Another look at principal curves and surfaces

Pedro Delicado*

Departament d'Economia i Empresa, Universitat Pompeu Fabra

Ramon Trias Fargas 25-27, 08005 Barcelona, SPAIN

February 1, 1999

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 characterization of which is the basis of the definition of principal curves.

We establish a new characterization of the first principal component and base our new definition of a principal curve on this property. We introduce the notion of *principal oriented points* and we prove the existence of principal curves passing through these points. We extend the definition of principal curves to multivariate data sets and propose an algorithm to find them. The new notions lead us to generalize the definition of total variance. Successive principal curves are recursively defined from this generalization. The new methods are illustrated on simulated and real data sets.

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

*The author is very grateful to Wilfredo Leiva-Maldonado for helpful conversations, suggestions and theoretical support. Comments of A. Kohatsu, G. Lugosi and K. Udina were also very useful. This work was partially supported by the Spanish DGES grant PB96-0300. Electronic address: `delicado@upf.es`, `http://www.econ.upf.es/%7Edelicado`

1 Introduction

Consider a multivariate random variable X in $\mathbb{R}^p$ with density function f and a random sample from X, namely $X_1, \ldots, X_n$. The first principal component can be viewed as the straight line which best fits the cloud of data (see, e.g., Johnson and Wichern 1992, pp. 386-387). When the distribution of X is ellipsoidal the population first principal component is the main axis of the ellipsoids of equal concentration.

In the last forty years many works have appeared proposing extensions of principal components to distributions with nonlinear structure. We cite Shepard and Carroll (1966), Gnanadesikan and Wilk (1966), Srivastava (1972), Etezadi-Amoli and McDonald (1983), Yohai, Ackermann, and Haigh (1985), Koyak (1987) and Gifi (1990), among others. Some of them look for nonlinear transformations of the observable variables into spaces admitting a usual principal component analysis. Others postulate the existence of a nonlinear link function between a latent lower dimensional linear space and the data space.

The work of Hastie and Stuetzle (1989) opens a new way to look at the problem. Its main distinguishing mark is that no parametric assumptions are made. The principal curves (of a random variable X) defined by Hastie and Stuetzle (1989) (hereafter, HSPC) are one-dimensional parameterized curves $\{x \in \mathbb{R}^p : x = \alpha(s), s \in I\}$ (where $I \subseteq \mathbb{R}$ is an interval and $\alpha : I \to \mathbb{R}$ is differentiable), having the property of *self-consistency*: every point $\alpha(s)$ in the curve is the mean (under the distribution of X) of the points x that project onto $\alpha(s)$. In this sense, HSPC passes through the "middle" of the distribution. It is not guaranteed that such a curve does exist. An appropriate definition of principal curves for data sets is also given. Nonparametric algorithms are used to approximate them. *Principal surfaces* are analogously defined.

In the 90's several works directly related with Hastie and Stuetzle (1989) have appeared. Banfield and Raftery (1992), mainly applied, modifies the Hastie and Stuetzle (1989) algorithm to reduce the estimation bias. Tibshirani (1992) provides a new definition of a principal curve such that if X is the result of adding a noise to a random point over a one-dimensional curve α, then α is a principal curve of X; HSPC does not have this property. LeBlanc and Tibshirani (1994) uses multivariate adaptive regression splines (see Friedman 1991) to develop estimation procedures of principal curves and surfaces. Duchamp and Stuetzle (1993, 95, 96) study principal curves in the plane. They prove the existence of (many) principal curves crossing each other for simple distributions and

they state a negative result: in general, principal curves are critical points of the expected squared distance from the data, but they are not extremal points of this functional. An application of HSPC in the clustering context is made by Stanford and Raftery (1997). Tarpey and Flury (1996) study in depth the self-consistency concept and extend it to more general settings.

Other recent papers on nonlinear multivariate analysis do not follow directly the line of Hastie and Stuetzle (1989). Kégl, Krzyżak, Linder, and Zeger (1997) introduce the concept of principal curves with a fixed length. They prove the existence and uniqueness of that curve for theoretical distributions, give an algorithm to implement their proposals, and calculate rates of convergence of the estimators. Salinelli (1998) studies nonlinear principal components as optimal transformations of the original variables, where the nonlinear admissible transformations belong to a functional space verifying certain properties. In the most recent years, several related works have appeared in the neural networks literature: Mulier and Cherkassky (1995), Tan and Mavarovouniotis (1995), Dong and McAvoy (1996), Bishop, Svensén, and Williams (1998), among others.

In this paper we give a new definition of principal curves. It is based on a 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 minimal when the hyperplane is orthogonal to the first principal component. The generalization of this result to nonlinear distributions leads us to define *principal oriented points* (as the fixed points of certain function from $\mathbb{R}^p$ to itself), and *principal curves of oriented points* (one-dimensional curves visiting only principal oriented points). The existence of principal oriented points is proved for theoretical distributions. It is also guaranteed that there exists a principal curve passing through each one of these points. Sample versions of these elements are introduced and illustrated with real and simulated data examples.

The new definition has some advantages over that of Hastie and Stuetzle (1989). For instance, in the normal multivariate case every principal component is a HSPC. However, only the first principal component satisfies our definition. In Remark 1 below we present an example in which the principle curve according to the new definition is more natural than that of Hastie and Stuetzle. In addition to that, our approach to principal curves involves the notion of principal oriented points, a concept with statistical interest in itself.

Our approach suggests a natural generalization of *total variance*, providing a good measure of the dispersion of a random variable distributed around a nonlinear principal curve. The generalized total variance allows us to define recursively local second (and higher order) principal curves.

Our proposals are close to Hastie and Stuetzle (1989) in spirit: no parametric assumptions are made, smoothing techniques are used in the proposed 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 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 presents 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. Appendix I presents the formal versions of the algorithms presented along the paper. The proofs of the results appearing in the paper are postponed to the Appendix II.

2 Definition of population principal curves

A well known property of the first principal component for normal distributions can be stated as follows: the projection of the normal random variable onto the hyperplane orthogonal to the first principal component has the lowest total variance among all the projected variables onto any hyperplane. Furthermore, 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. Our definition of principal curves is based on this property.

2.1 Definitions

Let X be a p-dimensional random variable with density function f and finite second moments. Consider $b \in S^{p-1} = \{w \in I\!R^p : ||w|| = 1\}$ and $x \in I\!R^p$. We call $H(x, b)$ the

hyperplane orthogonal to b passing through x: $H(x,b) = \{y \in I\!R^p : (y-x)^t b = 0\}$.

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 basis for $I\!R^p$. We define $b_\perp$ as the $(p \times (p-1))$ matrix $(b_2(b), \ldots, b_p(b))$. The total variance of a random variable Y (i.e., the trace of the variance matrix of Y) is denoted by $TV(Y)$. A parameterized curve α in $I\!R^p$, $\alpha: I \to I\!R^p$ where I is a possibly unbounded interval, is said to be *parameterized by the arc length* if the length of the curve from $\alpha(s_1)$ to $\alpha(s_2)$ is $|s_2 - s_1|$. This is equivalent to say that α is *unit-speed* paprameterized (i.e., $||\alpha'(s)|| = 1$ for all s) when it is differentiable. More properties about curves in $I\!R^p$ can be found, for instance, in Guggenheimer (1977).

With these definitions we introduce

$$f_1(x,b) = \int_{I\!R^{p-1}} f(x + b_\perp v) dv,$$

$$\mu(x,b) = E(X|X \in H(x,b)) = \frac{1}{f_1(x,b)} \int_{I\!R^{p-1}} (x + b_\perp v) f(x + b_\perp v) dv,$$

and

$$\phi(x,b) = TV(X|X \in H(x,b)) = \frac{1}{f_1(x,b)} \int_{I\!R^{p-1}} v^t v f(x + b_\perp v) dv - \mu(x,b)^t \mu(x,b),$$

for any x and b such that $f_1(x,b) > 0$. 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 functions μ and ϕ are well defined. Observe that $\mu(x,b) = \mu(x,-b)$ and $\phi(x,b) = \phi(x,-b)$. So we define in S^{p-1} the equivalence relation $\equiv$ by: $v \equiv w \iff v = w$ 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.

When ϕ is continuous, the infimum of $\phi(x,b)$ over b is achieved because $TV(X)$ is finite and because S^{p-1} is compact. We define the correspondence $b^*: I\!R^p \to S^{p-1}$ 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. Let $\phi^*(x) = \phi(x, b^*(x))$, be the minimum value. We also define the correspondence $\mu^*: I\!R^p \to I\!R^p$ as $\mu^*(x) = \mu(x, b^*(x))$. Smoothness properties of μ, ϕ, b^*, μ^* and ϕ^* are in accordance with the smoothness of f. Proposition 3 in the Appendix II summarizes these properties.

The result below formalizes the property we expressed at the beginning of the section. It characterizes the points of the first principal component line in terms of μ^* and b^*.

Proposition 1 *Consider a p-dimensional normal random variable X with mean value μ and variance matrix Σ. Let λ_1 be the largest eigenvalue of Σ and v_1 the corresponding unit length eigenvector. The following properties are verified.*

(i) For any $x_0 \in I\!R^p$ the correspondence b^* is in fact a function (i.e., the minimum of $\phi(x_0, b)$ as a function of b is unique) and $b^*(x_0) = v_1$, for all x_0.

(ii) For any $x_0 \in I\!R^p$, the point $x_1 = \mu^*(x_0)$ belongs to the first principal component line $\{\mu + sv_1 : s \in I\!R\}$.

(iii) A point $x_1 \in I\!R^p$ belongs to the first principal component line if and only if x_1 is a fixed point of μ^*.

Observe that only local information around a point x_1 is needed to verify whether x_1 is a fixed point of μ^* or not. This result also provides a mechanism to find points in the first principal component: the iteration of the function μ^* leads (in one step) from any arbitrary point x_0 to a point x_1 on the first principal component line. In the rest of this subsection we exploit this mechanism in order to generalize the first principal component to non-normal distributions.

A comment on the adequacy of conditioning on $H(x, b)$ is in order. As we are interested in defining valid concepts for non-ellipsoidal 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 for any $x \in$ Support(X) we define $H_c(x, b)$ as the connected component of $H(x, b) \cap$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 $H_c(x, b) \subset$ Support(X), and then μ^* maps Support(X) to itself. From now on, we condition always on $\{X \in H_c(x, b)\}$.

We are ready to introduce the notion of *principal oriented points* and then state our definition of *principal curves*.

Definition 1 *We define the set $\Gamma(X)$ of* principal oriented points *(POP) of X as the set of fixed points of μ^*:* $\Gamma(X) = \{x \in I\!R^p : x \in \mu^*(x)\}$.

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

Definition 2 *Consider a curve α from I to $\mathbb{R}^p$, where I is an interval in R and α is continuous and parameterized by the arc length. α is a* principal curve of oriented points *(PCOP or just principal curve) of X if $\{\alpha(s) : s \in I\} \subseteq \Gamma(X)$.*

Observe that Proposition 1 establishes 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.

Remark 1. Our definition of principal curve does not coincide in general with the definition of Hastie and Stuetzle (1989). For instance, Duchamp and Stuetzle (1995) prove that the circle with radius $r_{circ} = R + d^2/(3R)$ is one of the HSPCs for the uniform distribution on the annulus $\Omega_{R-d,R+d} = \{x \in \mathbb{R}^2 : R - d \leq ||x|| \leq R + d\}$, with $0 < d < R$. Moreover, there exists an infinite number of HSPC oscilating around this circle. However, it is easy to prove that the only PCOP for this distribution is the circle with radius R.

Remark 2. Consider a random vector X in $\mathbb{R}^p$ defined as the sum of a randomly chosen point on a given parametric curve α plus a noise term. This setting raises the question of whether the original curve α is a principal curve for X or not. Hastie and Stuetzle (1989) prove that the answer is negative for their principal curves definition, and Tibshirani (1992) defines an alternative concept overcoming this difficulty. In Delicado (1998) we show that the answer to this question is also negative for the PCOP, but there we argue that it is natural to have a negative answer and that it is not a so important awkwardness. So we do not worry about trying to recover a generating curve, and use the models given by *curve plus noise* only as appropriate mechanisms to generate data with nonlinear structure.

Next we define a distribution on $\mathbb{R}$ induced for a random vector X which has a PCOP α. This concept will play an important role in Section 4.

Definition 3 *Consider a random vector X with density function f and let α be a curve $\alpha: I \to \mathbb{R}^p$ parameterized by the arc length, where $I \subseteq \mathbb{R}$ is an interval. 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*

$$f_S(s) \propto f_1(\alpha(s), b^*(\alpha(s))), s \in I,$$

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 principal oriented points and principal curves

We consider the following conditions:

A1. Support(X) is a compact set.

A2. There exists a compact set $K \subset$ Support(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 $K \subset$ Support(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 $f_1(x, b)$ is positive, where the integral defining $f_1(x, b)$ is done over $\{v \in I\!R^{p-1} : x + b_\perp v \in H_c(x, b)\}$.

Observe that either A1 and A2 imply A3. Assumption A4(K) guarantees that conditional mean and variance are of class $\mathcal{C}^r$ at $x \in K$, provided that $f \in \mathcal{C}^{r+1}$ at x for $r \geq 1$. (A function g defined on an open subset U in $I\!R^p$ is said to be of class $\mathcal{C}^r$ if all partial derivatives of g of order r exist and are continuous.)

The following theorem deals with the existence of POPs.

Theorem 1 *Consider a random variable X with finite second moments and density function f 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 (i.e., $\#\{\mu^*(x)\} = 1$, for all $x \in$ Support(X)). Then the set $\Gamma(X)$ is a nonempty set.

Remark 3. The proof of this result is based on Brouwer's Fixed Point Theorem (see, e.g., Takayama 1985, p. 260). If μ^* is a correspondence, the natural extension of the preceding result would be done applying Kakutani's Theorem instead of Brouwer's (see, e.g., 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.

Remark 4. 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 satisfy this condition.

The existence of a principal curve in the neighborhood of any principal oriented point is guaranteed by the following theorem.

Theorem 2 *Consider a random variable X with finite second moments and density function f of class C^r, $r \geq 2$. Assume that the correspondence b^* is in fact a function (i.e., $\#\{b^*(x)\} = 1$, for all $x \in Support(X)$). Let x_0 be a POP for X in the interior of $Support(X)$, with principal direction $b^*(x_0)$. Then there exists a PCOP α in a neighborhood of x_0: there exists a positive ε and a curve $\alpha: (-\varepsilon, \varepsilon) \to I\!R^p$ 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 $\alpha'(0) = \lambda_0 K_0$, where*

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

and $\lambda_0 = b^*(x_0)^t \alpha'(0) \in I\!R$.

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

Corollary 1 *Let us assume that there exists a $\mathcal{C}^1$ curve $\alpha: I \to I\!R^p$ being a PCOP. Then $\alpha'(t) = \lambda(t) K(t)$ for all t in the interior of 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 5. 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 verify that $b^*(\alpha(t)) = \alpha'(t)$: the first principal component of a normal distribution, for instance).

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:

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

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

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 $0 < x < .5$. □

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 through 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 (the proof is direct).

Proposition 2 *Consider a random vector X with finite second moments and density function f in $\mathcal{C}^r$, $r \geq 2$. Assume that hypotheses* A3 *and* A4(K) *are verified for some compact set $K \subset \mathbb{R}^p$. Let $\Gamma(X)$ be the set of POPs for X inside K, which is assumed to be a nonempty set. Assume that for all $x \in \Gamma(X)$ there exists a positive ε, a continuous curve $\alpha_x\colon (-\varepsilon, \varepsilon) \to K$ with $\alpha_x(0) = x$, and an open set $V_x \subseteq K$ such that $V_x \cap \Gamma(X) = \{\alpha_x(s) : s \in (-\varepsilon, \varepsilon)\}$. Then there exists a finite number J of continuous curves $\alpha_j\colon I_j \to K$, $j = 1 \ldots, J$, such that $\Gamma(X) = \cup_{j=1}^{J} \alpha_j(I_j)$.*

3 Principal curves for data sets

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 through 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 $\widetilde{\mu}(x, b)$ 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

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

where $\mathrm{Var}_w(X_i^H, w_i)$ 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 $w = K_h$, the smoothness of $\widetilde{\mu}$ and $\widetilde{\phi}$ as functions of (x, b) depends 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 $\{X_i^H : w_i > 0\}$, assign x to one of these clusters, and use only the points in that cluster to compute $\widetilde{\phi}$ and $\widetilde{\mu}$. We have implemented this last procedure (see Algorithm 2 in Appendix I 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 principal oriented points

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 $\widetilde{\mu}^*$: $\tilde{\Gamma} = \{x \in I\!R^p : x \in \widetilde{\mu}^*(x)\}$. In order to approximate the set $\tilde{\Gamma}$ by a finite set of points, we propose the following procedure.

We randomly choose a point of the sample $X_1, \ldots, X_n$ and call it x_0. Then we 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. If convergence is attained then we include the last x_k in the set of sample POPs $\tilde{\Gamma}$. Repeating m times the previous steps (for a prefixed m) a finite set of sample POPs is obtained.

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. Nevertheless, in all the simulated and real data sets we have examined, we always quickly reached convergence.

Example 3.

We illustrate the performance of this procedure 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 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 elliptic. We use $m = 100$ and obtain the set of sample POPs represented in Figure 1 (upper panel) 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 $(.15, .1)$.

3.2 Finding a principal curve

In the population world, Theorem 2 guarantees that for any POP there exists a PCOP passing through this point. This result leads us to consider the following approach to build a sample 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.

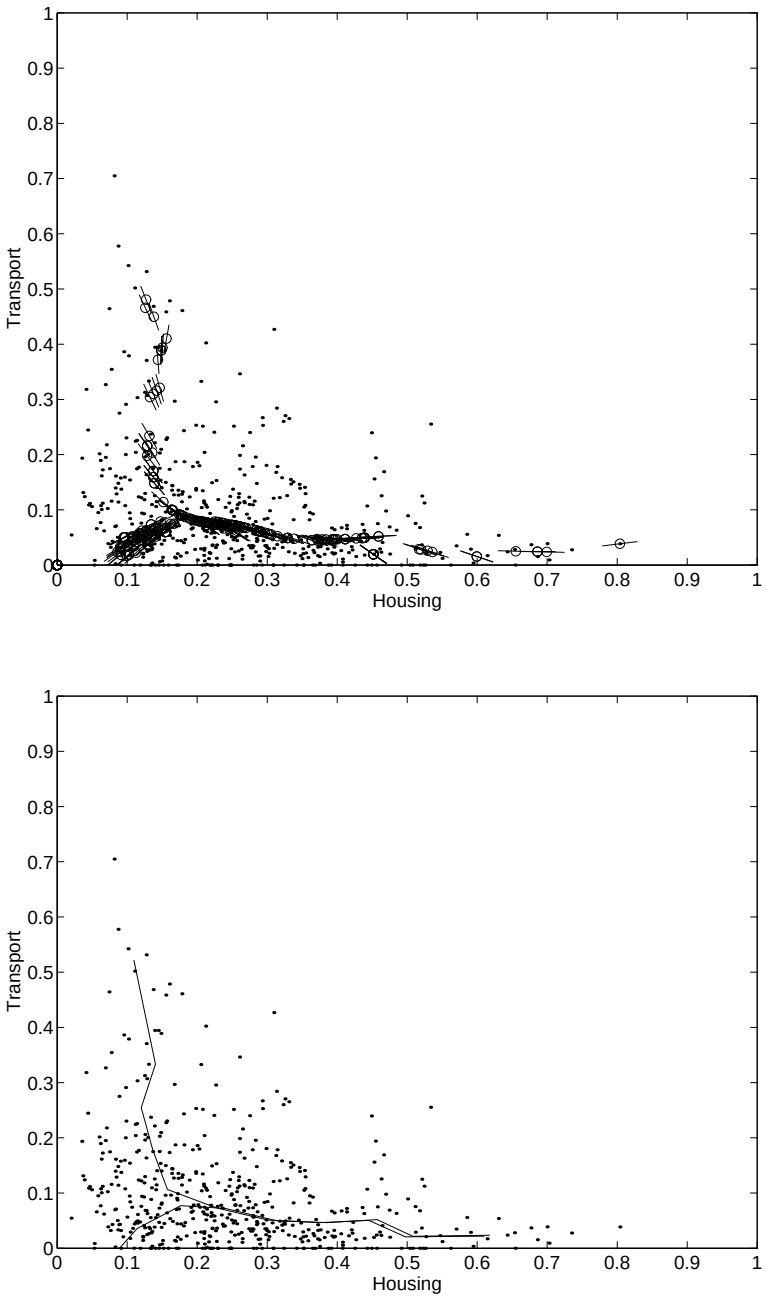

Figure 1: Example 3. Upper panel: principal oriented points for proportions of household expenditure data. Lower panel: two principal curves of oriented points.

We start following procedure described in the previous subsection, 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 of them). 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 $\{x_2^k = \widetilde{\mu}^*(x_2^{k-1}) : k \geq 1\}$, which eventually approaches 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 $-b_1$. Let K be the total number of sample POPs x_k visited by the procedure.

Algorithm 1 in the Appendix I formalizes the whole procedure. 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 estimate many important statistical objects. The density of the induced random variable S on 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 $\widehat{\mathrm{Var}}(S)$ the estimated variance of S. An estimation of the total variance in the normal hyperplane can also be recorded for each s_k: $\widetilde{\phi}(x_k, b_k)$.

Two more 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 $\hat{\alpha}(0)$. 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

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

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

Example 3 (Continuation).

We compute now PCOPs for the households' expenditures data. Some MATLAB routines have been written to implement the Algorithm 1. The Figure 1 (upper panel) suggests that there are more than one curve for this data set. We look for two of them by starting the Algorithm 1 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 1 (lower panel). 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 component does. □

Example 4.

To illustrate the algorithm 1, 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 $\alpha\colon [0,1] \to I\!R^2$, $x = \alpha_1(s) = 2\pi s + 1$, $y = \alpha_2(s) = 2\,(1/x - \cos(x-1))$, $S \sim U(0,1)$ and $Y \sim N(0, \sigma = .4)$. 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. 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

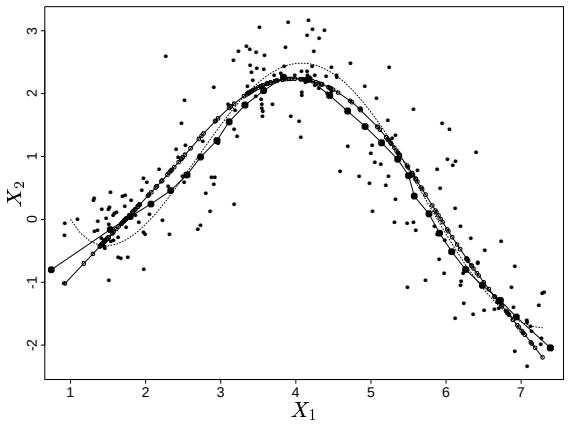

Figure 2: Example 4. PCOP and HSPC for a simulated data set. Dotted line: generating curve. Solid line with empty dots: HSPC. Solid line with big dot marks: PCOP.

can observe that the graphs of both principal curves are 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.

The bandwidth parameter h is 1 and δ is .33. 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 (the generating noise variance is 0.16). So the proportion of the total variability explained by the first principal curve is $p_1 = .98$. □

Example 5.

We replicate 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 3 summarizes the results of the estimation of the first principal curve by our methodology and also by using Hastie and Stuetzle (1989)

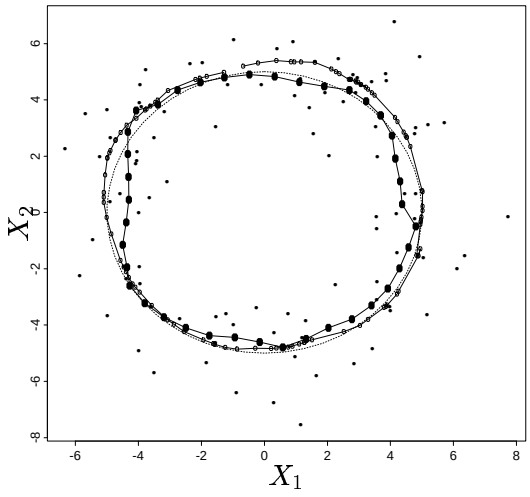

Figure 3: Example 5. Simulated data set around a circle: Original circle (dashed line), HSPC (solid line) and PCOP (solid bold line).

routines. Symbols are as in Figure 2.

The length of the original curve is 10π. When algorithm 1 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 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 4 from two points of view. The estimated curve explains a 92.19% of the total variability along the curve.

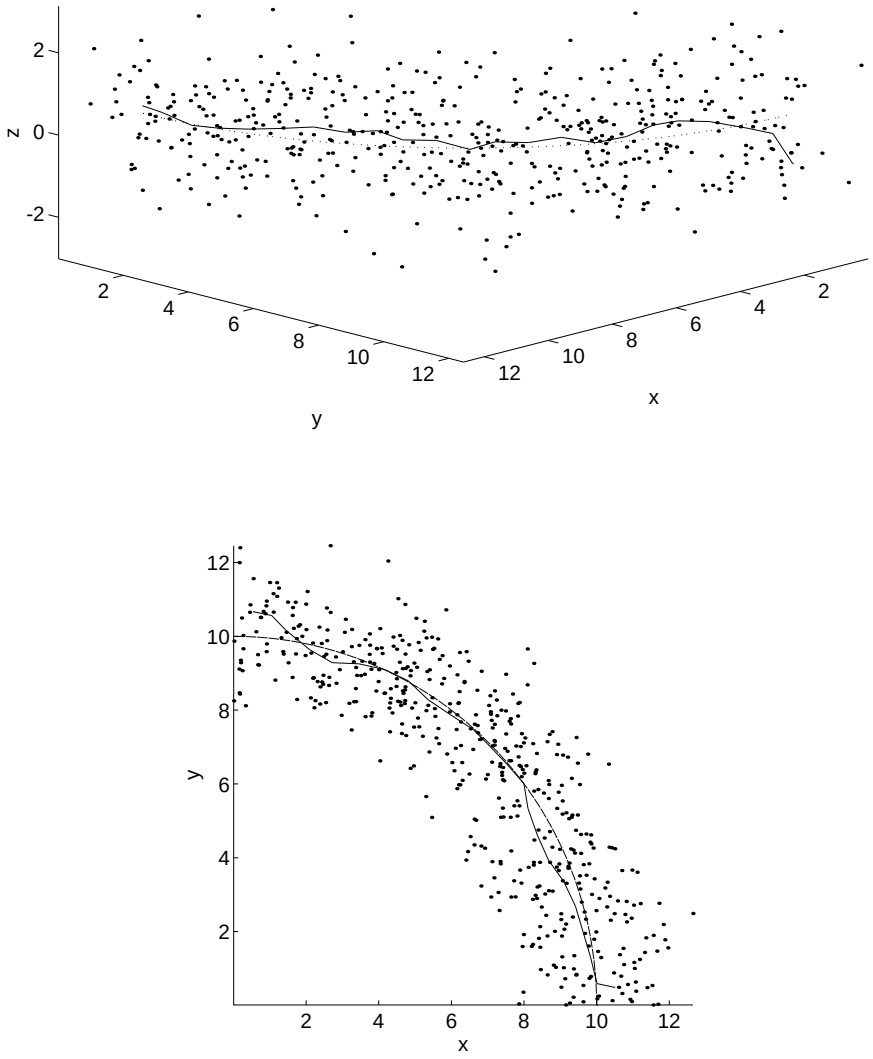

Figure 4: Example 6. Two perspectives of the estimated PCOP (solid line) for the three-dimensional data around a piece of circumference (dotted line).

4 Generalized total variance and higher order principal curves

In subsection 3.2 the total variability of a data set along an estimated curve was defined as $\widehat{TV}_{PCOP} = \widehat{\mathrm{Var}}(S) + \int_I \widetilde{\phi}^*(\alpha(s))\hat{f}_S(s)ds$. If a random variable X has the curve $\alpha: I \to \mathbb{R}^p$ 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 3).

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. 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 $TV[X|X \in H_c(\alpha(s), b^*(\alpha(s)))]$ appropriately measures 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 ellipsoidal, 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 replaced, 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 4

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 $\alpha: \{0\} \to \mathbb{R}$, 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^*, $\Gamma_G(X)$, 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 $\alpha(I) \subseteq \Gamma_G(X)$.*

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 $\bar{f}_S^G(s) = \int_{I\!R^{p-1}} f(\alpha(s) + (b_G^*)_\perp(\alpha(s))v)dv$. The third condition is:*

(iii) *The integral $\nu = \int_I \bar{f}_S^G(s)ds$ is finite and the random variable S with density function $f_S^G(s) = (1/\nu)\bar{f}_S^G(s)$ 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 whole 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 5 *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 5 illustrates 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 local second principal curves should smoothly vary along the first principal curve to allow the estimation. □

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 ellipsoidal 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 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 the second principal curve locally explains at 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.

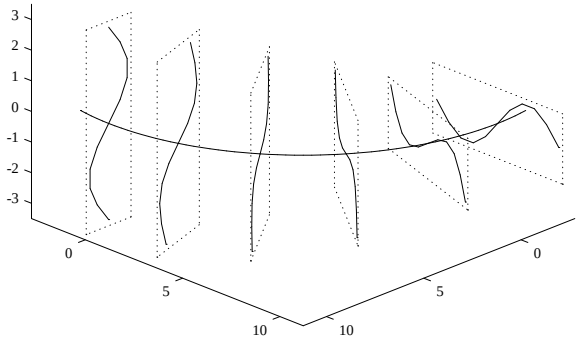

Figure 5: Example 7: Theoretical structure of local second principal curves along the first one.

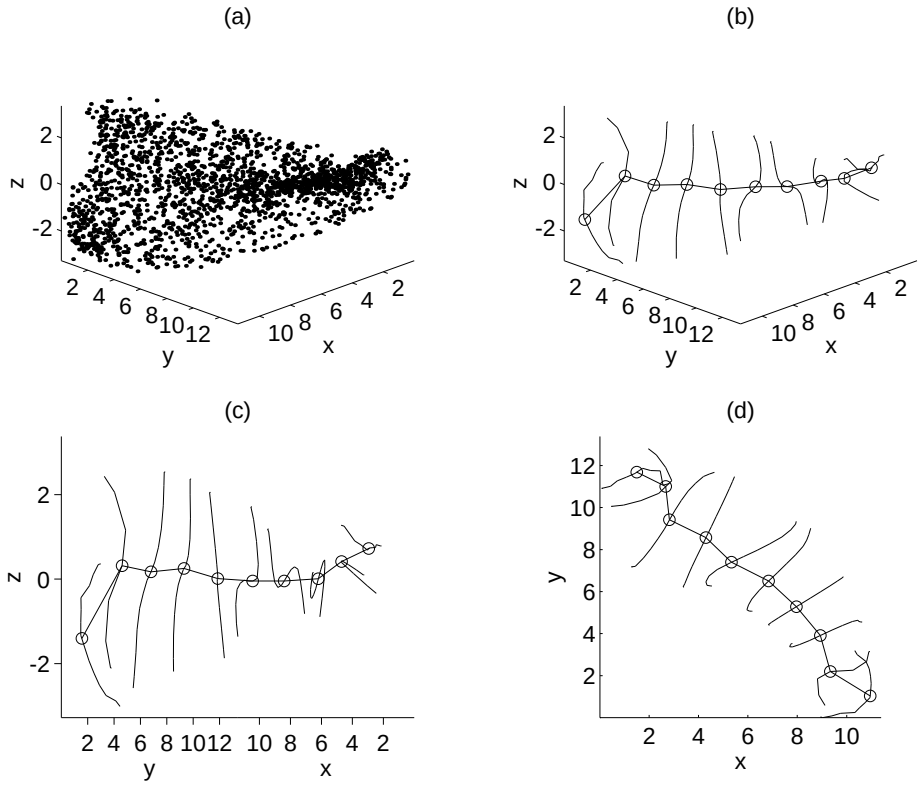

Figure 6: Example 7. Estimation of the first principal curve and the family of local second principal curves along the first one. (a) Data set; (b) first GPCOP and second GPCOPs; (c) same as (b) viewed from a point with zero degrees of elevation over the XY plane; (d) GPCOP system projected over the XY plane.

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 6.

Example 7. (Continuation)

Random data have been generated according to the structure shown in Figure 5. 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 perturbes 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 6 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 of the first principal curve has been introduced, based on the notion of principal oriented points.

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 marginal distributions, these new definitions allow us to define second a higher order local principal curves in a recursive way.

Appendix I: Algorithms

Algorithm 1 (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 $x_k^j = \widetilde{\mu}^*(x_k^{j-1})$ until *convergence*. 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).

□

We present now the algorithm we use to assign x to a cluster in $H(x,b)$. 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 2 (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 being 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 $\bar{d}$.

Appendix II: Proofs

The following result determines the smoothness of μ and ϕ, b^*, μ^* and ϕ^* in terms of the smoothness of f.

Proposition 3 *If f is of class $\mathcal{C}^r$ at x and $\int_{\mathbb{R}^{p-1}} f(x + b_\perp v)dv$ is not equal to zero at (x,b), then μ and ϕ are of class $\mathcal{C}^r$ at (x,b). If (x,b) verifies the previous hypothesis 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.*

Proof. Smoothness properties of μ and ϕ follow as a direct consequence of Fubini's Theorem (see, e.g., Corwin and Szczarba 1979, p. 524). The property concerning ϕ^* is a direct application of the Maximum Theorem (see, e.g., Takayama 1985, p. 254). The Sensitivity Theorem (a corollary of the Implicit Function Theorem; see, e.g., Bertsekas 1995, p.277) permits smoothness properties of b^* to be established, and then the smoothness of μ implies that of μ^*. □

Proof of Proposition 1. The proof follows directly from the next Lemma.

Lemma 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 through x_0. Consider the optimization problems*

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

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

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

Then the solutions to both optimization problems are, respectively,

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

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 $s_0 = (x_0 - \mu)^t v_1$.

Proof. 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

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

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

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

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

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 Theorem 1. The proof is direct because μ^* is a continuous function (Proposition 3) and Brouwer's Fixed Point Theorem applies (see, e.g., Takayama 1985, p. 260). □

Before proving Theorem 2, we need some lemmas.

Lemma 2 *Let $x \in I\!R^p$ and $b \in S^{p-1}$. The partial derivatives of μ are as follows.*

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

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 $(\partial \mu / \partial x)(x,b) = K_x^\mu(x,b) b^t$, where $K_x^\mu(x,b) = (\partial \mu / \partial x)(x,b) b$. 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 and obtain that $b^t \left(I - (\partial \mu / \partial x)(x,b)\right) = 0$. Then the result follows post-multiplying by b: $b^t b = 1 = b^t K_x^\mu(x,b)$.

(ii) 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 we obtain $(\partial\mu/\partial b)(x, b) = K_b^\mu(x, b)\,(I - bb^t)$, where $K_b^\mu(x, b) = (\partial\mu/\partial b)(x, b)$. □

Lemma 3 *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) We obtain that $b^*(x)^t\,((\partial b^*/\partial x)(x)) = 0$, deriving with respect to x the identity $b^*(x)^t b^*(x) = 1$. Therefore $(\partial b^*/\partial x)(x)$ is orthogonal to $b^*(x)$, and we can write that $(I - b^*(x)b^*(x)^t)\,((\partial b^*/\partial x)(x))$ equals $((\partial b^*/\partial x)(x))$.

(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 2, we can deduce that $(\partial b^*/\partial x)(x) = \tilde{K}(x)b^*(x)^t$ for some $\tilde{K}(x) \in I\!R^p$. Now, putting together (1) and (2) the result follows. □

Lemma 4 $(\partial\mu^*/\partial x)(x) = K_x^{\mu^*}(x)b^*(x)^t$, *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 2 and 3, 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 2. □

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

$$\begin{aligned} \Lambda\colon \mathbb{R} \times \mathbb{R}^{p-1} &\to \mathbb{R}^{p-1} \\ (x_1, x^2) &\to \mu^2\left(\binom{x_1}{x^2}, b^*\binom{x_1}{x^2}\right) - x^2 = (\mu^*)^2\binom{x_1}{x^2} - x^2, \end{aligned}$$

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

$$\begin{aligned} \Psi\colon (-\varepsilon, \varepsilon) \subset \mathbb{R} &\to \mathbb{R}^{p-1} \\ t &\to \Psi(t) \end{aligned}$$

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

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

or, equivalently,

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

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

$$\begin{aligned} \alpha\colon (-\varepsilon, \varepsilon) \subset \mathbb{R} &\to \mathbb{R}^p \\ t &\to \alpha(t) = \binom{t}{\Psi(t)} \end{aligned}$$

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

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

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, e.g., 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 4):

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

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 4). Then, $\partial \Lambda / \partial t = (K_x^{\mu^*}(x_0))^2$. Therefore,

$$\frac{\partial \alpha}{\partial t}(0) = \left(\frac{\partial}{\partial t}\begin{pmatrix} t \\ \Psi(t) \end{pmatrix}\right)(0) = \begin{pmatrix} 1 \\ (K_x^{\mu^*}(x_0))^2 \end{pmatrix}.$$

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 4. □

Proof of Corollary 1. 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 $\lambda(t) = b^*(\alpha(t))^t \alpha'(t) \in I\!R$. □

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 (1998). GTM: The generative topographic mapping. *Neural Computation*, **10**, 215–234.

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

Delicado, P. (1998). Principal curves and principal oriented points. Workig Paper 309, Department of Economics, Universitat Pompeu Fabra.

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 (1993). The geometry of principal curves in the plane. Technical Report 250, Department of Statistics, University of Washington.

Duchamp, T. and W. Stuetzle (1995). Geometric properties of principal curves in the plane. In *Robust Statistics, Data Analysis, and Computer Intensive Methods*, pp. 135–152. Springer, Berlin.

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.

Johnson, R. A. and D. W. Wichern (1992). *Applied Multivariate Statistical Analysis (Third Edition)*. Prentice-Hall.

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.

Salinelli, E. (1998). Nonlinear principal components i. absolutely continuous random variables with positive bounded densities. *The Annals of Statistics*, **26**, 596–616.

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.