Mixture of nonlinear models: a Bayesian fit for Principal Curves

Pedro Delicado and Marcelo Smrekar

***Abstract*— Principal curves are smooth parametric curves passing through the "middle" of a non-elliptical multivariate data set. We model the probability distribution of this kind of data as a mixture of simple nonlinear models and use MCMC techniques to fit the mixture model.**

I. Introduction

Principal curves were introduced by [1] as smooth parametric curves passing through the "middle" of a multidimensional data set. Several works on principal curves have appeared since then (see [2], [3], [4] and the references in there for a broader view to the principal curves literature).

A fruitful way to model principal curves is the mixture of multivariate normal random variables: [5] estimates a mixture of normals with as many components as observed data; [6] suggest a mixture with a fixed number of components and each of them is fitted by using principal component analysis; [7] generalize the work of [6] allowing the model noise to be orthogonal to the principal curve. A common feature of these papers is that they use the EM algorithm for parameter estimation.

We propose to model p-dimensional distributions around a curve as mixtures of simple nonlinear models. The main advantage of our proposal is that the number of required components is lower than when normal models are used in the mixture: a single nonlinear component may well produce a similar fitting than that given by the mixture of three or four normal components.

In Section II we introduce the simple model (that we call *single arch model*) and we propose to take a Bayesian approach to fit it. The Gibbs sampler algorithm is used to obtain samples from posterior distributions of the parameters given the data. A mixture model with nonlinear components is estimated using latent component indicator variables (see [8], for instance). This is developed in Section III, which finishes with an illustrative example. The paper ends with a list of open problems requiring additional attention.

II. Preliminaries. The single arch model

The simple model we put forward is that followed by random points X_i scattered around an arch of circumference with radius ρ. We call it *the single arch model*. Let $\alpha(s)$, $s \in [-\pi\rho, \pi\rho]$, be the usual parametric equation for that circumference with unit speed (that is, $\|\alpha'(s)\| = 1$). The value S_i such that the orthogonal projection of points to the circumference is $\alpha(S_i)$, is univariate normal, while the orthogonal differences $Y_i = X_i - \alpha(S_i)$ are $(p-1)$-dimensional normal, independent of S_i. Figure 1 helps to identify these elements.

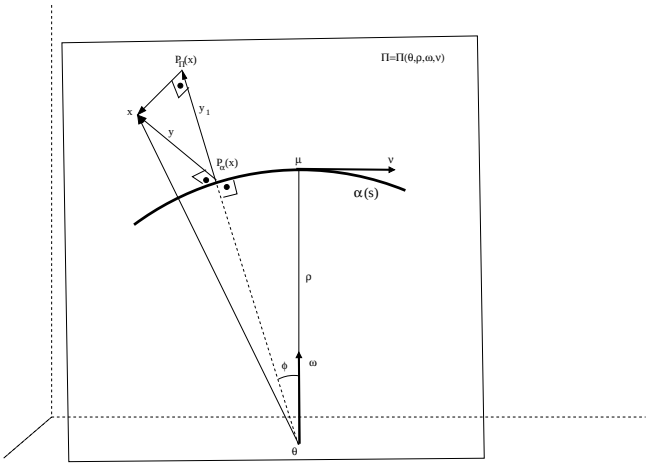

Fig. 1. Illustration of parameters and other elements involved in the single arch model.

A. The likelihood

This model can be obtained applying a one-to-one application χ from $\mathbb{R} \times \mathbb{R}^{p-1}$ to $\mathbb{R}^p$. Let $I = [-\pi\rho, \pi\rho] \subset \mathbb{R}$, where ρ is the radius of the circumference. Let S be a zero-mean one-dimensional random variable with $P(S \in I) = 1$ with distribution parameterized by a scale parameter $(\rho\sigma_S)$. Let Y be a zero-mean $(p-1)$-dimensional random variable such that $P(\|Y\| \leq \rho) = 1$ with spheric distribution parameterized by a scale parameter $(\rho\sigma_Y)$. We assume that S and Y are independent. Consider the joint distribution of (S, Y) on $\mathbb{R} \times \mathbb{R}^{p-1}$, with density $f_0(s, y) = f_S(s) f_Y(y)$.

Let α be a parameterized circumference in $\mathbb{R}^p$ with center θ and radius ρ,

$$\alpha \equiv \{\alpha(s) = (\alpha_1(s), \ldots, \alpha_p(s)) : s \in [-\pi\rho, \pi\rho)\},$$

such that $\|\alpha(s) - \theta\|^2 = \rho^2$ for all s and, for all s, $\alpha(s)$ belongs to the plane Π, defined by the points θ, $\mu = \alpha(0)$ and the speed vector $\nu = \alpha'(0)$. It is assumed that α is parameterized by the arc length (so $\|\alpha'(s)\| = 1$, for all s, and then $\|\nu\| = 1$). Let $\omega = (\mu - \theta)/\rho$. Then $\|\omega\| = 1$ and Π is also given by θ, ω and ν.

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

Pedro Delicado is with the Departament d'Estadstica i Investigaci Operativa, Universitat Politècnica de Catalunya Despatx 214, Edifici C5, Campus Nord, C/ Jordi Girona 1-3, 08034 Barcelona, Spain (phone: (34) 93 401 5698; fax: (34) 93 401 5855; email: pedro.delicado@upc.edu). Marcelo Smrekar is PhD student.

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, [9] 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 $\alpha(s)$; given that the curve we are considering here is a circumference, the radius of curvature is constant and equal to ρ), 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; in our case the center of curvature is the center of the circumference θ).

We consider the function χ defined by [1] 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 mapping $I \times I\!R^{p-1}$ into $\cup_{s\in I} H_s \subseteq I\!R^p$ so that $\chi(s, y) = \alpha(s) + A(s)(0, y^t)^t$. Thus χ put (s, y) in H_s in a differentiable way with respect to s and χ 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.

Consider the random variables in $I\!R^p$ obtained as $X = \chi(S, Y)$. The distribution of X has the following parameters: the center of the circumference θ (it is a *location parameter*); the radius of the circumference ρ (it is a *scale parameter*); the two unit vectors ω and ν that, jointly with θ, determine the plane Π where the circumference belongs to; the scale parameter of orthogonal projections, σ_S; the scale parameter of orthogonal distances, σ_Y. Observe that parameters σ_S and σ_Y are not scale parameters of the whole distribution, despite the fact that they are scale parameters of S and Y, respectively.

The expression for the density of X_i, given the parameters, is a special case of Lemma 1, stated at the Appendix. We need some notations to be able to write the density of X. Let χ^{-1} be the inverse function of χ. Let $s = \chi_s^{-1}(x) \in I\!R$ be the first component of $\chi^{-1}(x)$. Let $y = \chi_y^{-1}(x) \in I\!R^{p-1}$ be the remaining $(p-1)$ components of $\chi^{-1}(x)$, and let y_1 be the first component of y. Then

$$f_X(x|\theta, \rho, \omega, \nu, \sigma_S, \sigma_Y) = f_S(s|\rho\sigma_S)\frac{f_Y(y|\rho\sigma_Y)}{1-(y_1/\rho)}. \quad (1)$$

Observe that f_X at x depends on parameters $(\theta, \rho, \omega, \nu)$ because $(s, y) = \chi^{-1}(x)$ depends on them. Let us note that $f_Y(y|\sigma_Y^2)$ depends on y only by $\|y\|^2$ because Y is assumed to be spheric.

So we need to derive the explicit expressions of s, y_1 and $\|y\|^2$ to have the full expression of $f_X(x|\theta, \rho, \omega, \nu, \sigma_S^2, \sigma_Y^2)$. Figure 1 helps to find these expressions. Let x be a point in $I\!R^p$. Let $P_\Pi(x)$ and $P_\alpha(x)$ be the projections of x onto Π and α, respectively. Then $P_\Pi(x) = \theta + C(C'C)^{-1}C'(x-\theta)$, where $C = (\omega, \nu)$, and $P_\alpha(x) = \theta + \rho(P_\Pi(x) - \theta)/\|P_\Pi(x) - \theta\|$. Pythagoras' Theorem tells us that $(y_1+\rho)^2 + \|x - P_\Pi(x)\|^2 = \|x - \theta\|^2$, so

$$y_1 = -\rho + \sqrt{\|x-\theta\|^2 - \|x - P_\Pi(x)\|^2}.$$

Applying again Pythagoras' Theorem we have that $\|y\|^2 = y_1^2 + \|x - P_\Pi(x)\|^2$. Now we deal with s. Observe that s is the distance (over the curve α) between $P_\alpha(x)$ and $\mu = \theta + \rho\omega$. Then,

$$s = \rho\phi, \text{ where } \phi = \cos^{-1}\left(\frac{\omega'(P_\alpha(x) - \theta)}{\rho}\right).$$

For a set of n i.i.d. data $\mathbf{x} = \{x_i, \ldots, x_n\}$ the likelihood function is

$$f(\mathbf{x}|\theta, \rho, \omega, \nu, \sigma_S, \sigma_Y) = \prod_{i=1}^{n} f_S(s_i|\rho\sigma_S)\frac{f_Y(y_i|\rho\sigma_Y)1}{1-(y_{i1}/\rho)}, \quad (2)$$

where $s_i = \chi_s^{-1}(x_i)$, $y_i = \chi_y^{-1}(x_i)$ and y_{i1} is the first component of y_i.

To fix ideas, from now on we consider the model based on normality of S and Y. Let $S \sim N(0, (\rho\sigma_S)^2)$ and $Y \sim N_{p-1}(0, (\rho\sigma_Y)^2 I_{p-1})$ We assume that σ_S is such that $P(|S| \le \rho\pi) \approx 1$ and that σ_Y is such that $P(\|Y\| \le \rho) \approx 1$. To be specific, we impose $\sigma_S^2 < \pi/\chi^2_{1,.999} = \pi/(3.29)^2$ and $\sigma_Y^2 < 1/\chi^2_{p-1,.999}$. Observe that strictly speaking the assumptions stated at the beginning of this subsection on the distributions of S and Y are not satisfied by these normal distributions, and then equation (1) is an approximated result, but not the true expression of $f_X(x|\theta, \rho, \omega, \nu, \sigma_S, \sigma_Y)$ in this case. Nevertheless such approximation works well in practice.

B. Prior distribution

We consider the six parameters to be independent a priori. There are a location parameter (θ) and a scale parameter (ρ). Using Jeffreys priors (see [10], for instance) their joint contribution to the prior distribution is proportional to $1/\rho$. The two unit vectors ω and ν belongs to the compact set S_{p-1}. We take there a flat prior. We deal with parameters σ_S and σ_Y as if they really were scale parameters and, using again Jeffreys priors, they contribute to the prior with $1/\sigma_S$ and $1/\sigma_Y$, respectively. In summary, we take the prior

$$\pi(\theta, \rho, \omega, \nu, \sigma_S, \sigma_Y) \propto (\sigma_S\sigma_Y\rho)^{-1}. \quad (3)$$

C. Posterior and full conditional distributions

The posterior distribution can not be explicitly written, because the likelihood function (2) depends on the parameters in a very complicated way (through the inverse function χ^{-1}).

Nevertheless two of the full conditional distributions are easily derived. First,

$$\sigma_S^2|\theta, \rho, \omega, \nu, \sigma_Y, \mathbf{x} \equiv \sigma_S^2|\mathbf{s} \sim \mathcal{IG}\left(\frac{n}{2}, \frac{1}{2\rho^2}\sum_{i=1}^{n} s_i^2\right),$$

where $\mathbf{s} = \{s_1, \ldots, s_n\}$. Second, defining $\mathbf{y} = \{\|y_1\|^2, \ldots, \|y_n\|^2\}$,

$$\sigma_Y^2|\theta, \rho, \omega, \nu, \sigma_Y, \mathbf{x} \equiv \sigma_Y^2|\mathbf{y} \sim$$
$$\mathcal{IG}\left(\frac{n(p-1)}{2}, \frac{1}{2\rho^2}\sum_{i=1}^{n}\|y_i\|^2\right).$$

The remaining full conditional distributions are not so easily derived. We propose to use a Metropolis-Hastings algorithm

to approach $\pi(\theta,\rho,\omega,\nu|\sigma_S,\sigma_Y,\mathbf{x})$. We take a random walk proposal. After step m of the algorithm, we define $\mu^{(m)} = \theta^{(m)} + \rho^{(m)}\omega^{(m)}$ and generate

$$\xi_\theta = \theta^{(m)} + \varepsilon_\theta,\ \xi_\mu = \mu^{(m)} + \varepsilon_\mu,$$

$$\xi_\nu = (\nu^{(m)} + \varepsilon_\nu)/\|\nu^{(m)} + \varepsilon_\nu\|,$$

where ε_θ, ε_μ y ε_ν are p-dimensional independent normal random variables, centered at 0, and with variances $\sigma_\theta^2 I$, $\sigma_\mu^2 I$, $\sigma_\nu^2 I$, respectively. We define

$$\xi_\rho = \|\xi_\mu - \xi_\theta\|,\ \xi_\omega = (\xi_\mu - \xi_\theta)/\xi_\rho.$$

Then, the value δ is defined as

$$\delta = \min\left\{\frac{\pi(\xi_\theta,\xi_\rho,\xi_\omega,\xi_\nu|\sigma_S,\sigma_Y,\mathbf{x})}{\pi(\theta^{(m)},\rho^{(m)},\omega^{(m)},\nu^{(m)}|\sigma_S,\sigma_Y,\mathbf{x})},1\right\}.$$

Finally, we take

$$(\theta^{(m+1)},\rho^{(m+1)},\omega^{(m+1)},\nu^{(m+1)}) = \tag{4}$$

$$Z\times(\xi_\theta,\xi_\rho,\xi_\omega,\xi_\nu) + (1-Z)\times(\theta^{(m)},\rho^{(m)},\omega^{(m)},\nu^{(m)}),$$

where $Z\sim$ Bernoulli(δ), independent of other random variables. The values σ_θ, σ_μ and σ_ν are calibrated so that the average acceptance rate is roughly $1/4$, as recommended by [11]. Observe that

$$\pi(\theta,\rho,\omega,\nu|\sigma_S,\sigma_Y,\mathbf{x}) \propto$$

$$f(\mathbf{x}|\theta,\rho,\omega,\nu,\sigma_S,\sigma_Y)\pi(\theta,\rho,\omega,\nu|\sigma_S,\sigma_Y) =$$

$$f(\mathbf{x}|\theta,\rho,\omega,\nu,\sigma_S,\sigma_Y)\pi(\theta,\rho,\omega,\nu) \propto$$

$$f(\mathbf{x}|\theta,\rho,\omega,\nu,\sigma_S,\sigma_Y)/\rho.$$

Then

$$\delta = \min\left\{\frac{f(\mathbf{x}|\xi_\theta,\xi_\rho,\xi_\omega,\xi_\nu,\sigma_S,\sigma_Y)\rho^{(m)}}{f(\mathbf{x}|\theta^{(m)},\rho^{(m)},\omega^{(m)},\nu^{(m)},\sigma_S,\sigma_Y)\xi_\rho},1\right\}.$$

The likelihood values are computed as equation (2) indicates.

D. Estimating the single arch model by Gibbs sampling

The step t for the Gibbs sampler approaching the distribution $\pi(\theta,\rho,\omega,\nu,\sigma_S,\sigma_Y|\mathbf{x})$ is as follows:

1) Generate $\sigma_S^{2(t)} \sim \mathcal{IG}\left(\frac{n}{2}, \frac{1}{2(\rho^{(t-1)})^2}\sum_{i=1}^n s_i^2\right)$, where $\mathbf{s} = \{s_i,\ldots,s_n\}$ is computed using parameter values $(\theta^{(t-1)},\rho^{(t-1)},\omega^{(t-1)},\nu^{(t-1)})$.
2) Generate $\sigma_Y^{2(t)} \sim \mathcal{IG}\left(\frac{n(p-1)}{2}, \frac{1}{2(\rho^{(t-1)})^2}\sum_{i=1}^n \|y_i\|^2\right)$, where parameter values $(\theta^{(t-1)},\rho^{(t-1)},\omega^{(t-1)},\nu^{(t-1)})$ are used to compute $\mathbf{y} = \{\|y_1\|^2,\ldots,\|y_n\|^2\}$.
3) Use the Metropolis-Hastings algorithm described in (4) to generate $\theta^{(t)},\rho^{(t)},\omega^{(t)},\nu^{(t)} \sim \pi(\theta,\rho,\omega,\nu|\sigma_S^{(t)},\sigma_Y^{(t)},\mathbf{x})$.

Observe that in Step 3 it is enough to do only one iteration of the Metropolis-Hastings algorithm (4). The reason is that only one iteration of Step 3 is required to prove that the joint posterior distribution of parameters is the stationary distribution of the previous MCMC algorithm.

III. Main Results. The mixture model

Consider now a p-dimensional data set distributed around a one-dimensional curve. We propose to model these data as a mixture a k single arch models as those described in Section II. That is, we assume that the data come from a random variable X with density

$$\sum_{j=1}^k p_j f(x|\theta_j,\rho_j,\omega_j,\nu_j,\sigma_{S,j},\sigma_{Y,j}),$$

where $p_j > 0$ and $\sum p_j = 1$. Let $\mathbf{p} = (p_1,\ldots,p_k)$. In our analysis we follow the Section 6.4 in [8]. In particular, we assume that the number k of mixture components is known.

A set of latent vectors $z_i \in \{0,1\}^k$, $i = 1,\ldots,n$, are introduced. They are the so called *component indicator vectors*: for all $i = 1,\ldots,n$, $\sum_{j=1}^k z_{ij} = 1$ and $z_{ij} = 1$ if and only if $x_i \sim f(x|\theta_j,\rho_j,\omega_j,\nu_j,\sigma_{S,j},\sigma_{Y,j})$. The distribution of X can also be modeled as follows:

$$z_i|\{\theta_j,\rho_j,\omega_j,\nu_j,\sigma_{S,j},\sigma_{Y,j},p_j : j = 1\ldots k\}$$

$$\equiv z_i|\mathbf{p} \sim \mathcal{M}_k(1;\mathbf{p}),$$

$$x_i|\,(\{z_j, l = 1\ldots n\},$$

$$\{\theta_j,\rho_j,\omega_j,\nu_j,\sigma_{S,j},\sigma_{Y,j},p_j : j = 1\ldots k\}) \equiv$$

$$x_i|z_i,\{\theta_j,\rho_j,\omega_j,\nu_j,\sigma_{S,j},\sigma_{Y,j} : j = 1\ldots k\} \sim$$

$$f(x|\prod_{j=1}^k \theta_j^{z_{ij}},\prod_{j=1}^k \rho_j^{z_{ij}},\prod_{j=1}^k \omega_j^{z_{ij}},\prod_{j=1}^k \nu_j^{z_{ij}},\prod_{j=1}^k \sigma_{S,j}^{z_{ij}},\prod_{j=1}^k \sigma_{Y,j}^{z_{ij}}).$$

A. Prior distribution

We consider that parameter $\mathbf{p}$ is a priori independent of the others. Moreover we take $p \sim$ Dirichlet$(\alpha_1,\ldots,\alpha_k)$. We take equal α_j's with $\alpha_0 = \sum_{j=1}^k \alpha_j$ small, in order to not introduce much prior information. For instance, we can take $\alpha_j = 1/k$. The prior distributions of $\theta_j,\rho_j,\omega_j,\nu_j,\sigma_{S,j},\sigma_{Y,j}$ and $\theta_h,\rho_h,\omega_h,\nu_h,\sigma_{S,h},\sigma_{Y,h}$ are considered independent if $j \neq h$. Each of them is taken as indicated at Section II for the single arch model.

B. Full conditional distributions

For $j = 1,\ldots,k$ let $\mathbf{x}_j(\mathbf{z}) = \{x_i, i \in \{1,\ldots,n\} : z_{ij} = 1\}$, that is, the set of observations that, according to indicators z_{ij}, were generated by the j-th mixture component. Let $\Psi_j = (\theta_j,\rho_j,\omega_j,\nu_j,\sigma_{S,j},\sigma_{Y,j})$ the set of parameters identifying the j-th curve, $j = 1,\ldots,k$. With these definitions we have that

$$\Psi_j|(\Psi_h, h\neq j,\mathbf{p},\mathbf{z},\mathbf{x}) \equiv \Psi_j|\mathbf{x}_j(\mathbf{z}),\ \text{ for } j = 1,\ldots,k.$$

This distribution has been studied at Section II. Moreover, $z_i|(\Psi_j, j = 1,\ldots,k,\mathbf{p},\mathbf{x})$ is multinomial,

$$\mathcal{M}_k(1;p_1(x_i,\mathbf{p},\Psi_1),\ldots,p_k(x_i,\mathbf{p},\Psi_k)),$$

with

$$p_j(x_i,\mathbf{p},\Psi_j) = \frac{p_j f(x_i|\Psi_j)}{\sum_{h=1}^k p_h f(x|\Psi_h)}.$$

Finally,

$$\mathbf{p}|(\mathbf{x}, \mathbf{z}, \Psi_j, j = 1, \ldots, k) \equiv \mathbf{p}|\mathbf{z} \sim$$

$$\text{Dirichlet}(\alpha_1 + \sum_{i=1}^{n} z_{i1}, \ldots, \alpha_k + \sum_{i=1}^{n} z_{ik}).$$

C. Estimation of the mixture model by Gibbs sampling

The t-th step of a Gibbs sampler approaching the distribution of parameters (curve parameters, $\mathbf{z}$, $\mathbf{p}$) given the data $\mathbf{x}$ is as follows:

1) Generate $\Psi_j^{(t)} = (\theta_j^{(t)}, \rho_j^{(t)}, \omega_j^{(t)}, \nu_j^{(t)}, \sigma_{S,j}^{(t)}, \sigma_{Y,j}^{(t)})$ for $j = 1, \ldots, k$, following Sections II and III-B, and using $\mathbf{z}^{(t-1)}$ as component indicators.
2) Generate

$$z_i \sim \mathcal{M}_k(1; p_1(x_i, \mathbf{p}^{(t-1)}, \Psi_1^{(t)}), \ldots, p_k(x_i, \mathbf{p}^{(t-1)}, \Psi_k^{(t)}))$$

for $i = 1, \ldots, n$.
3) Generate $\mathbf{p}^{(t)} \sim$ Dirichlet$(\alpha_1 + \sum_{i=1}^{n} z_{i1}^{(t)}, \ldots, \alpha_k + \sum_{i=1}^{n} z_{ik}^{(t)})$.

IV. Simulation Results

We fit the mixture of arch models to a simulated data set with 100 two-dimensional points. The mixture model that generates the data has $k = 2$ components and these parameters:

$$p_1 = 0.5,\ \theta_1 = (0, -1),\ \rho_1 = 1,\ \omega_1 = (-1, 0),$$

$$\sigma_{S,1} = 0.75,\ \sigma_{Y,1} = 0.2,$$

$$p_2 = 0.5,\ \theta_2 = (0, 1),\ \rho_2 = 1,\ \omega_2 = (1, 0),$$

$$\sigma_{S,2} = 0.75,\ \sigma_{Y,2} = 0.2.$$

Observe that we do not need to specify parameters ν_1 and ν_2 because the plane Π is just the plane R^2. Figure 3 (upper panel) shows the data and the generating model.

The Gibbs sampling algorithm described above was used to simulate from the posterior distribution. A Matlab (Version 7) code was written for this purpose. The algorithm is left to run 6000 iterations. It takes 1 minute and 35 seconds in a Pentium 4 (CPU 2.8 GHz). Table I shows some posterior statistics calculated from the last third of simulated posterior values. Similar results were obtained using WinBUGS. The Gibbs sampling algorithm development is summarized at Figure 2, where the evolution of the cumulative mean for simulated parameters values is shown. Figure 3 compares the true model with the estimated model (under quadratic loss).

V. Conclusions and further research

We have presented a Bayesian framework for modelling data distributed around a principal curve. This approach reduces the required number of mixture components (compare with fitting a mixture of normal distributions). Moreover it allows us to face interesting statistical questions beyond the determination of point scores over the curve. The following aspects require additional attention:

TABLE I

Results from the estimation of the mixture of arch models. 6000 values from the posterior distribution were generated by Gibbs sampling. Posterior statistics were calculated using the last third of simulated values.

Parameter	True values	Posterior mean	Posterior Std. Dev.
p_1	.5	.5377	.0563
p_2	.5	.4623	.0563
θ_1	(0,-1)	(.1052,-1.1003)	(.1231,.0568)
θ_2	(0,1)	(.0620,1.0019)	(.1295,.0553)
ρ_1	1	1.0672	.0998
ρ_2	1	.9436	.1035
ω_1	(-1,0)	(-.9504,.2829)	(.0364,.1241)
ω_2	(1,0)	(.9796,-.1177)	(.0269,.1604)
μ_1	(-1,-1)	(-.9092,-.7990)	(.0582,.1272)
μ_2	(1,1)	(.9866,.8899)	(.0461,.1436)
$\sigma_{S_1} + \sigma_{S_2}$	1.5	1.5008	.1473
$.5(\sigma_{Y_1} + \sigma_{Y_2})$	.2	.1878	.0202

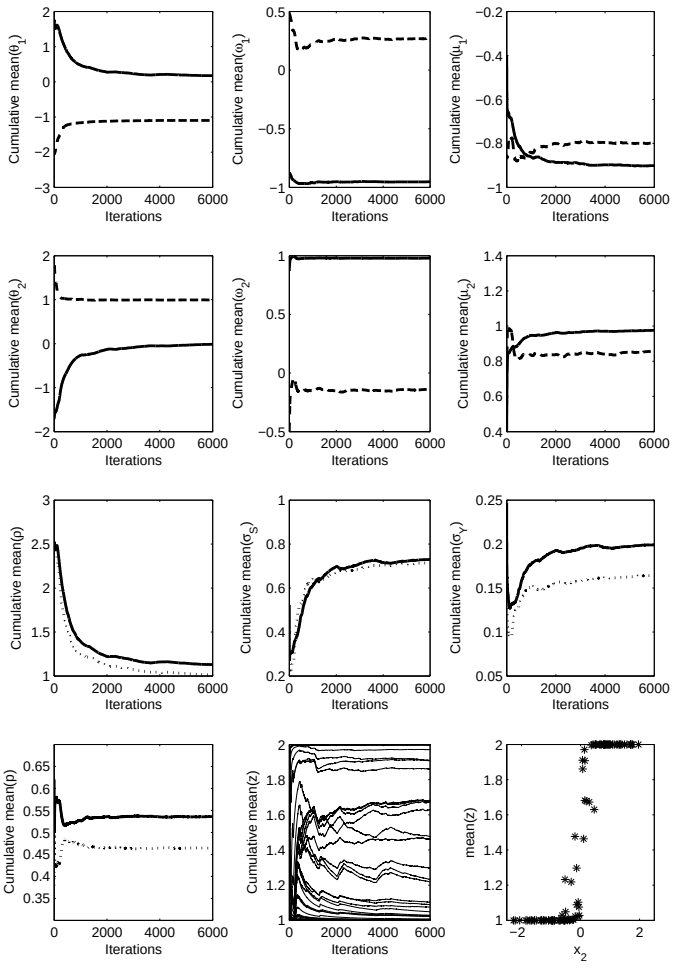

Fig. 2. Cumulative mean evolution for the simulated parameter values. In the first and second row, corresponding respectively to populations 1 and 2, solid and dashed lines represent the first and second (respectively) coordinates of the 2-dimensional vectors. In other panels, solid lines represent population 1 and dotted lines correspond to population 2.

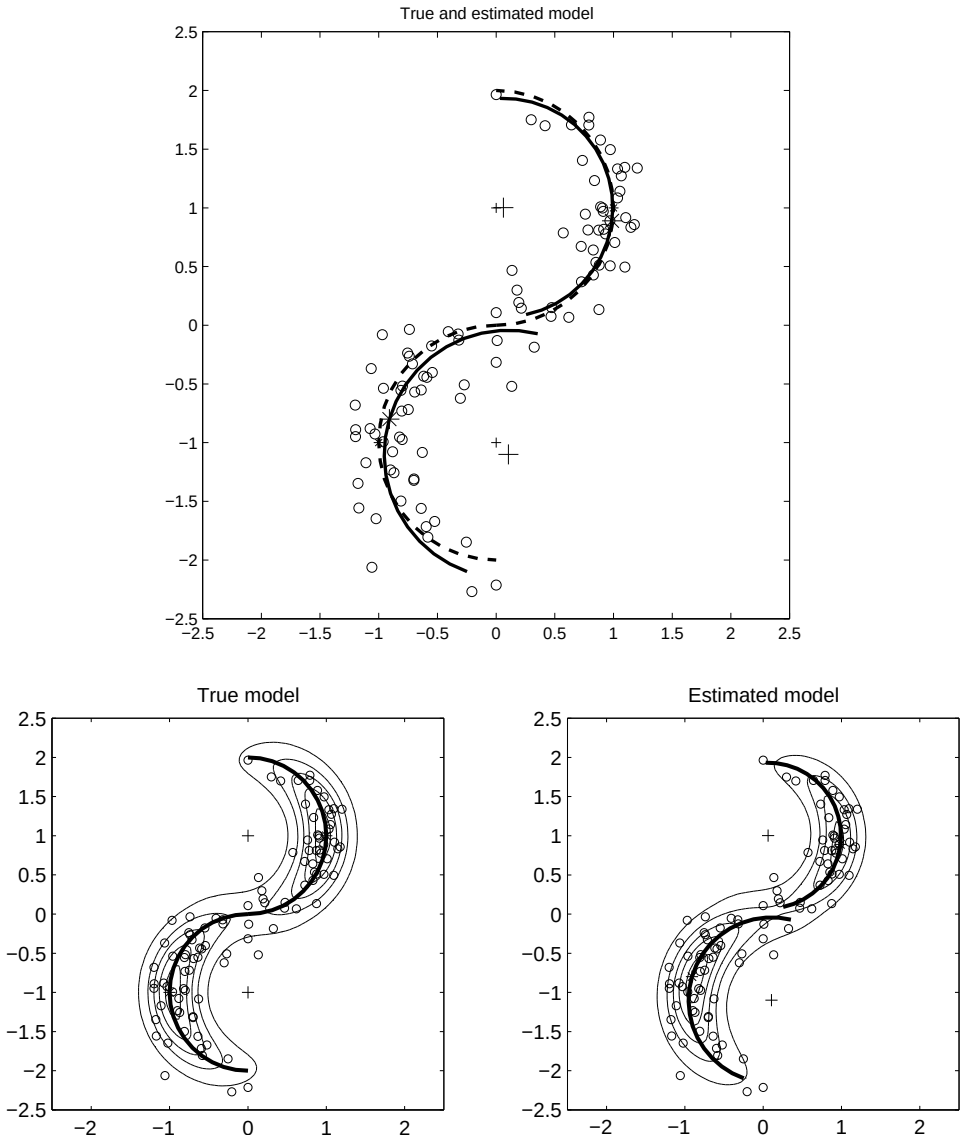

Fig. 3. True versus estimated model. The upper panel represents the true model (dashed lines, small + and * symbols) and the estimated model (solid lines, big + and * symbols). The observed data are represented as empty dots. The lower panels include the contour level curves for the true data density (left) and for the estimated data density (right).

- Analyzing more carefully the choice of the prior distribution. An open question is to find out whether or not the Jeffreys prior for the whole model is in fact proportional to our proposal in (3). The main difficulty (maybe insuperable) is the dependence of $f_X(x|\theta, \rho, \omega, \nu, \sigma_S, \sigma_Y)$ on $\theta, \rho, \omega, \nu$ through $\chi^{-1}(x)$.
- Allowing the number of mixture components k to be unknown. The ideas of reversible jump ([12]) or those of birth-and-death MCMC ([13]) are valid for deal with this problem. See also the Chapter 11 of [14].
- Allowing S to have distribution different from normal in each mixture component (we think that the uniform distribution would also be appropriate). The main effect of this change is that the full conditional distributions of σ_S^2 and σ_Y^2 are no longer Inverted Gamma. Steps 1 and 2 in the Gibbs sampling algorithm described in Subsection II-D might be replaced by Metropolis-Hastings steps.
- Extending the single arch model replacing the arch of circumference by a more flexible arch of ellipse. Some additional parameters would appear, but the level of difficulty is similar to the case we have studied in this paper.
- Extending the implemented application from dimension 2 to the general p-dimensional case. This extension is straightforward.
- Extracting a unifying principal curve from the density function estimated as a mixture of single arch model densities.

APPENDIX

Lemma 1: Assume that $I = \text{Support}(S)$ is a compact interval, and that the distributions $Y|S = s$ have convex compact support contained in the ball $B(0, \rho(s))$, where $\rho(s)$ is the curvature radius of α at the point $\alpha(s)$. Then the function $\chi\colon \text{Support}(S,Y) \to \text{Support}(X)$ is a homeomorfism. Moreover, the density function of X at a given point $x \in \text{Support}(X)$ is

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

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

Proof. The proof of this result is based on change of variable standard techniques. Observe that $\text{Support}(Y|S = s) \subseteq B(0, \rho(s))$ and that $H_c(\alpha(s), \alpha'(s)) \cap H_c(\alpha(t), \alpha'(t)) = \emptyset$ when $s \neq t$. Then χ is a 1-1 function form $\text{Support}(S,Y)$ to the image of this set. As χ is continuous and it is defined on a compact set, it follows that $\chi(\text{Support}(S,Y)) = \text{Support}(\chi(S,Y))$. Then χ is a homeomorfism because it is a 1-1 continuous function defined from a compact set to a metric space.

Remember that $\chi(s, y) = \alpha(s) + A(s)(0, y^t)^t$, 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, [9], pp. 158-160). As χ is 1-1, we call $(s(x), y(x)) = \chi^{-1}(x)$, for a given $x \in \text{Support}(X)$, where $X = \chi(X)$.

Applying change of variable standard techniques, the density function of X at a given x can be computed as $f_X(x) = f_{(S,Y)}(s(x), y(x))(\det(J_\chi(s(x), y(x))^{-1}$, where $J_\chi(s(x), y(x))$ is the Jacobian of χ at x, that is to say the $p \times p$ matrix

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

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 $\det(J_\chi(s, y)) = \det(\alpha'(s) + A_2'(s)y, A_2(s)) =$

$$\det(\alpha'(s), A_2(s)) + \det(A_2'(s)y, A_2(s)) =$$

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

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:

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

$k_j(s)$ 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 $\alpha''(s) = k_1(t)a_2(s)$, $a_2'(s) = -k_1(t)\alpha'(s) + k_2(s)a_3(s)$, $a_j'(s) = -k_{j-1}(t)a_{j-1}(s) + k_j(s)a_{j+1}(s)$ for $j = 3, \dots, p-1$, and $a_p'(s) = -k_{p-1}(t)a_{p-1}(s)$. Then, for $3 \leq j \leq p-1$ we have

$$\det(a_j'(s), A_2(s)) =$$

$$\det(-k_{j-1}(t)a_{j-1}(s) + k_j(s)a_{j+1}(s), (a_2(s), \dots, a_p(s)))$$

that is equal to 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 $j = 2$ $\det(a_2'(s), A_2(s)) = \det(-k_1(t)\alpha'(s) + k_2(s)a_3(s), (a_2(s), \dots, a_p(s))) =$

$$(-k_1(s)) \det(\alpha'(s), (a_2(s), \dots, a_p(s))) =$$

$$(-k_1(s)) \det(A(s)) = -k_1(s).$$

Moreover, $\det(A(s)) = 1$, because $A(s)$ is an orthonormal matrix. So we conclude that $\det(J_\chi(s, y)) = 1 - y_1 k_1(s)$, and the result is proved. □

ACKNOWLEDGMENT

Research partially supported by the Spanish Ministry of Education and Science and FEDER, MTM2006-09920, and by the EU PASCAL Network of Excellence, IST-2002-506778.

REFERENCES

[1] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, pp. 502–516, 1989.

[2] B. Kégl, A. Krzyżak, T. Linder, and K. Zeger, "Learning and design of principal curves," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 281–297, 2000.

[3] P. Delicado, "Another look at principal curves and surfaces," *Journal of Multivariate Analysis*, vol. 77, pp. 84–116, 2001.

[4] P. Delicado and M. Huerta, "Principal Curves of Oriented Points: Theoretical and computational improvements," *Computational Statistics*, vol. 18, pp. 293–315, 2003.

[5] R. J. Tibshirani, "Principal curves revisited," *Stats. & Comp.*, vol. 2, pp. 183–190, 1992.

[6] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, pp. 443–482, 1999.

[7] K. Chang and J. Ghosh, "A unified model for probabilistic principal surfaces," *IEEE Trans. on Pattern Anal. and Machine Intel.*, vol. 23, pp. 22–41, 2001.

[8] C. P. Robert, *The Bayesian choice*, 2nd ed. New York: Springer, 2001.

[9] H. W. Guggenheimer, *Differential Geometry*. Dover Publications, 1977.

[10] R. E. Kass and L. Wasserman, "The selection of prior distributions by formal rules," *Journal of the American Statistical Association*, vol. 91, pp. 1343–1370, 1996.

[11] G. O. Roberts, A. Gelman, and W. R. Gilks, "Weak convergence and optimal scaling of random walk Metropolis algorithms," *Ann. Appl. Probab.*, vol. 7, no. 1, pp. 110–120, 1997.

[12] P. Green, "Reversible jump MCMC computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1985.

[13] M. Stephens, "Bayesin analysis of mixture models with an unknown number of components– An alternative to reversible jumps methods," *Annals of Statistics*, vol. 28, pp. 143–151, 2000.

[14] C. P. Robert and G. Gasella, *Monte Carlo Statistical Methods*, 2nd ed. New York: Springer, 2004.