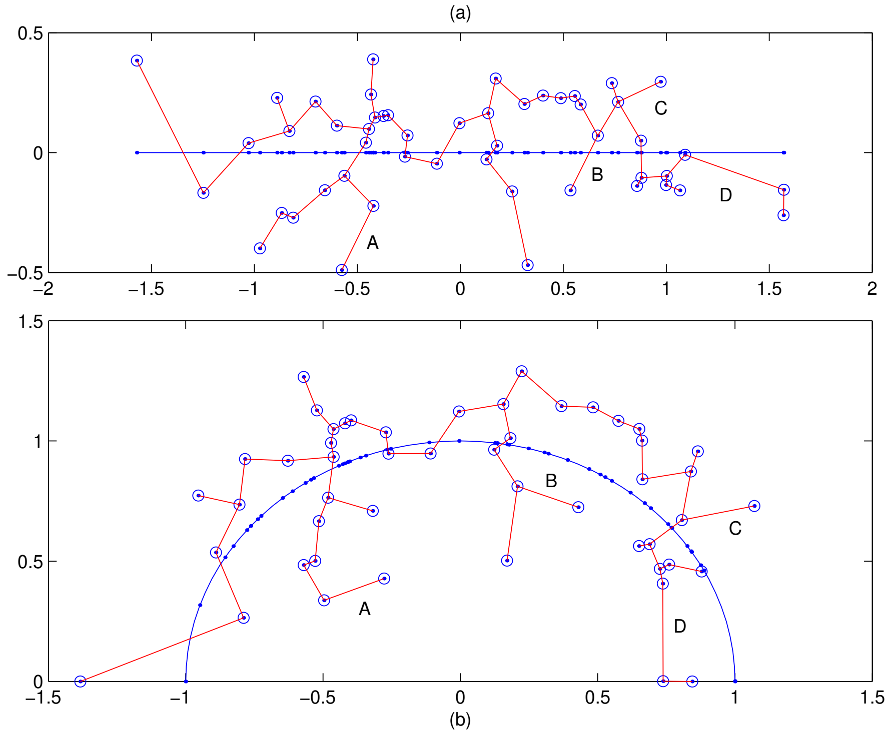

Figure 2: Changes in the MST when a data set is subjected to the action of a curvature. Only edges A, B, C and D change from one MST to the other

T, is lower than or equal to the cost of any other connected graph (E, T'), $T' \subseteq F$, in (E, F). It follows that a MST must be a tree.

A graph is a *random graph* if some of the elements defining it are random. Our interest concerns two ways to introduce randomness in a graph. The first one is as follows. Consider $X_1, \ldots, X_n$ a simple random sample of a random variable X. Let $E = \{X_1, \ldots, X_n\}$ and $W(\{X_i, X_j\}) = ||X_i - X_j||$, $i, j = 1 \ldots n$. The corresponding MST is called *euclidean MST*. Another possibility of having a random graph is to consider a set E of abstract points $\{a_1, \ldots, a_n\}$ and to define $W(\{a_i, a_j\})$, $i, j = 1 \ldots n$, as independent observations of a random variable D. The associated MST is called *independent MST*. When n is large and $D = ||X_1 - X_2||$, X_1, X_2 being i.i.d. copies of X, the probability distributions of the euclidean MST and the independent MST are very similar.

The main information derived from an euclidean MST is their set of $(n-1)$ distances. There exists a relationship between the total length of an euclidean MST (L_{MST}) and the probability distribution of the random variable X generating the p-dimensional observed points. Beardwood, Halton & Hammersley (1959) proved the following asymptotic result.

> *Let $X_1, \ldots, X_n, \ldots$ be a sequence of independent absolutely continuous p-dimensional random variables with compact support and common density $f(x)$, and let $L_{MST,n}$ be the length of the Minimum Spanning Tree determined by $X_1, \ldots, X_n$, then*
>
> $$\lim_{n \to \infty} \frac{L_{MST,n}}{n^{(p-1)/p}} = \beta_{MST}(p) \int_{\mathbb{R}^p} f(x)^{(p-1)/p} dx \text{ almost surely,} \quad (3)$$
>
> *where $\beta_{MST}(p)$ is a positive constant that depends on p but not on f.*

Similar probabilistic results in other combinatorial related problems can be found in Steele (1988), Steele (1993), and Avram & Bertsimas (1993).

The $\beta_{MST}(p)$ exact value determination was an open problem until the work of Avram & Bertsimas (1992). In that paper an exact but intricate expression is given (see Theorem 5, (Avram & Bertsimas 1993)). The authors also provide bounds for the constant: they prove that

$$\frac{\Gamma(1/p)}{p c_p^{1/p}} \leq \beta_{MST}(p) \leq \frac{2^{1/p} \Gamma(1/p)}{p c_p^{1/p}},$$

where $c_p = \pi^{p/2} / \Gamma((p/2) + 1)$ is the volume of the unit ball in dimension p. Moreover, they give the exact value for the corresponding constants in the case of the independent MST:

$$\widetilde{\beta}_{MST}(p) = \frac{1}{p c_p^{1/p}} \sum_{k=1}^{\infty} \frac{\Gamma(k + (1/p) - 1)}{k! k^{(1/p)+1}}.$$

These constants are always between the previous bounds for $\beta_{MST}(p)$ and are a good approximation to them when n is large. The computation of $\widetilde{\beta}_{MST}(p)$ (truncating the series) is much easier than that of the exact value of $\beta_{MST}(p)$.

3.2 Inference in a multivariate normal model through the MST

The significance of equation (3) is that it relates an observable figure (the length of the MST) with the unknown distribution of X_i (through $f(x)$). The following results is derived from equation (3):

$$\frac{L_{MST,n}}{n^{(p-1)/p}\beta_{MST}(p)} \longrightarrow_{n\to\infty} \int_{\mathbb{R}^p} f(x)^{(p-1)/p}dx \text{ almost surely.} \tag{4}$$

If an important feature of f can be found from $I_p(f) = \int_{\mathbb{R}^p} f(x)^{(p-1)/p}dx$, then the MST can help to make inference about it. Let us consider a particular multivariate normal model. The following proposition shows that the MST helps to estimate its main parameter.

Proposition 1 *Let X be a p-dimensional normal random variable with zero mean and diagonal variance matrix with all the elements equal to σ^2 except for the first component, the variance of which is $\sigma_S^2 > \sigma^2$. Let $\tau = \sigma_S^2/(\sigma_S^2 + (p-1)\sigma^2)$, the proportion of total variance explained for the first component of X (it is its first principal component). Let $f(x)$ be the density function of X. Then*

$$I_p(f) = \int_{\mathbb{R}^p} f(x)^{(p-1)/p}dx = \sqrt{2\pi}\left(\frac{p}{p-1}\right)^{p/2}\left(\frac{1-\tau}{p-1}\right)^{\frac{p-1}{2p}}\sigma_S.$$

The proof easily follows taking into account the independence of the components of X and on the fact that

$$\int_{\mathbb{R}} (\phi(u))^{(p-1)/p}du = (2\pi)^{1/(2p)}\left(\frac{p}{p-1}\right)^{1/2}\sigma^{1/p},$$

where ϕ is the density function of the univariate $N(0,\sigma^2)$.

Let $X_1,\ldots,X_n,\ldots$ be a sequence of independent random variables distributed as X in the preceding proposition, and let $L_{MST,n}$ be the length of the corresponding MST. Then

$$\hat{\sigma}_S = \frac{L_{MST,n}}{n^{(p-1)/p}\sqrt{2\pi}\left(\frac{p}{p-1}\right)^{p/2}\left(\frac{1-\tau}{p-1}\right)^{\frac{p-1}{2p}}\tilde{\beta}_{MST}(p)} \tag{5}$$

is a good estimate of σ_S.

Observe that the distribution of X has no compact support, and then the Theorem of Beardwood et al. (1959) is not directly applicable. So we can not write that $\hat{\sigma}_S$ converges to σ_S almost surely. Moreover, $\beta_{MST}(p)$ is replaced by $\tilde{\beta}_{MST}(p)$ because computational simplicity. Even though, the distribution of X can be approximated by a truncated normal distribution X_M as well as it is desired: for every $\varepsilon > 0$, it is possible to have $|I_p(f_X) - I_p(f_{X_M})| < \varepsilon$. Moreover, the distribution of $L_{MST,n}$ should not vary too much when data came from X or X_M. The result of Beardwood et al. (1959) applied to X_M justifies the definition of $\hat{\sigma}_S$ given in (5). Only the value of τ has still to be fixed. A consistent estimator of τ is the proportion of the variance explained by the first principal component of data coming from X.

3.3 Inference in the nonlinearly transformed normal model

Let us assume that a random vector X in $\mathbb{R}^p$ has been defined according to a theoretical model defined from a given curve α as follows. Let $\alpha: I \to \mathbb{R}^p$ be a curve, where $I \subseteq \mathbb{R}$ is an interval in $\mathbb{R}$. We assume that α is of class $\mathcal{C}^r$, $r \geq p$ and that α is parameterized by the arc length (i.e., $||\alpha'(s)|| = 1$). Physically, for $p = 3$ we can think on α as the result of deforming the segment I by both a torsion and a curvature.

At each point $\alpha(s)$ in the curve, an orthonormal coordinate system $A(s) = (a_1(s), \ldots, a_p(s))$ is defined, where $a_1(s) = \alpha'(s)$ and the other vectors a_i are a base of the normal hyperspace to α at $\alpha(s)$. The matrix $A(s)$ can be chosen as a differentiable function of s. Moreover, among others the following properties hold (see, for instance, (Guggenheimer 1977) for the details): the vector $\alpha''(s)$ is orthogonal to $\alpha'(s)$; the norm of the vector $\alpha''(s)$ is the *curvature* of α at $\alpha(s)$; one over the curvature is the *radius of curvature* (the radius of a circumference contained in the plane defined by the point $\alpha(s)$ and the vectors $\alpha'(s)$ and $\alpha''(s)$, passing by $\alpha(s)$ and having the same first and second derivatives as α at $\alpha(s)$); the second vector of $A(s)$, $a_2(s)$, can be chosen proportional to $\alpha''(s)$ and pointing at the *center of curvature* (defined as the center of the previously mentioned circumference); the third vector of $A(s)$ is related with the *torsion* of the curve.

We consider the function χ_α defined by Hastie & Stuetzle (1989) in the proof of their *Proposition 6*: let $H_s = H(\alpha(s), \alpha'(s))$ be the normal hyperplane to the curve α at $\alpha(s)$ and define χ_α as a mapping from $I \times \mathbb{R}^{p-1}$ into $\cup_{s \in I} H_s$ so that $\chi_\alpha(s, y) = \alpha(s) + A(s)(0, y^t)^t$. Thus χ_α put (s, y) in H_s in a differentiable way with respect to s, and it applies to $I \times \mathbb{R}^{p-1}$ the same torsion and curvature that α applies to I so that orthogonality is preserved in some sense.

Let (S, Y) be a random variable on $I \times \mathbb{R}^{p-1}$ with density $f(s, y) = f_S(s) f_{Y|S=s}(y)$

where S and s are in I, having zero conditional expectations $E(Y|S=s)$ for all s. We consider the random variable in $\mathbb{R}^p$ obtained as $X = \chi_\alpha(S,Y)$. The following result gives the density function of X under quite general conditions.

Proposition 2 *Assume that $I = Support(S)$ is a compact interval, and that the distributions $Y|S=s$ have convex compact support contained in the ball $B(0,\rho(s))$, where $\rho(s)$ is the curvature radius of α at the point $\alpha(s)$. Then the function χ_α: $Support(S,Y) \to Support(X)$ is a homeomorfism. Moreover, the density function of X at a given point $x \in Support(X)$ is*

$$f_X(x) = f_S(s) f_{Y|S=s}(y) \frac{1}{1 - y_1/\rho(s)},$$

where (s,y) is the inverse of x by χ_α and y_1 is the first component of y.

The proof of this result can be found in Delicado & Huerta (2002). It is not guaranteed that α is either a PCOP or a HSPC for X, even if X is defined as the curve α plus a noise.

Let us now consider a random vector V with distribution described in Proposition 1. Let S be the first component of V, and let Y be the remaining $(p-1)$ components. Let X be $X = \chi_\alpha(S,Y)$. This random variable follows a semiparametric model: parametric assumptions are made for (S,Y) but not for α.

Proposition 2 does not directly apply to X because the support compactness hypothesis is not verified by the multivariate normal distribution $V = (S,Y)$. Nevertheless, in practice the behavior of (S,Y) is not so different to that of a compact supported random vector (for instance, a univariate random variable $N(0,\sigma^2)$ is only very slightly affected if it is truncated to $[-5\sigma, 5\sigma]$). So it follows that the density of X can be well approximated as follows:

$$f_X(x) \approx f_S(s) f_Y(y) \frac{1}{1 - y_1/\rho(s)} = f_S(s) f_{Y_1}(y_1) \frac{1}{1 - y_1/\rho(s)} f_{Y_{(1)}}(y_{(1)}), \quad (6)$$

where the independence of S and Y has been taken into account, and the subscript (1) indicates that all the components of Y, except for the first one, are being considered.

Under the assumption that the curvature radius $\rho(s)$ is constant and equal to ρ, it follows from (6) that

$$\frac{I_p(f_X)}{I_p(f_V)} \approx \frac{\int_{\mathbb{R}} (1 - y_1/\rho)^{-(p-1)/p} (f_{Y_1}(y_1))^{(p-1)/p} dy_1}{\int_{\mathbb{R}} (f_{Y_1}(y_1))^{(p-1)/p} dy_1}.$$

Table 1 shows the value of this ratio for some combinations of dimension p and relative curvature radius $\gamma = \rho/\sigma$. They have been obtained by numerical integration.

$\gamma = \rho/\sigma$	p 2	3	4	6	10
3	.88064	.84846	.83443	.82116	.81078
5	.96254	.96082	.95972	.95850	.95745
7	.98329	.98184	.98106	.98027	.97963
9	.99028	.98932	.98883	.98833	.98793

Table 1: Values of the ratio $I_p(f_X)/I_p(f_V)$ for some combinations of dimension p and relative curvature radius $\gamma = \rho/\sigma$.

For moderate to big values of the curvature radius ($\rho \geq 5\sigma$) the ratio is near to 1. According to that, the following approximation is derived:

$$I_p(f_X) \approx I_p(f_V).$$

Taken it into account, remembering the relationship between the length of the MST and the value $I_p(f)$ stated in (4), and using Proposition 1, it follows that equation (5) still gives a good estimator of σ_S for data according to X.

An additional problem appears now: when only data X_i are available, it is difficult to estimate the proportion τ of the total variance of V explained by S, that is approximately the proportion of generalized total variance explained by the first principal curve of X. A lower bound for τ is $\hat{\tau}$, the proportion of variance explained by the first principal component of the data X_i, and a higher bound is 1. So we propose to use

$$\hat{\hat{\tau}} = (1 - \lambda)\hat{\tau} + \lambda$$

as an estimate to τ, with $\lambda \in [0, 1]$. The more the distribution of X is nonlinear, the more τ (and λ) is near to 1. As a default value, we put $\lambda = .5$. If prior information about the degree of nonlinearity of X is available, it should be used to accordingly choose λ.

The fact that $I_p(f_X) \approx I_p(f_V)$ and the equation (4) imply that a transformation like χ_α does not affect very much the total length of the MST. In fact, not only the length but all the MST remains almost the same when a χ_α is applied (given that the curvature radius is big enough). The following example illustrates this fact. Figure 2 shows a data set $V_i = (S_i, Y_i)$, $i = 1 \dots 50$, coming from $V = (S, Y)$, and the resulting data set coming from $X_i = \chi_\alpha((S_i, Y_i))$, when $p = 2$ and $\alpha(t) = (\sin(\pi/2 - t), \cos(t))$, $t \in I = [-\pi/2, \pi/2]$. The interval I, the half circumference representing $\alpha(I)$, the points $(S_i, 0)$, and the transformed points $\alpha((S_i, 0))$ are also drawn. The standard deviation of Y is $\sigma = 0.2$ and the curvature radius is constant and equal to $\rho = 1$, so the value of the relative curvature is $\gamma = \rho/\sigma = 5$. The MST calculated from both data sets are shown. It can be seen that only slight changes affect the MST: only edges A, B, C and D change from one MST to the other; the length of

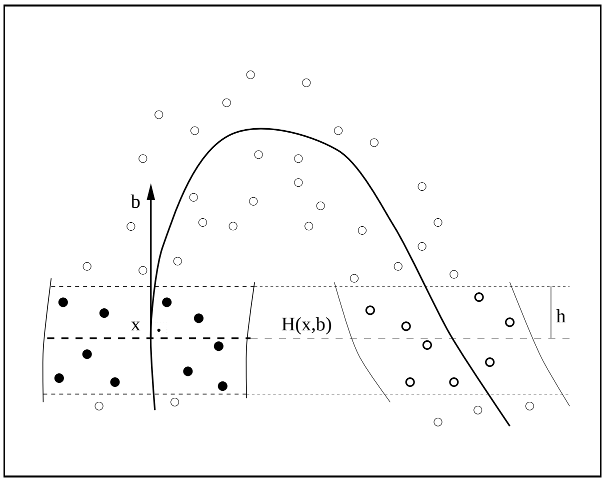

Figure 3: The same hyperplane $H(x, b)$ cut the data set in two different areas and then two clusters are defined.

any pair of corresponding edges in both MSTs are almost the same; the sum of lengths is 7.73 and 7.66 for the MST from V_i and from X_i, respectively.

4 Clusters detection

Sometimes $H(x, b)$ cut the data set in more than one area, as it is illustrated in Figure 3. Then the projected data cloud has more than one cluster. In Delicado (2001) this problem was approached as follows. A modified single linkage clustering method was used to evaluate the distances between pairs of adjacent points in the minimum path going from a given point x to any other in $H(x, b)$. Distances too large (the outliers in the distribution of distances) were considered as separation between different clusters in $H(x, b)$. Observe that the critical $d_{\max}$ value for considering a distance as an outlier depends on (x, b): $d_{\max} = d_{\max}(x, b)$.

According to Delicado (2001), the clustering method should be run every time a new hyperplane $H(x, b)$ was considered, that is, every time a new direction b was validate as a principal direction of any new candidate x for being a POP. Here we propose to fix a common critical distance $d_{\max}$ for detecting clusters in any hyperplane $H(x, b)$. We base the choice of $d_{\max}$ on the MST built on the whole data set. This common $d_{\max}$ provides the data set with a structure where each data point has a relative position.

Given a data set $\{x_1, \ldots, x_n\}$ in $I\!R^p$, let $\{d_j, j = 1, \ldots, n-1\}$ be the set of euclidean distances between adjacent points in the MST associated with $\{x_1, \ldots, x_n\}$. Compute the median m and the quartiles Q_1 and Q_3 of the data set $\{d_1, \ldots, d_{n-1}\}$. Define the *critical distance* as $d_{\max} = Q_3 + C(Q_3 - Q_1)$, where C is a positive number (usually, $C = 3$). Observe that extreme outlying distances d_j are identified as if a box-plot were used.

Let us note that the set of distances in the MST is not the set of the $(n-1)$ shorter inter points distances. Even though, this set contains a good sample of small distances. Big distances d_j are only needed in the MST to connect an outlier point to the data cloud, or to connect two or more well separated clusters (see (Rohlf 1975), and (Caroni & Prescott 1995), for information about the use of the set of MST distances in the detection of multivariate outliers).

We use $d_{\max}$ to detect clusters in $H(x, b)$ as follows. We define the *cluster given by the hyperplane $H(x, b)$ and the bandwidth h*, $\{y_1, \ldots, y_m\}$, as the set of data points y_j verifying two properties. First, y_j is closer than h to $H(x, b)$ (i.e., $|(x - y_j)^t b| \leq h$). And second, there exists a polygonal line from x to y_j with vertices in $\{x, y_1, \ldots, y_m\}$ and segments shorter than $d_{\max}$. Delicado & Huerta (2002) specify the way this definition has been implemented.

5 PCOP object oriented application

An object oriented application of the PCOP method has been developed. Its implementation exploits the recursive nature of the definitions of GTV and higher order PCOPs. Moreover, programming with abstract data types makes easy to represent the system of successive order principal curves and their relationships as a hierarchical structure of objects. The complete description of main classes and their relationships can be found in Delicado & Huerta (2002).

An implementation of the application is available at `http://www-eio.upc.es/~delicado/PCOP/`. There you can find the application C++ source code, a Matlab function file used to obtain graphical outputs, and some examples of them. Updated versions will be published at the same site.

6 Some examples

In this section we use the PCOP application presented at Section 5 to find principal curves in both simulated and real data.

6.1 Simulated data

We consider some experiments on synthetic data where the generating curve is known. We study the models presented by Kégl et al. (2000) (some of them also appeared in (Hastie & Stuetzle 1989)). Figure 4 shows the results for six of these data configurations that represent well the findings in the whole experiment. For each data set three algorithms with public implementations are used to find the principal curve: the PCOP algorithm implemented as in Section 5, the HSPC ((Hastie & Stuetzle 1989)) algorithm running as the `princurve` package in R, and the polygonal principal curve algorithm ((Kégl et al. 2000)) implemented as a JAVA application. Data were generated as a uniform over a curve in $\mathbb{R}^2$ plus a noise distributed as σ times a bivariate standard normal. The generating curve and the data also appear in the graphics.

In the first example the generating curve is a circle with radius equal to 1 and σ is equal to 0.2. The second one is a half of this circle. In both cases the sample size is 100. At these two examples the three algorithms behave similarly. When the noise variance is reduced (third example: circle with small noise, $\sigma = 0.01$ and 100 data points) PCOP and polygonal line algorithms are almost unbiased. It should be noted that in the circular cases the PCOP automatically detects that the curve is closed.

The fourth example (S-shape: two linked half circles of unit radii, $\sigma = .2$) presents a situation with large sample size (10.000 points, but only 2.000 are represented). As it was noted by Kégl et al. (2000), the polygonal line algorithm is much less biased than the HSPC algorithm. The PCOP is placed in the middle of the other, and it is closer to the polygonal principal curve at the extremes of the curve.

Kégl et al. (2000) described two difficult situations: zig-zagging curves and spiral curves (see that paper for for the precise definitions of these data configuration). For the zig-zag curves they consider three cases: 4, 8, and 16 segments. Here we only show the first one, but the results for the other are very similar. The curves given by the PCOP and polygonal line algorithms are almost identical, and they recover the generating curve very well, as it can be seen in Figure 4. The results from the HSPC algorithm are less satisfactory and they are worse when the number of segments increases.

The spiral curves presented in Kégl et al. (2000) are more difficult when the number of loops is greater. They consider spirals with one and a half, two, and three loops. Here we only present the intermediate case. Figure 4 shows that the polygonal principal curve is almost indistinguishable from the generating curve, and that the HSPC fails to recover the spiral shape. With regard to the PCOP algorithm, its estimation closely follows the generating curve except for the innermost loop. There the PCOP algorithm is not able to detect the end of the curve and it looks for more points until other part of the estimated curve is reached. Then the algorithm stops.

As a general conclusion, it can be said that the PCOP and the polygonal line algorithms behave quite similarly, and that both perform better than the HSPC algorithm.

6.2 A real data example

We illustrate the use of principal curves for data analysis with the following real data example. We consider the data matrix where individual are the countries in the world, and observations are their population pyramid. To be specific, for each of 223 countries we have their relative frequency table crossing sex and age, discretized in 17 intervals of five years (the last one corresponds to people older than 80 years). So the data set has 223 points in a 34-dimensional space.

Data were extracted from the International Data Base (IDB), produced by the International Programs Center, U.S. Census Bureau ((IPC 2000)). The IDB is a computerized data bank containing statistical tables of demographic and socio-economic data for 227 countries of the world (data for four countries were incomplete). On-line access to the IDB is available at `http://www.census.gov/ ipc/www/idbnew.html`.

As a first step, we reduced the dimensionality of the data by principal component analysis. Only the first four principal component have variances over the average variances value. They jointly explain the 94.57% of the total variance (80.42%, 6.61%, 4.64%, and 2.90%, respectively). The values of the first and second principal components are represented as a scatter plot in the upper panel of Figure 5. The joint distribution is clearly non-elliptical. This curved pattern recalls the *horseshoe effect*, a common feature of correspondence analysis maps (see, e.g., (Greenacre 1993)). In fact, our principal component analysis on relative frequency tables is closely related to the correspondence analysis of the same data.

In order to better deal with this nonlinear relationship, we find the first principal curve of the 4-dimensional data set containing the score of each country in the four principal component axis. We assume that the conditional distributions on the hyperplanes orthogonal to this curve are elliptical. So we do not look for principal curves of higher order.

The fitted PCOP is shown in the upper panel of Figure 5. This curve explains the 91.62% of the generalized total variance of the data set (see equation (2)). The density function estimated over the PCOP is almost constant, indicating that countries are uniformly distributed along the curve.

In order to give an interpretation for the PCOP we mark seven equispaced

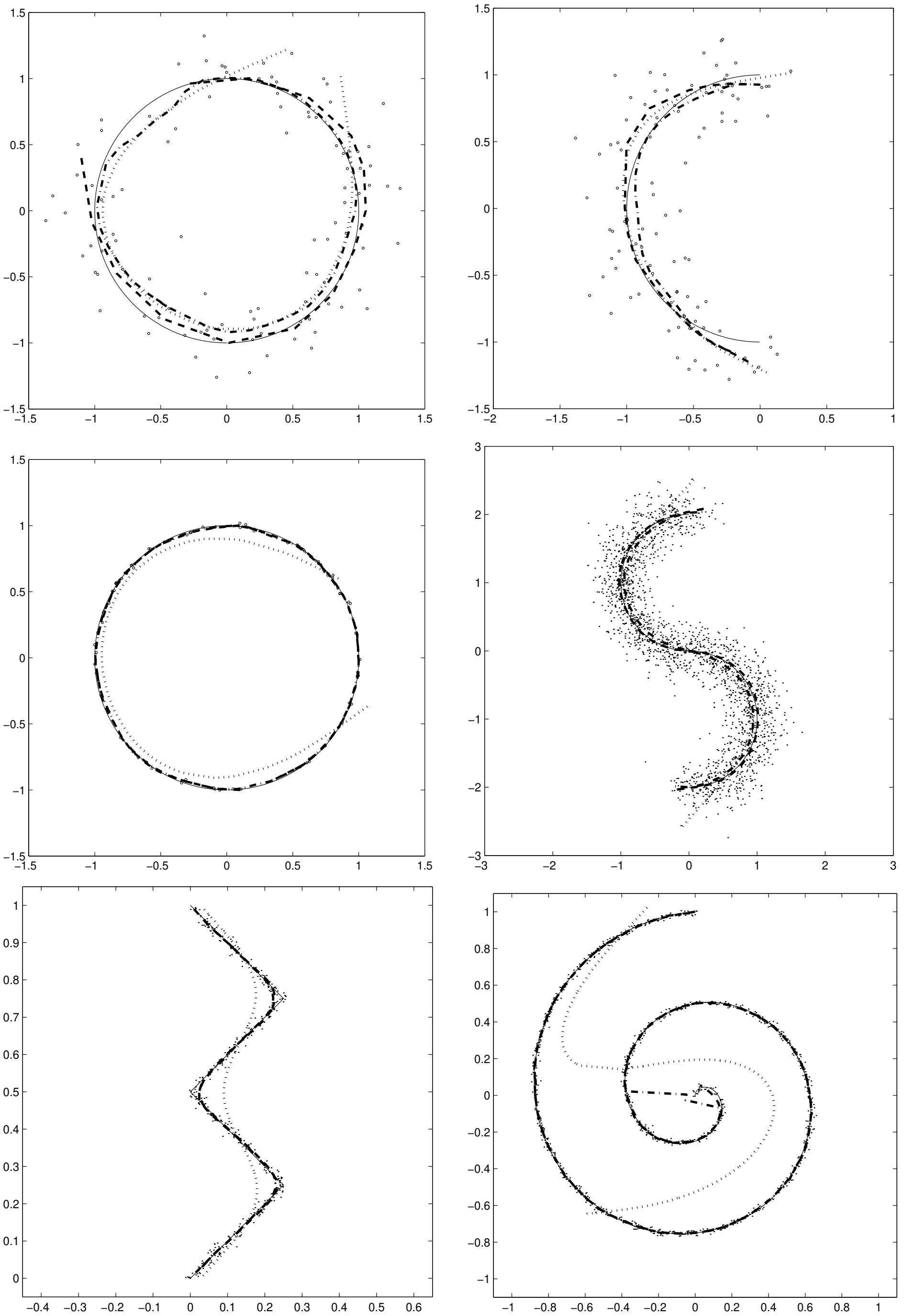

Figure 4: Some simulated data. The generating curve (solid line), the HSPC (dotted line), the polygonal principal curve (dashed line), and the PCOP (dash-dot line) are represented, as well as the data points.

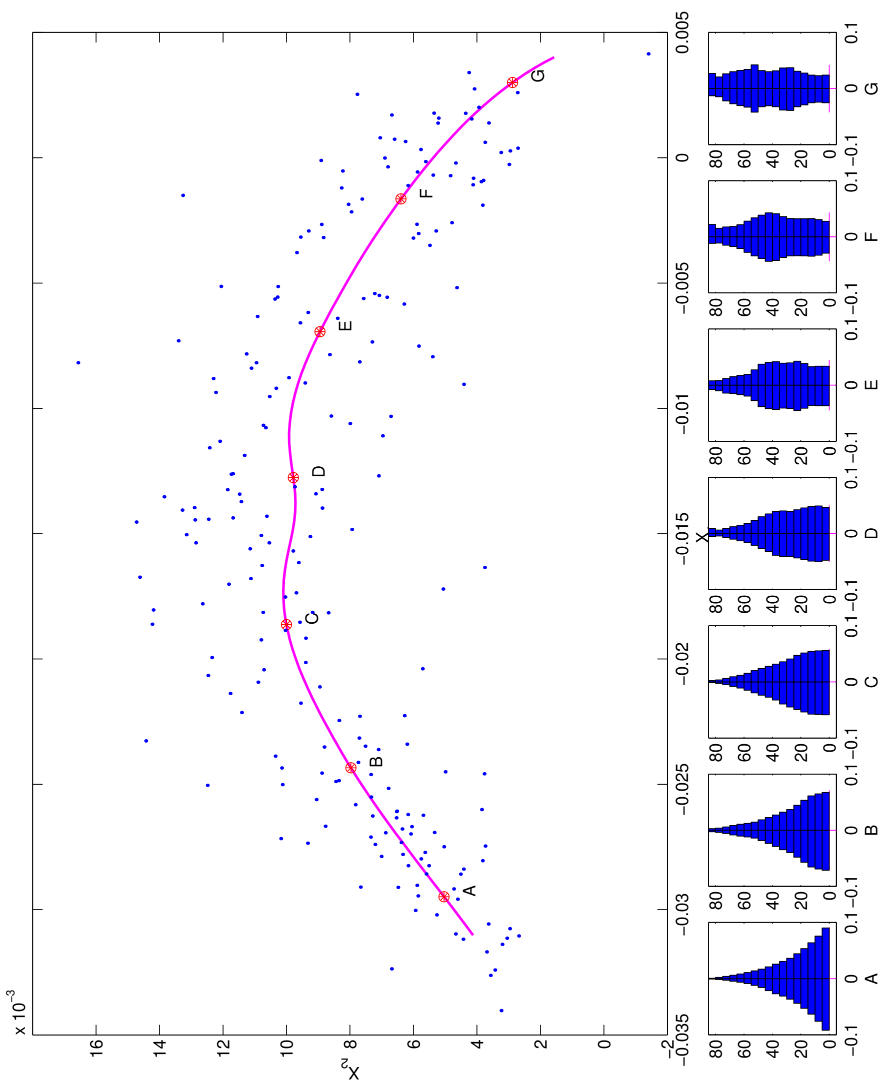

Figure 5: Population pyramids data set. The PCOP for the first four principal components has been projected onto the first and second PC plane. Population pyramids A to G in the lower row of graphics correspond to Mauritania, Pakistan, Peru, The Bahamas, Taiwan, Canada, and Japan, respectively. Each pyramid is associated with one of the seven equispaced points marked over the PCOP.

points over it (labeled as A to G in the graphic) and we represent the population pyramid of their nearest country at the bottom of the figure. Pyramids A and B are typical for underdeveloped countries (they corresponds to Mauritania and Pakistan) where the proportion of people of a given age rapidly decreases with the age. Usually these countries have low expected life and positive population growth rate. The pyramids gradually change until having barrel form (E, Taiwan, and F, Canada) and then being rectangular (G, Japan). These forms corresponds to well developed countries with a big proportion of elderly people, and negative population growth rate. It can be concluded that the principal curve goes from underdeveloped young countries to developed elderly ones.

In this example the main contribution of the first principal curve is not providing much more information than the first principal component (in fact the interpretations given to one and another are very similar). The point of interest is that PCOP analysis shows that the second principal component has not valuable information. It appears to deal with the data configuration nonlinearity. The PCOP is able to detect this problem and to manage it.

References

Avram, F. & Bertsimas, D. (1992), 'The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach', *The Annals of Applied Probability* **2**(1), 113–130.

Avram, F. & Bertsimas, D. (1993), 'On central limit theorems in geometrical probability', *The Annals of Applied Probability* **3**(4), 1033–1046.

Banfield, J. D. & Raftery, A. E. (1992), 'Ice floe identification in satellite images using mathematical morphology and clustering about principal curves', *Journal of the American Statistical Association* **87**, 7–16.

Beardwood, J., Halton, H. J. & Hammersley, J. M. (1959), 'The shortest path through many points', *Proc. Cambridge Philos. Soc.* **55**, 299–327.

Bishop, C. M. & Tipping, M. E. (1998), 'A hierarchical latent variable model for data visualization', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 281–293.

Bishop, C., Svensén, M. & Williams, C. K. I. (1998), 'GTM: The generative topographic mapping', *Neural Computation* **10**(1), 215–234.

Caroni, C. & Prescott, P. (1995), 'On Rohlf's method for the detection of outliers in multivariate data', *Journal of Multivariate Analysis* **52**, 295–307.

Chang, K. & Ghosh, J. (2001), 'A unified model for probabilistic principal surfaces', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, 22–41.

Delicado, P. (2001), 'Another look at principal curves and surfaces', *Journal of Multivariate Analysis* **77**, 84–116.

Delicado, P. & Huerta, M. (2002), Principal Curves of Oriented Points: Theoretical and computational improvements, Technical Report DR 2002/06, EIO, UPC. (Available at `http://www-eio.upc.es/~delicado/PCOP/`).

Dong, D. & McAvoy, T. J. (1996), 'Nonlinear principal component analysis based on principal curves and neural networks', *Computers Chem. Engng.* **20**, 65–78.

Friedman, J. H. (1991), 'Multivariate adaptive regression splines', *The Annals of Statistics* **19**, 1–141. (With discussion).

Greenacre, M. J. (1993), *Correspondence analysis in practice*, Academic Press.

Guggenheimer, H. W. (1977), *Differential Geometry*, Dover Publications.

Hastie, T. (1984), Principal curves and surfaces, Laboratory for Computational Statistics Technical Report 11, Stanford University, Dept. of Statistics.

Hastie, T. & Stuetzle, W. (1989), 'Principal curves', *Journal of the American Statistical Association* **84**, 502–516.

IPC (2000), '*International Data Base*', International Programs Center, U.S. Census Bureau. `http://www.census.gov/ipc/www/idbnew.html`.

Kégl, B., Krzyżak, A., Linder, T. & Zeger, K. (2000), 'Learning and design of principal curves', *IEEE Trans. Pattern Analysis and Machine Intelligence* **22**(3), 281–297.

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

Mulier, F. & Cherkassky, V. (1995), 'Self-organization as an iterative kernel smoothing process', *Neural Computation* **7**(6), 1165–1177.

Rohlf, F. J. (1975), 'Generalization of the gap test for the detection of multivariate outliers', *Biometrics* **31**, 93–102.

Sandilya, S. & Kulkarni, S. R. (2000), Principal curves with bounded turn, *in* 'IEEE International Symposium on Information Theory', p. 321.

Scott, D. (1992), *Multivariate Density Estimation*, Wiley.

Simonoff, J. (1995), *Smoothing Methods in Statistics*, Springer, New York.

Smola, A. J., Mika, S., Schölkopf, B. & Williamson, R. C. (2001), 'Regularized principal manifolds', *Journal of Machine Learning Research* **1**, 179–209.

Steele, M. (1988), 'Growth rates of euclidean minimal spanning trees with power weighted edges', *The Annals of Probability* **16**(4), 1767–1787.

Steele, M. (1993), 'Probability and problems in euclidean combinatorial optimization', *Statistical Science* **8**(1), 48–56.

Tan, S. & Mavarovouniotis, M. L. (1995), 'Reducing data dimensionality through optimizing neural network inputs', *AIChE Journal* **41**, 1471–1480.

Tarpey, T. & Flury, B. (1996), 'Self-consistency: A fundamental concept in statistics', *Statistical Science* **11**, 229–243.

Tibshirani, R. J. (1992), 'Principal curves revisited', *Statistics and Computing* **2**, 183–190.

Tipping, M. E. & Bishop, C. M. (1999), 'Probabilistic principal component analysis', *Journal of the Royal Statistical Society, Series B, Methodological* **61**, 611–622.

Verbeek, J., Vlassis, N. & Krse, B. (2002), 'A k-segments algorithm for finding principal curves', *Pattern Recognition Letters* **23**, 1009–1017.

Wand, M. & Jones, M. (1995), *Kernel Smoothing*, Chapman and Hall, London.