Some asymptotics for localized principal components and curves

Jochen Einbeck*,[a]

Mohammad A. Zayed[a,b]

[a] Durham University, Department of Mathematical Sciences, Durham City DH1 3LE, UK.

[b] Applied Statistics & Insurance Department, Mansoura University, Egypt.

11th March 2013

Abstract

The asymptotic behavior of localized principal components applying kernels as weights is investigated. In particular, we show that the first-order approximation of the first localized principal component at any given point only depends on the bandwidth parameter(s) and the density at that point. This result is extended to the context of local principal curves, where the characteristics of the points at which the curve stops at the edges are identified. This is used to provide a method which allows the curve to proceed beyond its natural endpoint if desired.

Key Words: Kernels, mean shift, PCA, local principal curves.

*corresponding author, jochen.einbeck@durham.ac.uk

1 Introduction

In this work, we consider *localized PCA* in the sense of *locally weighted PCA*, where the weighting enters through multivariate kernel functions. More specifically, we are given a multivariate random vector $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}) : S \longrightarrow \mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$, which maps elements from a sample space S into $d-$variate space. (The sample space S may be considered as latent and does not play a role henceforth.) The *global* first principal component line would be that line through the data cloud which minimizes the expected squared distances between data and their projections onto the line. It is well known that the solution to this problem is the line through $\boldsymbol{\mu}$ which points into the direction of the eigenvector $\boldsymbol{\gamma}_1$ of $\boldsymbol{\Sigma}$ corresponding to the largest eigenvalue λ_1 of $\boldsymbol{\Sigma}$. Turning from the probabilistic to the empirical setting, i.e. given n independent replicates of $\boldsymbol{X}$, say $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, then $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ need to be replaced by consistent estimators, for instance the ML estimators $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$.

This concept is straightforwardly extended to a scenario in which, given a (non-random) vector $\boldsymbol{x} \in \mathbb{R}^d$, and weights $w^{\boldsymbol{x}}(\boldsymbol{x}_i)$ centered at $\boldsymbol{x}$, we aim to minimize the *weighted* squared distances between data and their projections onto the line. If the weights are of bell–shaped and symmetric shape, their role is effectively to *localize* the estimation problem at $\boldsymbol{x}$. Weight functions of this type are known as kernels, with the prominent example of the Gaussian kernel. As we will verify later, it turns out that, unsurprisingly, the solution to this problem is the line through the locally weighted

mean, or short, *local mean*[1]

$$\boldsymbol{\mu}^{\boldsymbol{x}} = \frac{\sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)} \tag{1}$$

which points into the direction of the first eigenvector[2] of the *local covariance matrix*

$$\boldsymbol{\Sigma}^{\boldsymbol{x}} = \frac{\sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}^{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}^{\boldsymbol{x}})^T}{\sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)}. \tag{2}$$

That is, the "locally weighted" first principal component is given by a vector pointing into the direction which explains most of the "local variance" around $\boldsymbol{x}$, or, in simpler terms, which locally gives the best fit. Localized principal components, in this kernel-weighted sense, have found their way into the statistical literature only relatively recently. Schaal, Vijayakumar & Atkeson (1998) present locally weighted principal component analysis as a tool for local dimensionality reduction. Einbeck, Tutz & Evers (2005) used iterative localized PCA in a kernel-based approach to principal curve estimation. A variant of this technique was developed in Wang, Assadi & Spalding (2008) for adaptive tracing of curvilinear structures. Charlton, Brunsdon, Demsar, Harris & Fotheringham (2010) used localized PCA to implement geographically weighted principal components. Zayed & Einbeck (2010) used localized principal components to track the contribution of sub–indices to a summary index over time.

The question of the "asymptotic" behavior of the method, for small neighborhoods and large sample sizes, has not been investigated yet. This is in great contrast to

[1] For denotational convenience, we will from now on omit all 'hats' on symbols denoting estimators – it is clear that $\boldsymbol{\mu}^{\boldsymbol{x}}$ etc. are empirical and not theoretical quantities.

[2] When using the term 'first eigenvector', we mean the eigenvector corresponding to the largest eigenvalue.

the huge literature on kernel-based asymptotics for nonparametric regression, which exploit the nice theoretical properties of the kernel approach in much depth and detail. In this paper we will fill this gap; in particular we will show that the first-order approximation of the first localized principal component at $\boldsymbol{x}$ only depends on the bandwidth parameter(s) and the local topology of the data cloud (in terms of its density $f(\boldsymbol{x})$ and its derivatives). We will use this result to understand the convergence behavior of local principal curves (Einbeck, Tutz & Evers, 2005). We take advantage of this understanding in order to provide a method which allows the curve to 'delay' convergence if desired, i.e. to converge further into the tails than usual.

It should be noted that the term "localized" has been used in further, different meanings and facets in the context of principal component analysis. Historically, *localized* PCA meant *cluster-wise PCA* (E. Diday et Collaborateurs, 1979), which evoluted to powerful recursive partitioning algorithms during the last decades (for instance, Breiman, Friedman, Olshen & Stone (1984), Hawkins (1995), Liu, Chiu & Xu (2003)). A further family of methods for nonlinear principal component analysis is known under the term *kernel PCA* (Schölkopf & Smola, 1998). Good accounts of these developments, which are not of interest for the present paper, are found in Gorban, Kégl, Wunsch & Zinovyev (2008).

This paper is structured as follows. In Section 2, we formalize localized principal components and recall some known theoretical results. In Section 3 we provide some asymptotics for localized PCA, which we apply on local principal curves in Section 4. Section 5 exploits these results for boundary extension of local principal curves, before the paper is concluded in Section 6.

2 Localized principal components

As stated earlier, we consider the term "localized" to be equivalent to "locally kernel-weighted". To fix terms, let $\boldsymbol{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ and $\kappa(\cdot)$ be a bounded symmetric univariate function which integrates to 1 (we do not strictly require it to be non-negative, but usually this will be the case). From this, a $d-$variate kernel function K can be defined by either taking the product kernel $K(\boldsymbol{x}) = \kappa(x_1) \cdot \ldots \cdot \kappa(x_d)$ or a radial kernel function $K(\boldsymbol{x}) = \kappa(||\boldsymbol{x}||)$. The two formulations are equivalent if the base kernel κ is the Gaussian probability density function, $\kappa(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. The following applies to either construction of K. Now, let $\boldsymbol{H} \in \mathbb{R}^{d \times d}$ denote a positive definite bandwidth matrix, employing the usual notation as set out in Wand & Jones (1993) (for instance, if we localize only in the directions of the coordinate axes, then $\boldsymbol{H} = \text{diag}(h_1^2, \ldots, h_d^2)$, where $h_j, j = 1, \ldots, d$ are the individual bandwidths; and if we smooth equally strong in all directions, then $\boldsymbol{H} = h^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.) Then we can define

$$K_{\boldsymbol{H}}(\cdot) = |\boldsymbol{H}|^{-1/2} K\left(\boldsymbol{H}^{-1/2} \cdot\right)$$

which is a d-variate probability density function in itself. Given any line in $\mathbb{R}^d$, say $\boldsymbol{g}(t) = \boldsymbol{m} + t\boldsymbol{\gamma} \in \mathbb{R}^d$, with $t \in \mathbb{R}$ and suitable vectors $\boldsymbol{m}$ and $\boldsymbol{\gamma}$ with $||\boldsymbol{\gamma}|| = 1$, denote the coordinate of $\boldsymbol{X}$ projected orthogonally onto $\boldsymbol{g}$ by $\boldsymbol{X}^{\boldsymbol{g}}$, where

$$\boldsymbol{X}^{\boldsymbol{g}} = \boldsymbol{m} + \boldsymbol{\gamma}\boldsymbol{\gamma}^T(\boldsymbol{X} - \boldsymbol{m}) = (\boldsymbol{I} - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)\boldsymbol{m} + \boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X} \equiv \boldsymbol{A}_{\boldsymbol{\gamma}}\boldsymbol{m} + \boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}.$$

The matrix $\boldsymbol{A}_{\boldsymbol{\gamma}} = (\boldsymbol{I} - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)$ is positive semi-definite, which is evident by noting that ${\boldsymbol{A}_{\boldsymbol{\gamma}}}^T \boldsymbol{A}_{\boldsymbol{\gamma}} = \boldsymbol{A}_{\boldsymbol{\gamma}}$, and hence $||\boldsymbol{A}_{\boldsymbol{\gamma}}\boldsymbol{u}||^2 = \boldsymbol{u}^T \boldsymbol{A}_{\boldsymbol{\gamma}} \boldsymbol{u}$, for $\boldsymbol{u} \in \mathbb{R}^d$. However, it is not positive definite, since $\det(\boldsymbol{A}_{\boldsymbol{\gamma}}) = 1 - \boldsymbol{\gamma}^T\boldsymbol{\gamma} = 0$.

Now, at point $\boldsymbol{x}$, we seek to find $\boldsymbol{m}$ and $\boldsymbol{\gamma}$ such that the line $\boldsymbol{g}$ locally minimizes the weighted squared distances between the data and their projected counterparts

$\boldsymbol{x}_i^{\boldsymbol{g}} = \boldsymbol{A}_{\boldsymbol{\gamma}}\boldsymbol{m} + \boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{x}_i$. Using weights $w^{\boldsymbol{x}}(\boldsymbol{x}_i) = K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})$, and taking into account the restriction $||\boldsymbol{\gamma}|| = 1$, the expression to minimize is

$$\begin{aligned} Q(\boldsymbol{m}, \boldsymbol{\gamma}) &= \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})||\boldsymbol{x}_i - \boldsymbol{x}_i^g||^2 - \lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1) \quad (3) \\ &= \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})||\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}_i - \boldsymbol{m})||^2 - \lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1) \\ &= \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})(\boldsymbol{x}_i - \boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}_i - \boldsymbol{m}) - \lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1) \quad (4) \end{aligned}$$

Then,

$$\frac{\partial Q(\boldsymbol{m}, \boldsymbol{\gamma})}{\partial \boldsymbol{m}} = 2\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}_i - \boldsymbol{m}) \quad (5)$$

which, when equated to zero, yields the equation

$$\boldsymbol{A}_{\boldsymbol{\gamma}}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})\boldsymbol{x}_i = \boldsymbol{A}_{\boldsymbol{\gamma}}\boldsymbol{m}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x}). \quad (6)$$

Further, using the fact that $\frac{\partial}{\partial \boldsymbol{\gamma}}\boldsymbol{u}^T\boldsymbol{A}_{\boldsymbol{\gamma}}\boldsymbol{u} = -2(\boldsymbol{u}\boldsymbol{u}^T)\boldsymbol{\gamma}$,

$$\frac{\partial Q(\boldsymbol{m}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = -2\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})(\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T\boldsymbol{\gamma} - 2\lambda\boldsymbol{\gamma}, \quad (7)$$

and setting this equal to zero yields

$$\left[\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})(\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T\right]\boldsymbol{\gamma} = -\lambda\boldsymbol{\gamma} \quad (8)$$

One immediate solution to equation (6) is found through

$$\boldsymbol{m} = \frac{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})\boldsymbol{x}_i}{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})} \equiv \boldsymbol{\mu}^{\boldsymbol{x}}. \quad (9)$$

in which case (8) takes the shape

$$\boldsymbol{\Sigma}^{\boldsymbol{x}}\boldsymbol{\gamma} = -\lambda\boldsymbol{\gamma},$$

with $\boldsymbol{\Sigma}^{\boldsymbol{x}}$ defined as in (2), implying that $\boldsymbol{\gamma}$ is an eigenvector of $\boldsymbol{\Sigma}^{\boldsymbol{x}}$. By multiplying the latter equation to the left with $\boldsymbol{\gamma}^T$ it is clear that it needs to be the first eigenvector, which we denote by $\boldsymbol{\gamma}^{\boldsymbol{x}}$ henceforth.

However, since $\det(\boldsymbol{A}_{\boldsymbol{\gamma}}) = 0$, the solution to (6) is not unique. This is not different to the case of linear (unweighted) PCA (Hastie, Tibshirani & Friedman, 2001, Exercise 14.7). It is easily verified that the family of equivalent solutions of (6) is given by $\boldsymbol{\mu}^{\boldsymbol{x}} + t\boldsymbol{\gamma}^{\boldsymbol{x}}$, for $t \in \mathbb{R}$, and that $\boldsymbol{\gamma}^{\boldsymbol{x}}$ is the solution to (8) whatever member of the family is used for $\boldsymbol{m}$.

Summarizing, the localized first principal component line at $\boldsymbol{x}$ is given by

$$\boldsymbol{g}^{\boldsymbol{x}}(t) = \boldsymbol{\mu}^{\boldsymbol{x}} + t\boldsymbol{\gamma}^{\boldsymbol{x}},$$

i.e. a line through the local mean in direction of the first eigenvector of the local covariance matrix.

An important quantity that should also be introduced at this point, and which will become relevant later on, is the *mean shift*

$$\boldsymbol{s}^{\boldsymbol{x}} = \boldsymbol{\mu}^{\boldsymbol{x}} - \boldsymbol{x} = \frac{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})(\boldsymbol{x}_i - \boldsymbol{x})}{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})}. \tag{10}$$

This is a simple tool originating from the computer science literature (Cheng, 1995) which computes the "shift" necessary to move a certain point $\boldsymbol{x} \in \mathbb{R}^d$ towards the local mean of all data points in a neighborhood of $\boldsymbol{x}$. Note that both $\boldsymbol{x}$ and $\boldsymbol{\mu}^x$ are vector-valued, so the mean shift is vector-valued as well.

3 Asymptotics

Next, an asymptotic version of the result from Section 2 is provided. Define the function $Q(\boldsymbol{m}, \boldsymbol{\gamma})$ as in (3), and let f denote the density function of $\boldsymbol{X}$ with support $\text{supp}(f)$. Denote $\mathbf{0}$ and $\mathbf{1}$ vectors of appropriate dimension which only consist of 0's and 1's, respectively, and let $o_p(1)$ denote a sequence which tends to zero in probability as $n \longrightarrow \infty$. We assume

(A1) The kernel K is a bounded and compactly supported probability density function such that $\int \boldsymbol{u} K(\boldsymbol{u})\, d\boldsymbol{u} = \mathbf{0}$ and $\int \boldsymbol{u}\boldsymbol{u}^T K(\boldsymbol{u})\, d\boldsymbol{u} = \mu_2(K)\boldsymbol{I}$, with $\mu_2(K) \in \mathbb{R}$.

(A2) At $\boldsymbol{x} \in \text{supp}(f)$, f is continuously differentiable and $f(\boldsymbol{x}) > 0$.

(A3) The sequence of bandwidth matrices $\boldsymbol{H}$ is such that $n^{-1}|\boldsymbol{H}|^{-1/2}$ and each entry of $\boldsymbol{H}$ tends to zero as $n \longrightarrow \infty$, with $\boldsymbol{H}$ remaining symmetric and positive definite.

We firstly provide an approximation of the mean shift. We make use of the well known results

$$\begin{aligned} \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})(\boldsymbol{x}_i - \boldsymbol{x}) &= n[\mu_2(K)\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}) + o_p(\boldsymbol{H}\mathbf{1})] \\ \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x}) &= n[f(\boldsymbol{x}) + o_p(1)] \end{aligned}$$

which were established in a different context by Ruppert & Wand (1994) but are equally true here. The quotient of these two expressions gives the asymptotic mean shift,

$$\tilde{s}^{\boldsymbol{x}} = \mu_2(K)\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x})/f(\boldsymbol{x}) + o_p(\boldsymbol{H}\mathbf{1}) \tag{11}$$

Hence, the operator $\tilde{s}^{\boldsymbol{x}}$ shifts a given point $\boldsymbol{x}$ into a direction in which the data tend to be more dense, with the step size being the larger the less dense the data are at $\boldsymbol{x}$. This implies that, asymptotically, the gradient is zero when the mean shift is zero, which makes the mean shift a suitable tool for density mode detection (Cheng, 1995).

Equation (11) provides an asymptotic version of (10), hence, implicitly, of (9). In order to find an asymptotic version of (8), the strategy is to find an asymptotic version of Q, based on which the minimization problem is solved. Considering the first term of (4),

$$
\begin{aligned}
& E\left(\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i-\boldsymbol{x})(\boldsymbol{x}_i-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}_i-\boldsymbol{m})\right) \\
&= n|\boldsymbol{H}|^{-1/2}\int K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{s}-\boldsymbol{x})\right)(\boldsymbol{s}-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{s}-\boldsymbol{m})f(\boldsymbol{s})\,d\boldsymbol{s} \\
&= n\int K(\boldsymbol{u})(\boldsymbol{H}^{1/2}\boldsymbol{u}+\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{H}^{1/2}\boldsymbol{u}+\boldsymbol{x}-\boldsymbol{m})f(\boldsymbol{x}+\boldsymbol{H}^{1/2}\boldsymbol{u})\,d\boldsymbol{u} \\
&= n\int K(\boldsymbol{u})\left\{(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}-\boldsymbol{m})+O(\mathbf{1}^T\boldsymbol{H}^{1/2}\boldsymbol{u})\right\}\left\{f(\boldsymbol{x})+O(\mathbf{1}^T\boldsymbol{H}^{1/2}\boldsymbol{u})\right\}\,d\boldsymbol{u} \\
&= nf(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}-\boldsymbol{m})+o(n),
\end{aligned}
$$

where $\int K(\boldsymbol{u})\,d\boldsymbol{u}=1$. Similarly, one can show that

$$
\mathrm{Var}\left(\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i-\boldsymbol{x})(\boldsymbol{x}_i-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}_i-\boldsymbol{m})\right)=o(n^2),
$$

so that, in summary,

$$
\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i-\boldsymbol{x})(\boldsymbol{x}_i-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}_i-\boldsymbol{m})=nf(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}-\boldsymbol{m})+o_P(n)
$$

holds. We arrive at the penalized asymptotic minimization problem

$$
\tilde{Q}(\boldsymbol{m},\boldsymbol{\gamma})=nf(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{A}_{\boldsymbol{\gamma}}(\boldsymbol{x}-\boldsymbol{m})-\lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma}-1).
$$

Taking again the derivative,

$$
\frac{\partial\tilde{Q}(\boldsymbol{m},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}=-2\left[nf(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{\gamma}+\lambda\boldsymbol{\gamma}\right], \tag{12}
$$

and equating this to zero,

$$
nf(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{\gamma}=-\lambda\boldsymbol{\gamma} \tag{13}
$$

i.e. $\boldsymbol{\gamma}$ is eigenvector of $nf(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^T$. Now, note that for any matrix of type $\boldsymbol{\Sigma} = c\boldsymbol{\psi}\boldsymbol{\psi}^T$, with $c \in \mathbb{R}, \boldsymbol{\psi} \in \mathbb{R}^d$, the only eigenvector of $\boldsymbol{\Sigma}$ is (in standardized form) $\boldsymbol{\gamma} = \boldsymbol{\psi}/||\boldsymbol{\psi}||$, with eigenvalue $\lambda = c||\boldsymbol{\psi}||^2 = \mathrm{Tr}(\boldsymbol{\Sigma})$. Hence, the only eigenvector of $nf(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})(\boldsymbol{x}-\boldsymbol{m})^T$ is given by

$$\frac{\boldsymbol{x}-\boldsymbol{m}}{||\boldsymbol{x}-\boldsymbol{m}||}.$$

Using that the local estimate of $\boldsymbol{m}$ is $\boldsymbol{\mu}^x$, we can replace $\boldsymbol{x}-\boldsymbol{m}$ by (11), yielding the asymptotic version of $\boldsymbol{\gamma}^{\boldsymbol{x}}$,

$$\tilde{\boldsymbol{\gamma}}^{\boldsymbol{x}} \stackrel{a}{=} \frac{-\mu_2(K)\boldsymbol{H}\boldsymbol{\nabla}f(\boldsymbol{x})/f(\boldsymbol{x})}{\mu_2(K)||\boldsymbol{H}\boldsymbol{\nabla}f(\boldsymbol{x})||/f(\boldsymbol{x})} = -\frac{\boldsymbol{H}\boldsymbol{\nabla}f(\boldsymbol{x})}{||\boldsymbol{H}\boldsymbol{\nabla}f(\boldsymbol{x})||},$$

where the denotation $\stackrel{a}{=}$ means that in the expression succeeding this symbol all terms of an asymptotically higher order than the leading term are omitted. This shows that, asymptotically, the first local principal component always steers into the direction of the density gradient.

4 Local principal curves

The nonparametric equivalent to a (globally fitted) principal component line is a principal curve, which can be descriptively defined as a 'smooth curve through the middle of the data cloud'. There do exist several competing algorithms for fitting principal curves. Most of these are not directly based on a 'localized' or 'nonparametric' version of PCA, but rather start with some globally fitted line (which may be the first principal component line), and iteratively bend this line (or concatenate other lines to it), until it fits satisfactorily through the middle of the data cloud (in some sense). In contrast, the local principal curve algorithm (Einbeck, Tutz & Evers, 2005) is explicitly based on iterating local PCA steps. To fix terms, let $\boldsymbol{x}_{(0)} \in \mathbb{R}^d$ be a starting point (chosen at random or by hand). At j-th iteration, the mean shift brings us from

$\boldsymbol{x}_{(j)}$ to $\boldsymbol{\mu}_{(j)} \equiv \boldsymbol{\mu}^{\boldsymbol{x}_{(j)}}$, and stepping from there a predetermined distance, say t, into the direction of $\boldsymbol{\gamma}_{(j)} \equiv \boldsymbol{\gamma}^{\boldsymbol{x}_{(j)}}$ yields $\boldsymbol{x}_{(j+1)}$. Condensing these two steps into one row, one has

$$\boldsymbol{x}_{(j+1)} = \boldsymbol{\mu}_{(j)} \pm t\boldsymbol{\gamma}_{(j)} \tag{14}$$

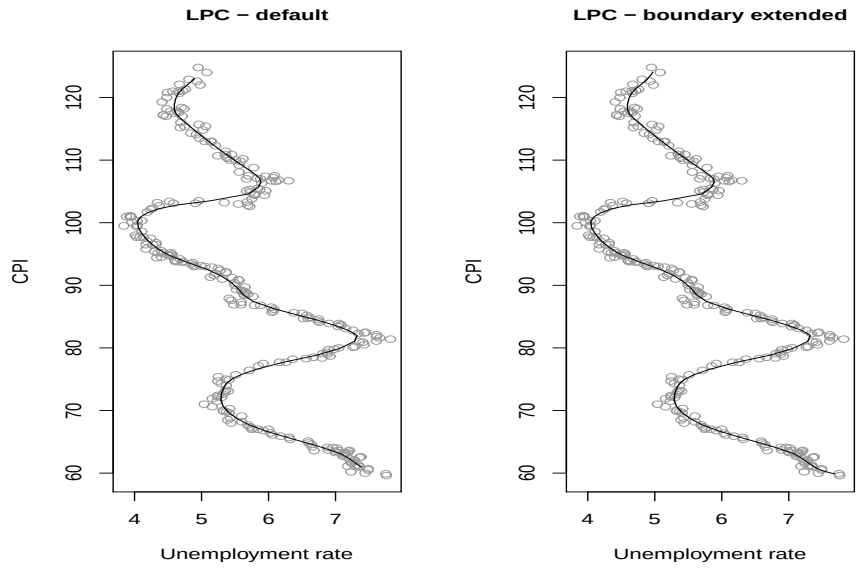
where the sign in '$\pm$' is given by $\text{sign}(\boldsymbol{\gamma}_{(j)} \circ \boldsymbol{\gamma}_{(j-1)})$ (this 'signum flipping' ensures that the curve maintains their direction). The step size t needs to be specified by the data analyst, and is usually set equal to h if $\boldsymbol{H} = h^2\boldsymbol{I}$ (this is a very reasonable assumption if the data are previously scaled, for instance by dividing through their range or standard deviation). The resulting principal curve is constituted through the series of local means, $\{\boldsymbol{\mu}_{(j)}\}_{j\geq 0}$, and the iteration is stopped when the difference between neighboring $\boldsymbol{\mu}_{(j)}$'s falls below a given threshold. We will refer to this state in what follows as 'convergence', though this does not necessarily imply convergence in a strict, mathematical sense. Next, one proceeds from the starting point in the opposite direction, i.e. one changes the signum preceding $\boldsymbol{\gamma}_{(0)}$ in the computation of $\boldsymbol{x}_{(1)}$, and continues as before until convergence is reached. There are some additional technicalities involved in the curve fitting, which are not relevant for this presentation; the interested reader is referred to the paper mentioned above.

Figure 1 (left) shows a local principal curve fitted to time series data for monthly unemployment and inflation rates in the Unites States from March 1984 until April 2008. The horizontal axis is the monthly rate of unemployment and the vertical axis is the monthly consumer price index, which is considered the most commonly used measure of inflation[3].

The developments in Section 3 enable us to study the asymptotic, local behavior of the local principal curve algorithm. Firstly, we look at the difference between two

[3]In Economics, the curve representing the relationship between unemployment and inflation is well known as 'Phillips Curve'(Phillips, 1958).

Figure 1: local principal curve for unemployment-inflation data



neighboring points $\boldsymbol{x}_{(j)}$ and $\boldsymbol{x}_{(j+1)}$. From (14), one has

$$\begin{aligned}
\boldsymbol{x}_{(j+1)} - \boldsymbol{x}_{(j)} &= \boldsymbol{\mu}^{\boldsymbol{x}_{(j)}} - \boldsymbol{x}_{(j)} \pm t\boldsymbol{\gamma}^{\boldsymbol{x}_{(j)}} \\
&\stackrel{a}{=} \mu_2(K)\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})/f(\boldsymbol{x}_{(j)}) \pm t\frac{\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})||} \\
&= \left(\frac{\mu_2(K)}{f(\boldsymbol{x}_{(j)})} \pm \frac{t}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})||}\right)\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)}).
\end{aligned}$$

Note that, defining $\boldsymbol{S}(\boldsymbol{x}) \equiv \boldsymbol{\nabla} f(\boldsymbol{x})/f(\boldsymbol{x})$, the Taylor expansion of $\boldsymbol{S}$ at $\boldsymbol{x}$ is given by

$$\boldsymbol{S}(\boldsymbol{x} \pm \boldsymbol{\delta}) = \boldsymbol{S}(\boldsymbol{x}) \pm \left[\frac{\boldsymbol{H}_f(\boldsymbol{x})}{f(\boldsymbol{x})} - \boldsymbol{S}(\boldsymbol{x})\boldsymbol{S}(\boldsymbol{x})^T\right]\boldsymbol{\delta} + O(\boldsymbol{\delta}^2)$$

where $\boldsymbol{\delta} \longrightarrow 0$ (component-wise), and $\boldsymbol{H}_f(\boldsymbol{x})$ is the Hessian of f at $\boldsymbol{x}$.

This implies that, in first order approximation, the difference between two neigh-

boring local centers of mass $\boldsymbol{\mu}_{(j)}$ and $\boldsymbol{\mu}_{(j+1)}$ is given by

$$
\begin{aligned}
&\boldsymbol{\mu}_{(j+1)} - \boldsymbol{\mu}_{(j)} \qquad (15)\\
&= (\boldsymbol{\mu}_{(j+1)} - \boldsymbol{x}_{(j+1)}) - (\boldsymbol{x}_{(j+1)} - \boldsymbol{\mu}_{(j)})\\
&\overset{a}{=} \mu_2(K)\boldsymbol{H}\frac{\boldsymbol{\nabla} f(\boldsymbol{x}_{(j+1)})}{f(\boldsymbol{x}_{(j+1))}} \pm t\tilde{\boldsymbol{\gamma}}^{\boldsymbol{x}_{(j)}}\\
&= \mu_2(K)\boldsymbol{H}\boldsymbol{S}(\boldsymbol{x}_{(j)} + (\boldsymbol{x}_{(j+1)} - \boldsymbol{x}_{(j)})) \pm t\frac{\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})||}\\
&= \mu_2(K)\boldsymbol{H}\left\{\boldsymbol{S}(\boldsymbol{x}_{(j)}) + \left[\frac{\boldsymbol{H}_f(\boldsymbol{x}_{(j)})}{f(\boldsymbol{x}_{(j)})} - \frac{\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})^T}{f(\boldsymbol{x}_{(j)})^2}\right]\underbrace{(\boldsymbol{x}_{(j+1)} - \boldsymbol{x}_{(j)})}_{O(\boldsymbol{H}\boldsymbol{1})}\right\} \pm t\frac{\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})||}\\
&\overset{a}{=} \mu_2(K)\boldsymbol{H}\frac{\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})}{f(\boldsymbol{x}_{(j)})} \pm t\frac{\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})||}\\
&= \left[\frac{\mu_2(K)}{f(\boldsymbol{x}_{(j)})} \pm \frac{t}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})||}\right]\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)}) \qquad (16)
\end{aligned}
$$

In this result, the two-step character of the algorithm is still visible: The first term inside the squared bracket corresponds to the contribution of the mean shift, while the second term corresponds to the local PCA step. In order to gain more insight, we assume from now on that $\boldsymbol{H} = h^2\boldsymbol{I}$, and $t = h$, which is the recommended default setting according to Einbeck, Tutz & Evers (2005). Then,

$$\boldsymbol{\mu}_{(j+1)} - \boldsymbol{\mu}_{(j)} \overset{a}{=} h\left[\frac{h\ \mu_2(K)}{f(\boldsymbol{x}_{(j)})} \pm \frac{1}{||\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)}||}\right]\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)}) \qquad (17)$$

Now, if the curve is proceeding uphill (i.e., towards higher densities), then both the mean shift step and the local PCA step will steer the curve in the same direction, so the term $\boldsymbol{\mu}_{(j+1)} - \boldsymbol{\mu}_{(j)}$ will be quite large. If the curve is moving downhill, then the mean shift step will pull the curve backwards – towards higher densities –, so the total step size (17) will be rather small. This means the curve will get stuck if the two contributions are exactly the same. This is the case when

$$\frac{h\mu_2(K)}{f(\boldsymbol{x})} = \frac{1}{||\boldsymbol{\nabla} f(\boldsymbol{x})||}$$

(subscript j omitted for notational ease), implying

$$f(\boldsymbol{x}) = h\ \mu_2(K)||\boldsymbol{\nabla} f(\boldsymbol{x})||. \tag{18}$$

Hence, the position at which the curve gets stuck only depends on: the density of the random vector $\boldsymbol{X}$, the kernel function K and the bandwidth h. If we use a Gaussian kernel K, then $\mu_2(K) = 1$, and (18) becomes

$$f(\boldsymbol{x}) = h\ ||\boldsymbol{\nabla} f(\boldsymbol{x})||. \tag{19}$$

Next, we wish to gain some understanding of where the points with property (18) or (19), respectively, are situated. Therefore, let us assume that the random vector under consideration is given by

$$\boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{20}$$

with $\boldsymbol{0} \in \mathbb{R}^d$ being a vector of $0's$, and $\boldsymbol{I} \in \mathbb{R}^{d \times d}$ being the identity matrix. One has then

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{x}\right\}$$

and

$$||\boldsymbol{\nabla} f(\boldsymbol{x})|| = \frac{1}{(2\pi)^{d/2}\sigma^{d+2}} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{x}\right\} ||\boldsymbol{x}||$$

so that (19) boils down to

$$||\boldsymbol{x}|| = \frac{\sigma^2}{h}. \tag{21}$$

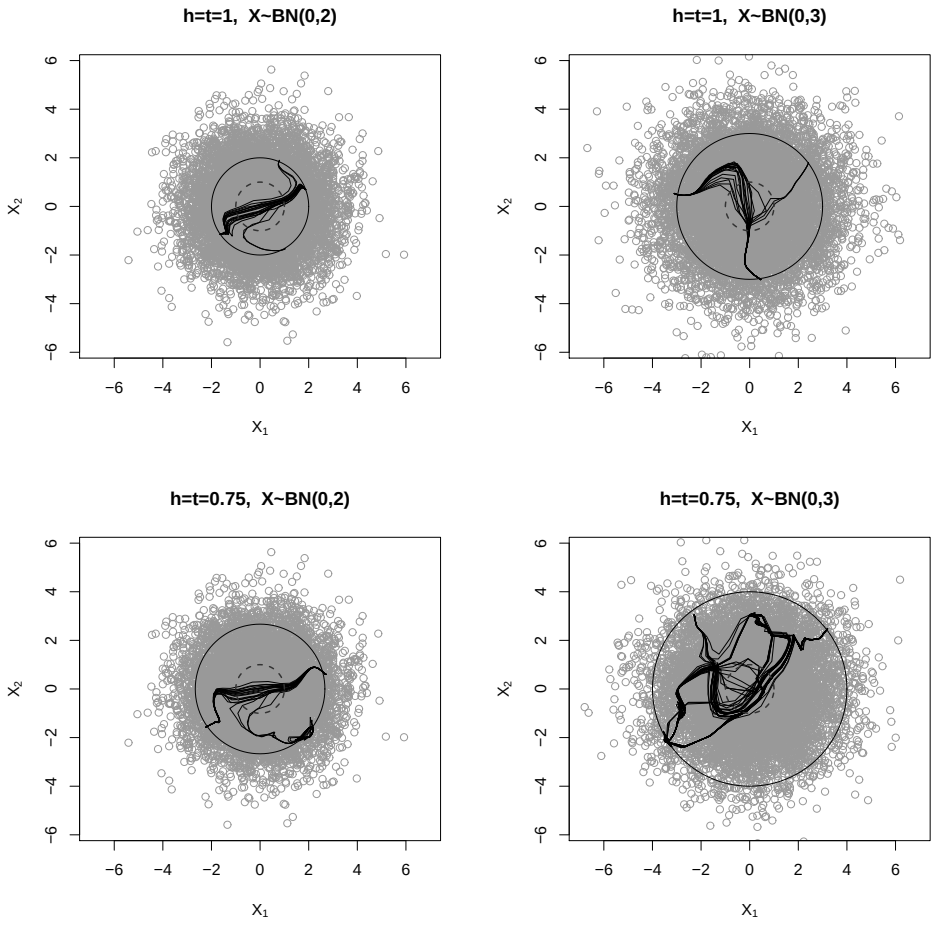
We will verify result (21) by experiment. Therefore, we assume that $\boldsymbol{X}$ is bivariate normal, i.e. of type (20) with $d = 2$. We simulate $n = 10000$ replicates from $\boldsymbol{X}$, for each $\sigma^2 = 2$ and $\sigma^2 = 3$. Next, we fit 20 local principal curves (with $t = h = 1$) to each of both data clouds, where the starting points are randomly chosen among all those observations $\boldsymbol{x}_i$ which satisfy $||\boldsymbol{x}_i|| \leq 1$. For the principal curve fitting, we use a Gaussian base kernel $\kappa(\cdot)$ which is truncated at ± 5. This does not actually affect

the results of the simulation, but ensures compatibility with assumption (A1); see also Ruppert & Wand (1994). The resulting curves are displayed in Figure 2 (top row). In these plots, the dashed and solid circle symbolize the radii $||\boldsymbol{x}|| = 1$ and $||\boldsymbol{x}|| = \sigma^2$, respectively, so according to the theory the curves should get stuck close to the solid circle. We see that this is always the case. For both $\sigma^2 = 2$ (left) and $\sigma^2 = 3$ (right), all principal curves converge to endpoints which are very close to the solid circle. In the bottom panels of this figure we repeat the analysis for $h = t = 0.75$. We observe that, according to (21), by decreasing h the curves will visit a larger area of the data than if $h = 1$, as the radius σ^2/h gets larger. Similarly, when $h > 1$, the curves are expected to visit a smaller area of the data as the radius gets bigger (not shown). [Of course, the example itself is a little bit contrived, as one, realistically, would not be interested in fitting curves to bivariate normal data, but the theory is convincingly confirmed].

5 Boundary extension

We have seen towards the end of Section 4 that by reducing the bandwidth one obtains curves which proceed further into the boundary region of the data. Access to these boundary regions can be of a special importance, for instance for time series data where the endpoints correspond to the most current observations. Furthermore, curves which are "too short" in the boundaries will result in projections clustered at the endpoints, which impacts negatively on the usability of the curve as a data compression tool, a problem which was observed by Einbeck, Evers & Hinchliff (2010) in the context of nonlinear compression of high-dimensional spectrographic data. In such situations, one may attempt extending the local principal curve beyond its natural endpoint in order to reach more data points at boundaries. Obviously, decreasing the bandwidth

Figure 2: 20 local principal curves with bandwidths $h = t = 1$ (top) and $h = t = 0.75$ (bottom) through multivariate Gaussian data with $\sigma^2 = 2$ (left) and $\sigma^2 = 3$ (right). The dashed circle indicates the radius $||\boldsymbol{x}|| = 1$, while the radius of the solid circle is equal to $||\boldsymbol{x}|| = \sigma^2/h$ according to (21).



arbitrarily will not be the solution, as this will result in a curve which gets stuck even sooner.

To find a practicable way of dealing with this problem, note firstly that the simulation in Section 4 was run assuming $t = h$. If we allow these two parameters to decouple, then we see from (16) that the difference between two successive centers of mass can be written as

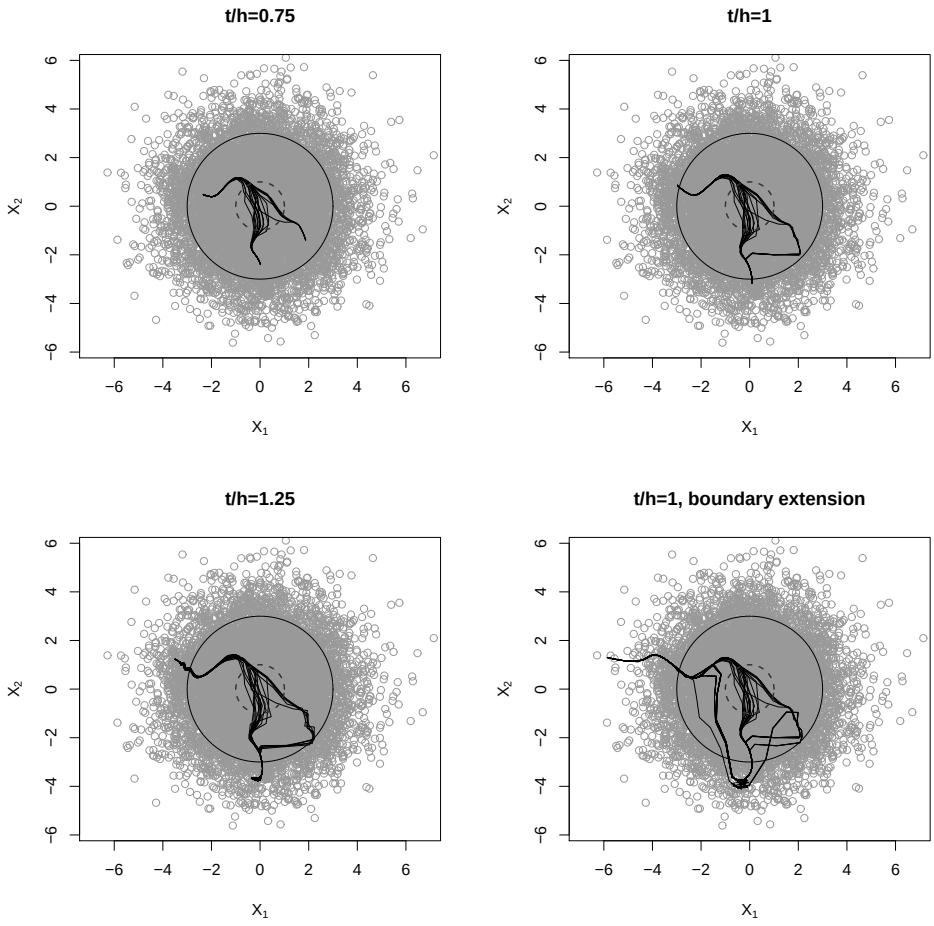
$$\boldsymbol{\mu}_{(j+1)} - \boldsymbol{\mu}_{(j)} \overset{a}{=} h \left[\frac{h\mu_2(K)}{f(\boldsymbol{x}_{(j)})} \pm \frac{t}{h} \frac{1}{||\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})||} \right] \boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})$$

Compared to (17), one observes that now the term corresponding to the principal component step is multiplied by t/h. Hence, if t is increased relative to h, the PCA contribution increases relative to the mean shift contribution, and the principal curve will proceed *beyond* the limit given by (21).

We illustrate this effect again through simulation. Using Gaussian data $\boldsymbol{X}$ with $\sigma^2 = 3$, 20 local principal curves have been fitted with different ratios of t and h. The resulting curves are displayed in the first three panels of Figure 3, and one observes that, for $t/h < 1$, the curve will stop inside the circle defined by (21), while for $t/h > 1$, it will stop outside (in fact, the radius at which the curves converge is now $\sigma^2 t/h^2$). However, in practice it is impractical to increase t beyond h, as this would impact detrimentally onto large parts of the curve, and cause erratic behavior especially in the boundary region. Therefore, it is recommended to keep the default setting $t = h$, which has proven to work generally well, for the non-boundary part of the principal curve, and reduce h (but not t) adaptively as soon as the curve begins to converge to its endpoint.

In the implementation of the LPC algorithm (Einbeck & Evers, 2010), this is achieved by defining a threshold, say T_1, and reducing the bandwidth adaptively as

Figure 3: 20 local principal curves, all with $h = 1$, and $t = 0.75$ (top left), $t = 1$ (right), and $t = 1.25$ (bottom left) through a multivariate Gaussian sample of size $n = 10000$ with $\sigma^2 = 3$. The bottom right plot uses the boundary extension proposed in Section 5. The outer (solid) circles have radius σ^2, and the inner (dashed) circles radius 1.



$h_{(j+1)} = (1-\delta)h_{(j)}$, for some small constant $\delta > 0$, as soon as

$$\frac{||\boldsymbol{\mu}_{(j+1)} - \boldsymbol{\mu}_{(j)}||}{||\boldsymbol{\mu}_{(j+1)} + \boldsymbol{\mu}_{(j)}||} \leq T_1.$$

A second threshold, $0 < T_2 < T_1$, determines when the state of convergence is reached and the algorithm is stopped. The performance of this technique is demonstrated in the bottom right panel of Figure 3. Compared to the non-extended fit, it is clear that, after applying the boundary extension, the local principal curves reach further into the boundary region of the data cloud.

This technique extends straightforwardly to situations where different bandwidths h_j are used in each direction $j = 1, \ldots, d$, by multiplying each bandwidth individually by $1 - \delta$. This technique was used for the US unemployment-inflation data example, where the "default" curve displayed in Figure 1 (left) used the bandwidth vector $h = (0.5, 1.5)$. The boundary extended version is provided in Figure 1 (right), and it is obvious that this curve deals better with the two boundaries than the default curve.

6 Conclusion

In this work we have explored some asymptotics for localized principal components using multivariate kernels as weights. It was shown that for small neighborhoods and large sample sizes, at any point $\boldsymbol{x}$, the first eigenvector of the local covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{x}}$ can be approximated in terms of the density function and the bandwidth matrix $\boldsymbol{H}$. For local principal curves (LPCs), this result implied that the LPC always steers into the direction of the gradient of the density function, which means in practice that it will closely follow the density ridge. The previous approximation was extended to explore the behaviour of the local principal curve in terms of the difference between neighboring local centers of mass which compose the fitted curve. Using the first order approximation of the latter, the stopping criteria for the LPC was further investigated.

It was concluded that the position at which the curve stops only depends on: the topology of data in the neighborhood in terms of the density function and its derivative, the multivariate kernel function used and the bandwidth matrix. This was verified experimentally, and it was confirmed that the smaller is the bandwidth, the larger is the area of the data visited by the curves. It was shown that by reducing the bandwidth adaptively relative to the step size as soon as the local principal curve begins to converge to its endpoint, the curve reaches further into the boundaries. This is of particular importance for data with time series character, even if time is not used for the multivariate analysis, as in the real data example provided. For data of this type, the current time point, which is likely to be the point of interest, is by construction a boundary point. In addition, this technique avoids the projections being clustered at the curve endpoints, which would impact negatively on the usability of the curve as a data compression tool.

Acknowledgements

The second author is supported through a scholarship by the Egyptian Government. The authors are grateful to Prof. Tutz, LMU Munich, for drawing their attention to the application of Phillips curves.

References

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. New York, NY, USA: Wadsworth.

Charlton, M., Brunsdon, C., Demsar, U., Harris, P., and Fotheringham, S. (2010). Principal component analysis: from global to local. In *13th AGILE International Conference on Geographic Information Science*. Guimarães, Portugal.

Cheng, Y. (1995). Mean shift, mode seeking and clustering. *IEEE Trans. Pattern Anal. Machine Intell.* **17**, 790–799.

E. Diday et Collaborateurs (1979). *Optimisation en Classification Automatique*. Le Chesnay, France: INRIA.

Einbeck, J. and Evers, L. (2010). *LPCM: Local principal curve methods*. R package version 0.41-6.

Einbeck, J., Evers, L., and Hinchliff, K. (2010). Data compression and regression based on local principal curves. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in Data Analysis, Data Handling and Business Intelligence*, Heidelberg, pp. 701–712. Springer.

Einbeck, J., Tutz, G., and Evers, L. (2005). Local principal curves. *Statistics and Computing* **15**, 301–313.

Gorban, A., Kégl, B., Wunsch, D., and Zinovyev, A. (2008). *Principal Manifolds for Data Visualization and Dimension Reduction*. Heidelberg: Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Hawkins, D. M. (1995). *FIRM Formal Inference-based Recursive Modeling*. St. Paul, MN, USA: University of Minnesota.

Liu, Z.-Y., Chiu, K.-C., and Xu, L. (2003). Improved system for object detection and star/galaxy classification via local subspace analysis. *Neural Networks* **16**, 437–451.

Phillips, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom. *Economica* **25**, 283–299.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

Schaal, S., Vijayakumar, S., and Atkeson, C. (1998). Local dimensionality reduction.

In M. Jordan, M. Kearns, & S. Solla (Eds.), *Advances in Neural Information Processing Systems 10*. MIT Press, Cambridge, MA.

Schölkopf, B. and Smola, A. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**, 1299–1319.

Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parametrizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88**, 520–528.

Wang, L., Assadi, A. H., and Spalding, E. P. (2008). Tracing branched curvilinear structures with a novel adaptive local pca algorithm. In *Proceedings of the 2008 International Conference on Image Processing, Computer Vision, and Pattern Recognition*, Volume 17, pp. 557–563. CSREA Press, Athens, GA.

Zayed, M. and Einbeck, J. (2010). Constructing economic summary indexes via principal curves. In *COMPSTAT 2010 Proceedings (e-book)*, pp. 1709–1716.