Exploring Multivariate Data Structures with Principal Curves

Jochen Einbeck

Department of Mathematical Sciences, Durham University

jochen.einbeck@durham.ac.uk

Durham — 4th April 2007

joint work with

Gerhard Tutz (University of Munich), Ludger Evers (University of Bristol),

and Coryn Bailer-Jones (MPIA Heidelberg).

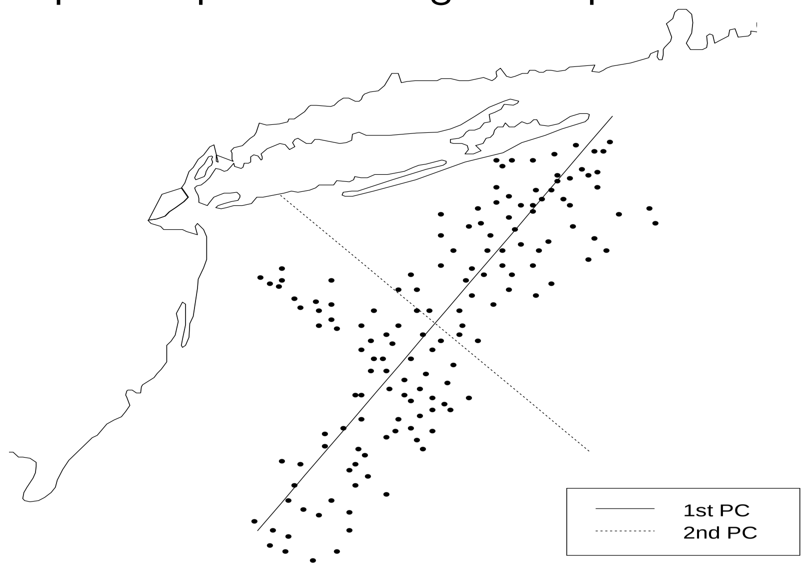
Principal components and principal curves

- Principal components are a sequence of best linear approximations to a data cloud

  $X = (X_1, \ldots, X_n)$, where $X_i \in \mathbb{R}^d$.

  Example:

  First and second principal component through scallops near the NE coast of the USA:

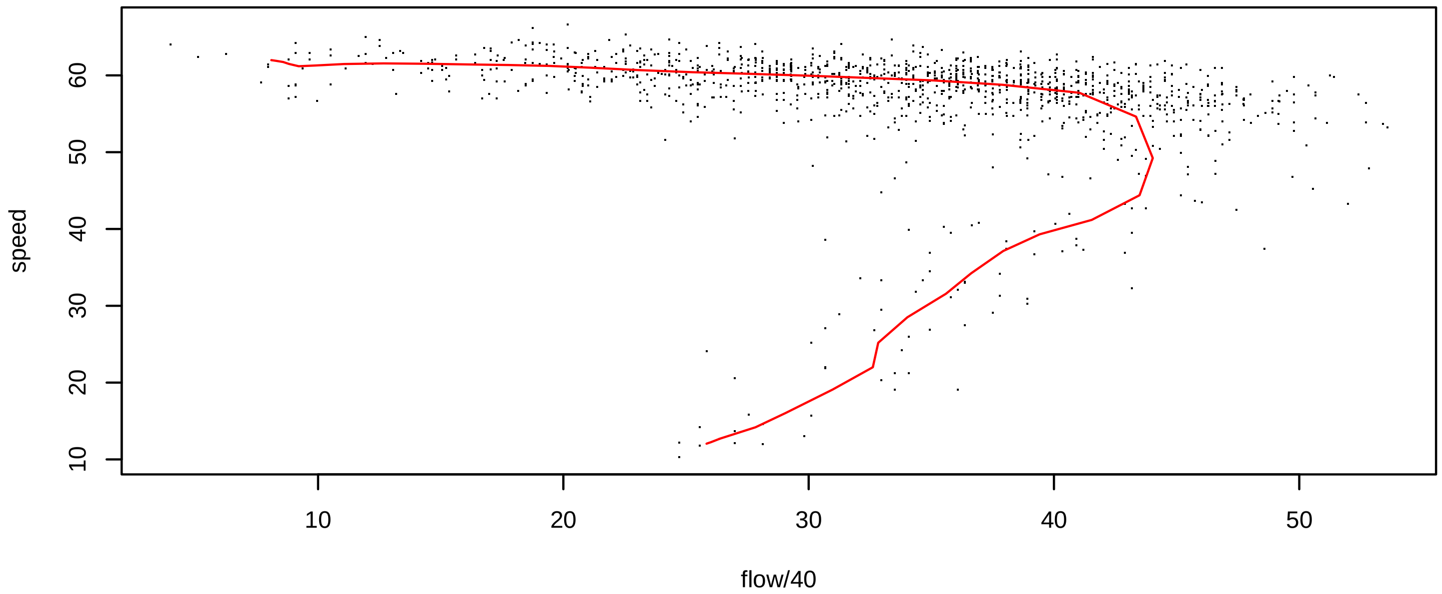


- Principal curves can be considered as nonparametric versions of principal components.

Descriptive Definition

Principal Curves are smooth curves passing through the ‘middle’ of a multidimensional data cloud.

Example: Speed-Flow diagram.



(data from: R package hdrcde)

X: traffic flow in cars/hour, Y: speed in Miles/hour recorded on a Californian “freeway”.

Types of principal curves

There exist a variety of definitions of principal curves, which essentially vary in what is understood of the "middle" of a data cloud. The algorithms associated to these definitions can be divided into two major groups:

- Global ('top-down') algorithms start with the first linear principal component line and try to dwell out this line or concatenate other lines to the initial line until the resulting curve fits well through the data cloud.
  - Hastie & Stützle (HS, 1989), Kégl, Krżyzak, Linder & Zeger (KKLZ, 2000).
  - Quite fast and computationally stable.
  - Dependence on an initial line leads to a lack of flexibility (particularly for HS).

- Local (‘bottom-up’) algorithms estimate the principal curve locally moving step by step through the data cloud.

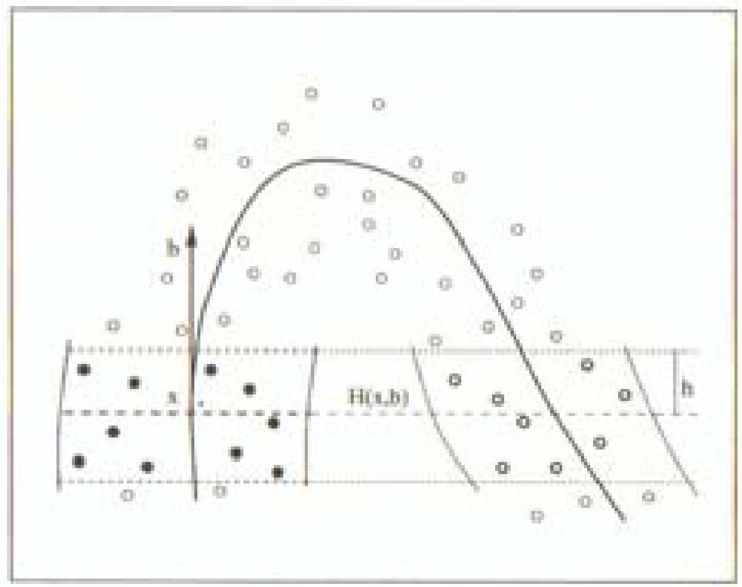
  Specifically, Delicado (2001) defines principal curves as a sequence of fixed points of the function $\mu^*(x) = E(X|X \in H)$, where H is the hyperplane through x minimizing locally the variance of the data points projected on it.

  – Works fine for most (not too complex) data sets.

  – Mathematically elegant

  – However, quite complicated and computationally demanding.

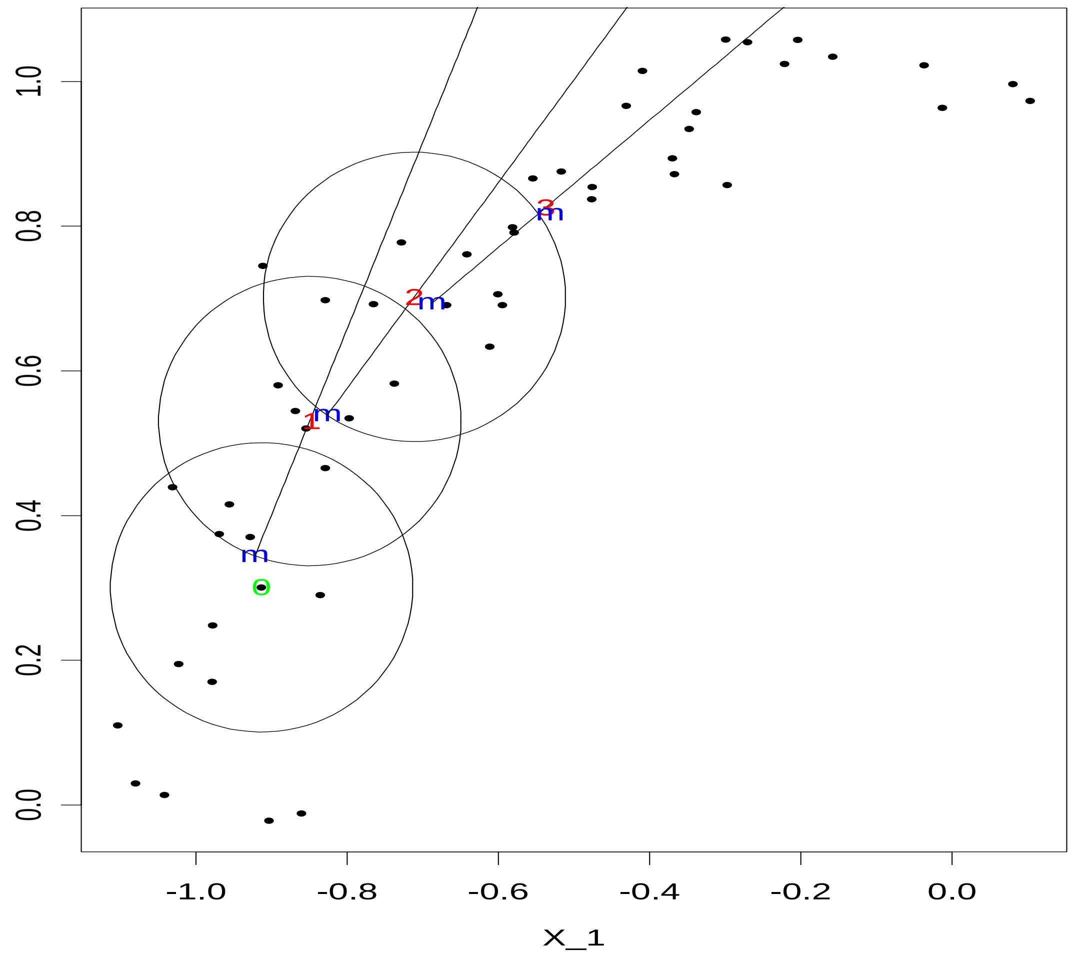
  – Requires a cluster analysis at every point of the principal curve.



(Picture from: Delicado, 2001)

Simple alternative: Local principal curves (LPC; Einbeck, Tutz & Evers, 2005)

Idea: Calculate alternately a local center of mass and a first local principal component.



0: starting point,

m: points of the LPC,

$1, 2, 3$: enumeration of steps.

Algorithm for LPCs

Given: A data cloud $X = (X_1, \ldots, X_n)$, where $X_i = (X_{i1}, \ldots, X_{id})$.

1. *Choose a starting point* x_0. *Set* $x = x_0$.

2. *At* x, *calculate the local center of mass* $\mu^x = \sum_{i=1}^n w_i X_i$, *where*

   $w_i = K_H(X_i - x)X_i / \sum_{i=1}^n K_H(X_i - x)$, *with bandwidth matrix* H.

3. *Compute the* 1^{st} *local eigenvector* γ^x *of* $\Sigma^x = (\sigma^x_{jk})_{(1 \le j,k \le d)}$, *where*

   $$\sigma^x_{jk} = \sum_{i=1}^n w_i (X_{ij} - \mu^x_j)(X_{ik} - \mu^x_k).$$

4. *Step from* μ^x *to* $x := \mu^x + t_0 \gamma^x_1$.

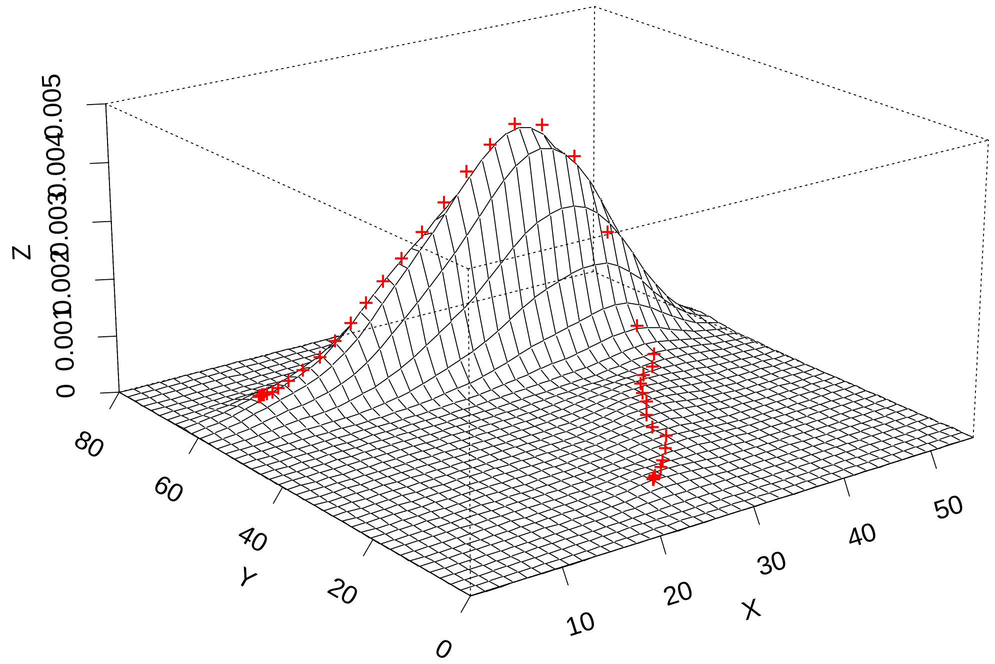
5. *Repeat steps 2. to 4. until the* μ^x *remain constant. Then set* $x = x_0$, *set* $\gamma^x := -\gamma^x$ *and continue with 4.*

The sequence of the local centers of mass μ^x makes up the local principal curve (LPC).

Background

- LPCs can be seen as a simplified version of Delicado's approach. Both algorithms can be shown to differ essentially by the type of weighting and centering used in Σ^x. But Delicado's Σ^x depends on the 'principal direction' b, ruling out a simple eigenanalysis as for LPCs.

- A local principal curve approximates the density ridge. For instance, speed-flow data:



Kernel density estimate:

$$\hat{f}_K(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

Comaniciu & Meer (2002):

'Mean Shift' $\mu^x - x \sim \nabla \hat{f}_K(x)$

Technical Details

- "Signum flipping": Check in every cycle if

$$\gamma^x_{(i-1)} \circ \gamma^x_{(i)} > 0.$$

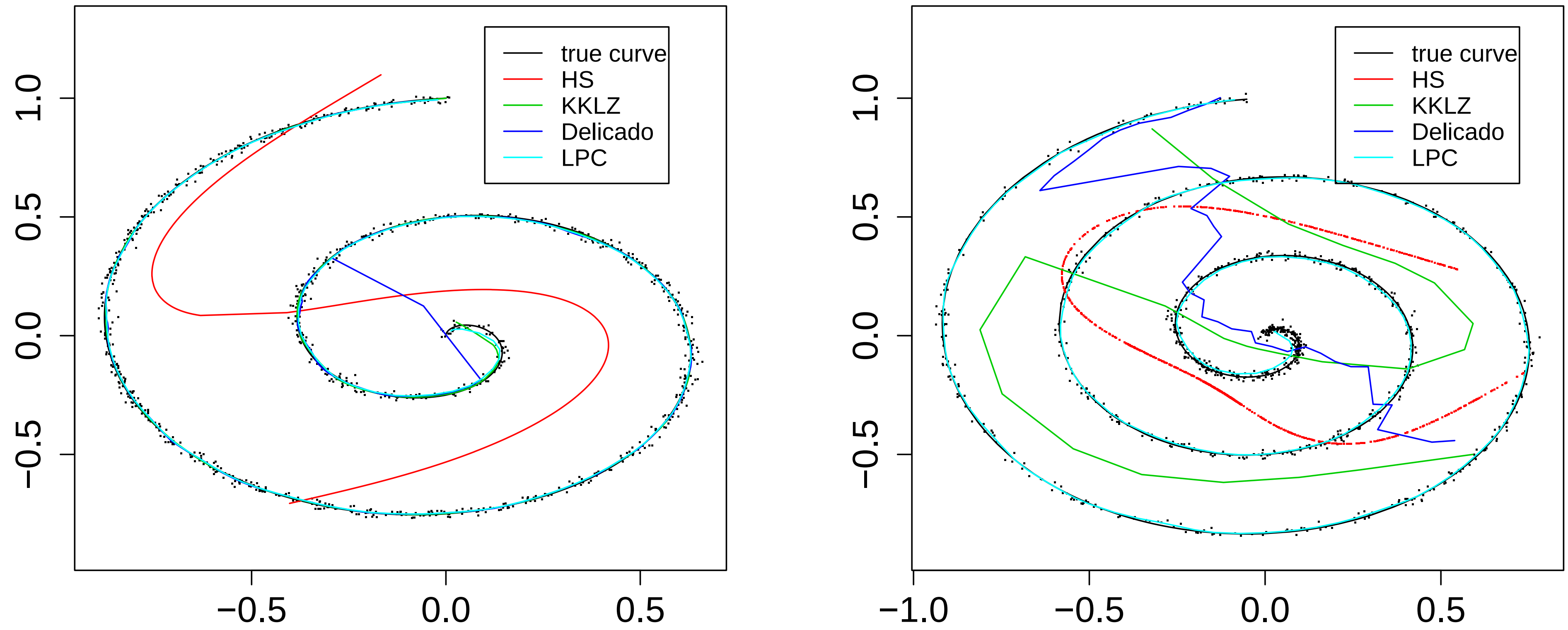
  Otherwise, set $\gamma^x_{(i)} := -\gamma^x_{(i)}$.

- Angle penalization, to hamper the principle curve from bending off at crossings.

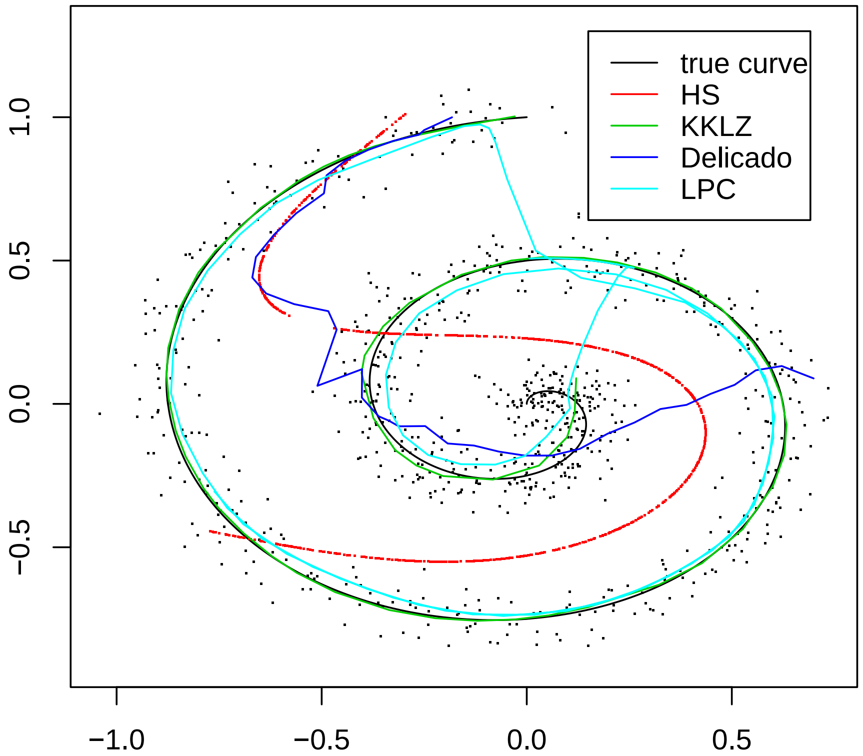
- Use multiple initializations if data cloud consists of several branches (e.g. using a random generator).

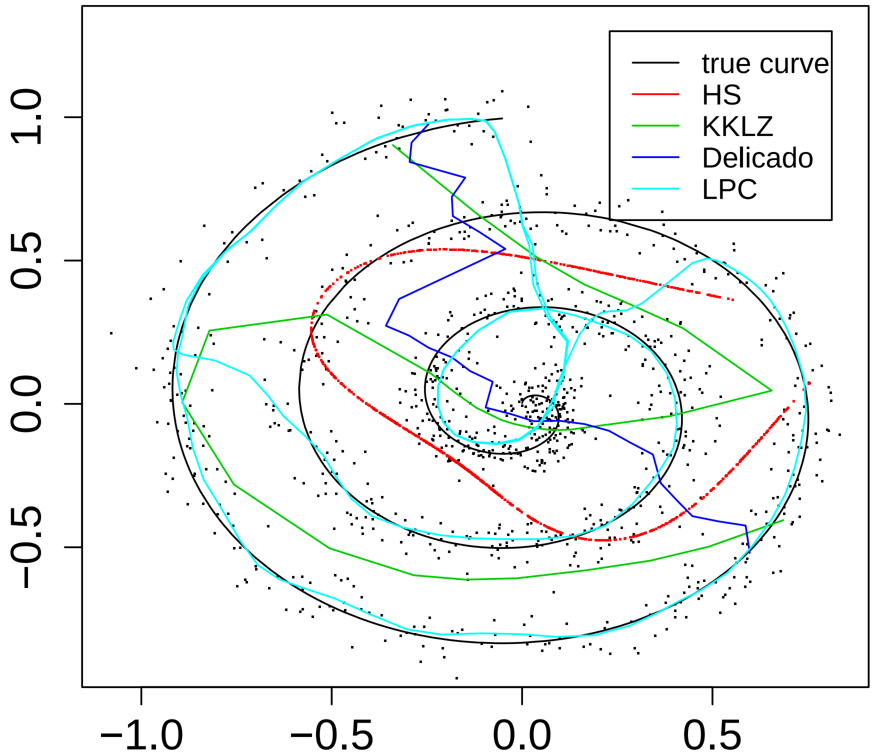
Simulated Examples

Spirals with small noise



Spirals with large noise





Measuring performance: Coverage

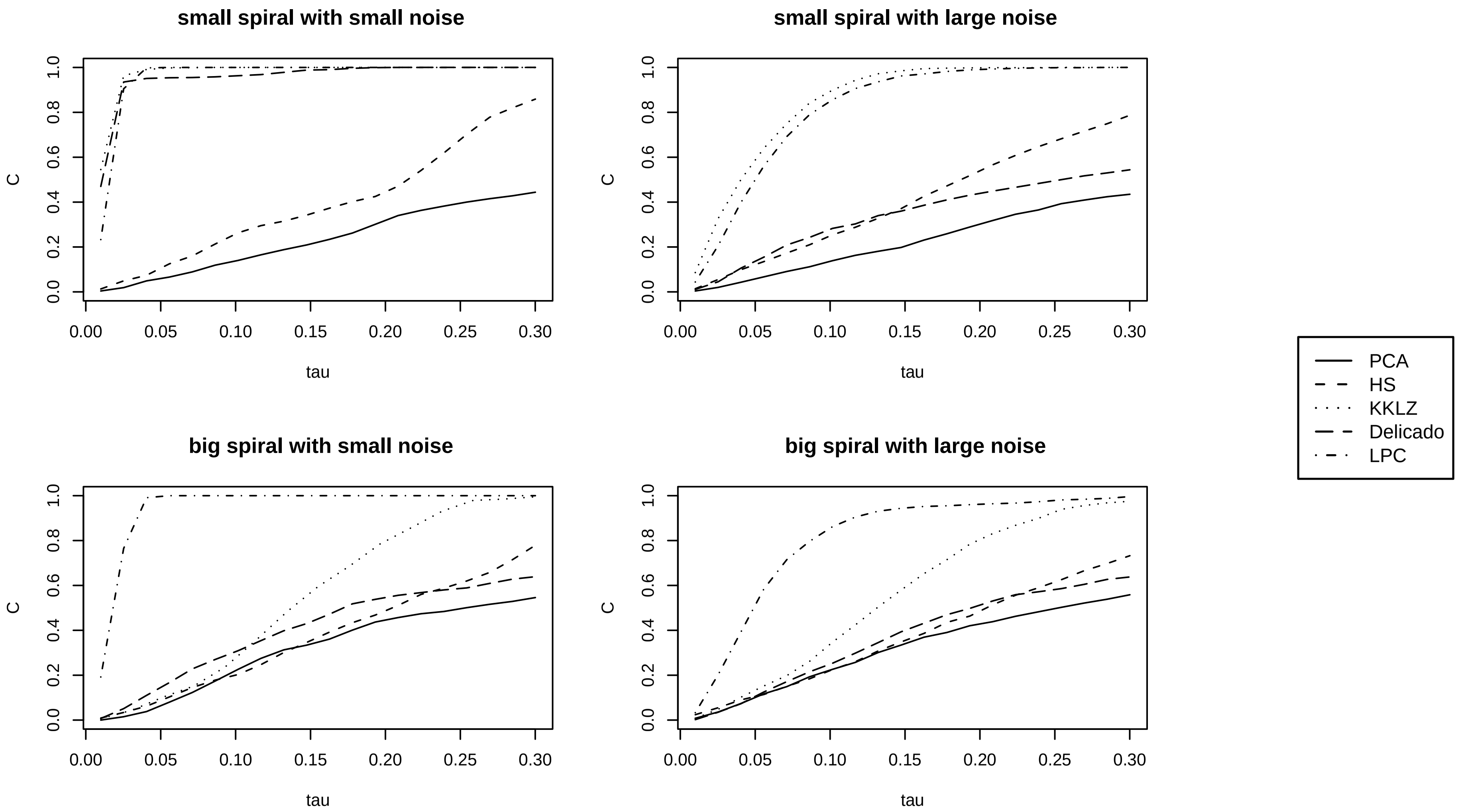
The coverage of a principal curve is the fraction of all data points found in a certain neighborhood of the principal curve.

Formally, for a principal curve m consisting of a set P_m of points, the coverage is given by

$$C_m(\tau) = \#\{x \in X | \exists p \in P_m \text{ with } ||x - p|| \leq \tau\}/n$$

- The coverage can also be interpreted as empirical distribution function of the residuals.
- The area between $C_m(\tau)$ and the constant 1 corresponds to the mean length of the observed residuals.

Coverage for spiral-data



Residual mean length relative to principal components (A_C):

A_C	small spiral		big spiral	
	small noise	large noise	small noise	large noise
HS	0.72	0.77	0.92	0.92
KKLZ	0.03	0.20	0.50	0.65
Delicado	0.05	0.85	0.87	0.92
LPC	0.05	0.24	0.08	0.29

- The closer to 0, the better the performance
- the quantity $R_C = 1 - A_C$ can be interpreted in analogy to R^2 used in regression analysis

Bandwidth selection with self-coverage

Idea: A bandwidth suitable for computation of a principal curve m should also be able to cover adequately the data cloud. This motivates to define the self-coverage,

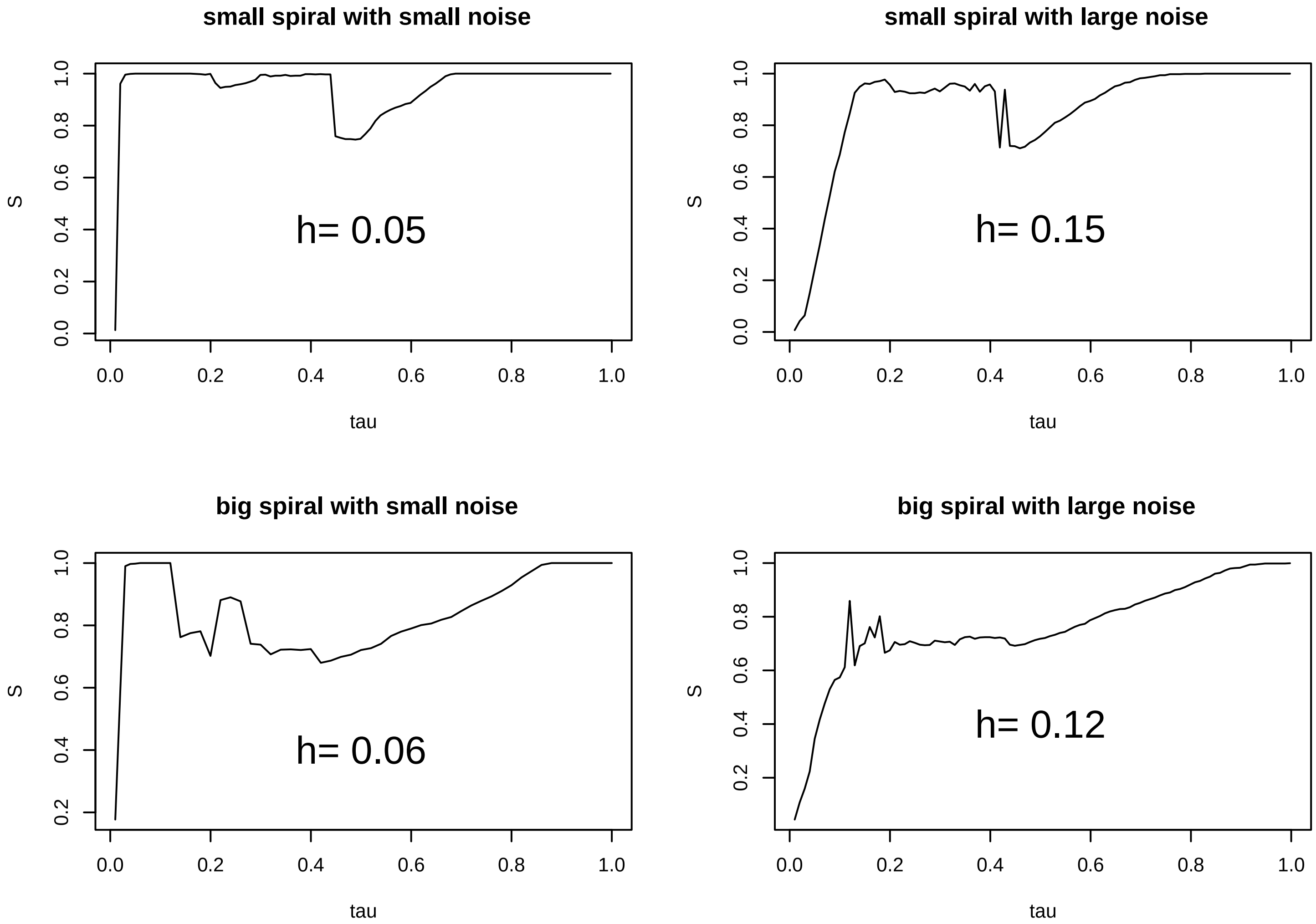
$$S(\tau) = C_{m(\tau)}(\tau) = \frac{\#\{x \in X | \exists p \in P_{m(\tau)} \text{ with } ||x - p|| \leq \tau\}}{n},$$

where $P_{m(\tau)}$ is the set of points belonging to a principal curve $m(\tau)$ calculated with bandwidth τ. Then

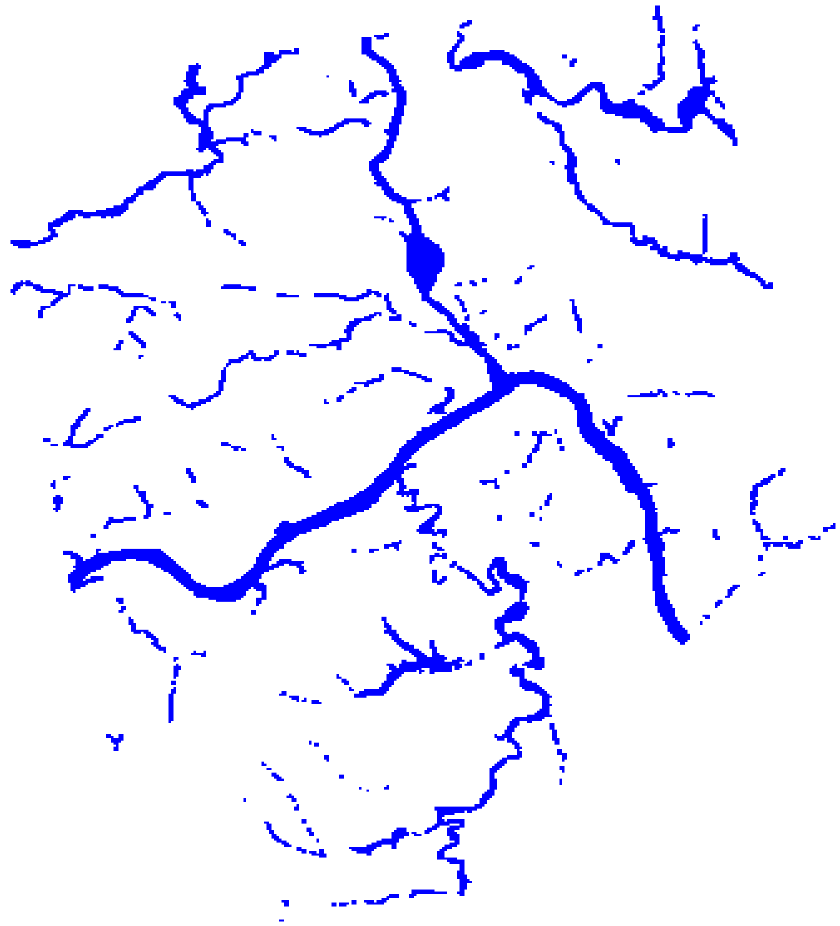
$$h = \text{first local maximum of} \quad S(\tau)$$

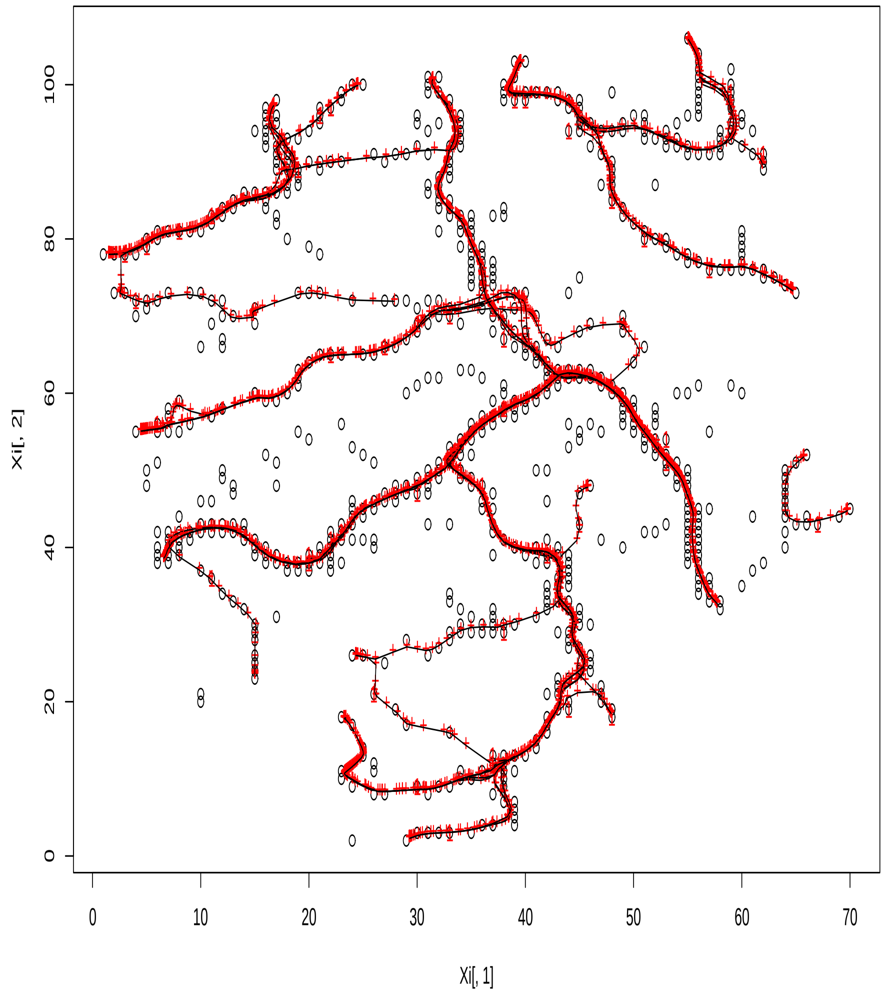
is a suitable bandwidth.

Self-coverage for spiral-data



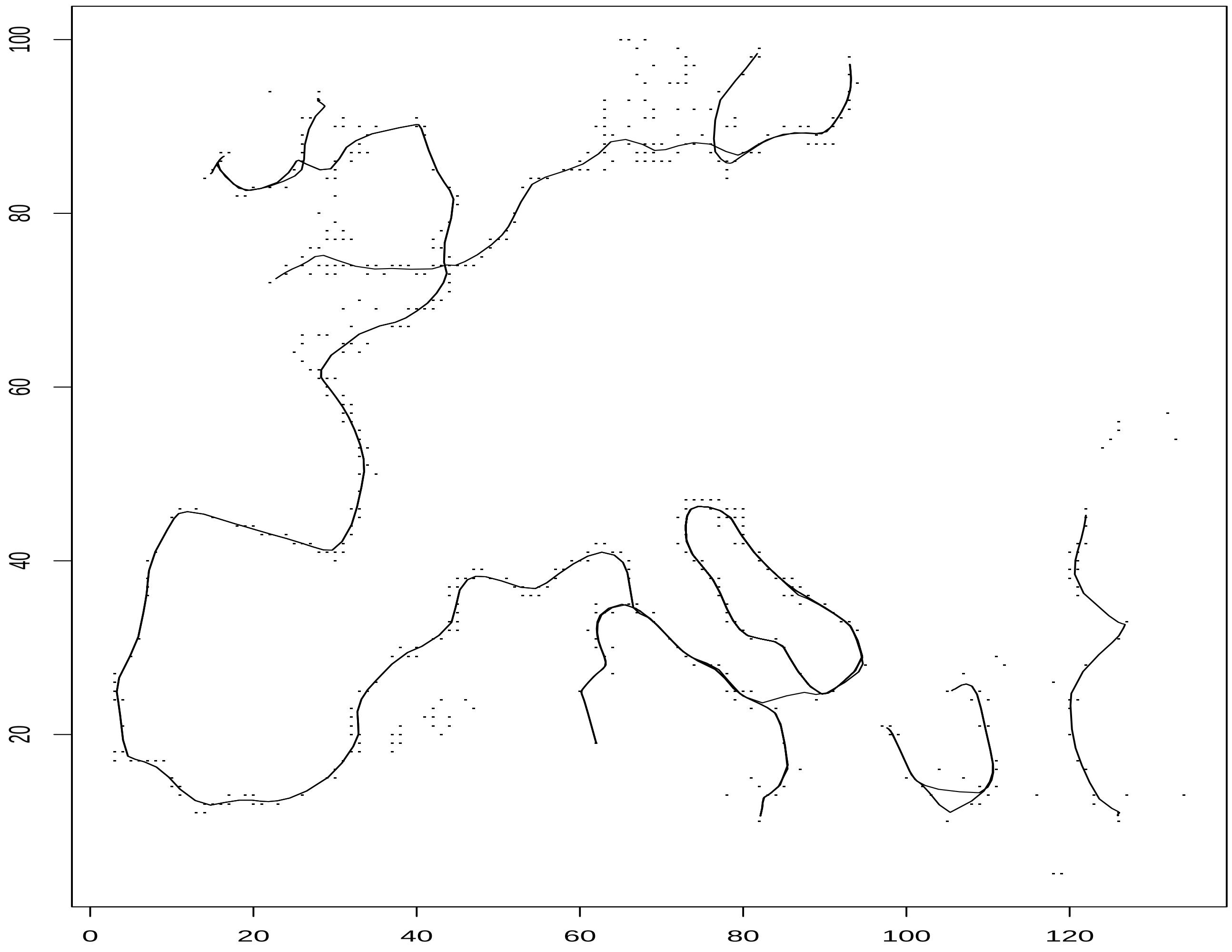
Real data example: Floodplains in Pennsylvania





LPC with multiple (50) initializations.

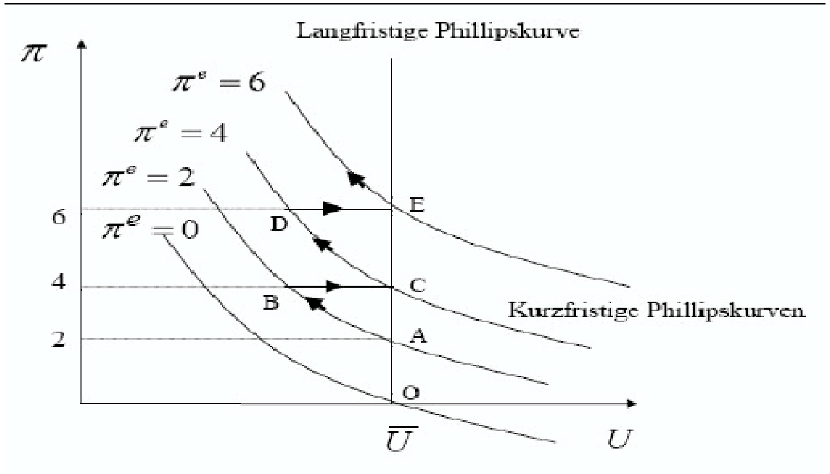
Further example: Coastal Resorts in Europe



3D example: Phillips curves

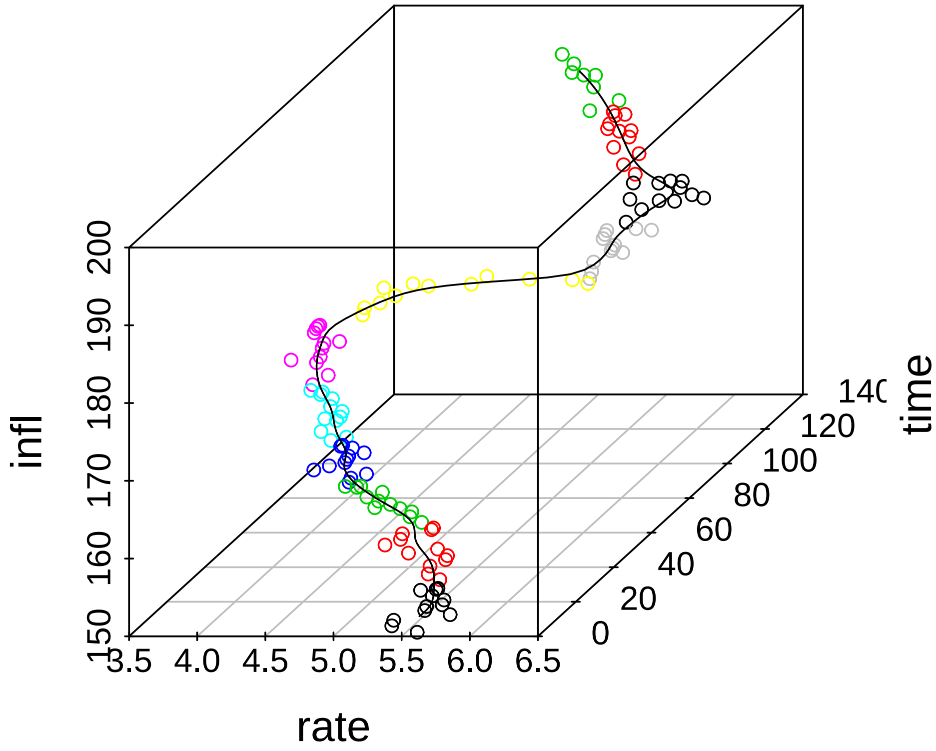
Dependence between inflation (price index) and unemployment rate over time.

Usually just seen as a two-dimensional problem (infl/rate):



(Picture from: Prof. Eisen, University of Frankfurt)

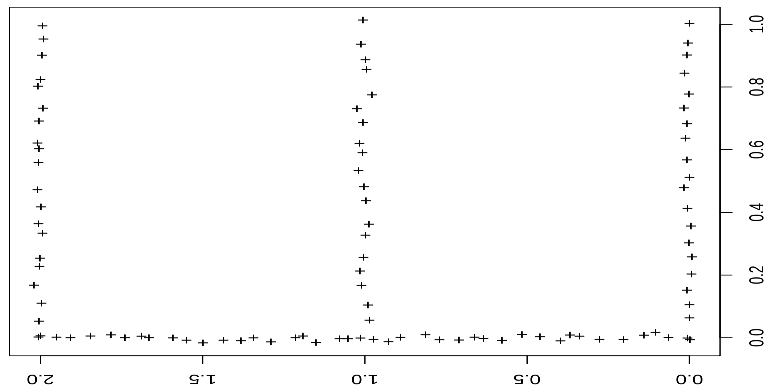
Price index and unemployment in the USA, 1995-2005, with LPC:

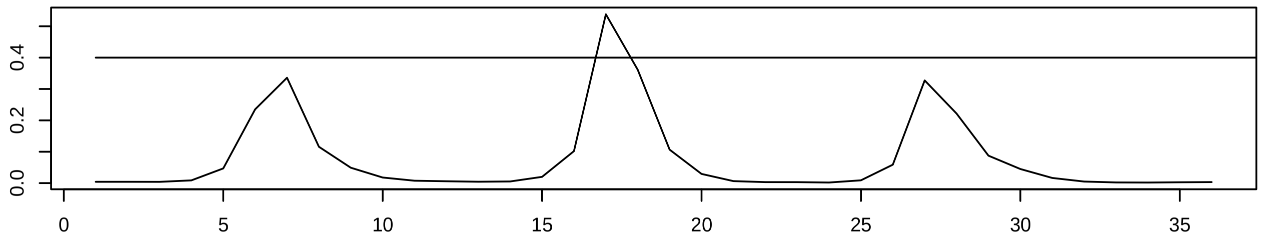


Higher-order-LPC's

Consider the second local eigenvalue λ_2^x, i.e. the second largest eigenvalue of Σ^x: If this value is large at a certain point of the original LPC, a new LPC is launched in direction of the second local eigenvector γ_2^x. Every bifurcation raises the depth of the LPC tree.

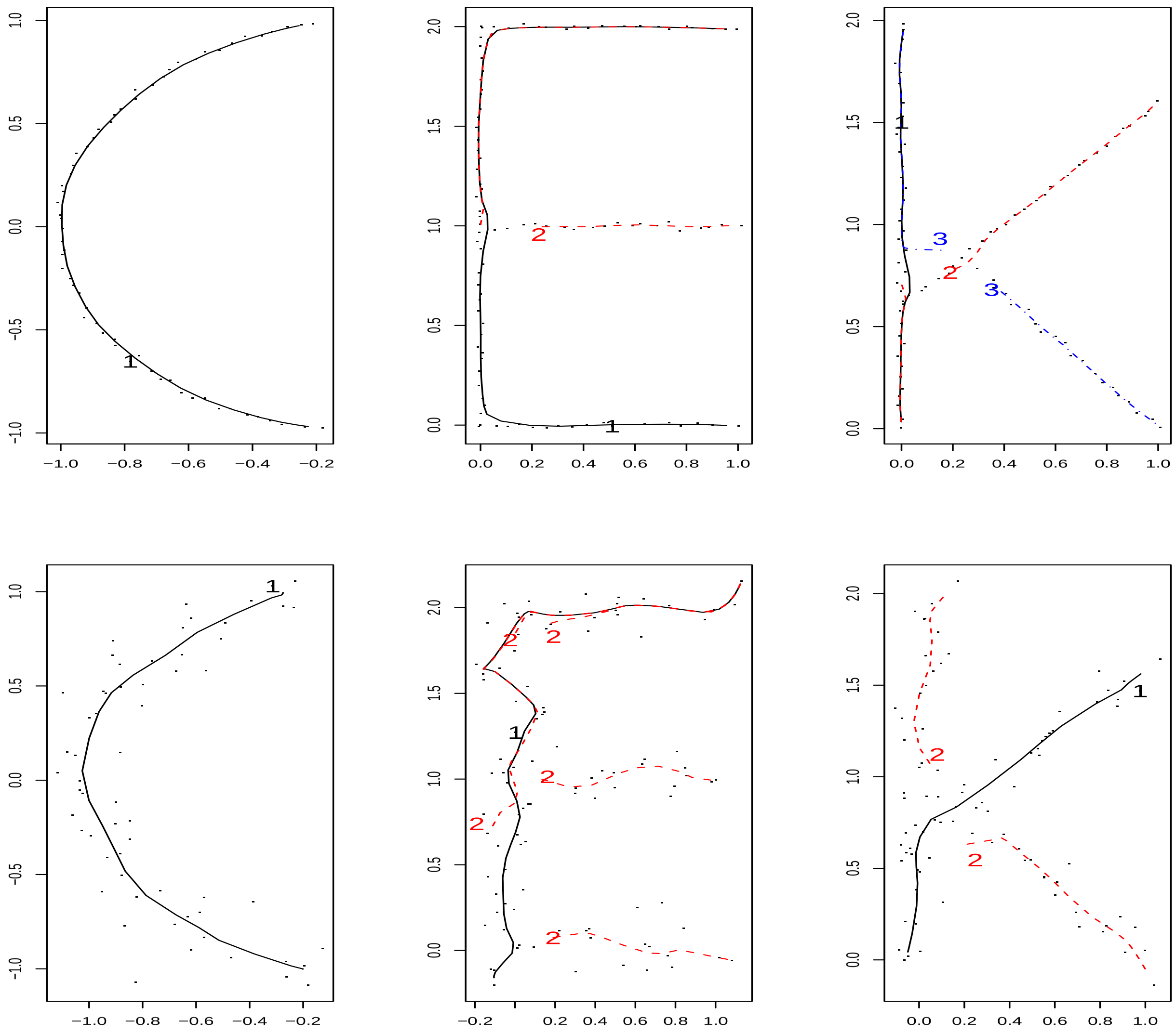
Example





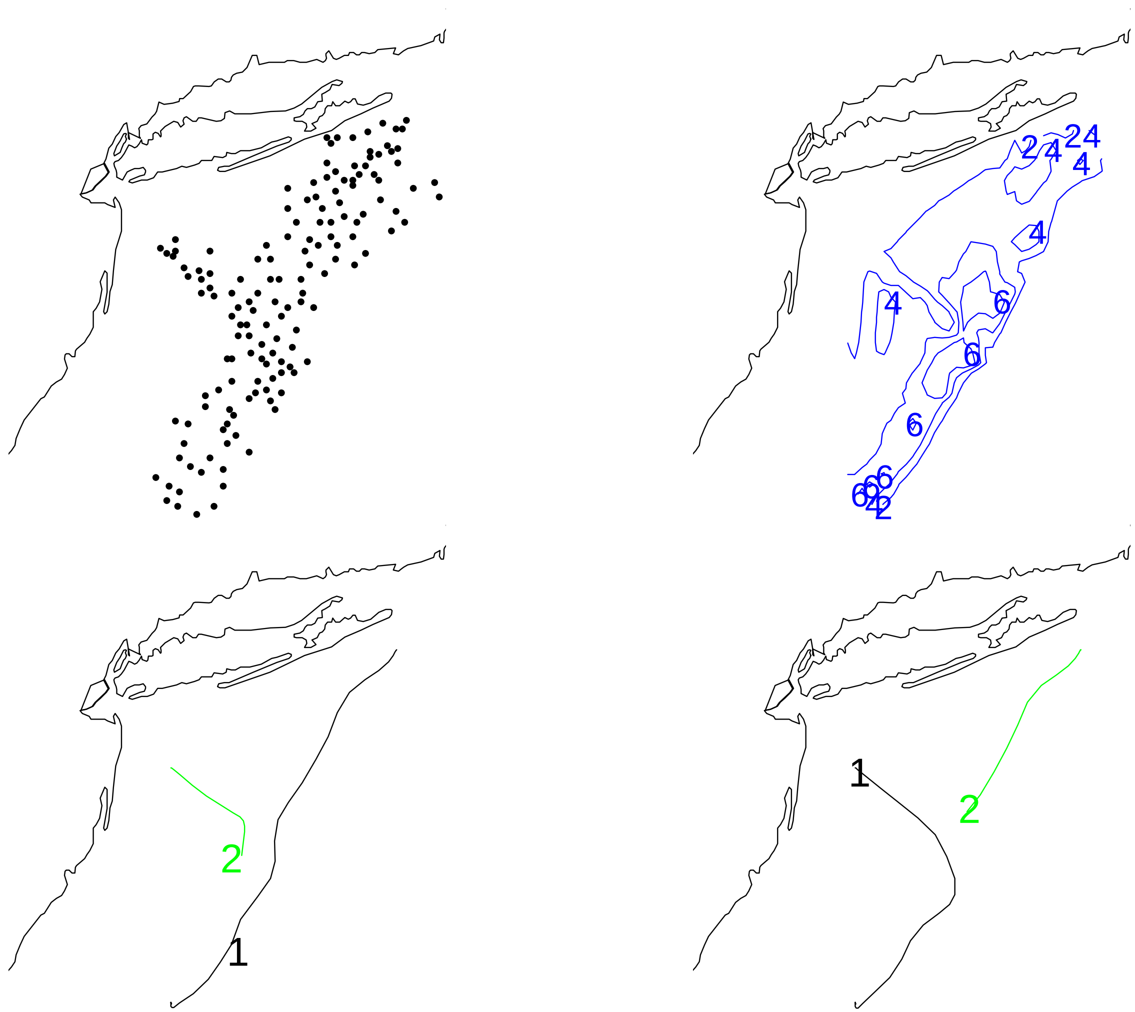
Simulated E and flow diagram of relation $\lambda_2^x / \lambda_1^x$.

LPC's through simulated letters (C,E,K)



LPC's and corresponding starting points with depth 1, 2, 3.

Example: Scallops



Top left:	Scallops
Top right:	Water depth
Bottom left, right:	Two LPC's
1, 2:	Branches of depth 1, 2.

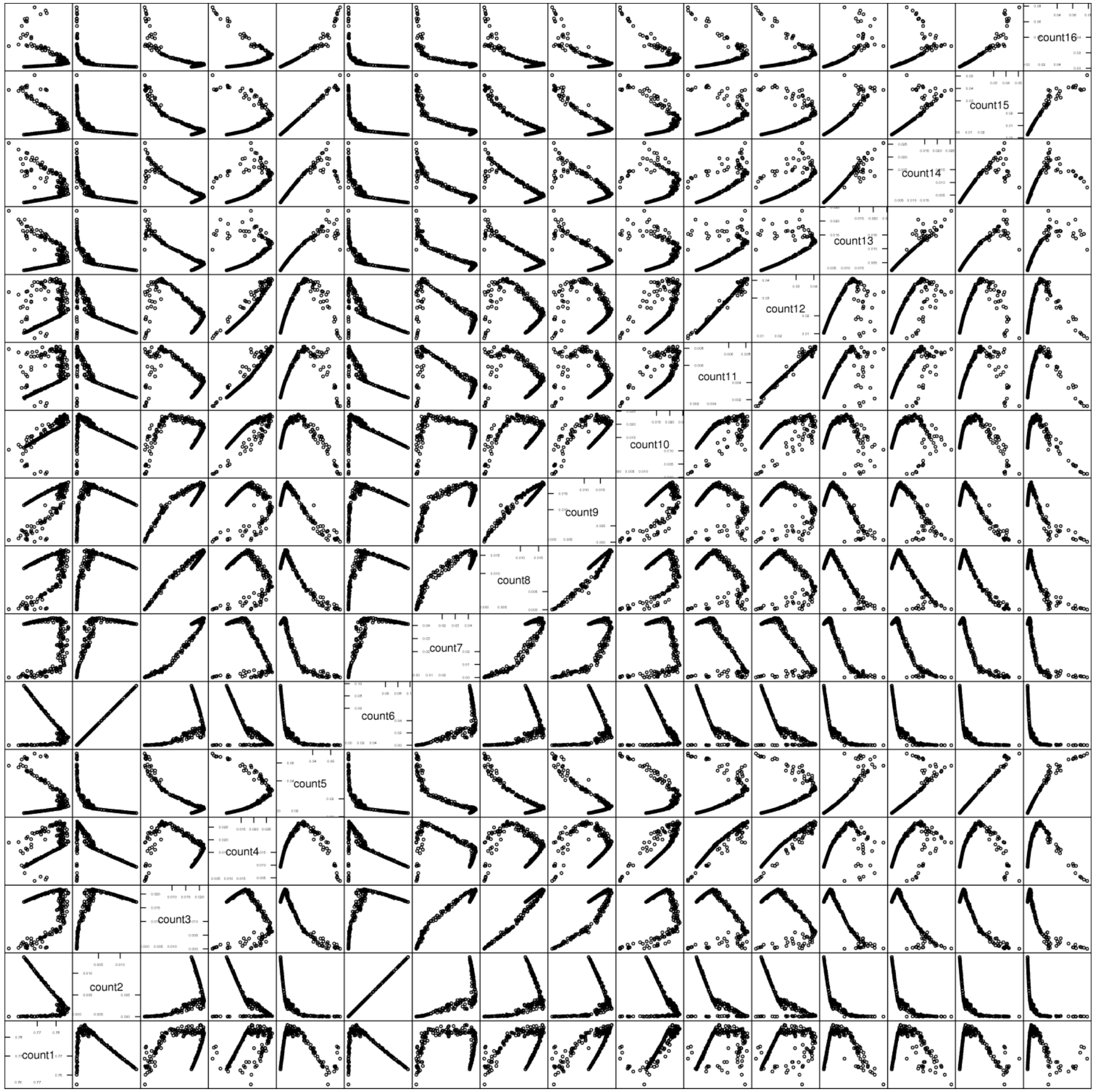
Principal curves as a dimension reduction tool

Consider a nonparametric regression problem with multivariate predictor, i.e.

$$y = f(x_1, \ldots, x_d) + \epsilon$$

- For dimensions $d > 2$ or 3, nonparametric multivariate regression is often infeasible due to the "curse of dimensionality".
- Generally, regression methods do not take an inherent structure of the covariate space into account.
- If the information contained in the covariate space can be approximated by a smooth principal curve (or manifold), this curve can be used as a low-dimensional predictor.
- Application: The GAIA mission. Estimate physical properties of stars based on photometric data (photon counts for 16 frequency/colour bands). One of the predictors is the stellar temperature.

GAIA data: The covariate space



Conclusions

- LPCs work well in a variety of data situations, and seem to be more suitable for some noisy complex structures than its competitors.
- The price to be paid for the increase in flexibility is an increase in variability. Always compute several LPCs to confirm the first run!
- Bandwidth selection works by means of a coverage measure.
- LPCs are not based on a statistical model and hence there is no 'true' principal curve.
- R Code and all data examples available at:

`http://www.maths.dur.ac.uk/~dma0je/lpc/lpc.htm`

Literature

Comaniciu & Meer (2002): Mean shift: A robust approach towards feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.* **24**, *603-619*.

Hastie & Stuetzle (1989): Principal curves. *JASA* **84**, 502–516.

Kégl, Krzyżak, Linder & Zeger (2000): Learning and design of principal curves. *IEEE Transactions Patt. Anal. Mach. Intell.* **22**, 281–297.

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

Einbeck, Tutz & Evers (2005): Local principal curves. *Statistics and Computing* **15**, 301–313.

Einbeck, Tutz & Evers (2005b): Exploring multivariate data structures with local principal curves. In: Weihs, C. and Gaul, W. (Eds.): *Classification - The Ubiquitous Challenge*. Springer, Heidelberg.

Einbeck, Evers & Bailer-Jones (2007): Representing complex data using localized principal components with application to astronomical data. *Lecture Notes in Comp. Science and Engineering*, to appear.