Principal Curves and Principal Oriented Points

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 29, 1998

*The author is very grateful to Wilfredo Leiva-Maldonado for helpful conversations, suggestions and theoretical support. Comments from other members of the Department were also very useful. This work was partially supported by the Spanish DGES grant PB96-0300.

Abstract

Principal curves have been defined (Hastie and Stuetzle 1989) as smooth curves passing through the middle of a multidimensional data set. They are nonlinear generalizations of the first principal component, a charazterization of which is the basis for the principal curves definition.

In this paper we propose an alternative approach based on a different property of principal components. Consider a point in the space where a multivariate normal is defined and, for each hyperplane containing that point, compute the total variance of the normal distribution conditioned to belong to that hyperplane. Choose now the hyperplane minimizing this conditional total variance and look for the corresponding conditional mean. The first principal component of the original distribution passes by this conditional mean and it is orthogonal to that hyperplane. This property is easily generalized to data sets with nonlinear structure. Repeating the search from different starting points, many points analogous to conditional means are found. We call them *principal oriented points*. When a one-dimensional curve runs the set of these special points it is called *principal curve of oriented points*. Successive principal curves are recursively defined from a generalization of the total variance.

Key Words: Fixed points; Generalized Total Variance; nonlinear multivariate analysis; principal components; smoothing techniques.
JEL: C10; C14.

Contents

1 Introduction

Consider a multivariate random variable X in $\mathbb{R}^p$ with density function f_X and a random sample from X, namely $X_1, \ldots, X_n$. When the distribution of X is nearly elliptical the first principal component is a good way to summarize the information of that sample. As a data analysis tool, the first principal component is the straight line that "better" passes through the cloud of data. As a distributional concept, the first principal component is roughly speaking the straight line that runs through the highest density areas of X in $\mathbb{R}^p$.

Hastie and Stuetzle (1989) introduce principal curves as an extension of the first principal component to distributions with nonlinear structure. They define the principal curve of a random variable as a one-dimensional parameterized curve $\{x \in \mathbb{R}^p : x = \alpha(s), s \in I \subseteq \mathbb{R}, \alpha \text{ differentiable}\}$ having a property of *self-consistency* in the following sense: for every point $\alpha(s)$ in the curve, the conditional mean of X given that $\alpha(s)$ is the closest point to X in the curve, is just the original point $\alpha(s)$. They give an appropriate definition of the principal curve of a multivariate data set and they present nonparametric algorithms to obtain it.

The main objective of this paper is to give an alternative definition of principal curves. It is based on the generalization of a local property of principal components for a multivariate normal distribution X: the total variance of the conditional distribution of X given that X belongs to a hyperplane, is minimum when the hyperplane is orthogonal to the first principal component. The generalization of this result to nonlinear distribution leads us to define *principal oriented points* (the fixed points of a certain function from $\mathbb{R}^p$ to itself, and *principal curves of oriented points*. Our approach to principal curves suggests a generalization of the concept of *total variance*. This extension provides a good measure of the dispersion of a random variable distributed around a nonlinear principal curve, and it permits *local* second (and higher order) principal curves to be recursively defined.

Before starting to introduce the new elements, a short review of related work is helpful to set the present paper in a context. In the last forty years many works have appeared in the statistical literature proposing extensions of the simple and powerful concept of principal components to more general setting than the multivariate linear world. Starting with an observed data matrix $X_{n \times p}$, some proposals look for a non-observable data matrix $Y_{n \times r}$, $r < p$, such that the configuration of the n points in $\mathbb{R}^r$ and that in $\mathbb{R}^p$

are as much *similar* as possible. The particular definition of *similarity* leads to the works of Shepard and Carroll (1966) or Srivastava (1972) (that pays special attention on the selection of r), to multidimensional scaling and to techniques compiled in Gifi (1990), among others.

Other authors propose to increase the dimension of the data matrix including known functions of the observed data, and then to apply the usual principal components technique on the enlarged matrix for detecting and describing nonlinear relations among the data. Gnanadesikan and Wilk (1966) use powers and crossed products of the original data.

A different approach is developed in Etezadi-Amoli and McDonald (1983) and Yohai, Ackermann, and Haigh (1985). A nonlinear factorial model is proposed: $X_{n\times p} = \Phi(Y_{n\times r}) + \text{noise}$, where r and Y are unknown and Φ is assumed to belongs to a parametric family of functions. Usually r is fixed in 1 or 2, and an alternating optimization procedure is repeated until convergence: for a fixed Y, the best parameters for Φ are obtained; then for these parameters, the best configuration $Y_{n\times r}$ is taken. In Etezadi-Amoli and McDonald (1983), the parametric family of functions is the second degree polynomials in Y. Yohai, Ackermann, and Haigh (1985) choose $r = 1$ and use the nondecreasing segments of quadratic parabolae as family of functions. The residuals of these models are used to find a second principal component. At this step, the class of functions may contain any monotonic segments of parabolae.

Koyak (1987) looks the transformation $\Psi: I\!R^p \to I\!R^r$, $r < p$ such that the r-dimensional transformed data matrix $\Psi(X)$ has a good linear representation. No parametric form is assumed for Ψ. The proposed algorithm of estimation is based on nonparametric smoothers.

The work of Hastie and Stuetzle (1989) opens a new way to look at the problem. Among the above mentioned references, the most related with Hastie and Stuetzle (1989) could be Etezadi-Amoli and McDonald (1983) and Yohai, Ackermann, and Haigh (1985). The main difference with these papers is that now no parametric assumptions about the *link* function Φ are made. The values $r = 1$ (*principal curves*) and $r = 2$ (*principal surfaces*) are used.

The principal curves defined by Hastie and Stuetzle (1989) (hereafter, HSPC) pass through the "middle" of the distribution and they are self-consistent (in the same sense as Tarpey and Flury (1996) define self-consistent set of points for a random variable). In Section 2 we offer a rigorous defini-

tion of HSPCs. It is not guaranteed that a HSPC does exist. The concept of *principal surface* is analogous.

In 1992 two works directly related with Hastie and Stuetzle (1989) appeared. Banfield and Raftery (1992) is mainly applied. It includes a modification of the Hastie and Stuetzle (1989) algorithm in order to reduce the bias of the original procedure. They report empirical examples of estimation of closed principal curves where the bias reduction they get is important. Tibshirani (1992) is rather theoretical. The main point of that paper is to provide a new definition of principal curve such that if X is transformed from a one-dimensional random variable S by α plus noise, then α is a principal curve of X (HSPC does not have this property). The existence of Tibshirani's principal curve for a given random variable X is not guaranteed. The author proposes a method for the estimation of α based on the EM algorithm, under the assumption that noise distributions are normal. Thus, this approach leaves the nonparametric spirit of Hastie and Stuetzle (1989) methodology.

LeBlanc and Tibshirani (1994) faces again the problem as in Etezadi-Amoli and McDonald (1983) and Yohai, Ackermann, and Haigh (1985), but now the family of link functions is flexible enough to consider it as nonparametric. Multivariate adaptive regression splines (Friedman 1991) are used to develop procedures that allow the successful estimation of principal curves and surfaces defined as in Hastie and Stuetzle (1989). Recently, the paper Duchamp and Stuetzle (1996) states that the principal curves defined as in Hastie and Stuetzle (1989) are critical points of the expected squared distance from the data, but they are not extremal of this functional. An application of HSPC in the clustering context is made by Stanford and Raftery (1997).

Kégl, Krzyżak, Linder, and Zeger (1997) introduces the concept of principal curve with a fixed length. They prove the existence and uniqueness of that curve for theoretical distributions and propose an algorithm to implement their proposals. So this paper means the first proof of existence of principal curves, having however the limitation that the length of the curve has to be previously fixed.

In the most recent years, many related work is being done in neural networks literature: Mulier and Cherkassky (1995), Tan and Mavarovouniotis (1995), Dong and McAvoy (1996), Bishop, Svensén, and Williams (1996), Bishop, Svensén, and Williams (1997), among others.

The present paper is close to Hastie and Stuetzle (1989) in spirit: no parametric assumptions are made, smoothing techniques are used in the pro-

posed algorithms for estimation, and the conceptual idea of *principal curve* we have in mind is very similar to that introduced by Hastie and Stuetzle (1989). Nevertheless, there exist significant differences in definitions and in the implemented algorithms. On the other hand, our approach to second and higher order principal curves does not recall directly any of the previously cited works.

The structure of the rest of the paper is as follows. Section 2 deals with our proposal of definition for principal oriented points and principal curves of oriented points, as distributional concepts. The definition of sample counterparts is postponed to section 3, where algorithmic aspects and some examples are examined. The generalization of the total variance and the definitions of local higher order principal curves are the core of section 4. Section 5 contains some concluding remarks. The proofs of the results appearing in the paper are postponed to a final Appendix.

2 Alternative definition of distributional principal curves

The definition of principal curves given by Hastie and Stuetzle (1989) is based on the generalization to a nonlinear context of a known property of the first principal component: the conditional mean of an elliptical random variable given that the variable is in the orthogonal hyperplane to the first principal component, is the point of intersection of that hyperplane and the first principal component.

The rigorous definition is as follows. Consider a p-dimensional random variable X with density function f_X. A parameterized curve α in $\mathbb{R}^p$

$$\alpha: I \subseteq \mathbb{R} \to \mathbb{R}^p$$

is said to be *parameterized by the arc length* when the length of the curve from $\alpha(s_1)$ to $\alpha(s_2)$ is $|s_2 - s_1|$. This is equivalent to be *unit-speed* paprameterized when α is differentiable (i.e., $|\alpha'(s)| = 1$ for all s). Hastie and Stuetzle 1989 consider a differentiable curve α that does not intersect itself (if $s_1 \neq s_2$ then $\alpha(s_1) \neq \alpha(s_2)$), that is a unit-speed curve for all $s \in I$) and that has finite length in finite balls. They define the projection index $s_\alpha: \mathbb{R}^p \to \mathbb{R}$ as

$$s_\alpha(x) = \sup_s \{s : \|x - \alpha(s)\| = \inf_t \|x - \alpha(t)\|\},$$

and therefore $\alpha(s_\alpha(x))$ is the closest point to x in the curve α. They define a principal curve (and we denote it as HSPC) as any curve α that is self-consistent in the following sense.

Definition 1 (HSPC, Hastie and Stuetzle (1989))
Given the random variable X in $\mathbb{R}^p$, the curve α is called self-consistent for X, or principal curve of X in the sense of Hastie and Stuetzle (HSPC) *if*

$$E(X|s_\alpha(X) = s) = \alpha(s).$$

Hastie and Stuetzle (1989) prove that when the HSPC is linear, then it is the first principal component. So the HSPC is a generalization of the first principal component.

In the present paper we generalize another well known property of the first principal component when the underlying distribution is normal: the projection of the normal random variable over the orthogonal hyperplane to the first principal component has the lowest total variance among all the projected variables over any hyperplane. Moreover, this is true not only for the marginal distribution of the projected variable but also for its conditional distribution given any value of the first principal component. The following proposition establishes that property.

Proposition 1 *Consider $X \sim N_p(\mu, \Sigma)$. Take $x_0 \in \mathbb{R}^p$ and for each $b \in \mathbb{R}^p$ such that $b^t \Sigma b = 1$, let $H(x_0, b) = \{x \in \mathbb{R}^p : (x - x_0)^t b = 0\}$ the orthogonal hyperplane to b passing by x_0. Consider the problems*

$$(\mathbf{P1}) \quad \min_{b: b^t \Sigma b = 1} \{TV(X|X \in H(x_0, b))\},$$

where for any random variable Y, $TV(Y) = Trace(Var(Y))$ is the <u>total variance</u> of Y, and

$$(\mathbf{P2}) \quad \max_{h: h^t h = 1} \{Var(h^t X)\}.$$

Then the solutions to both optimization problems are, respectively,

$$b^* = \frac{1}{\lambda_1^{1/2}} h^* \quad \text{and} \quad h^* = v_1,$$

where λ_1 is the largest eigenvalue of Σ and v_1 the corresponding unit length eigenvector. Moreover, $E(X|X \in H(x_0, b^*)) = \mu + s_0 v_1$, with $s_0 = (x_0 - \mu)^t v_1$.

The straight line $\{\mu + s v_1 : s \in \mathbb{R}\}$ is the first principal component of X. The corollary bellow characterizes the points of this line.

Corollary 1 *Consider $X \sim N_p(\mu, \Sigma)$. A point $x_0 \in \mathbb{R}^p$ belongs to the first principal component line $\{\mu + sv_1 : s \in \mathbb{R}\}$ if and only if x_0 is a fixed point of the function G_X defined as*

$$G_X(x) = E(X|X \in H(x, b_X(x))),$$

where

$$b_X(x) = \arg \min_{b:b^t b=1} TV(X|X \in H(x,b))$$

Corollary 1 characterizes points on the first principal component as the fixed points of a function going from $\mathbb{R}^p$ to $\mathbb{R}^p$. Observe that only local information around a point x_0 is needed to verify whether x_0 is a such fixed point or not. Proposition 1 provides a mechanism to find points in the first principal component: the iteration of the function G_X leads (in one step) from an arbitrary point x_0 to a point on the first principal component line. In the next subsection we exploit this mechanism in order to generalize the first principal component to non-normal distributions.

2.1 Main definitions

Let X be a p-dimensional random variable with density function f_X and finite second moments. Consider $b \in S^{p-1} = \{w \in \mathbb{R}^p : ||w|| = 1\}$ and $x \in \mathbb{R}^p$. We call $H(x,b)$ the hyperplane orthogonal to b passing by x: $H(x,b) = \{y \in \mathbb{R}^p : (y-x)^t b = 0\}$.

Given $b \in S^{p-1}$, it is possible to find vectors $b_2(b), \ldots, b_p(b)$ such that $T(b) = (b, b_2(b), \ldots, b_p(b))$ is an orthonormal base of $\mathbb{R}^p$. We define $b_\perp$ as the $(p \times (p-1))$ matrix $(b_2(b), \ldots, b_p(b))$.

With these definitions we have

$$\begin{aligned} E(X|X \in H(x,b)) &= \frac{\int_{\mathbb{R}^{p-1}} (x + b_\perp v) f_X(x + b_\perp v) dv}{\int_{\mathbb{R}^{p-1}} f_X(x + b_\perp v) dv}, \text{ and} \\ TV(X|X \in H(x,b)) &= \frac{\int_{\mathbb{R}^{p-1}} v^t v f_X(x + b_\perp v) dv}{\int_{\mathbb{R}^{p-1}} f_X(x + b_\perp v) dv} - \\ & \quad E(X|X \in H(x,b))^t E(X|X \in H(x,b)), \end{aligned}$$

for any x and b such that $\int_{\mathbb{R}^{p-1}} f_X(x + b_\perp v) dv > 0$.

Observe that $E(X|X \in H(x,b))$ and $TV(X|X \in H(x,b))$ do not depend on the choice of $b_\perp$, but only on x and b. Therefore the following functions are well defined:

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

and

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

The following result determines the smoothness of μ and ϕ in accordance with the smoothness of f_X. It is a direct consequence of Fubini's Theorem (see, for instance, Corwin and Szczarba 1979, p. 524).

Proposition 2 *If f_X is of class $\mathcal{C}^r$ at x and $\int_{\mathbb{R}^{p-1}} f_X(x+b_\perp v)dv$ is not equal to zero at (x,b), then μ and ϕ are of class $\mathcal{C}^r$ at (x,b).*

Observe that $\mu(x,b) = \mu(x,-b)$ and the same thing happens for ϕ. So we define in S^{p-1} the equivalence relation $\equiv$ by,

$$v \equiv w \iff v = w \text{ or } v = -w.$$

Let $S^{p-1}_{\equiv}$ be the quotient set. From now on, we write S^{p-1} instead of $S^{p-1}_{\equiv}$ even if we want to refer to the quotient set.

Now we extend the functions b_X and G_X we introduced for the normal case. Remember that, when X is normal, for all x, the function $b_X(x)$ returns the direction of the first principal component, and G_X gives a point in that line.

Definition 2 *We define the correspondence $b^*: \mathbb{R}^p \to S^{p-1}$ by*

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

We say that each element of $b^*(x)$ is a principal direction *of x. We also define the correspondence $\mu^*: \mathbb{R}^p \to \mathbb{R}^p$ as*

$$\mu^*(x) = \mu(x, b^*(x)).$$

The infimum of $\phi(x,b)$ over b is achieved because $TV(X)$ is finite and because S^{p-1} is compact. Let $\phi^*(x) = \phi(x,b), b \in b^*(x)$, be the minimum value.

The next result summarizes the smoothness properties of b^*, μ^* and ϕ^*.

Proposition 3 *If (x,b) verifies the hypothesis of Proposition 2 for all $b \in b^*(x)$, the function $\phi^*: \mathbb{R}^p \to \mathbb{R}$ is of class $\mathcal{C}^r$ at x. Moreover, if $r \geq 2$ and b^* is a function in a neighborhood of x (i.e., $\#\{b^*(y)\} = 1$ for y near x), then μ^* is also a function in a neighborhood of x, and μ^* and b^* are of class $\mathcal{C}^{r-1}$ at x.*

A comment on the adequacy of conditioning on $H(x,b)$ is in order. As we are interested in defining valid concepts for non-elliptical distributions, random variables with non convex support have to be considered. If the support of X is not convex, the intersection of a fixed hyperplane with this support can be a non connected set. So we define $H_c(x,b)$ as the connected component of $H(x,b) \cap \text{Support}(X)$ where x lies in. It is more natural defining conditional concepts based on $H_c(x,b)$ than on $H(x,b)$. Moreover, if $H_c(x,b)$ is convex then $E(X|X \in H_c(x,b))$ always belongs to $H_c(x,b) \subset \text{Support}(A)$, and then μ^* is going from $\text{Supp}(X)$ to itself. From now on, we assume that we are conditioning always to $H_c(x,b)$.

We are ready to extend the definition of the first principal component for normal random variable as fixed points of μ^* (remember Corollary 1).

Definition 3 *We define the set* $, (X)$ *of* principal oriented points *(POP) of* X *as the set of fixed points of* μ^**:*

$$, (X) = \{x \in I\!R^p : x \in \mu^*(x)\}.$$

When we refer to a POP x we also make implicit reference to its principal directions: the elements of $b^*(x)$.

At this step we can precisely establish our concept of *principal curves*.

Definition 4 *Consider a curve* α *from* $I \subseteq I\!R$ *to* $I\!R^p$*, where* I *is an interval and* α *is continuous and it is parameterized by the arc length. It is a* principal curve of oriented points *(PCOP) of* X *if*

$$\{\alpha(s) : s \in I\} \subseteq , (X).$$

Observe that the first principal component line is a PCOP for a multivariate normal distribution. The question of existence of POPs and PCOPs for an arbitrary p-dimensional random variable is considered in the next subsection.

We finish this subsection by defining the distribution on $I\!R$ induced for a random variable X who has a PCOP α. This concept will play an important role in Section 4.

Definition 5 *Consider a random vector* X *with density function* f_X *and let* α *be a curve* $\alpha: I \subseteq I\!R \to I\!R^p$ *parameterized by the arc length. Assume that* α *is PCOP for* X*. The* probability distribution on I induced by X and α *is the distribution of a random variable* S *having probability density function*

$$f_S(s) \propto \int_{I\!R^{p-1}} f_X(\alpha(s) + b^*_\perp(\alpha(s))v)dv,$$

provided that $\int_I f_S(s)ds < \infty$. *Moreover, if* $E(S) < \infty$, *we reparameterize* α *adding the constant* $(-E(S))$ *to the values of* I, *in order to have an induced random variable* S *with zero mean.*

2.2 Existence of POPs and PCOPs

We consider the following scenarios:

A1. Supp(X) is a compact set.

A2. There exists a compact set $K \subset$ Supp(X) such that for all $x \in K$ and all $b \in S^{p-1}$, $\mu(x, b) \in K$.

A3. There exists a compact set $K \subset$ Supp(X) such that for all $x \in K$, $\mu^*(x) \subseteq K$.

A4(K). For all $x \in K$ and all $b \in S^{p-1}$ the integral $\int_{H_c(x,b)} f(u)d(u)$ is positive.

Observe that either A1 and A2 imply A3. Assumption A4(K) guarantees that conditional mean and variance (and then μ^* too) are of class $\mathcal{C}^r$ at $x \in K$ (if $f_X \in \mathcal{C}^{r+1}$ at x, $r \geq 1$).

The following theorem deals with the existence of POPs. Its proof is direct because Brouwer's Fixed Point Theorem applies (see, for instance, Takayama 1985, p. 260).

Theorem 1 *Consider a random variable* X *with finite second moments and density function* f_X *of class* C^r, $r \geq 2$. *Assume that A3 is verified for a compact set* K, *that* A4(K) holds and that μ^* is a function. Then the set , (X) is a non empty set.

Remark 1. If μ^* is a correspondence, the natural extension of the preceding result would be done applying Kakutani's Theorem instead of Brouwer's one (see, for instance, Takayama 1985, p. 259). Nevertheless, Kakutani's result needs the set $\mu^*(x)$ to be convex, and in general this is not true in our case. So, we must require μ^* to be a function in order to have a not empty set , (X).

Remark 2. The existence of a compact set K verifying A2 implies that there is a kind of *attractive core* in the support of X (the compact set K): the mean of any hyperplane crossing K is inside K. For instance, if X

is normal with zero mean and variance matrix Σ, then the compact sets $K_c = \{x \in I\!R^p : x^t\Sigma^{-1}x \leq c\}$ verify condition A2. In general it looks sensible to think that sets of the form $\{x : f(x) > \epsilon\}$, for small $\epsilon > 0$, should hold this condition.

The existence of a PCOP in the neighborhood of any POP is guaranteed by the following theorem.

Theorem 2 *Consider a random variable X with finite second moments and density function f_X of class C^r, $r \geq 2$. Assume that the correspondence b^* is in fact a function ($\#\{b^*(x)\} = 1$, for all $x \in Supp(X)$). Let x_0 be a POP for X with principal direction $b_0 = b^*(x_0)$. Then there exists a PCOP α in a neighborhood of x_0 (i.e., there exists a positive ε and a curve $\alpha\colon(-\varepsilon,\varepsilon) \to I\!R^p$ such that $\alpha(0) = x_0$ and $\alpha(t)$ is a POP of X for all $t \in (-\varepsilon,\varepsilon)$). Moreover α is continuously differentiable and $\alpha'(t_0) = \lambda_0 K_0$, where*

$$K_0 = \frac{\partial \mu^*}{\partial x}(x_0)b^*(x_0) \in I\!R^p$$

and $\lambda_0 = b^*(x_0)^t\alpha'(t_0) \in I\!R$.

Because of this result, it is possible to compute the value of the tangent vector to a PCOP at a given point:

Corollary 2 *Let us assume that there exists a $\mathcal{C}^1$ curve $\alpha\colon I \subseteq I\!R \to I\!R^p$ being a PCOP. Then $\alpha'(t) = \lambda(t)K(t)$ for all $t \in I$, where*

$$K(t) = \frac{\partial \mu^*}{\partial x}(\alpha(t))b^*(\alpha(t)) \in I\!R^p$$

and $\lambda(t) = b^*(\alpha(t))^t\alpha'(t) \in I\!R$.

Remark 3. At that point, the question about whether $\alpha'(t)$ coincides with $b^*(\alpha(t))$ or not arises in a natural way. The answer to that question is in general negative. Here we have a simple example. (Other examples can be constructed where $b^*(\alpha(t)) = \alpha'(t)$; see the example in the next subsection for a particular case).

Example 1.
Consider the set

$$A = \{(x,y) \in I\!R^2 : x < 0, y > 1\} \cup \{(x,y) \in I\!R^2 : 0 \leq y \leq 1\} \cup$$

$$\cup\{(x,y) \in I\!R^2 : x > 0, y < 0\} \subset I\!R^2$$

and let X be a uniform random variable in $K = A \cap B((0,.5), r)$, for some large enough r. Then, it is not difficult to verify that near the point $(0,.5)$ the following set is a principal curve of oriented points:

$$\alpha = \{(x,y) : y = -x, x \leq -.5\} \cup \{(x,y) : y = .5, -.5 \leq x \leq .5\} \cup$$

$$\cup\{(x,y) : y = 1 - x, x \geq .5\}.$$

We can parameterize this curve by $t = x$. Observe that for all $(x,y) \in \alpha$ with $-.5 < x < 0$ the tangent direction to the curve α is parallel to the vector $(1,0)$. Moreover, for these points the principal direction of (x,y), say $b^*(x,y)$, is such that its orthogonal hyperplane (line, in this example) $H((x,y), b^*(x,y))$ is the line determined by (x,y) and the point $(0,1)$. So $b^*(x,y)$ is not parallel to $(1,0)$ and we conclude that in general $\alpha'(t) \neq b^*(\alpha(t))$. A similar reasoning can be done for (x,y) with $0 < x < .5$. $\square$

Some comments about the uniqueness of the PCOP are in order. It is easy to find examples of random vectors with a unique PCOP (e.g., the first principal component is the unique PCOP for a non spherical multivariate normal) or many (even infinite) PCOP (e.g., any line passing by the mean is a PCOP for a spherical multivariate normal). Theorem 2 establishes the existence of principal curves in a neighborhood of any POP. So the uniqueness question regards when these pieces of local curves can be joined to form a unique PCOP (or a finite number of them). The following result is based on compactness arguments and gives an intuition about when a PCOP is unique (its proof is direct).

Proposition 4 *Consider a random vector X with finite second moments and density function f_X in $\mathcal{C}^r$, $r \geq 2$. Assume that hypotheses* A3 *and* A4(K) *are verified for some compact set K. Let , (X) be the set of POPs for X, which is assumed to be a non empty set. Assume that for all $x \in$, X there exists a positive ε, a continuous curve $\alpha_x\colon (-\varepsilon, \varepsilon) \to I\!R^p$ with $\alpha_x(0) = x$, and an open set $V_x \subseteq I\!R^p$ such that*

$$V_x \cap \text{ , } (X) = \{\alpha_x(s) : s \in (-\varepsilon, \varepsilon)\}.$$

Then there exists a finite number J of continuous curves $\alpha_j\colon I_j \to I\!R^p$, $j = 1 \ldots, J$, such that

$$\text{ , } (X) = \cup_{j=1}^{J} \alpha_j(I_j).$$

2.3 PCOP of a distribution defined from a given curve

In this subsection we deal with the following problem. Assume that a random vector X in $\mathbb{R}^p$ has been defined as the sum of a randomly chosen point on a given parametric curve α plus a noise term. The question is whether the original curve α is a PCOP for X or not. Hastie and Stuetzle (1989) prove that the answer is negative for the principal curves they define (HSPC), and Tibshirani (1992) defines an alternative concept of principal curve overcoming this (using his words) *unsettling* difficulty.

We show here that the answer to the question mentioned above is also negative for the PCOP, but we argue that it is natural to have a negative answer. So we do not worry about trying to recover a generating curve, and we purely use the models given by *curve plus noise* as appropriate mechanisms to generate data with nonlinear structure.

According to our approach to principal curves, a theoretical model for a multivariate distribution defined from a given curve is as follows. Consider a curve in the p-dimensional space $\alpha: I \to \mathbb{R}^p$, where $I \subseteq \mathbb{R}$ is a possibly non-bounded 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 of α as the result of subjecting the segment I to 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 *frame 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 and 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 mapping $I \times \mathbb{R}^{p-1}$ 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 χ_α applies to $I \times R^{p-1}$ 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 $I \times R^{p-1}$ with density $f_0(s,y) = f_S(s) f_{Y|S=s}(y)$ where S and s are in I, having zero conditional expectations $E(Y|S=s)$ for all s. Consider the random variables in $I\!R^p$ obtained as $X = \chi_\alpha(S,Y)$.

As we mentioned bellow, it is not guaranteed that α is a PCOP for X. The following proposition shows that under quite general conditions it is natural that X has not α as a PCOP.

Proposition 5 *Assume that $I = Supp(S)$ 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: $Supp(S,Y) \to Supp(X)$ is a homeomorfism. Moreover, the density function of X at a given point $x \in Supp(X)$ 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.

Besides the previous assumptions, consider now that the random variables Y_1 and $(Y_2,\ldots,Y_{p-1})$ are conditional independent, given that $S=s$. Then

$$E(X|X \in H_c(\alpha(s),\alpha'(s))) = \alpha(s) + a_2(s) E\left[\left(\frac{Y_1}{1 - Y_1/\rho(s)}\right)\middle| S=s\right],$$

where $a_2(s) = \alpha''(s)\rho(s)$.

The proof of this result is based on change of variable standard techniques, and it is deferred to the Appendix.

Observe that the larger the values of y_1 are, the closer to the center of curvature the points x are. Then the density of the transformed variable is higher for points near the center of curvature, as it is expected: near the center of curvature the probability is compressed and it fills less room, so the density of probability raises. Consequently, the conditional expected value of the transformed variable X given that X belongs to a hyperplane orthogonal to α at $\alpha(s)$ is closer to the center of curvature than $\alpha(s)$. In addition to that, in many examples it is easy to verify (by symmetry arguments) that $b^*(\alpha(s))$ is just $\alpha'(s)$, and then it can be concluded that $E(X|X \in$

$H_c(\alpha(s), b^*(\alpha(s)))) \neq \alpha(s)$ and then α is not a PCOP. The next example is one of these cases.

Example 2.
Consider two independent random variables $S \sim U([-\pi R, \pi R])$, $R > 1$, and $Y \sim U([-1, 1])$. Let α be the parametric curve $\alpha(s) = (R\cos(s/R), R\sin(s/R))$, $s \in [-\pi R, \pi R]$. The transformation χ_α mentioned above transforms the rectangular region $[-\pi R, \pi R] \times [0, 1]$ to the annulus centered at the origin of $I\!R^2$ with radius $R-1$ and $R+1$. We transform (S, Y) according to χ_α and obtain the random vector X,

$$X = \alpha(S) + \alpha'_\perp(S)Y,$$

where $\alpha'_\perp(t)$ is the unit length vector orthogonal to $\alpha(t)$ oriented to the center of the annulus (i.e., $\alpha'_\perp(t) \propto \alpha''(t)$).

The bivariate density of X is not uniform over the transformed region. The density is larger in points closer to the center. For instance, the conditional distribution of X given that $X \in H_c(x = \alpha(0) = (R, 0), b = \alpha'(0) = (0, 1))$ has density function

$$f(u) = k\frac{1}{u}I_{[R-1,R+1]}(u),$$

where $k = (\log((R+1)/(R-1)))^{-1}$ and expected value $E(X|X \in H_c(x, b)) = 2k < R$, for all $R > 1$. We conclude that α is not a HSPC because $\mu(\alpha(0), \alpha'(0)) \neq \alpha(0)$ (i.e., α is not self consistent). Besides that fact, it can be shown that $b^*(\alpha(0) = (R, 0))$ is precisely $\alpha'(0)$ (see the Appendix for a justification) difficult) and concluded that α neither is a PCOP. □

The previous example and the Proposition 5 indicate that the concepts we are handling with (HSPC, PCOP) are not invariant against nonlinear deformations of the spaces they live in. The reason of this fact is that principal curves are defined by statistical properties (conditional expectation and variance, mainly) that are not invariant against this sort of transformations (the transformed density function strongly depends on the involved nonlinear deformation). We conclude that when we manage nonlinear curves with statistical tools it must be admitted that invariant objects are not likely to be reached.

3 Principal curves for data sets

Now we consider a random sample $X_1, \ldots, X_n$ from a multivariate random variable X. We assume that a non linear curve is a good summary of the structure of the distribution of X and we try to recover such a curve from the observed data X_i. In general, the hyperplanes passing by a given x_0 contain a very few (usually, only zero or one) observed X_i. So we need to include some smoothing procedure to calculate both conditional expected values and conditional total variances.

To define smoothed expectation and variance corresponding to a hyperplane $H = H(x, b)$, we project observations X_i orthogonally to the hyperplane and we denote the projections by X_i^H. A weight is associated to each projected observation,

$$w_i = w\left(|(X_i - x)^t b|\right) = w(\|X_i - X_i^H\|),$$

where w is any decreasing positive function.

The smoothed expectation of the sample corresponding to H is defined as the weighted expectation of $\{X_i^H\}$ with weights $\{w_i\}$. Let $\widetilde{\mu}(x, b) = \widetilde{\mu}(H(x, b))$ be such a value that, by definition, belongs to $H(x, b)$. The way we define the smoothed variance corresponding to a hyperplane $H(x, b)$ is

$$\widetilde{\mathrm{Var}}(x, b) = \widetilde{\mathrm{Var}}(H(x, b)) = Var_w(X_i^H, w_i; i = 1, \ldots, n),$$

where $Var_w(X_i^H, w_i)$ denotes the weighted variance of the projected sample with weights $\{w_i\}$. The smoothed total variance is $\widetilde{\phi}(x, b) = \mathrm{Trace}(\widetilde{\mathrm{Var}}(x, b))$.

Several definitions are available for w. For instance, we can use $w(d) = K_h(d) = K(d/h)$, where K is a univariate kernel function used in nonparametric density or regression estimation and h is its bandwidth parameter. If we use $w = K_h$, we can denote the smoothed total variance by $\widetilde{\phi}_h(x, b)$. The smoothness of $\widetilde{\phi}_h$ as a function of (x, b) will depend on h, as well as it happens in univariate nonparametric functional estimation.

In Section 2 the convenience on conditioning on $H_c(x, b)$, instead of $H(x, b)$, was pointed out. Translated to the sample smoothed world, conditioning to $H(x, b)$ is equivalent to using all the projected observations X_i^H with positive weights w_i. On the other hand, conditioning to $H_c(x, b)$ implies to look for clusters on the projected data configuration $\{X_i^H : w_i > 0\}$, to assign x to one of these clusters, and to use only the points in that cluster

to compute $\widetilde{\phi}$ and $\widetilde{\mu}$. We have implemented this last procedure (see subsection 3.3 for details). So, when we write $\widetilde{\phi}$ and $\widetilde{\mu}$ we assume that care for the eventual existence of more than one cluster in $H(x, b)$ has been taken.

Once the main tools for dealing with data sets $(\widetilde{\mu}, \widetilde{\phi})$ have been defined, we can look for sample POPs (subsection 3.1) and afterwards sample PCOPs (subsection 3.2).

3.1 Finding POPs

The sample version of b^* and μ^* are defined from $\widetilde{\mu}$ and $\widetilde{\phi}$ in a direct way. We call them $\tilde{b}^*$ and $\widetilde{\mu}^*$, respectively. So the set of sample POPs is the set of invariant points for $\widetilde{\mu}^*$:

$$\tilde{,} = \{x \in I\!R^p : x \in \widetilde{\mu}^*(x)\}.$$

In order to approximate the set $\tilde{,}$ by a finite set of points, we propose the following algorithm.

Algorithm 1 (Finite set of POPs)

Step 1. Draw randomly a point of the sample $X_1, \ldots, X_n$. Call it x_0. Make $k = 0$.

Step 2. Iterate the function $\widetilde{\mu}^*$ and define $x_k = \widetilde{\mu}^*(x_{k-1})$ until convergence (i.e., $||x_k - x_{k-1}|| \leq \epsilon$, for some prefixed ϵ) or until a prefixed maximum number of iterations is reached.

Step 3. If Step 2 finishes with convergence, include x_k in the set of sample POPs $\tilde{,}$.

Step 4. Repeat m times the steps 1, 2 and 3, for a prefixed m.

There is no theoretical guarantee about the convergence of the sequence $\{x_k = \widetilde{\mu}^*(x_{k-1}) : k \geq 1\}$, for a given x_0, but in all the simulated and real data sets we have examined, the step 2 of Algorithm 1 always reached quickly the convergence.

Example 3.
We illustrate the performance of Algorithm 1 with a real data set. Data came from the Spanish household budget survey (EPF, *Encuesta de Presupuestos Familiares*) corresponding to year 1991. We select randomly 500

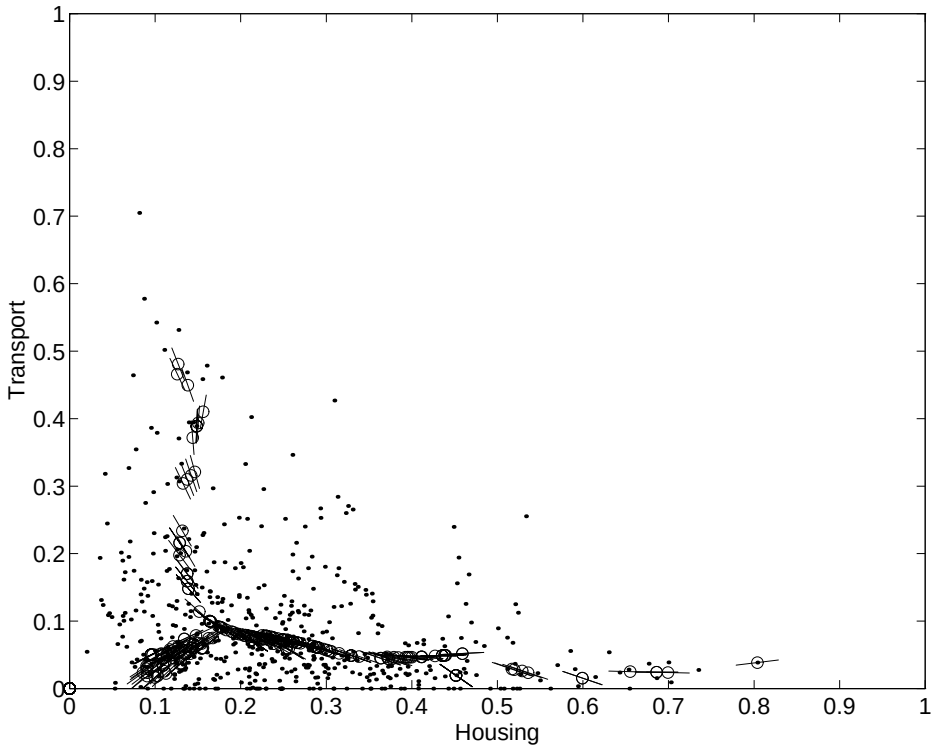

Figure 1: Example 3. Principal oriented points for proportions of household expenditure data.

households from the 21.155 observations of the EPF, and for each of them we annotate proportions of the total expenditure dedicated to housing (variable P_1) and transport (variable P_2). Our data are the 500 observations of the two-dimensional variable $P = (P_1, P_2)$. By definition, values of P fall inside the triangle defined by the points $(0,0)$, $(0,1)$ and $(1,0)$. A graphic representation indicates that data are non elliptical. We apply Algorithm 1 for $m = 100$ and obtain the set of sample POPs represented in Figure 1 as big empty dots. The principal direction of each one of these points is also represented as a short segment. Observe that the pattern of the POPs suggests that there are three principal curves joining at a point around $(.2, .1)$. □

3.2 Estimating a PCOP

In this subsection we deal with the extraction of a PCOP from a data set. In the population world, Theorem 2 guarantees that for any POP there exists a PCOP passing by this point. This result leads us to consider the following approach to build a PCOP: starting with a sample POP, we look for other POPs close to the first one, and placed in a way such that they recall a piece of a curve.

We start following Algorithm 1 until the first point considered as a POP appears. We call this point x_1 and denote by b_1 the principal direction of x_1 (if there are more than one element in $\tilde{b}^*(x_1)$, we choose one). We take $s_1 = 0$ and define $\alpha(s_1) = x_1$. Now we move a little bit from x_1 in the direction of b_1 and define $x_2^0 = x_1 + \delta b_1$, for some $\delta > 0$ previously fixed. The point x_2^0 serves as the seed of the sequence $\{x_2^k = \tilde{\mu}^*(x_2^{k-1}) : k \geq 1\}$, which eventually approach to a new point x_2. Define b_2 as $b^*(x_2)$, s_2 as $s_1 + \|x_2 - x_1\|$ and $\alpha(s_2) = x_2$.

We iterate that procedure until no points X_i can be considered "near" the hyperplane $H(x_k^0, b_k)$. Then we return to (x_1, b_1) and complete the principal curve in the direction of $-b_1$. The following algorithm formalizes the hole procedure.

Algorithm 2 (First Principal Curve)

Step 1. Make $k = 1$, $j = 0$ and $F = 1$. Choose $x_1^0 \in I\!R^p$ (for instance, the observed data closest to the sample mean). Choose $b_1^0 \in S^{p-1}$ (for instance, $b_1^0 = v_1$, where v_1 is the director vector of the first principal component of the sample). Choose $h > 0$, $\delta > 0$ and $p_t \in [0, 1]$. Let n be the sample size.

Step 2. Iterate in $j \geq 1$ the expression $x_k^j = \tilde{\mu}^*(x_k^{j-1})$ until *convergence* (see Algorithm 1 for details). Let x_k the final point of the iteration. Let $b_k = b^*(x_k)$. If $(b_k^0)^t b_k < 0$, then assign $-b_k$ to b_k.

Step 3. If $k = 1$ define $s_1 = 0$, and if $k > 1$ define $s_k = \text{Prec}(s_k) + F\|x_k - \text{Prec}(x_k)\|$. Define a new point in the principal curve $\alpha(s_k) = x_k$.

Step 4. Define $x_{k+1}^0 = x_k + F\delta b_k$, $b_{k+1}^0 = b_k$.

Step 5. First stopping rule.
If $\#\{i : (X_i - x_{k+1}^0)^t b_k^0 > 0\} < p_t n$ (i.e., there are less than a proportion p_t of the remaining points in the present direction of the principal curve) then go to **Step 7**.

Step 6. Define $\text{Prec}(s_{k+1}) = s_k$ and $\text{Prec}(x_{k+1}) = x_k$. Let $k = k + 1$ and $j = 0$. Return to **Step 2**.

Step 7. Second stopping rule.
If $F = 1$ (i.e., only one tail of the principal curve has been explored) then make $\text{Prec}(s_{k+1}) = s_1 = 0$, $\text{Prec}(x_{k+1}) = x_1$, $k = k+1$, $F = -1$, $x_k^0 = x_1^0 + F\delta b_1$ and $b_{k+1}^0 = b_1$. Go to **Step 2**.

Step 8. Final step. Let $K = k$. Order the values $\{(s_k, x_k), k = 1, \ldots, K\}$ according to the values $\{s_k\}$. The ordered sequence of pairs is the estimated *principal curve of oriented points* (PCOP).

In principle, only open principal curves are allowed by this algorithm but minor changes are needed to permit the estimation of a closed curve.

To obtain a curve $\hat{\alpha}$ from $I \subseteq I\!R$ to $I\!R^p$ we define $I = [s_1, s_k]$ and identify the curve with the polygonal $\{x_1, \ldots, x_k\}$. Observe that this curve is parameterized by the arc length. Spline techniques can also be used to find a smooth curve in $I\!R^p$ visiting all the points x_k.

During the algorithm completion it is possible to obtain estimation of many important statistical objects. The density of the induced random variable S in I can be estimated by

$$\hat{f}_S(s_k) = C_1 \frac{1}{nh} \sum_{i=1}^{n} K_h \left(|(X_i - x_k)^t b_k| \right),$$

where the constant C_1 is chosen to have integral of $\hat{f}_S$ equal to one. We also can assign a mass to each s_k:

$$\hat{p}_S(s_k) = C_2 \hat{f}_S(s_k) \left(\frac{s_{k+1} - s_{k-1}}{2} \right),$$

where C_2 is such that the sum of $\hat{p}_S(s_k)$ is one. Then we could consider $s_1, \ldots, s_k$ as a weighted sample of S. The mean and variance of this sample can be computed and subtracting the mean to the values s_k we obtain that S has estimated zero mean. Let us call $\widehat{Var}(S)$ the estimated variance of S.

Also, an estimation of the total variance in the normal hyperplane can be recorded for each s_k: $\widetilde{\phi}(x_k, b_k)$.

At this stage in the exposition, two definitions appear as natural. The first one is the *central point of the data set along the curve*. As S has estimated zero mean, this central point is defined as

$$\hat{E}_{PCOP} = \hat{\alpha}(0).$$

The second is a measure of total variability consistent with the estimated structure around a curve. Our proposal is to define the *total variability of the data along the curve* as

$$\widehat{TV}_{PCOP} = \widehat{Var}(S) + \int_I \widetilde{\phi}^*(\alpha(s))\hat{f}_S(s)ds \simeq \widehat{Var}(S) + \sum_k \widetilde{\phi}(x_k, b_k)\hat{p}_S(s_k).$$

From these numbers we define the *proportion of total variability* explained by the estimated curve as $p_1 = \widehat{Var}(S)/\widehat{TV}_{PCOP}$. This quantity plays the role of the proportion of variance explained by the first principal component in the linear world.

Example 4.

To illustrate the algorithm 2, we apply it to a simulated data set. The data are generated as

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \alpha_1(S) \\ \alpha_2(S) \end{pmatrix} + \frac{1}{\|\alpha'(S)\|} \begin{pmatrix} -\alpha_2'(S) \\ \alpha_1'(S) \end{pmatrix} Y$$

where $\alpha: [0,1] \to I\!R^2$, $x = \alpha_1(s) = 2\pi s + 1$, $y = \alpha_2(s) = 2\,(1/x - \cos(x-1))$, $S \sim U(0,1)$ and $Y \sim N(0, \sigma = .2)$. The sample size in our example is $n = 200$.

Figure 2 shows the data set (small dots) and the graph of α (dashed curve). For that data set two principal curve methodologies have been applied: our own algorithm and that of Hastie and Stuetzle (1989). The S-plus public domain routines written by Trevor Hastie and available on STATLIB (`http://www.stat.cmu.edu/S/principal.curve`) are used to implement the HSPC methodology. Default parameters of these routines have been used. Some MATLAB routines have been written to implement the Algorithm 2. The HSPC has been represented in Figure 2 by a solid line with empty dot marks. The bold solid curve with big dot marks corresponds to the resultant PCOP. We can observe that the graphs of both principal curves are

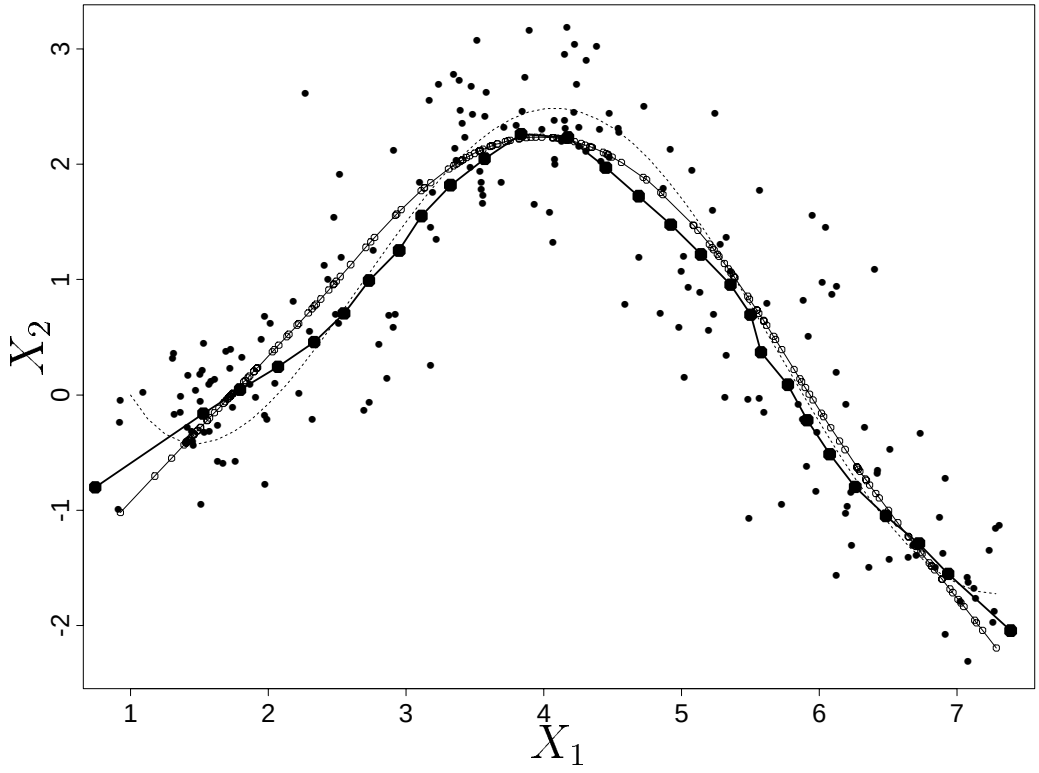

Figure 2: Example 4. PCOP and HSPC for a simulated data set. Dotted line: generating curve. Solid line with empty dots: HSPC. Solid line with big dot marks: PCOP.

very similar in almost all their range of parameters. They differ for values of (X_1, X_2) near the extreme $(1, 0)$ of the scatter plot. Both procedures present a bias when the curvature of the original parametric curve α is important (near the point $(4, 2)$). Techniques proposed in Banfield and Raftery (1992) should be applied.

Now we report some details of the implementation of our algorithm in that particular data set. The bandwidth parameter h is 1 and δ is .33. The estimated parameterization of the principal curve goes trough the data from right to left. The estimated interval I is $I = [-5.10, 4.37]$, so the length of the PCOP is 9.47 (the length of the HSPC is 10.23 and the length of the generating curve is 10.39). The total variability along the curve is 6.97. The estimation of the variance of the random variable S defined on I is 6.82 and the average value of the variance along the orthogonal lines to the principal curve is 0.15 (remember that the generating noise variance used to generate the data was 0.16). So the proportion of the total variability explained by the first principal curve is $p_1 = .98$.

Figure 3 shows the estimated density of S on $I = [-5.10, 4.37]$. Remember that the principal curve goes from right to left. Observe that the density of S is not uniform, although the original random variable used as parameter to generate the data set was uniform. The reason is that the original parameterization was not of unit-speed and the estimated principal curve is unit-speed. In fact, the density function in Figure 3 looks like $(\|\alpha'(s)\|)^{-1}$, as it should be. □

Example 5.

We replicate now the example contained in section 5.3 of Hastie and Stuetzle (1989). We generate a set of 100 data points from a circle in $\mathbb{R}^2$ with independent normal noise:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 5\ \sin(S) \\ 5\ \cos(S) \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix},$$

with $S \sim U[0, 2\pi]$ and $\epsilon_i \sim N(0, 1)$. Figure 4 summarizes the results of the estimation of the first principal curve by our methodology and also by using Hastie and Stuetzle (1989) routines. Panel (a) is similar to Figure 2. Panel (b) shows the HSPC estimation. In panel (c) they can be seen the POPs (and its principal directions) obtained during the application of algorithm 2. They are the base for the PCOP represented in panel (d).

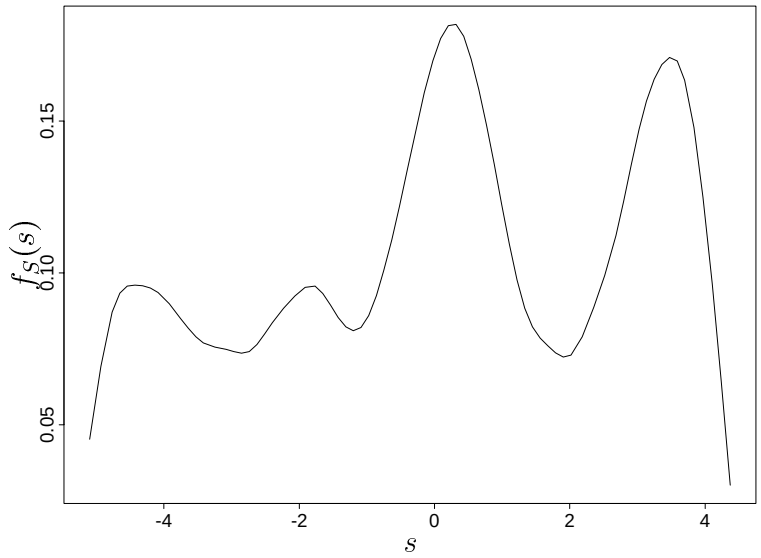

Figure 3: Example 4. Density estimation of S.

The length of the original curve is 10π. When algorithm 2 is used, the estimated curve has length 30.8342 and the length for the estimated HSPC is 33.41086. The estimated total variability along the curve is 87.65, the estimated $\mathrm{Var}(S)$ is 86.58 (the value for the generating distribution is $100\pi^2/12 = 82.25$) and the average residual variance in the orthogonal directions is 1.06 (this value should not be compared directly with $\mathrm{Var}(\epsilon_i)$). Density estimation of variable S and local orthogonal variance estimation are approximately constant over the estimated support of S. These facts are according to the data generating process, which original parameterization was unit-speed in this example. □

Example 3 (Continuation).
We compute now PCOPs for the households' expenditures data. The Figure 1 suggests that there are more than one curve for this data set. We look for two of them by starting the Algorithm 2 with two different points $x_1^0 = (.1, .05)$ and $x_1^0 = (.15, .2)$, and respective values of the starting vectors $b_1^0 = (1, 1)$ and $b_1^0 = (0, -1)$. The resulting curves are drawn in Figure 5. The total variability along the curves are, respectively, .0201 and .0306, with percentages of variability explained by the correspondent PCOP equal to 78.24% and 84.25%. For this data set, the total variance is .0302, and the first principal component explains the 70.6% of it. So we conclude that any of the two estimated PCOPs summarizes the data better than the first principal

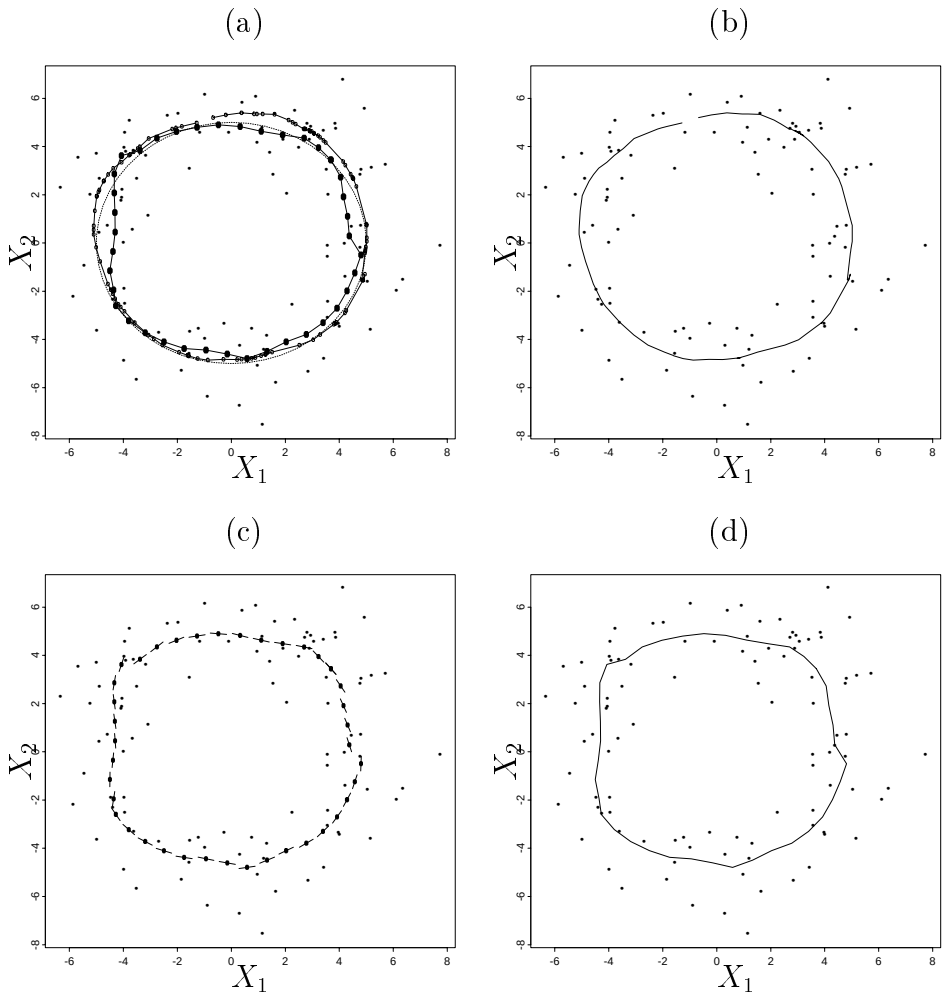

Figure 4: Example 5. Simulated data set around a circle. (a) Original circle (dashed line), HSPC (solid line) and PCOP (solid bold line). (b) HSPC. (c) Some POPs. (d) PCOP.

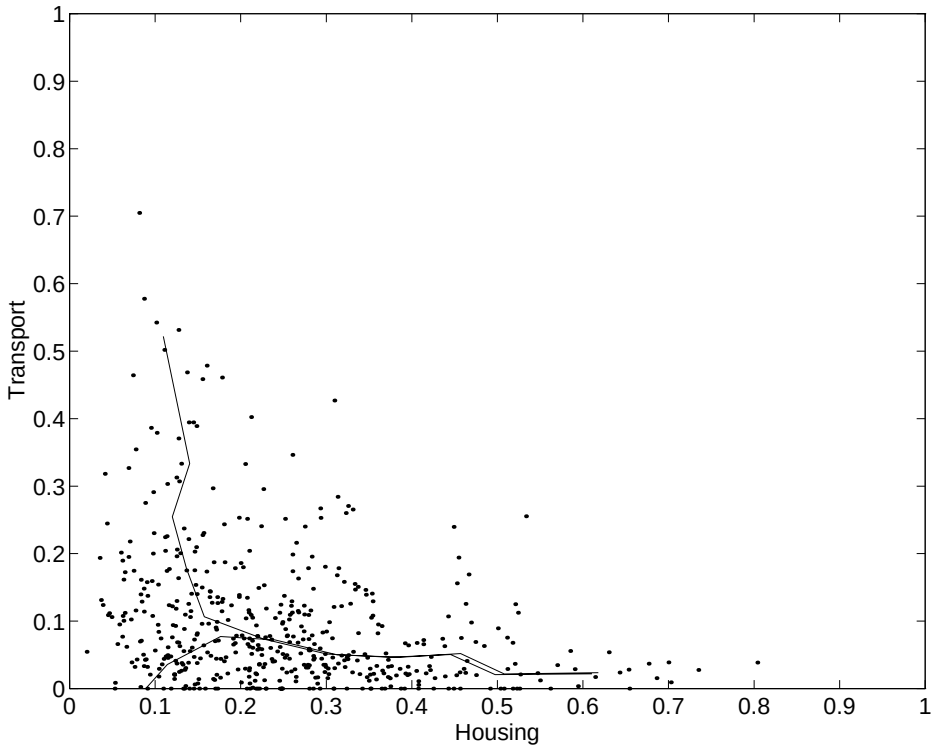

Figure 5: Example 3. Two principal curves of oriented points for proportions of households' expenditures data.

component does. □

Example 6. Data in $I\!R^3$

A simulated data set in $I\!R^3$ is considered. Data are around the piece of circumference $\{(x, y, z) : x^2 + y^2 = 10^2, z = 0\}$. A uniform random variable S over this set was generated, and then a noise Y was added to it so that $(Y|S = s)$ fall in the orthogonal plane to the circumference at the point s, and has bidimensional normal distribution with variance matrix equal to the 2×2 identity matrix. We used the parameters $h = 1$ and $\delta = .75$. The resulting PCOP is represented in Figure 6 from two points of view. The estimated curve explains a 92.19% of the total variability along the curve.
□

3.3 Assigning x to a cluster in $H(x,b)$

In Section 2 was argued that, when we deal with nonlinearities, conditioning on $H_c(x,b)$ (the convex component of $H(x,b) \cup \text{Supp}(X)$ containing x) has more sense that conditioning on $H(x,b)$. In the sample world, conditioning to $H(x,b)$ is equivalent to using all the projected observations X_i^H over $H(x,b)$ with positive weights w_i, and conditioning to $H_c(x,b)$ is like using only some of them: those laying in the same cluster as x lies. So we need an algorithm that identifies the points X_i^H belonging to the same cluster as x does. Our proposal is inspired on the agglomerative hierarchical clustering methods based on single linkage.

Consider a set of points $\{y_0, y_1, \ldots, y_n\}$ in $\mathbb{R}^d$. The objective is to identify what points $y_i, i \geq 1$ belong to the same cluster as y_0. The algorithm is as follows.

Algorithm 3 (Clustering around a given point)

Step 1. Define the sets $C = \{y_0\}$ and $D = \{y_1, \ldots, y_n\}$. Set $j = 1$. Choose a positive real number λ (for instance, $\lambda = 3$).

Step 2. While $j \leq n$, repeat:

2.1 Define $d_j = d(C,D) = \min\{d(x,y) : x \in C, y \in D\}$ and let y_j^* the point $y \in D$ where this minimum is achieved.

2.2 Set $C = C \cup \{y_j^*\}$ and $D = D - \{y_j^*\}$. Set $j = j + 1$.

Step 3. Compute the median m and quartiles Q_1 and Q_3 of the data set $\{d_1, \ldots, d_n\}$. Define the *distance barrier* as $\bar{d} = Q_3 + \lambda(Q_3 - Q_1)$.

Step 4. Let $j^* = \min\{\{j : d_j > \bar{d}\} \cup \{n+1\}\} - 1$. The final cluster is $C^* = \{y_1^*, \ldots, y_{j^*}^*\}$.

Observe that the algorithm identifies extreme outlying distances d_j as we would do it by using a box-plot, and it only accepts a point y_i as in the same cluster as y_0 when there is a polygonal line from y_0 to y_i with vertex in $\{y_0, \ldots, y_n\}$ and segments shorter than $\bar{d}$.

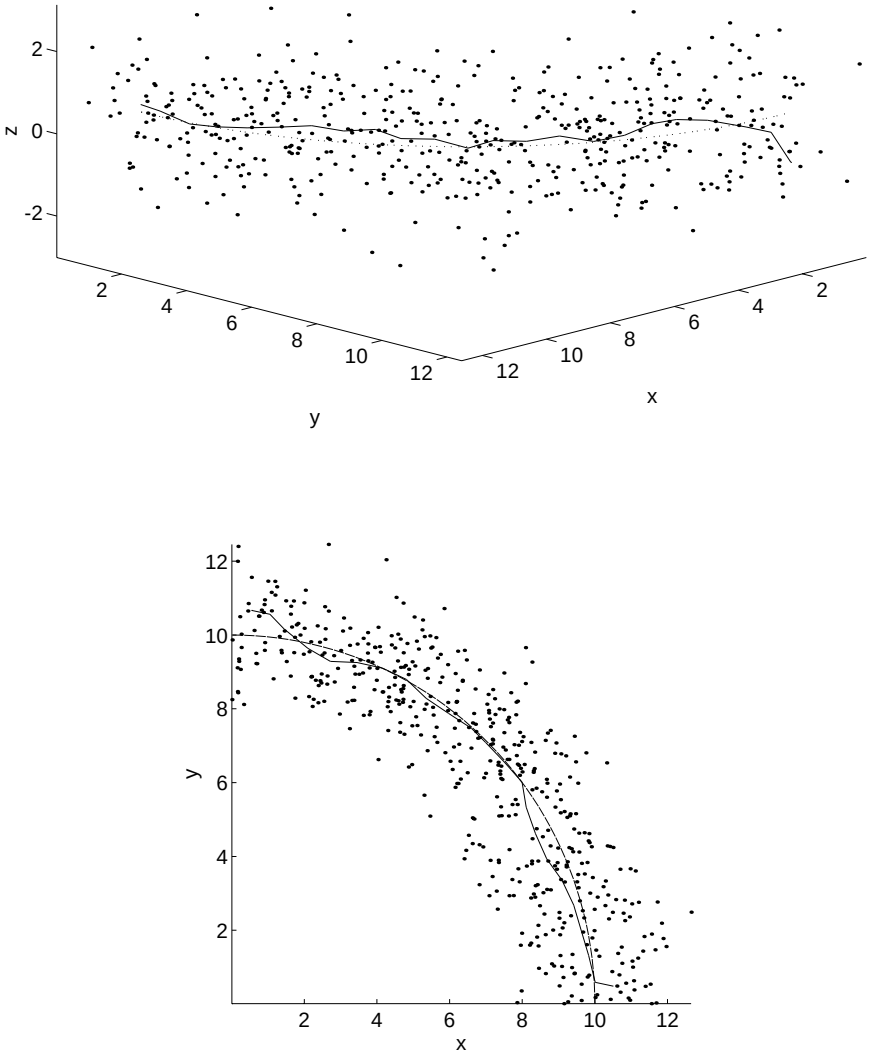

Figure 6: Example 6. Two perspectives of the estimated PCOP (solid line) for the three-dimensional data around a piece of circumference (dotted line).

4 Generalized total variance and higher order principal points and curves

In subsection 3.2 the total variability of a data set along an estimated curve was defined as $\widehat{TV}_{PCOP} = \widehat{Var}(S) + \int_I \widetilde{\phi}^*(\alpha(s))\hat{f}_S(s)ds$. If a random variable X has the curve $\alpha: I \to I\!R^p$ as a principal curve of oriented points, the sample measure $\widehat{TV}_{PCOP}$ corresponds to the population quantity

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

where S is a random variable on I having probability distribution induced by X and α (see Definition 5).

Observe that when X has normal distribution and α is the first principal component line, $TV_\alpha(X)$ is precisely the total variance of X because $TV[X|X \in H_c(\alpha(s), b^*(\alpha(s)))]$ is constant in s and equals the total variance of the joint distribution of the remaining $(p-1)$ principal components of X. We conclude that $TV_\alpha(X)$ is a good way to measure the variability of a p-dimensional random vector X having a PCOP α, provided that $TV[X|X \in H_c(\alpha(s), b^*(\alpha(s)))]$ would appropriately measure the dispersion of the $(p-1)$-dimensional conditional random vector $(X|X \in H_c(\alpha(s), b^*(\alpha(s))))$. When these $(p-1)$-dimensional distributions are elliptical the total variance is a well-suited measure, but when non-linearities also appear in $(X|X \in H_c(\alpha(s), b^*(\alpha(s))))$, the total variance is no longer advisable and it should be changed, in the definition of $\widehat{TV}_{PCOP}$, by a measure of the variability along a nonlinear curve.

The former arguments lead us to define the *generalized total variance* (hereafter GTV) of a p-dimensional random variable by induction in the dimension p. The definition is laborious because many concepts have to be simultaneously and recursively introduced.

Definition 6

For any one-dimensional random variable X with finite variance we say that X recursively admits a generalized principal curve of oriented points *(GPCOP). We say that $x = E(X)$ is the only* generalized principal oriented point *(GPOP) for X, that $\alpha: \{0\} \to I\!R$, with $\alpha(0) = E(X)$ is the only GPOP for X. We define the* generalized expectation *of X (along α) as $GE_1(X) = \alpha(0) = E(X)$, and the* generalized total variance *of X (along α) as $GTV_1(X) = Var(X)$.*

Now we consider $p > 1$. We assume that for $k < p$ we know whether a k-dimensional random variable recursively does admit or not GPCOPs, and what GPOPs, GPCOPs, GE_k and GTV_k are for k-dimensional random variables that recursively admit GPCOP.

Consider a p-dimensional random variable X with finite second moments. We say that X recursively admits GPCOPs *if the following conditions* **(i)**, **(ii)** *and* **(iii)** *are verified The first one is as follows:*

(i) *For all $x \in I\!R^p$ and all $b \in S^{p-1}$ the $(p-1)$-dimensional distribution $(X|X \in H_c(x,b))$ recursively admits principal curves.*

If this condition holds, we define

$$\mu_G(x,b) = GE_{p-1}(X|X \in H_c(x,b)),\ \phi_G(x,b) = GTV_{p-1}(X|X \in H_c(x,b)),$$

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

The set of fixed points of μ_G^*, , $_G(X)$, is called the set of generalized principal oriented points *of X. Given a curve $\alpha: I \subseteq I\!R \to I\!R^p$ parameterized by the arc length, we say that it is a* generalized principal curve of oriented points *for X if $\alpha(I) \subseteq$, $_G(X)$.*

Now we can express the second condition for X recursively admitting GPCOPs:

(ii) *There exists a unique curve such that α is GPCOP for X.*

When conditions (i) and (ii) apply, we define for any $s \in I$ the value $\bar{f}_S^G(s) = \int_{I\!R^{p-1}} f_X(\alpha(s) + (b_G^*)_\perp(\alpha(s))v)dv$. The third condition is:

(iii) *The integral $k = \int_I \bar{f}_S^G(s)ds$ is finite and the random variable S with density function $f_S^G(s) = (1/k)\bar{f}_S^G(s)$ has finite variance and zero mean (may be a translation of S is required to have $E(S) = 0$).*

If condition (iii) holds, we say that the distribution of S has been induced *by X and α.*

Now we define GE_p as

$$GE_p(X) = \alpha(0),$$

and the GTV_p by

$$GTV_p(X) = Var(S) + \int_I GTV_{p-1}(X|X \in H_c(\alpha(s), b_G^*(\alpha(s)))f_S(s)ds =$$

$$= Var(S) + \int_I \phi_G^*(\alpha(s)))f_S(s)ds.$$

Observe that the concept of second (and higher order) principal curves is involved in the former definition. Our approach implies that there is not a common second principal curve for the hole distribution of X, but that there is a different second principal curve for each point in the first one. So the concept of second principal curve (and higher order) is a *local* concept.

Definition 7 *If X recursively admits GPCOPs and α is GPCOP for X, we say that α is* the first GPCOP *of X. We say that the first GPCOPs for the $(p-1)$-dimensional distributions $(X|X \in H_c(\alpha(s), b_G^*(\alpha(s)))$ are* the family of second GPCOPs *for X, and so on.*

Example 7.
Figure 7 illustrate these ideas. The first GPCOP is a curve in $\mathbb{R}^3$: $\{(x, y, z) : x^2 + y^2 = 10^2, z = 0\}$. For each point $p_0 = (x_0, y_0, z_0)$ in this curve, there exists a specific second GPCOP $\beta_{p_0}: \mathbb{R} \to H_{p_0}$, where H_{p_0} is the orthogonal hyperplane to the first principal curve at p_0. In this case, β_{p_0} is

$$\beta_{p_0}(v) = \begin{pmatrix} -x_0/10 & 0 \\ -y_0/10 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_0/10 & x_0/10 \\ x_0/10 & -y_0/10 \end{pmatrix} \begin{pmatrix} v \\ \sin(v) \end{pmatrix},$$

for $v \in [-\pi, \pi]$. The way the local second principal curve varies along the first principal curve should be smooth enough to permit the estimation by smoothing techniques. □

Observe that the definition of GPCOPs coincides with that of PCOP for $p = 2$. For any p, both definitions coincide if the conditional distributions to $X \in H(x, b)$ are elliptical for all x and all b. In this case, the second principal curves are the first principal component of these conditional distributions, and so on.

When second principal curves are considered, we can say that the quantity

$$p_1 = \frac{\mathrm{Var}(S)}{\mathrm{GTV}_p(X)}$$

is the proportion of generalized total variance explained by the first principal curve. As for each $s \in I$, the local second principal curve is the first principal curve for a $(p-1)$-dimensional random variable, we can compute the proportion $p_1(s)$ of the generalized total variance that locally explain the second

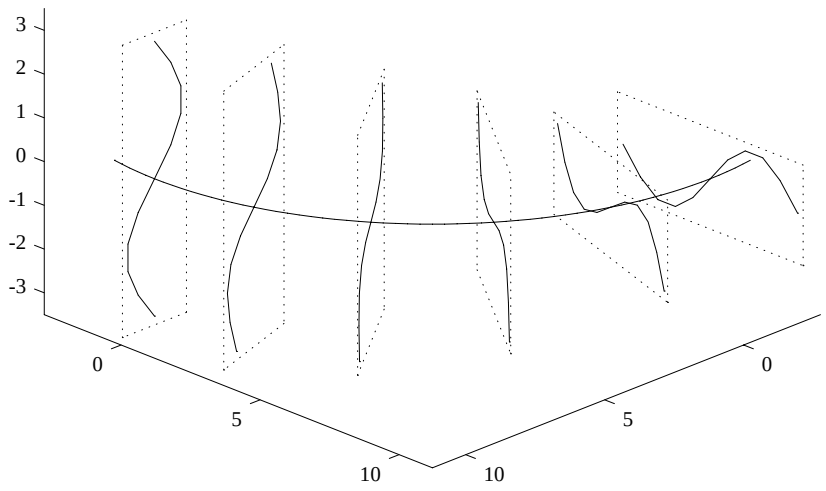

Figure 7: Example 7: Theoretical structure of local second principal curves along the first one.

Source of variability	GTV	$\%GTV$	Cum.GTV	Cum. $\%GTV$
First Principal Curve	22.18	88.45%	22.18	.88.45%
Local 2nd. Ppal. Cvs.	2.71	10.80%	24.89	99.25%
Local 3rd. Ppal. Cvs.	.19	.75%	25.08	100.00%
Total	25.08	100%		

Table 1: Example 7. Proportion of the generalized total variance due to the first principal curve and to local second principal curves, for data set of Figure 8.

principal curve in the point $\alpha(s)$. We calculate the expected proportion of explained GTV by the local second principal curves, define

$$p_2 = (1 - p_1) \int_I p_1(s) f_S(s) ds$$

and interpret it as the proportion of the GTV explained by the second principal curves. We can iterate the process and obtain $p_j,\ j = 1, \ldots, p$, adding up to 1.

Example 7. (Continuation)
Random data have been generated according to the structure shown in Figure 7. Uniform data were generated over the piece of circumference that constitutes the first principal curve. Then, each of these data (namely, p_0) was (uniformly) randomly moved along the sinusoidal second principal curve laying on p_0, to a new position p_1. Finally, a univariate random noise perturbed the point p_1 inside the line orthogonal to the second curve at p_1, also contained H_{p_0}. The resulting point, p_3, is one of the simulated points. The normal noise has standard deviation $\sigma = .2$.

Figure 8 shows the results of the estimation procedure for a sample of size equal to 1000, offering three different perspectives of the estimated object. Table 1 indicates what percentages of the generalized total variance are due to the first GPCOP and to the family of second GPCOPs. □

5 Discussion

In the present work the concept of principal curve introduced by Hastie and Stuetzle (1989) is approached from a different perspective. A new definition

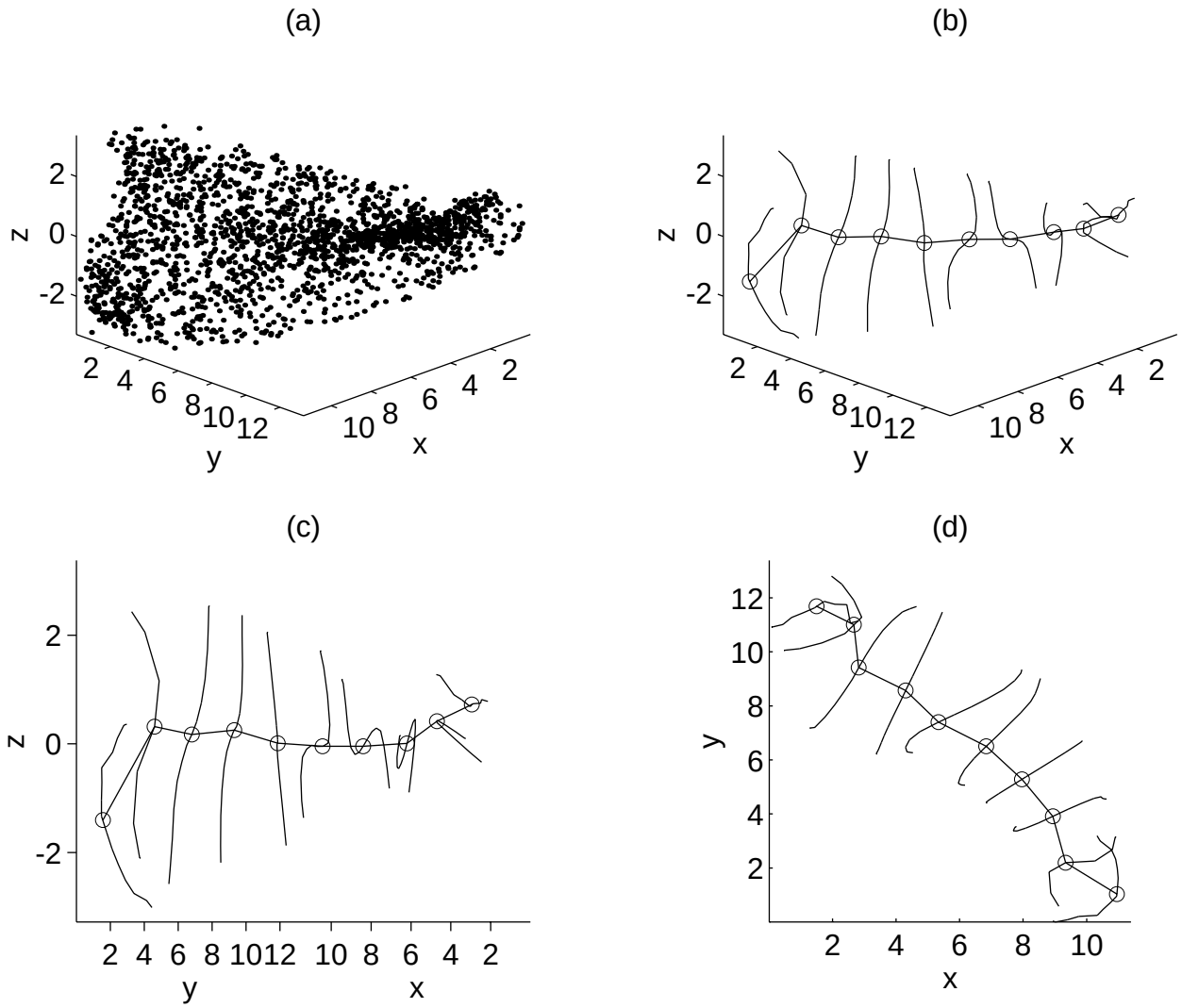

Figure 8: Example 7. Estimation of the first principal curve and the family of local second principal curves along the first one. (a) Data set; (b) first GPCOP and second GPCOPs; (c) same as (b) viewed from a point with zero degrees of elevation over the XY plane; (d) GPCOP system projected over the XY plane.

of the first principal curve has been introduced, based on the concept of *principal oriented point*.

All the arguments are based on conditional expectation and variance, given that a p-dimensional random variable lies in the hyperplane defined by a point x and the orthogonal direction b, but different measures of conditional location and dispersion could be used, as far as they are smooth function of x and b. More robust procedures could be obtained in that way.

In the last part of the paper we introduce generalized definitions of expectation and total variance along a principal curve. For random variables having principal curves for all its lower dimensional marginals, these new definitions allow us to define second a higher order local principal curves in a recursive way.

Appendix: Proofs

Proof of Proposition 1. Defining $Y = b^t X$, the joint distribution of $(X^t, Y)^t$ is $(p+1)$-dimensional normal. So standard theory on conditional normal distributions tells us that

$$(X|X \in H(x_0, b)) \equiv (X|Y = b^t x_0) \sim N_p \left(\mu + \frac{b^t(x_0 - \mu)}{b^t \Sigma b} \Sigma b,\ \Sigma - \frac{\Sigma b b^t \Sigma}{b^t \Sigma b} \right). \tag{1}$$

So the conditional total variance is

$$TV(X|X \in H(x_0, b)) = \text{Trace}(\Sigma) - \frac{1}{b^t \Sigma b} \text{Trace}(\Sigma b b^t \Sigma),$$

and the problem **(P1)** is

$$\min_{b: b^t \Sigma b = 1} \{TV(X|X \in H(x_0, b))\} = \text{Trace}(\Sigma) - \max_{b: b^t \Sigma b = 1} (b^t \Sigma \Sigma b) =$$

$$= \text{Trace}(\Sigma) - \max_{h: h^t h = 1} (h^t \Sigma h) = \text{Trace}(\Sigma) - \max_{h: h^t h = 1} \text{Var}(h^t X),$$

where $h = \Sigma^{1/2} b$. So the solution of **(P1)** is given by the solution of **(P2)**, which is the classical problem of principal components, with optimal solution $h^* = v_1$, the eigenvector associated with the largest eigenvalue λ_1 of Σ. The corresponding solution of **(P1)** is

$$b^* = \Sigma^{-1/2} h^* = \frac{1}{\lambda_1} \Sigma^{-1/2} \Sigma h^* = \frac{1}{\lambda_1} \Sigma^{1/2} h^* = \frac{1}{\lambda_1} \lambda^{1/2} h^* = \lambda^{-1/2} h^*,$$

and the main part of the proposition is proved. Two facts were used in this chain of equalities: first, h^* is eigenvector of Σ, and second, that if v is eigenvector of Σ with associate eigenvalue λ, then v is eigenvector of $\Sigma^{1/2}$ with associate eigenvalue $\lambda^{1/2}$. To prove the last sentence of the result, it suffices to replace $b = b^*$ in (1). □

Proof of Proposition 3. The property concerning ϕ^* is a direct application of the Maximum Theorem (see, for instance, Takayama 1985, p. 254). The Sensitivity Theorem (a corollary of the Implicit Function Theorem; see Bertsekas 1995, p.277, for instance) permits smoothness properties of b^* to be established, and then the smoothness of μ implies that of μ^*. □

Before proving Theorem 2, we need somme lemmas.

Lemma 1 *Let $x \in I\!R^p$ and $b \in S^{p-1}$. The partial derivatives of μ and ϕ are as follows.*

$$
\begin{aligned}
(i)\ \frac{\partial \mu}{\partial x}(x,b) &= K_x^{\mu}(x,b)b^t,\ K_x^{\mu}(x,b) \in I\!R^p,\ \textit{and } b^t K_x^{\mu}(x,b) = 1. \\
(ii)\ \frac{\partial \phi}{\partial x}(x,b) &= k_x^{\phi}(x,b)b^t,\ k_x^{\phi}(x,b) \in I\!R. \\
(iii)\ \frac{\partial \mu}{\partial b}(x,b) &= K_b^{\mu}(x,b)\left(I_p - bb^t\right),\ K_b^{\mu}(x,b) \in I\!R^{p\times p}, \\
(iv)\ \frac{\partial \phi}{\partial b}(x,b) &= K_b^{\phi}(x,b)^t\left(I_p - bb^t\right),\ K_b^{\phi}(x,b) \in I\!R^{p},
\end{aligned}
$$

Proof. *(i):* As $\mu(x,b)$ (as a function of x) is constant on $H_c(x,b)$, then $\mu(x + (I - bb^t)v), b)$ is constant in v, so its derivative with respect to v is equal to 0:

$$0 = \frac{\partial}{\partial v}\left(\mu(x + (I - bb^t)v), b)\right) = \frac{\partial \mu}{\partial x}\left(x + (I - bb^t)v, b\right)(I - bb^t).$$

That can be written as

$$\frac{\partial \mu}{\partial x}\left(x + (I - bb^t)v, b\right) = \left[\frac{\partial \mu}{\partial x}\left(x + (I - bb^t)v, b\right)b\right]b^t,$$

and when v goes to 0, we obtain that

$$\frac{\partial \mu}{\partial x}(x,b) = K_x^{\mu}(x,b)b^t$$

where

$$K_x^{\mu}(x,b) = \frac{\partial \mu}{\partial x}(x,b)b.$$

In order to see that $K_x^{\mu}(x,b)^t b = 1$ we derive the identity

$$(x - \mu(x,b))^t b = 0$$

with respect to x:

$$b^t \left(I - \frac{\partial \mu}{\partial x}(x,b) \right) = 0$$

and post-multiplying by b the result follows:

$$b^t b = 1 = b^t K_x^{\mu}(x,b).$$

(ii) As a function of its first argument, $\phi(x,b)$ is constant in $H_c(x,b)$. Then, proceeding as above we obtain the result for

$$k_x^{\phi}(x,b) = \frac{\partial \phi}{\partial x}(x,b)b \in I\!R.$$

(iii) Observe that $\mu(x, b + vb)$ is constant for $v \in I\!R$, so

$$0 = \frac{\partial}{\partial v}\mu(x, b + vb) = \frac{\partial \mu}{\partial b}(x, b + vb)b,$$

and then the rows of $(\partial\mu/\partial b)(x, b + vb)$ are orthogonal to b. Therefore,

$$\frac{\partial \mu}{\partial b}(x, b + vb) \left(I - bb^t \right) = \frac{\partial \mu}{\partial b}(x, b + vb).$$

When v goes to zero,

$$\frac{\partial \mu}{\partial b}(x,b) = K_b^{\mu}(x,b) \left(I - bb^t \right),$$

where $K_b^{\mu}(x,b) = (\partial\mu/\partial b)(x,b)$.

(iv) A similar reasoning leads to prove that point. □

Lemma 2 *For all x such that $(x, b^*(x))$ is a POP, it is verified that*

$$\frac{\partial b^*}{\partial x}(x) = \left(I_p - b^*(x)b^*(x)^t \right) \tilde{K}(x)b^*(x)^t.$$

Proof. We divide the proof in two parts.
(1) As $b^*(x)^t b^*(x) = 1$, deriving with respect to x we obtain that

$$b^*(x)^t \frac{\partial b^*}{\partial x}(x) = 0,$$

therefore $(\partial b^*/\partial x)(x)$ is orthogonal to $b^*(x)$, and we can write that

$$(I - b^*(x)b^*(x)^t)\frac{\partial b^*}{\partial x}(x) = \frac{\partial b^*}{\partial x}(x).$$

(2) As $b^*(x)$ is constant on $y \in H_c(x, b^*(x))$, by similar arguments to those used in the proof of Lemma 1, we can deduce that

$$\frac{\partial b^*}{\partial x}(x) = \tilde{K}(x)b^*(x)^t$$

for some $\tilde{K}(x) \in I\!R^p$. Now, putting together (1) and (2) we obtain the desired result. □

Lemma 3

$$\frac{\partial \mu^*}{\partial x}(x) = K_x^{\mu^*}(x)b^*(x)^t,$$

where $K_x^{\mu^*}(x) \in I\!R^p$. *Moreover,*

$$b^*(x)^t K_x^{\mu^*}(x) = 1,$$

Proof. We derive the identity

$$\mu^*(x) = \mu(x, b^*(x))$$

with respect to x, and we obtain that

$$\frac{\partial \mu^*}{\partial x}(x) = \frac{\partial \mu}{\partial x}(x, b^*(x)) + \frac{\partial \mu}{\partial b}(x, b^*(x))\frac{\partial b^*}{\partial x}(x).$$

Now, from Lemmas 1 and 2, it follows that

$$\frac{\partial \mu^*}{\partial x}(x) = K_x^{\mu}(x, b^*(x))b^*(x)^t +$$

$$+K_b^{\mu}(x, b^*(x))(I - b^*(x)b^*(x)^t)\tilde{K}(x)b^*(x)^t = K_x^{\mu^*}(x)b^*(x)^t$$

for some $K_x^{\mu^*}(x) \in I\!R^p$.

To prove the last sentence, we derive with respect to x the identity $(x - \mu^*(x))^t b^*(x) = 0$, as we did in the proof of Lemma 1. $\square$

Proof of Theorem 2. The proof is based on the Implicit Function Theorem. For the point x_0, we have that $x_0 = \mu(x_0, b^*(x_0))$. Without loss of generality, we can assume that $x_0 = 0 \in I\!R^p$ and that $b_0 = b^*(x_0) = e_1 = (1, 0, \ldots, 0)^t \in I\!R^p$. For any $x \in I\!R^P$ we call x_1 its first component and denote by x^2 its remaining $(p-1)$ components. Analogous notation is used for defining μ_1 and μ^2 from function μ (we do the same thing also for μ^* and α).

Consider the function

$$\begin{array}{rcl} \Lambda: I\!R \times I\!R^{p-1} & \rightarrow & I\!R^{p-1} \\ (x_1, x^2) & \rightarrow & \mu^2\left(\binom{x_1}{x^2}, b^*\binom{x_1}{x^2}\right) - x^2 = (\mu^*)^2\binom{x_1}{x^2} - x^2, \end{array}$$

and observe that $\Lambda(0, \mathbf{0}) = \mathbf{0}$, where $\mathbf{0}$ is the zero of $I\!R^{p-1}$. If the Implicit Function Theorem could be applied here, we would obtain that there exists a positive ε and a function Ψ

$$\begin{array}{rcl} \Psi: (-\varepsilon, \varepsilon) \subset I\!R & \rightarrow & I\!R^{p-1} \\ t & \rightarrow & \Psi(t) \end{array}$$

such that $\Psi(0) = \mathbf{0}$, and

$$\Lambda(t, \Psi(t)) = \mathbf{0}$$

or, equivalently,

$$\Psi(t) = \mu^2\left(\binom{t}{\Psi(t)}, b^*\binom{t}{\Psi(t)}\right)$$

for all $t \in (-\varepsilon, \varepsilon)$. We now define

$$\begin{array}{rcl} \alpha: (-\varepsilon, \varepsilon) \subset I\!R & \rightarrow & I\!R^p \\ t & \rightarrow & \alpha(t) = \binom{t}{\Psi(t)} \end{array}$$

Observe that the properties of Ψ guarantee that $\alpha^2(t) = \mu^2(\alpha(t), b^*(\alpha(t)))$. So if we prove that $\mu_1(\alpha(t), b^*(\alpha(t))) = t$ then we will have that α is the PCOP we are looking for. But indeed that is true. Observe that always $\mu(x, b)$ belongs to $H(x, b)$, so $(x - \mu(x, b))^t b = 0$. In our case, this fact implies that

$$\left(\alpha(t) - \mu(\alpha(t), b^*(\alpha(t)))\right)^t b^*(\alpha(t)) = 0.$$

As $\alpha^2(t) = \mu^2(\alpha(t), b^*(\alpha(t)))$, the last equation is equivalent to write

$$(t - \mu_1(\alpha(t), b^*(\alpha(t)))) \, b_1^*(\alpha(t)) = 0.$$

Remember that $b^*(x_0) = e_1$, so $b_1^*(x_0) = 1$. Continuity of b^* implies that $b_1^*(x) > .5$ if x is close enough to x_0. So, ϵ can be chosen in order to have $b_1^*(\alpha(t)) \neq 0$, and then we deduce that $(t - \mu_1(\alpha(t), b^*(\alpha(t))))$ must be zero, and we conclude that α is a PCOP.

Only checking the assumptions for the Implicit Function Theorem (see, for instance, Corwin and Szczarba 1979, p.277) remains to complete the proof of the Theorem. We need to show that the last $(p-1)$ columns of the Jacobian of Λ at $x_0 = (0, \mathbf{0})$ are independent. These columns are

$$\frac{\partial \Lambda}{\partial x^2}(x_0) = \left(\frac{\partial}{\partial x^2} \left(\mu^2(x, b^*(x)) \right) \right)(x_0) - I_{p-1}.$$

Observe that the first term in this sum is the matrix obtained by dropping out the first row and the first column of the following Jacobian matrix (see Lemma 3):

$$\frac{\partial \mu^*}{\partial x} = \left(\frac{\partial}{\partial x} \left(\mu(x, b^*(x)) \right) \right)(x) = K_x^{\mu^*}(x) b^*(x)^t.$$

As $b^*(x_0) = b_0 = e_1$, the product $K_x^{\mu^*}(x_0) b^*(x_0)^t$ has its last $(p-1)$ rows equal to zero. Therefore,

$$\frac{\partial \Lambda}{\partial x^2}(x_0) = \mathbf{0}_{(p-1)\times(p-1)} - I_{p-1} = -I_{p-1}$$

and it has complete rank. So Implicit Function Theorem applies and the first part of the Theorem is proved.

Let us compute $\alpha'(0)$. Again, the Implicit Function Theorem determines the derivative of Ψ with respect to t:

$$\frac{\partial \Psi}{\partial t} = \left(\frac{\partial \Lambda}{\partial \Psi} \right)^{-1} \frac{\partial \Lambda}{\partial t}.$$

In our case,

$$\frac{\partial \Lambda}{\partial \Psi} = I_{p-1}$$

and

$$\frac{\partial \Lambda}{\partial t} = \frac{\partial}{\partial x_1} \left(\mu^2(x, b^*(x)) \right) = \frac{\partial}{\partial x_1} \left((\mu^*)^2(x) \right)$$

and this is the first column of $(\partial\mu^*/\partial x)(x_0) = K_x^{\mu^*}(x_0)b_0^t$ (i.e., $K_x^{\mu^*}(x_0)$), without its first element (we have used Lemma 3). Then, $\partial\Lambda/\partial t = (K_x^{\mu^*}(x_0))^2$. Therefore,

$$\frac{\partial\alpha}{\partial t}(t_0) = \left(\frac{\partial}{\partial t}\begin{pmatrix} t \\ \Psi(t)\end{pmatrix}\right)(t_0) = \begin{pmatrix} 1 \\ (K_x^{\mu^*}(x_0))^2\end{pmatrix}.$$

The result would be proved if we can show that $(K_x^{\mu^*}(x_0))_1$ is equal to 1. But this is true because $(K_x^{\mu^*}(x_0))_1 = K_x^{\mu^*}(x_0)^t b_0 = 1$, by Lemma 3. □

Proof of Corollary 2. As $\alpha(t) = \mu^*(\alpha(t))$, deriving with respect to t, we have

$$\alpha'(t) = \left(\frac{\partial\mu^*}{\partial x}(\alpha(t))\right)\alpha'(t) = K_x^{\mu^*}(\alpha(t))b^*(\alpha(t))^t\alpha'(t).$$

Then $\alpha'(t) = \lambda(t)K_x^*(\alpha(t))$ for all $t \in I$, and $\lambda(t) = b^*(\alpha(t))^t\alpha'(t) \in I\!R$. □

Proof of Proposition 5. Because $\mathrm{Supp}(Y|S = s) \subseteq B(0, \rho(s))$ and $H_c(\alpha(s), \alpha'(s)) \cap H_c(\alpha(t), \alpha'(t)) = \emptyset$ when $s \neq t$, then χ_α is a 1-1 function form $\mathrm{Supp}(S, Y)$ to the image of this set. As χ_α is continuous and it is defined on a compact set, it follows that $\chi_\alpha(\mathrm{Supp}(S, Y)) = \mathrm{Supp}(\chi_\alpha(S, Y))$. Then χ_α is a homeomorfism because it is a 1-1 continuous function defined from a compact set to a metric space.

Remember that $\chi_\alpha(s, y) = \alpha(s) + A(s)(0, y^t)^t$, where the frame matrix $A(s)$ is an orthonormal matrix, it is differentiable as a function of s, and its first column is $\alpha'(s)$. Moreover, $A(s)$ can be chosen so that the corresponding Cartan matrix $C(A) = A^{-1}A' = A^tA'$ is skew-symmetric ($C^t = -C$) having elements $c_{ij}(s) = 0$ for $|i - j| \neq 1$, where A' is the matrix whose elements are the derivatives of the elements of matrix A (for details see, for instance, Guggenheimer 1977, pp. 158-160). As χ_α is 1-1, we call $(s(x), y(x)) = \chi_\alpha^{-1}(x)$, for a given $x \in \mathrm{Supp}(X)$, where $X = \chi_\alpha(X)$.

Applying change of variable standard techniques, the density function of X at a given x can be computed as

$$f_X(x) = f_{(S,Y)}(s(x), y(x))(\det(J_{\chi_\alpha}(s(x), y(x)))^{-1},$$

where $J_{\chi_\alpha}(s(x), y(x))$ is the Jacobian of χ_α at x, that is to say the $p \times p$ matrix

$$J_{\chi_\alpha}(s, y) = \frac{\partial\chi_\alpha}{\partial s\partial y}(s, y) = (\alpha'(s) + A_2'(s)y, A_2(s)),$$

where $A_2(s)$ is the $p \times (p-1)$ matrix containing the last $(p-1)$ columns of $A(s)$ (so $A(s) = (\alpha'(s), A_2(s))$). Then

$$\det(J_{\chi_\alpha}(s,y)) = \det(\alpha'(s) + A_2'(s)y, A_2(s)) =$$

$$= \det(\alpha'(s), A_2(s)) + \det(A_2'(s)y, A_2(s)) = \det(A(s)) + \sum_{j=2}^{p} y_{j-1} \det(a_j'(s), A_2(s)),$$

where $a_j(s)$ is the j-th column of $A(s)$. Remember that $(A(s))^t A'(s) = C(A(s))$ (so $A'(s) = A(s)C(A(s))$) and that the Cartan Matrix $C(A(s))$ has the following structure:

$$\begin{pmatrix} 0 & -k_1(s) & 0 & 0 & \dots & 0 & 0 & 0 \\ k_1(s) & 0 & -k_2(s) & 0 & \dots & 0 & 0 & 0 \\ 0 & k_2(s) & 0 & -k_3(s) & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & k_{p-2}(s) & 0 & -k_{p-1}(s) \\ 0 & 0 & 0 & 0 & \dots & 0 & k_{p-1}(s) & 0 \end{pmatrix}.$$

$k_j(s)$ is the j-th *curvature* of $\alpha(s)$. In particular, $k_1(s) = ||\alpha''(s)||$ is the curvature of α at $\alpha(s)$. From $A'(s) = A(s)C(A(s))$, it follows that

$$\begin{aligned} \alpha''(s) &= k_1(t)a_2(s) \\ a_2'(s) &= -k_1(t)\alpha'(s) + k_2(s)a_3(s) \\ a_3'(s) &= -k_2(t)a_2(s) + k_3(s)a_4(s) \\ &\vdots \quad \vdots \quad \vdots \\ a_j'(s) &= -k_{j-1}(t)a_{j-1}(s) + k_j(s)a_{j+1}(s) \\ &\vdots \quad \vdots \quad \vdots \\ a_p'(s) &= -k_{p-1}(t)a_{p-1}(s) \end{aligned}$$

Then, for $3 \leq j \leq p-1$ we have

$$\det(a_j'(s), A_2(s)) = \det(-k_{j-1}(t)a_{j-1}(s) + k_j(s)a_{j+1}(s), (a_2(s), \dots, a_p(s))) = 0,$$

for $j = p$,

$$\det(a_p'(s), A_2(s)) = \det(-k_{p-1}(t)a_{p-1}(s), (a_2(s), \dots, a_p(s))) = 0,$$

and for $j = 2$,

$$\det(a_2'(s), A_2(s)) = \det(-k_1(t)\alpha'(s) + k_2(s)a_3(s), (a_2(s), \dots, a_p(s))) =$$

$$= (-k_1(s)) \det(\alpha'(s), (a_2(s), \ldots, a_p(s))) = (-k_1(s)) \det(A(s)) = -k_1(s).$$

Moreover, $\det(A(s)) = 1$, because $A(s)$ is an orthonormal matrix. So we conclude that $\det(J_{\chi_\alpha}(s,y)) = 1 - k_1(s)$, and the first part of the result is proved.

For the second part, without loss of generality we can assume that $s = 0$, $\alpha(0) = 0$ and $A(0) = I_p$. Defining $e_1 = (1, 0, \ldots, 0)^t$, we have that

$$E(X|X \in H_c(\alpha(0), \alpha'(0))) = E((X_1, \ldots, X_p)^t | X \in H_c(0, e_1)) =$$

$$= (0, \int_{\mathbf{R}^{p-1}} (y_1, \ldots, y_{p-1}) \frac{1}{1 - y_1/\rho(0)} f_{Y|S=0}(y_1, \ldots, y_{p-1}) dy_1 \ldots dy_{p-1})^t,$$

and under conditional independence of Y_1 and $(Y_2, \ldots, Y_{p-1})$ that equals

$$(0, \int_I R \frac{y_1}{1 - y_1/\rho(0)} f_{Y_1|S=0}(y_1) dy_1, 0, \ldots, 0)^t,$$

and the proof finishes. □

Justification of the last point in Example 2.
The argument is based in the behavior of the example when R goes to infinity. For R large, the distribution on the annulus resembles that of the uniform over the rectangle $\{(u,v) : |u - R| < 1, |v| < r\}$ for a very large r, so the variance of $((U,V)|V = m(U - R))$ is $V(U)(1 + m^2)$ and takes its minimum value for $m = 0$. That corresponds to the orthogonal direction to $(0, 1)$, as we pointed out in the example. □

References

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

Bertsekas, D. P. (1995). *Nonlinear Programming*. Athenea Scientific.

Bishop, C.M., M. Svensén, and C. K. I. Williams (1996). GTM: The generative topographic mapping. Technical Report NCRG/96/015 (To appear in *Neural computation*), Neural Computing Research Group, Aston University.

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. Cambridge, MA: The MIT Press.

Corwin, L. and R. Szczarba (1979). *Calculus in Vector Spaces*. Marcel Dekker, Inc.

Dong, D. and T. J. McAvoy (1996). Nonlinear principal component analysis based on principal curves and neural networks. *Computers chem. Engng.*, **20**, 65–78.

Duchamp, T. and W. Stuetzle (1996). Extremal properties of principal curves in the plane. *The Annals of Statistics*, **24**, 1511–1520.

Etezadi-Amoli, J. and R.P. McDonald (1983). A second generation nonlinear factor analysis. *Psychometrika*, **48**, 315–342.

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

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. New York: John Wiley.

Gnanadesikan, R. and M.B. Wilk (1966). Data analytic methods in multivariate statistical analysis. In P.R. Krisnaiah (Ed.), *Multivariate analysis, Vol. II*.

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

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.

Koyak, R. (1987). On measuring internal dependence in a set of random variables. *Annals of Statistics*, **15**, 1215–1228.

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

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

Shepard, R.N. and J.D. Carroll (1966). Parametric representation of nonlinear data structures. In P.R. Krisnaiah (Ed.), *Multivariate analysis, Vol. II*.

Srivastava, J.N. (1972). An information approach to dimensionality analysis and curved manifold clustering. In P.R. Krisnaiah (Ed.), *Multivariate analysis, Vol. III*.

Stanford, D. and A. E. Raftery (1997). Principal curve clustering with noise. Technical Report 317, Department of Statistics, University of Washington.

Takayama, A. (1985). *Mathematical Economics (Second Edition)*. Cambridge University Press.

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

Tarpey, T. and B. Flury (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.

Yohai, V.J., W. Ackermann, and C. Haigh (1985). Nonlinear principal components. *Quality and Quantity*, **19**, 53–69.