Local Principal Curves

Jochen Einbeck

Department of Mathematics, NUI Galway

jochen.einbeck@nuigalway.ie

Dublin - 22nd November 2005

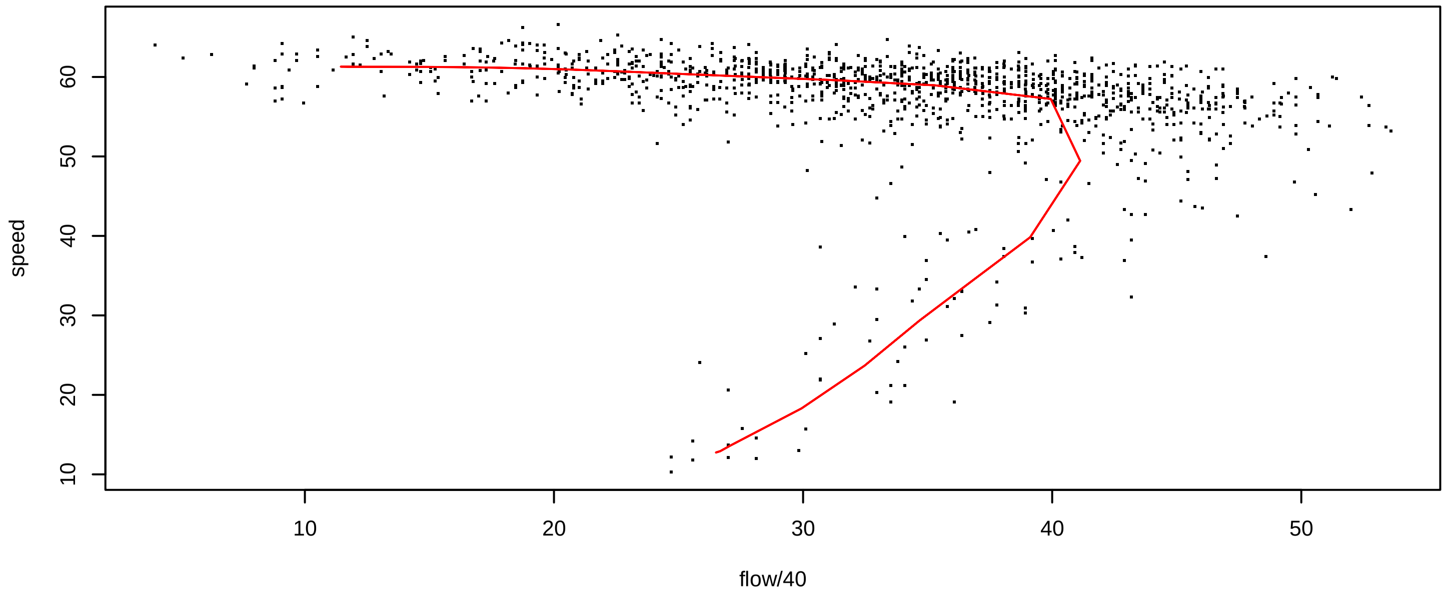
joint work with

Gerhard Tutz, University of Munich, and Ludger Evers, University of Oxford.

Descriptive Definition

Principal Curves are smooth curves passing through the 'middle' of a multidimensional data cloud $X = (X_1, \ldots, X_n)$, where $X_i \in \mathbb{R}^d$.

Example: Speed-Flow diagram.



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 an initial 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.
- Local ('bottom-up') algorithms estimate the principal curve locally moving step by step through the data cloud.

Principal curve definitions associated to ‘top-down’ - approaches

Hastie & Stützle (HS, 1989) define a point on the principal curve as the average of all points which project there (‘self-consistency’).

Self-consistent curves $m : I \longrightarrow \mathbb{R}^d$ are obtained as critical points of the distance function

$$\triangle(m) = E\left(\inf_t ||X - m(t)||^2\right) \qquad (1)$$

and generalize linear principal components in a natural way.

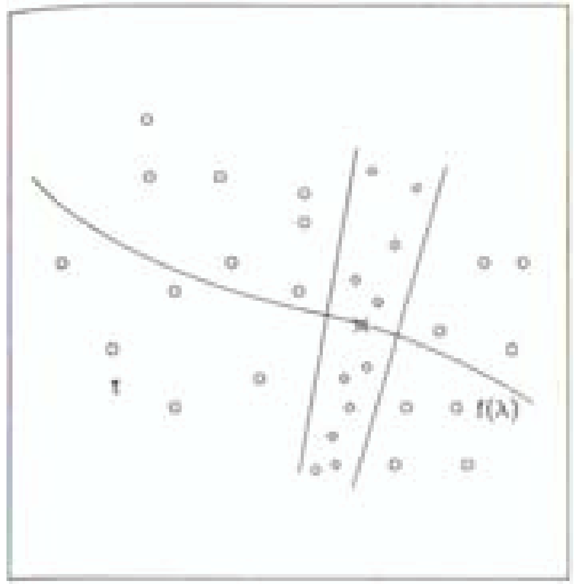


Figure 3. Each point on a principal curve is the average of the points that project there.

(Picture from: Hastie & Stützle, 1989)

Kégl, Krzyzak, Linder & Zeger (KKLZ, 2000)

define a principal curve as the curve minimizing the average squared distance (1) over all curves with bounded length L.

Tibshirani (1992) defines principal curves such that for data generated as

$$X = m(t) + \epsilon \qquad \text{with} \quad E(\epsilon) = 0$$

curve m is also principal curve of the data cloud X.

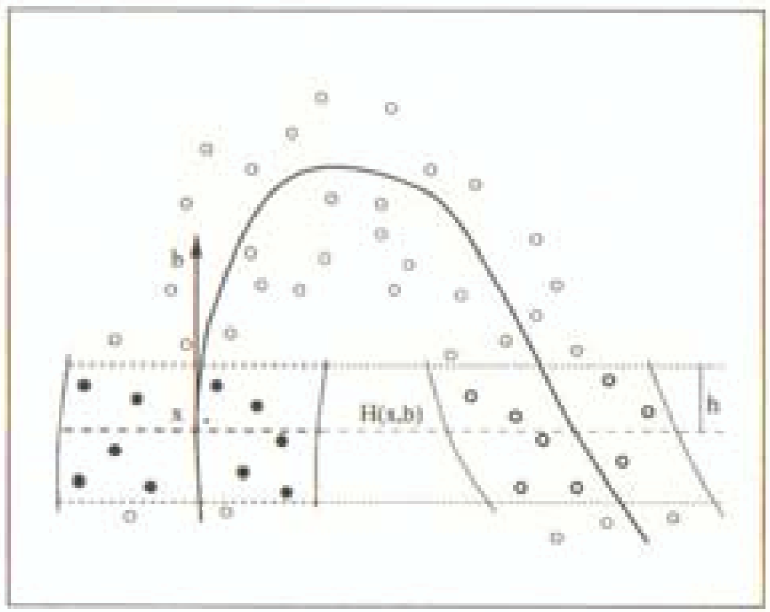
Properties of ‘top-down’ algorithms:

- Starting from the first principal component line of the whole data set, the principal curve is estimated iteratively with EM-like algorithms.
- Dependance on an initial line leads to a lack of flexibility, as an initial unsuitable assignment of projection indices can often not be correct in the further run of the algorithm
- Estimation of branched or disconnected data clouds not (directly) possible.

Alternative: ‘Bottom-up’ algorithms

Delicado (2001) defines principal curves as a sequence of fix 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. He estimates 'PCOPs' using a fix point algorithm moving smoothly through the data cloud.

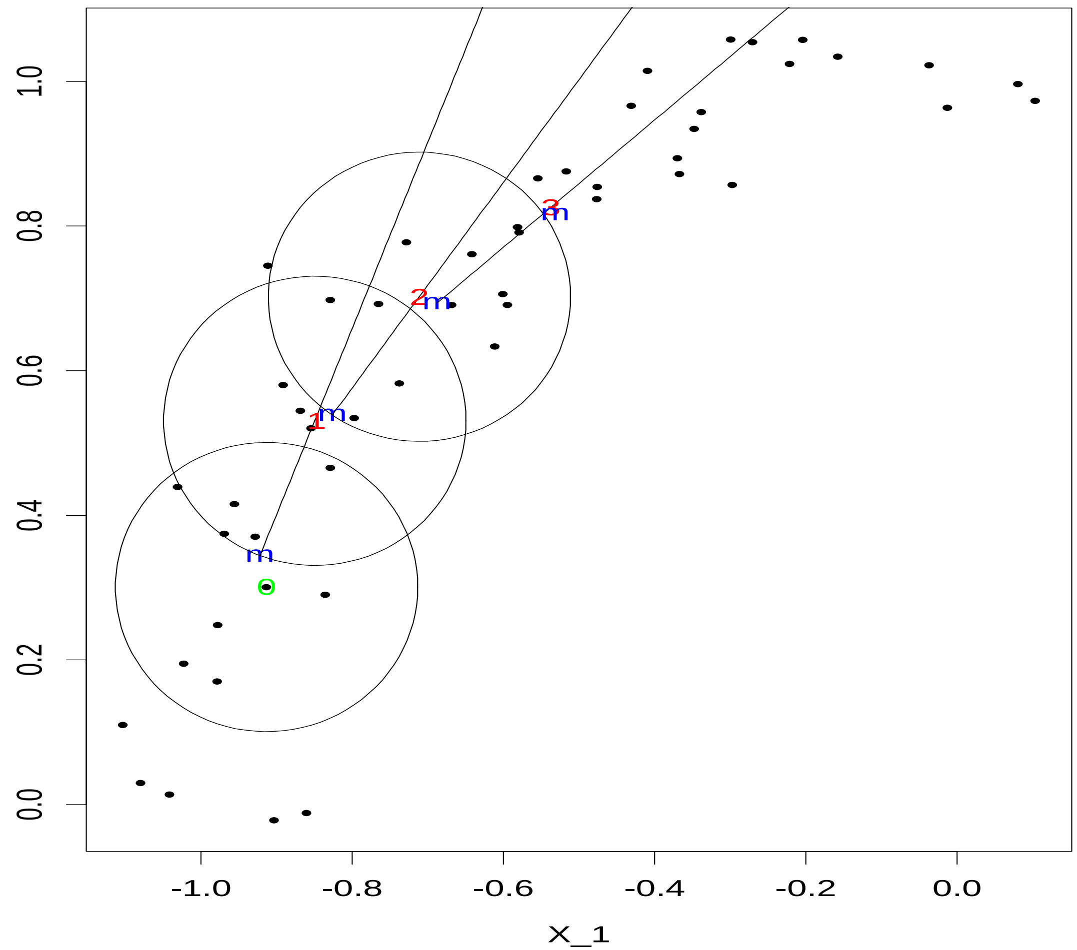
- Works fine for most (not too complex) data sets.
- Mathematically elegant
- However, quite complicated and computationally demanding.
- Requires a cluster analysis at every point of the principal curve.



(Picture from: Delicado, 2001)

A simple alternative ‘bottom-up’ approach: Local principal curves (LPC)

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 LPC's

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*

   $w_i = K_H(X_i - x)X_i / \sum_{i=1}^n K_H(X_i - x)$.
3. *Compute the* 1^{st} *local eigenvector* γ^x *of* $\Sigma^x = (\sigma_{jk}^x)_{(1 \le j,k \le d)}$, *where*

$$\sigma_{jk}^x = \sum_{i=1}^n w_i (X_{ij} - \mu_j^x)(X_{ik} - \mu_k^x).$$

4. *Step from* μ^x *to* $x := \mu^x + t_0 \gamma_1^x$.
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.*

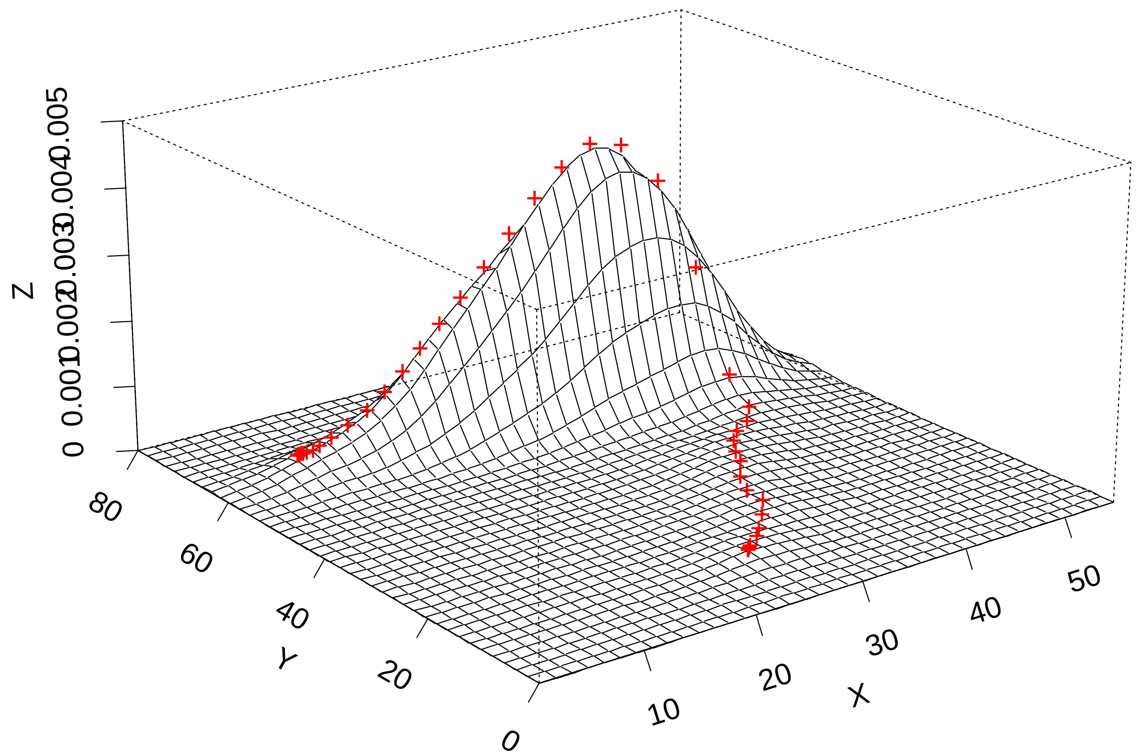
The sequence of the local centers of mass μ^x makes up the local principal curve (LPC).

Background

Kernel density estimate:

$$\hat{f}_K(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \qquad (2)$$

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



Comaniciu & Meer (2002): 'Mean Shift' $\mu^x - x \sim \nabla \hat{f}_K(x)$

Technical Details

- "Signum flipping": Check in every cycle if

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

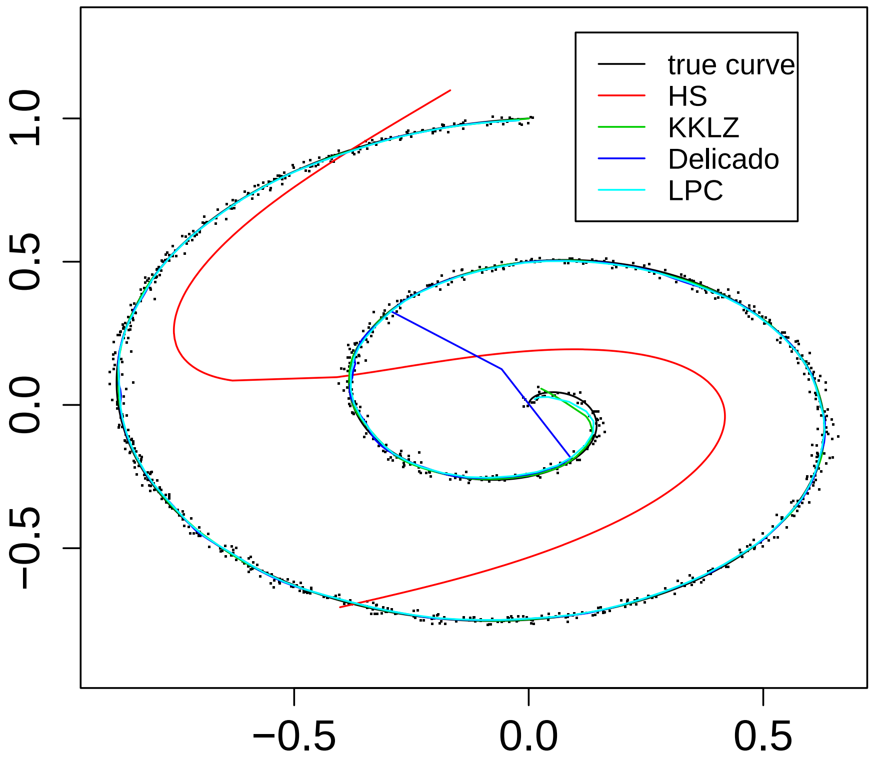
  Otherwise, set $\gamma^x_{(i)} := -\gamma^x_{(i)}$.

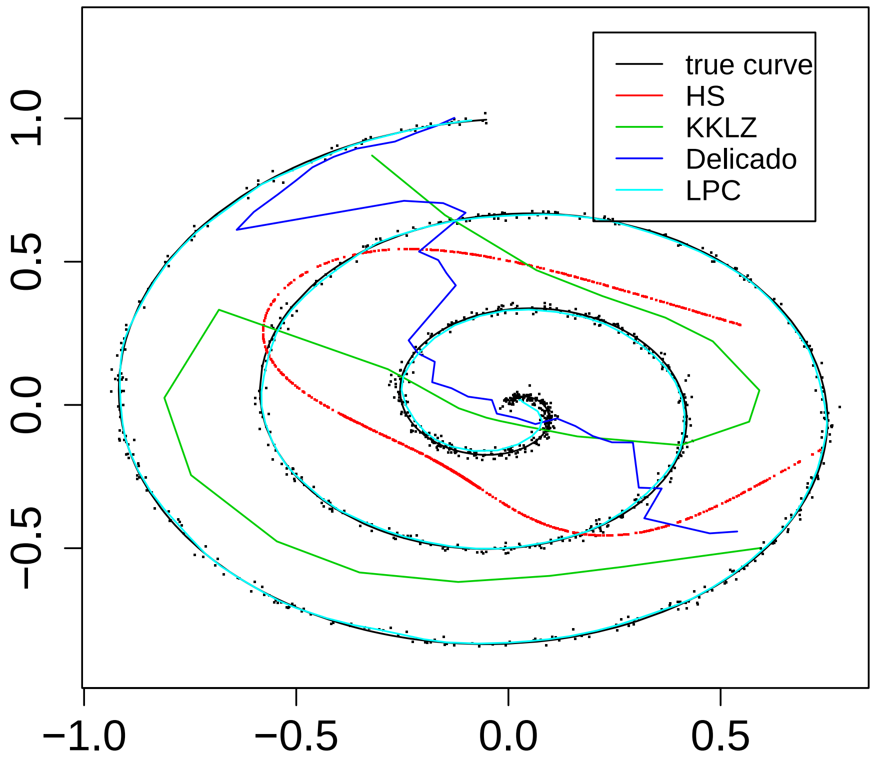
- Angle penalization, to hamper the principle curve from bending off at crossings.

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

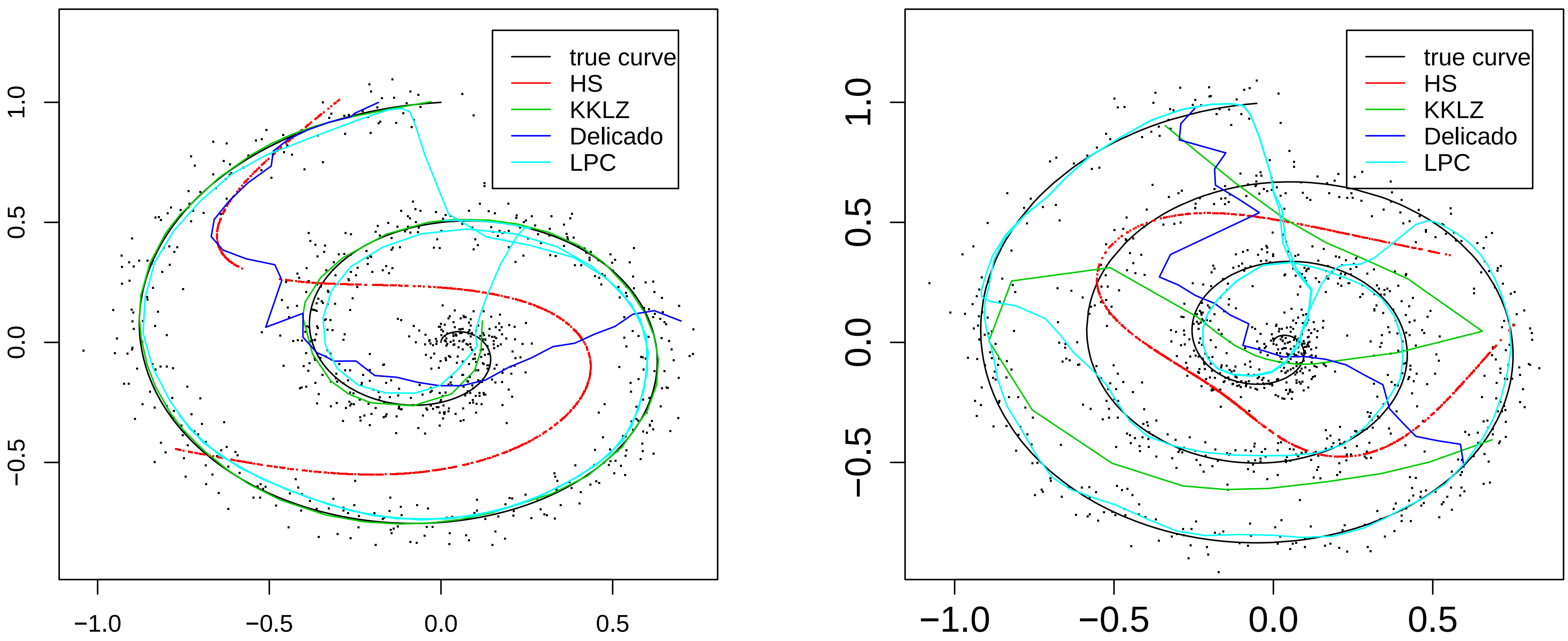
Simulated Examples

Spirals with small noise





Spirals with large noise



Measuring performance: Coverage

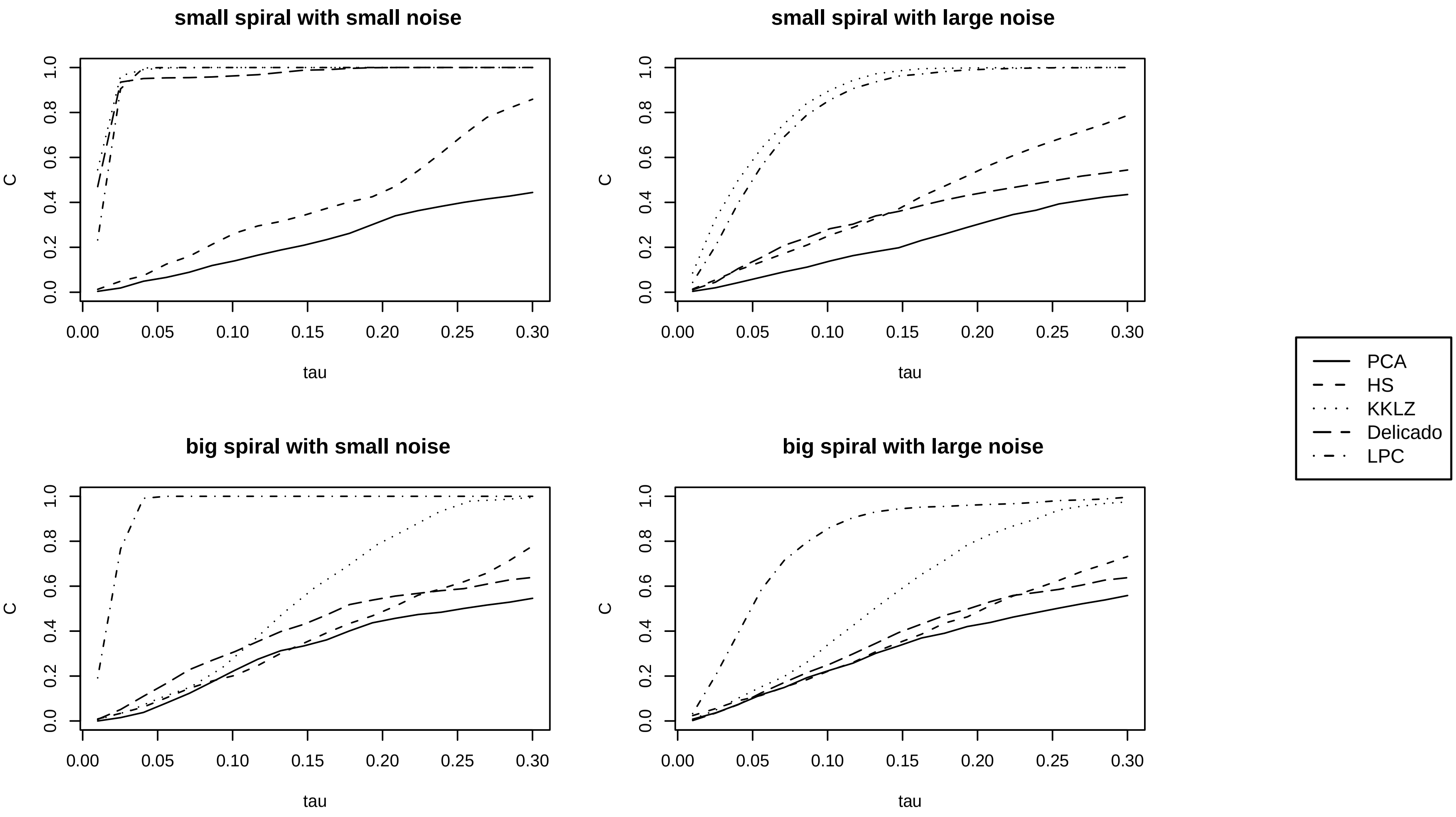
The coverage of a principal curve is the fraction of all data points found in a certain neighborhood of the principal curve.

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

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

- 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,

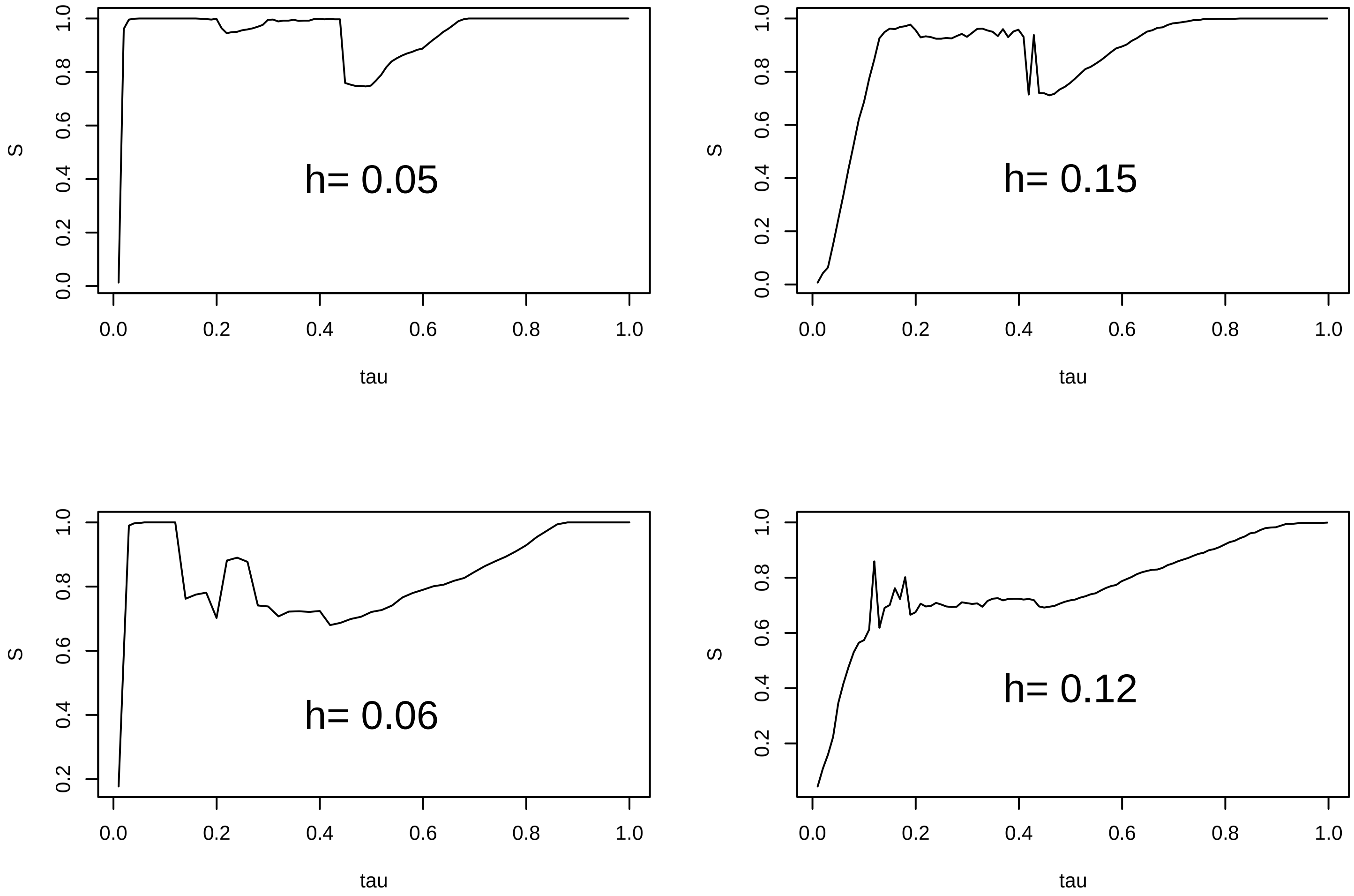
$$S(\tau) = C_{m(\tau)}(\tau) = \frac{\#\{x \in X | \exists p \in P_{m(\tau)} \text{ mit } ||x - p|| \leq \tau\}}{n},$$

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

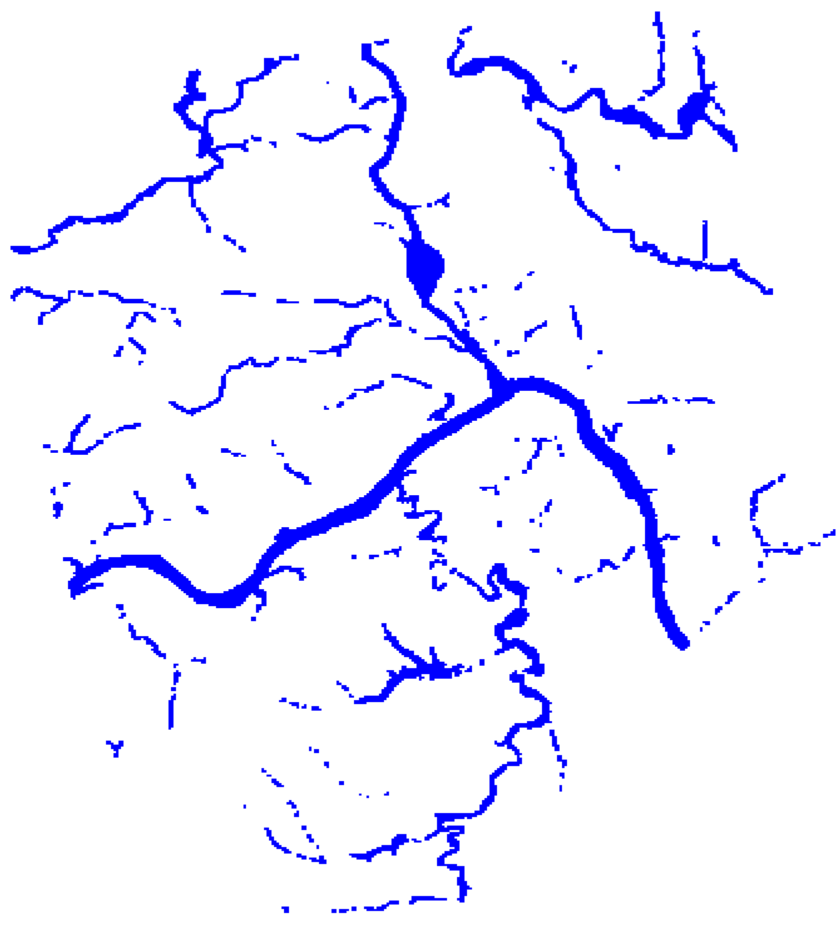
$$h = \text{first local maximum of} \quad S(\tau)$$

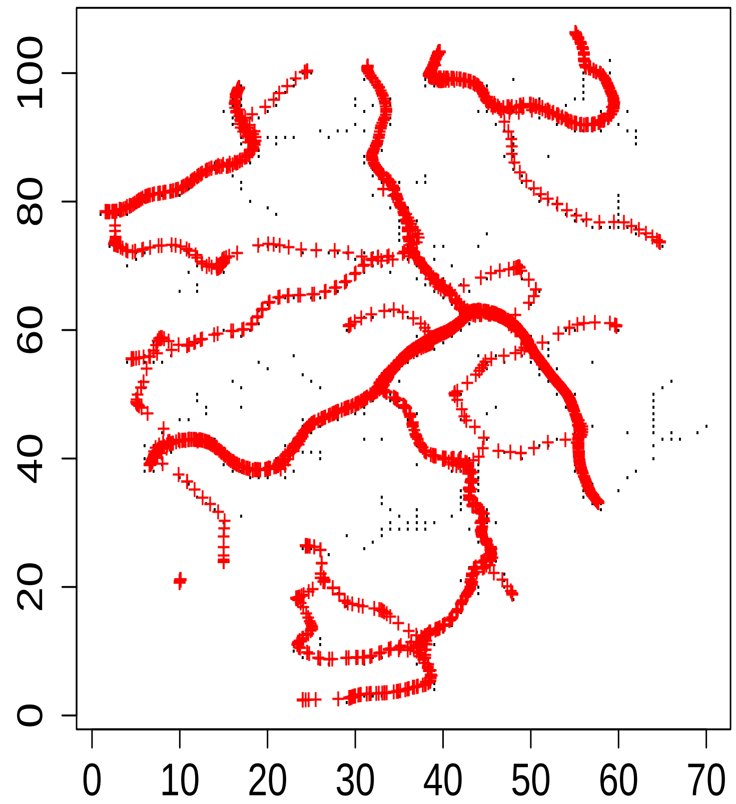
is a suitable bandwidth.

Self-coverage for spiral-data



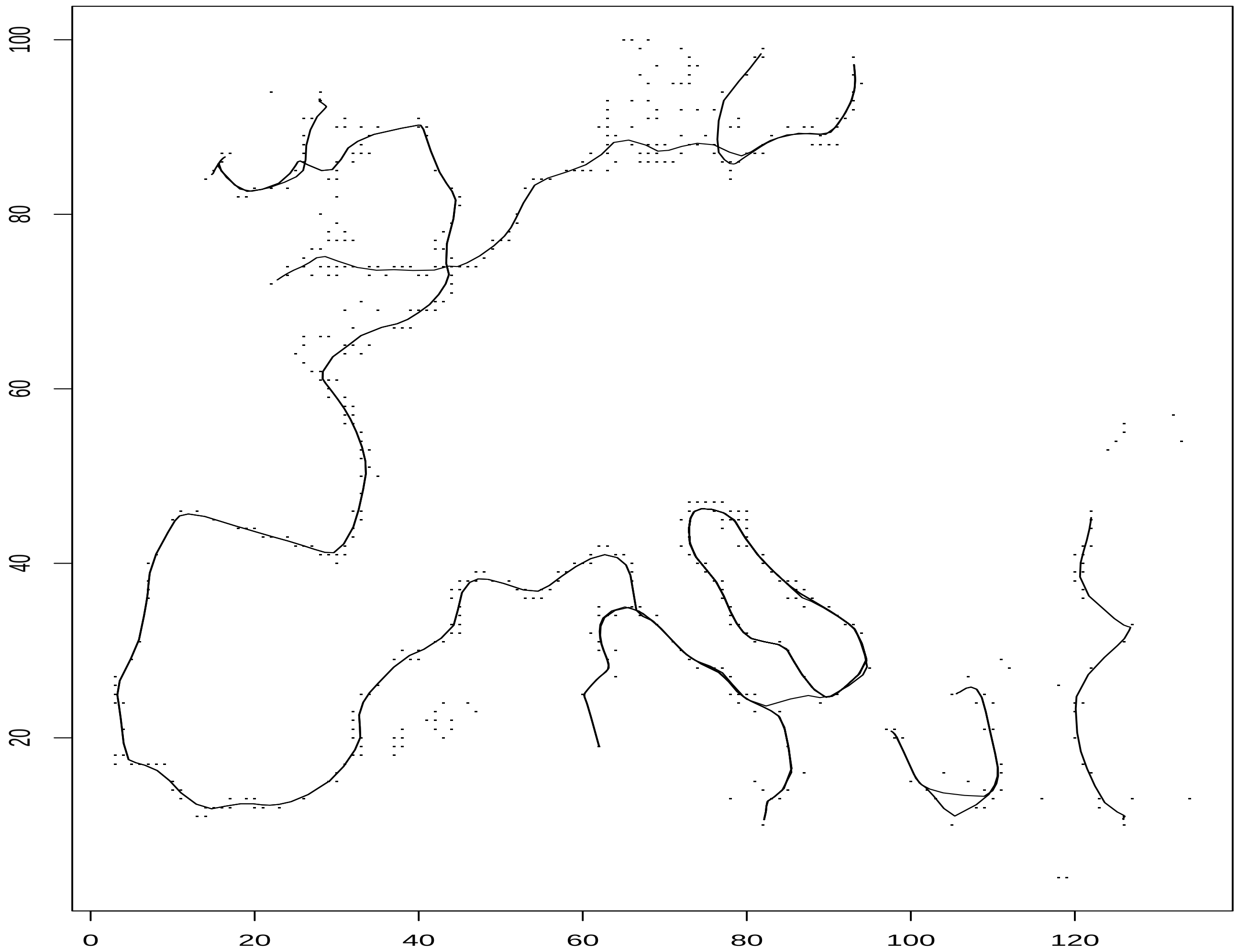
Real data example: Floodplains in Pennsylvania





LPC with multiple (50) initializations.

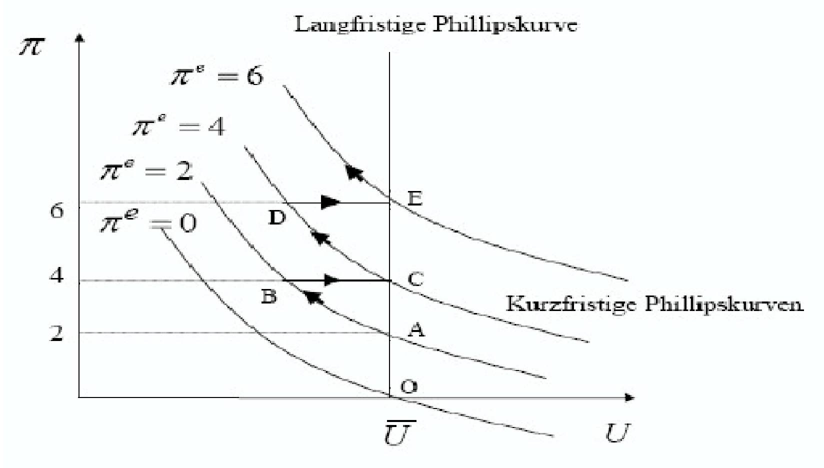
Further example: Coastal Resorts in Europe



3D example: Philips curves

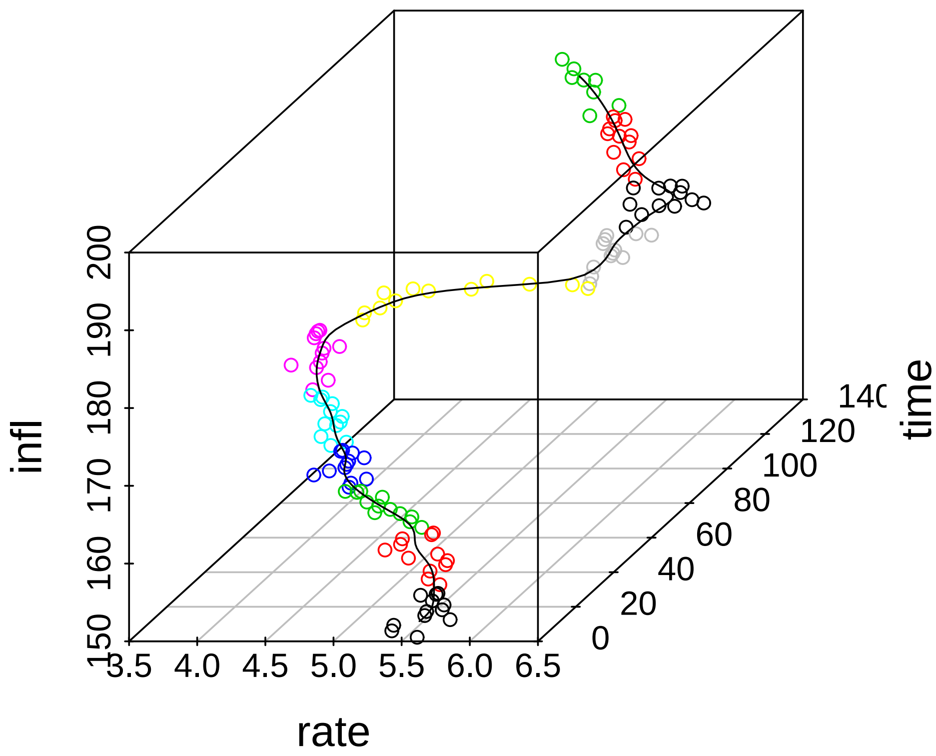
Dependance between inflation (price index) and unemployment rate over time.

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



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

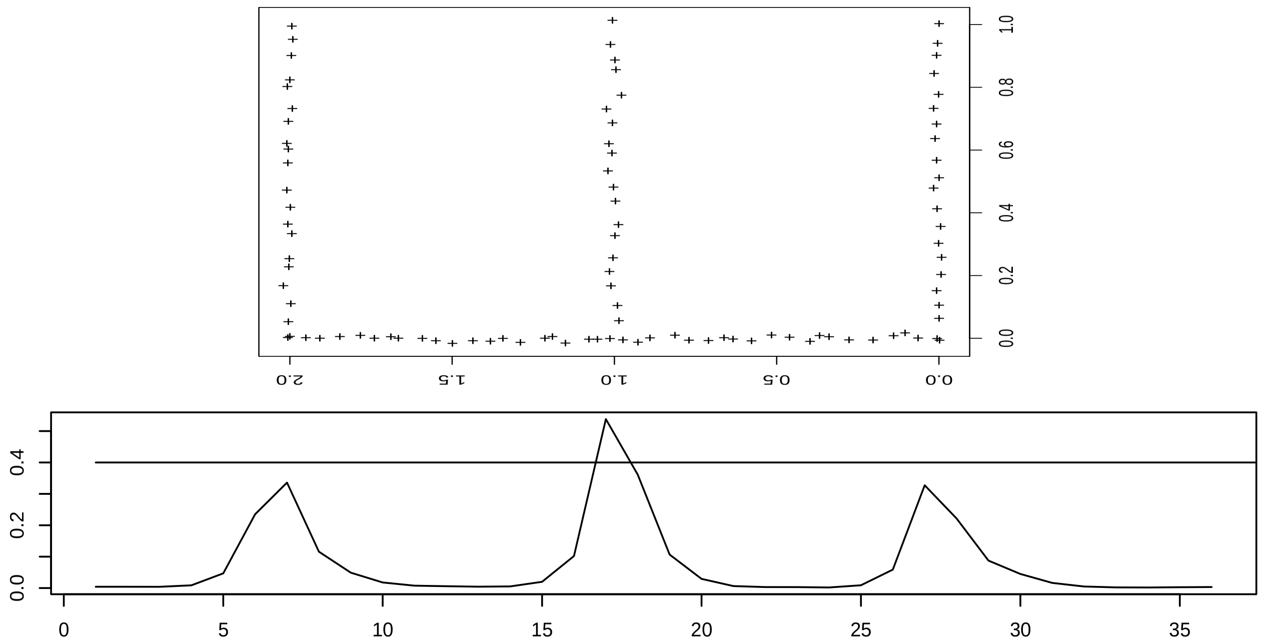
Price index and unemployment in the USA, 1995-2005, with LPC:



Higher-order-LPC's

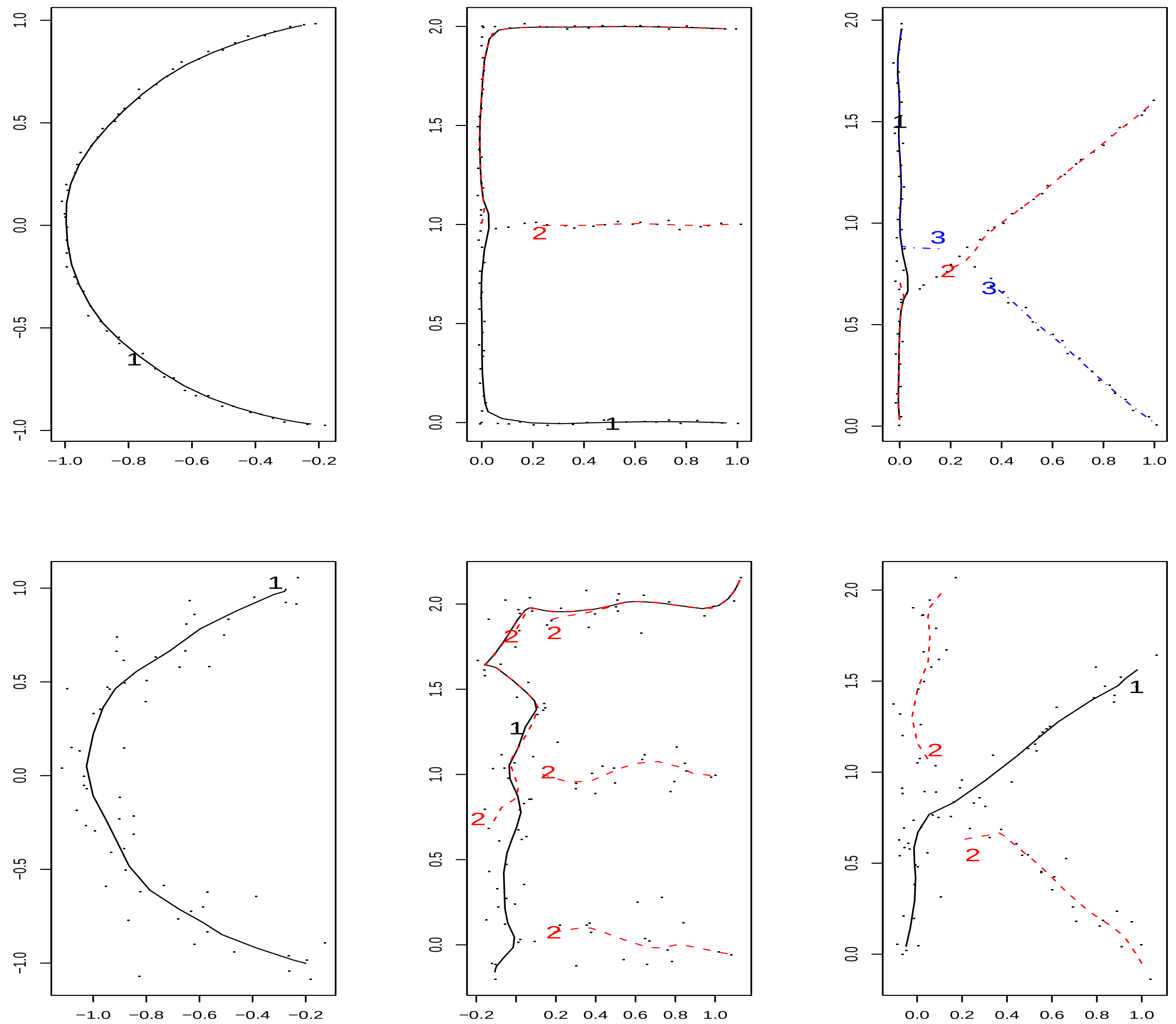
Consider the second local eigenvalue λ_2^x, i.e. the second largest eigenvalue of Σ^x: If this value is large at a certain point of the original LPC, a new LPC is launched in direction of the second local eigenvector γ_2^x. Every bifurcation raises the depth of the LPC tree.

Example



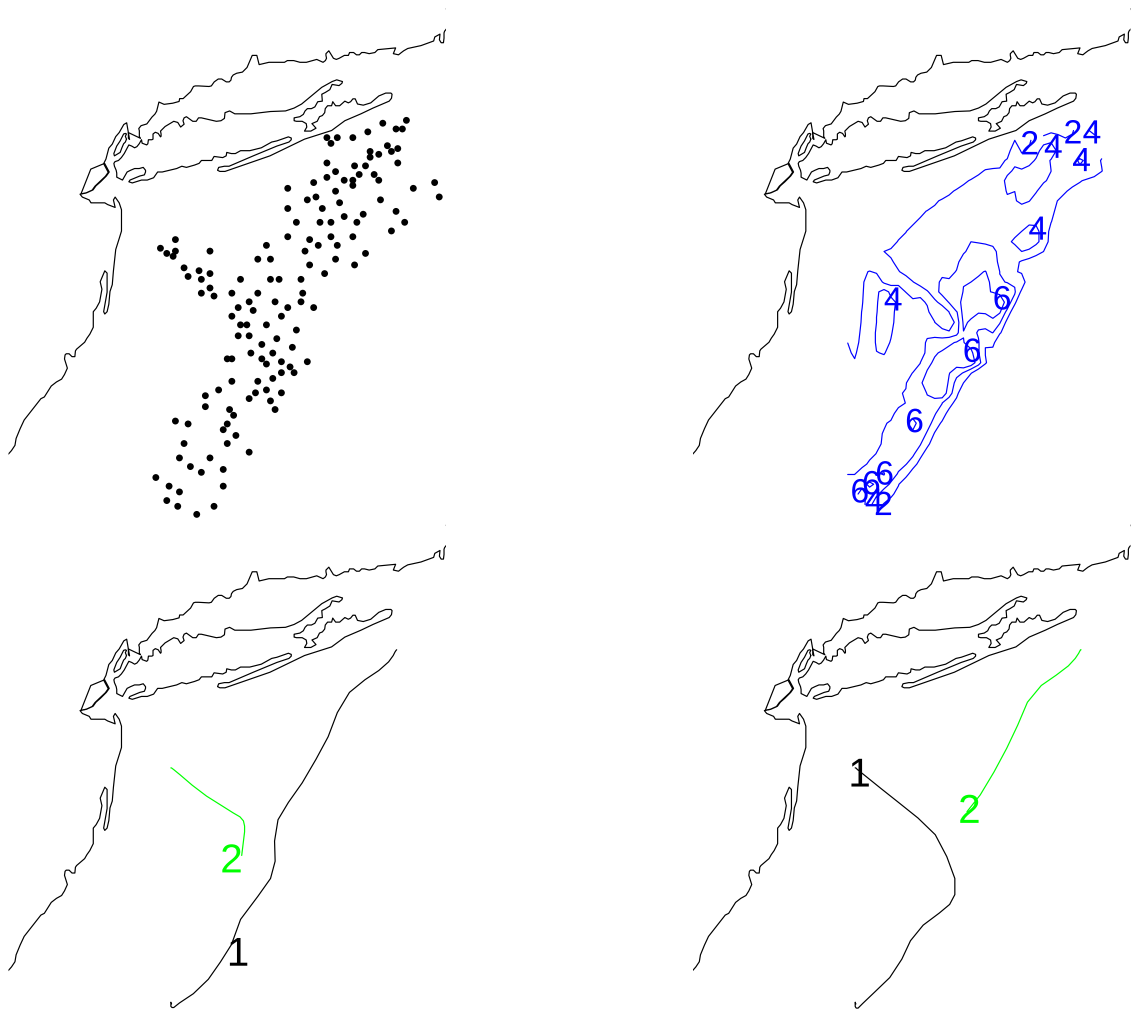
Simulated E and flow diagram of relation λ_2^x/λ_1^x .

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.

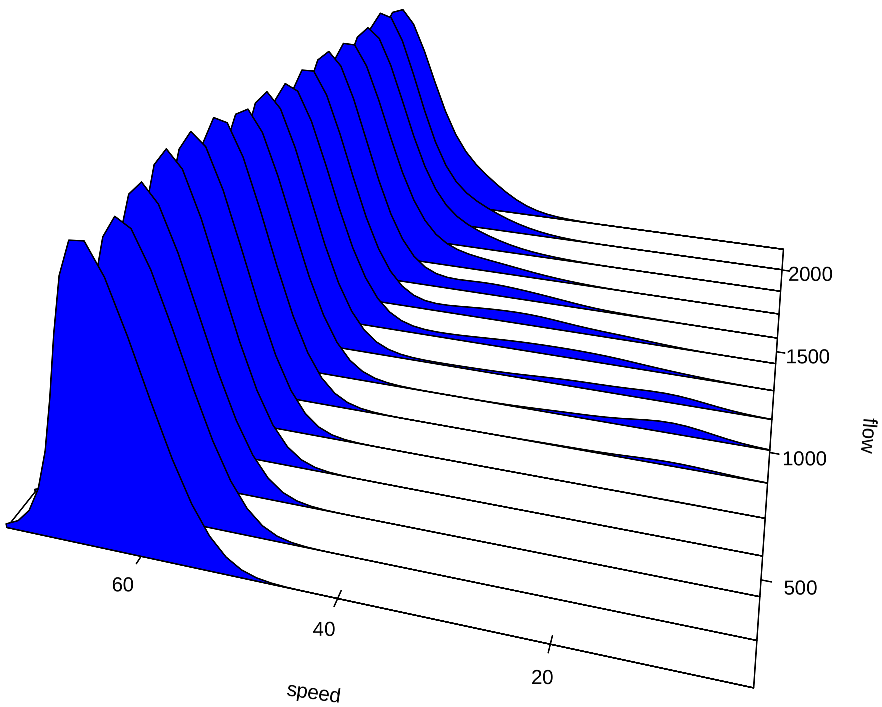
Summary

- LPCs work well in a variety of data situations, and are particularly for noisy complex structures more suitable than the die “global” algorithms developed by Hastie & Stuetzle (1989) and Kegl et al. (2000) .
- LPC’s can be seen as a simplified version of Delicado’s ’PCOPs’, but seem to work better than Delicado’s algorithm for complex or branched data.
- Bandwidth selection works by means of a coverage measure.
- Drawbacks of LPCs: No statistical model and hence no ’true’ principal curve; estimated principal curves depends strongly on selected starting point(s).
- General drawback of principal curves: Principal curves are not suitable for prediction of Y for given $X = x$.

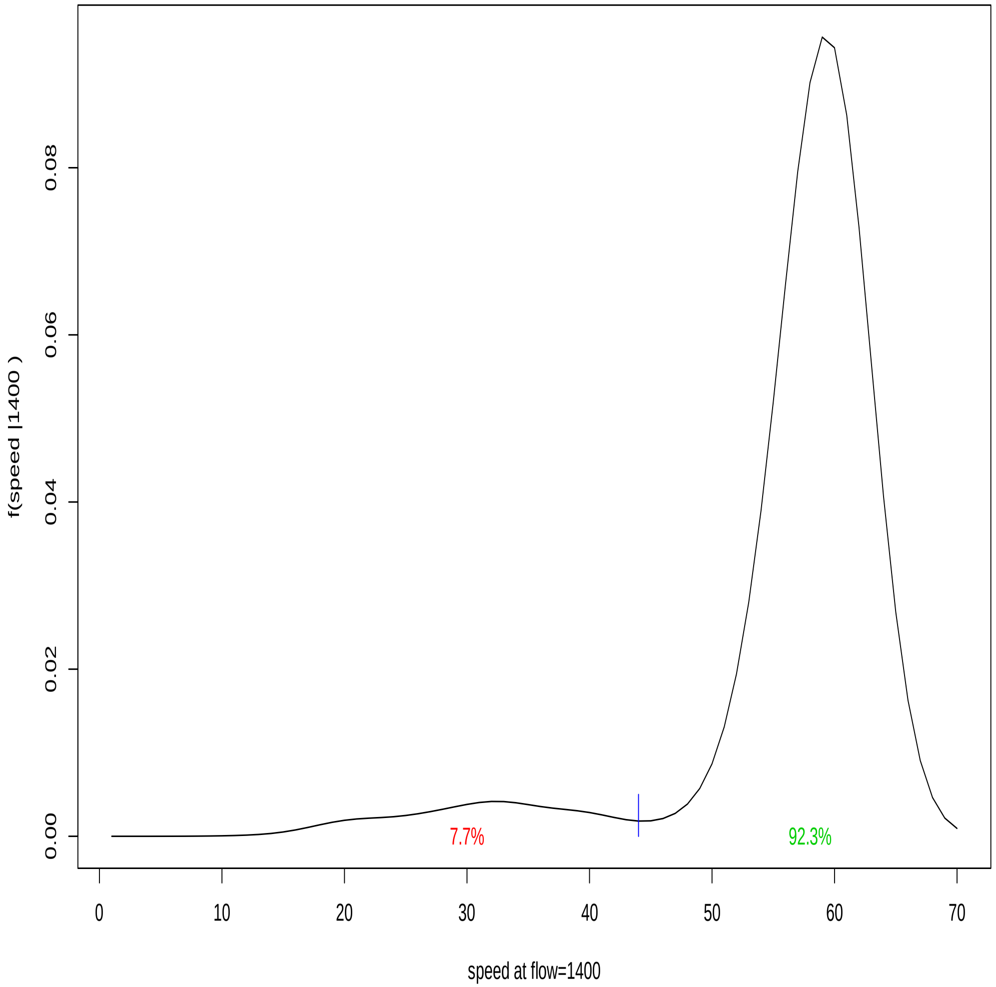
Outlook: Multi-valued regression

Goal: Estimate a multifunction $r : \mathbb{R} \longrightarrow \mathbb{R}$ rather than a regression function

Idea: Consider the conditional densities, e.g. for speed-flow data:

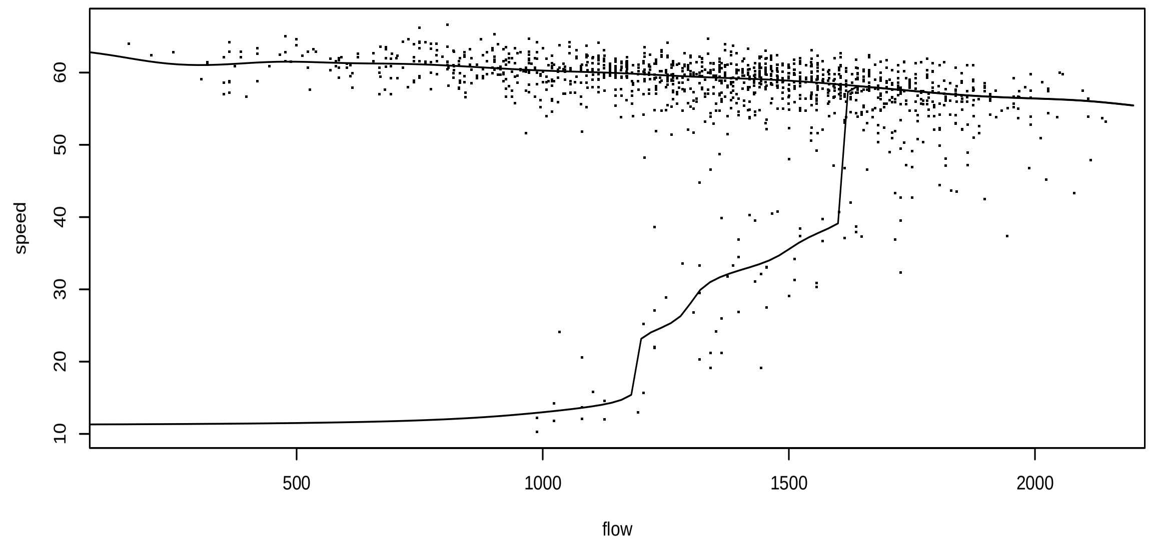


For instance, conditional density at a flow $= 1400$.

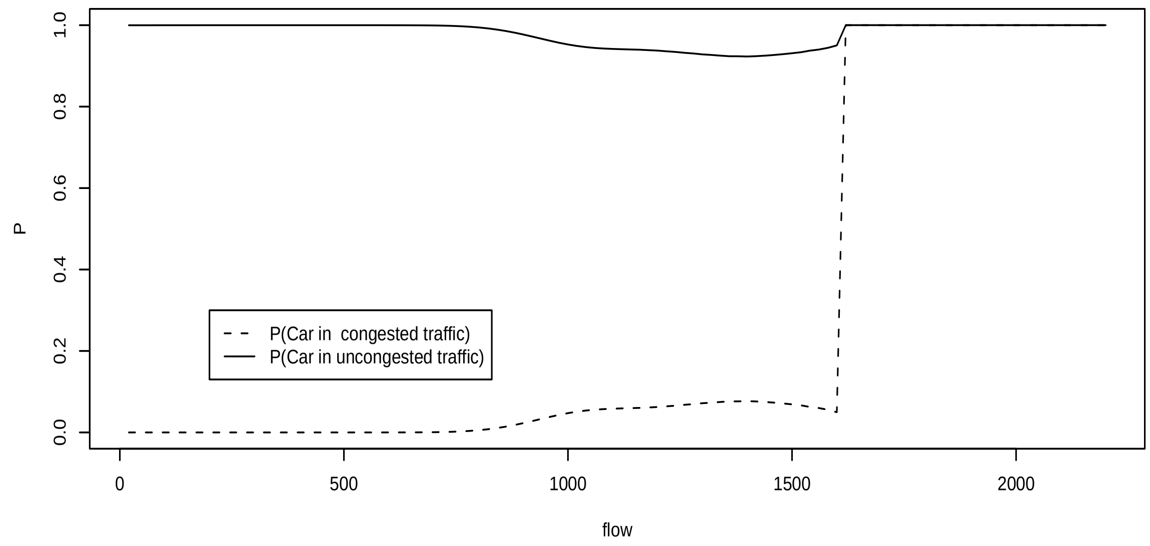


- For estimation of $r(x)$, compute the modes of the estimated conditional densities $\hat{f}(y|x)$.
- The area between a mode and the neighboring 'antimode' serves as estimated probability, that, given x, a value on the corresponding branch is attained.

Multi-valued regression curve



Relevance assessment



Estimation of conditional modes

We are interested in all local maxima of the estimated conditional densities

$$\hat{f}(y|x) = \frac{\hat{f}(x,y)}{\hat{f}(x)} = \frac{\sum_{i=1}^{n} K_1\left(\frac{x-X_i}{h_1}\right) K_2\left(\frac{y-Y_i}{h_2}\right)}{h_2 \sum_{i=1}^{n} K_1\left(\frac{x-X_i}{h_1}\right)}$$

with kernels K_1, K_2 and bandwidths h_1, h_2. We assume that a profile $k(\cdot)$ for kernel K_2 exists

such that $K_2(\cdot) = c_k k((\cdot)^2)$ holds. One calculates

$$\frac{\partial \hat{f}(y|x)}{\partial y} = \frac{2c_k}{h_2^3} \sum_{i=1}^{n} K_1\left(\frac{x-X_i}{h_1}\right) k'\left(\left(\frac{y-Y_i}{h_2}\right)^2\right)(y-Y_i) \stackrel{!}{=} 0$$

and obtains

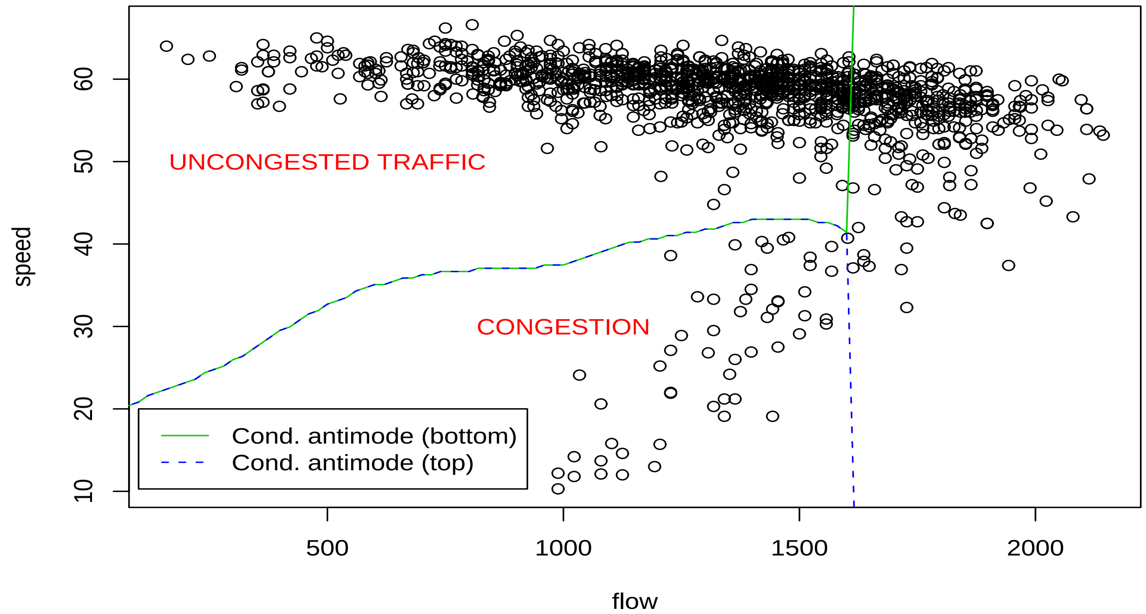
$$y = \frac{\sum\limits_{i=1}^{n} K_1\left(\frac{x-X_i}{h_1}\right) G\left(\frac{y-Y_i}{h_2}\right) Y_i}{\sum\limits_{i=1}^{n} K_1\left(\frac{x-X_i}{h_1}\right) G\left(\frac{y-Y_i}{h_2}\right)}. \tag{3}$$

with $G(\cdot) = -k'((\cdot)^2)$.

- Gives conditional mean shift procedure.
- The right side of (3) is just the “Sigma-Filter” used in digital image smoothing.

Antiregression and Classification

If one plots the antimodes, which are obtained as a by-product of the computation of the relevances, one obtains an antiprediction or antiregression curve.



This curve serves as a separator between the branches, and thus as a tool to classify observations to the uncongested or congested regime.

Literature on Principal Curves

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

Tibshirani, R. (1992): Principal Curves Revisited. *Statistics and Computing* **2**, 183–190.

Kégl, B., Krzyzak, A., Linder, T. & Zeger, K. (2000): Learning and Design of Principal Curves. *IEEE Transactions Patt. Anal. Mach. Intell.* **24**, 59–74.

Delicado, P. (2001): Another Look at Principal Curves and Surfaces, *Journal of Multivariate Analysis* **77**, 84–116.

⋮

Einbeck, J, Tutz, G. & Evers, L. (2005): Local Principal Curves. *Statistics and Computing* **15**, 301–313.

Einbeck, J, Tutz, G. & Evers, L. (2005): Exploring Multivariate Data Structures with Local Principal Curves. In: Weihs, C. and Gaul, W. (Eds.): *Classification - The Ubiquitous Challenge*. Springer, Heidelberg.

Literature on Multi-valued Regression

Einbeck, J, & Tutz, G.(2004): Modelling Beyond Regression Functions: An Application of Multimodal Regression to Speed-Flow Data. *SFB386 Discussion Paper No. 395*, LMU Munich.