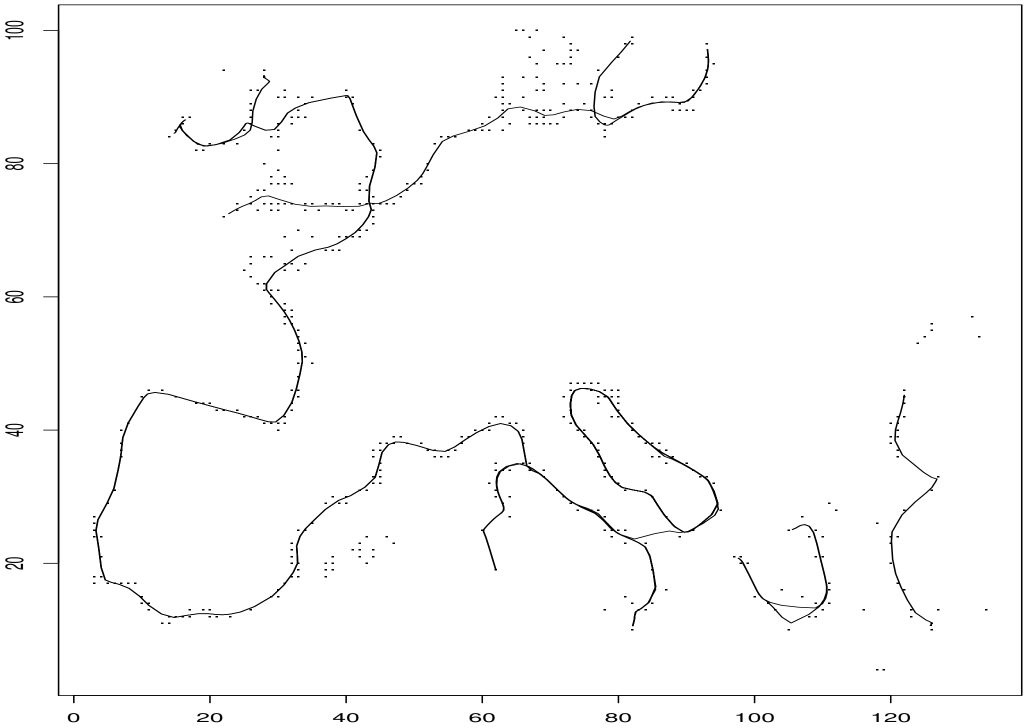
Exploring Multivariate Data Structures with Local Principal Curves

Jochen Einbeck
Gerhard Tutz
Ludger Evers

University of Munich,
Department of Statistics



Introduction: Principal Components

Given a random variable $\mathbf{X}$ in $\mathbb{R}^d$ with covariance matrix Σ, the principal components decomposition is given by

$$\Sigma = \Gamma \Lambda \Gamma^T, \tag{1}$$

where Λ is a diagonal matrix containing the ordered eigenvalues of Σ. The columns of

$$\Gamma = (\gamma_1, \ldots, \gamma_d)$$

are the eigenvectors of Σ. The j-th principal component lines are given by

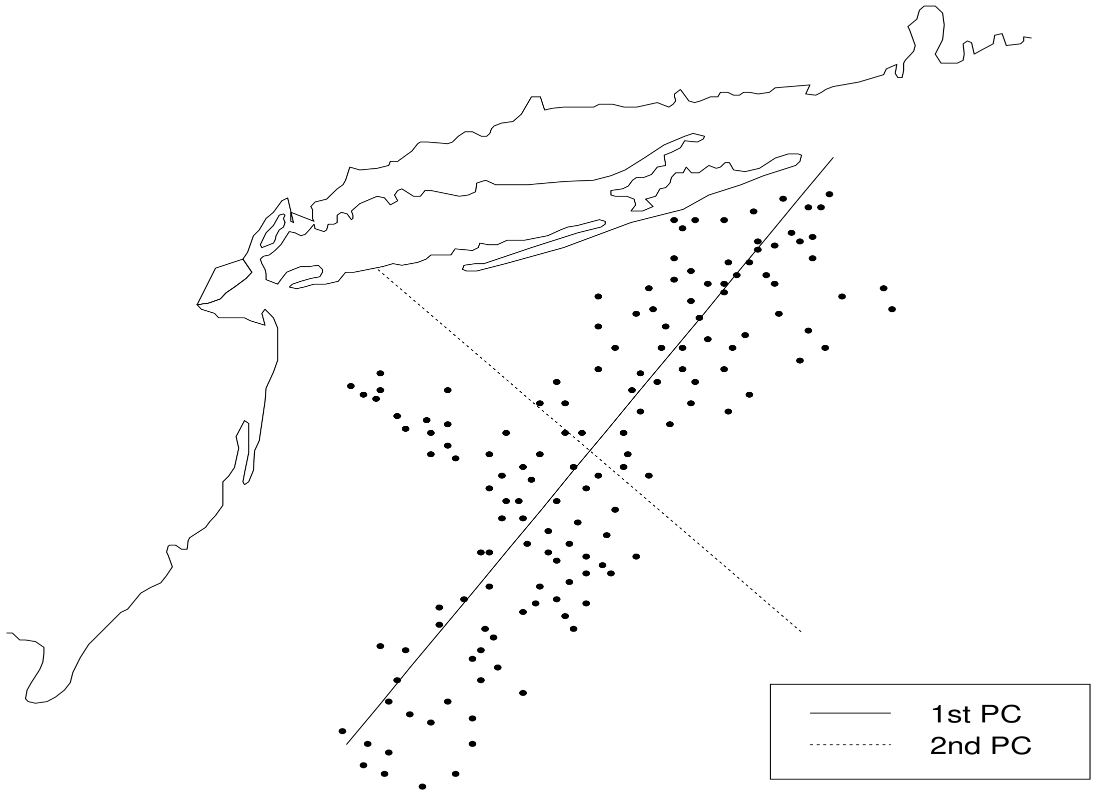
$$g_j(t) = \mu + t\gamma_j \qquad (j \leq d),$$

with $\mu = E(\mathbf{X})$. For real data

$$\mathbf{X} = (X_1, \ldots, X_n)^T$$

replace μ and Σ by consistent estimators $\hat{\mu}$ and $\hat{\Sigma}$.

Illustration: Scallops Data



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

Local Principal Curves

Assume a data cloud $\mathbf{X} = (X_1, \ldots, X_n)^T$, where $X_i = (X_{i1}, \ldots, X_{id})^T$. We propose the following

Algorithm (Local principal curves)

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*

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

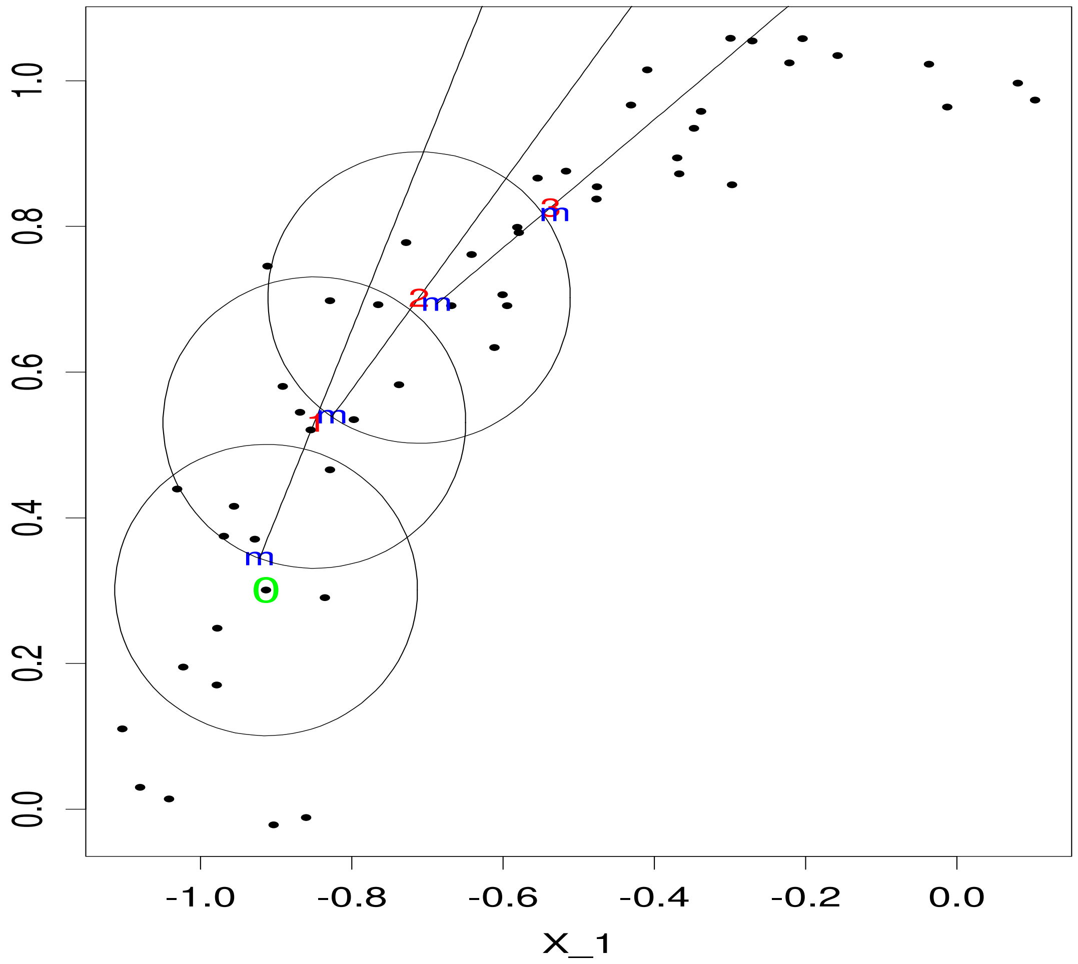
4. *Step from* μ^x *to*

$$x := \mu^x + t_0 \gamma_1^x.$$

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.*

Illustration

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



0	Starting point
m	Points (μ^x) of the LPC
$1, 2, 3$	Position of x after 1,2, resp. 3 steps.
Circles	symbolize the bandwidth matrix $H = h^2 \cdot I$.

Theoretical Background

Assume a data cloud $\mathbf{X}$ with density f. Let $\hat{f}_K$ a kernel density estimate of f. Comaniciu & Meer (2002) examine the *mean shift vector*

$$M_K(x) = \mu^x - x$$

and show:

(A) $M_K(x) \sim \hat{f}'_K(x)$

(B) The sequence

$$\begin{aligned} m_{(0)} &= x \\ m_{(k+1)} &= \mu^{m_{(k)}} \end{aligned}$$

converges to a neighboring maximum of $\hat{f}_K$.

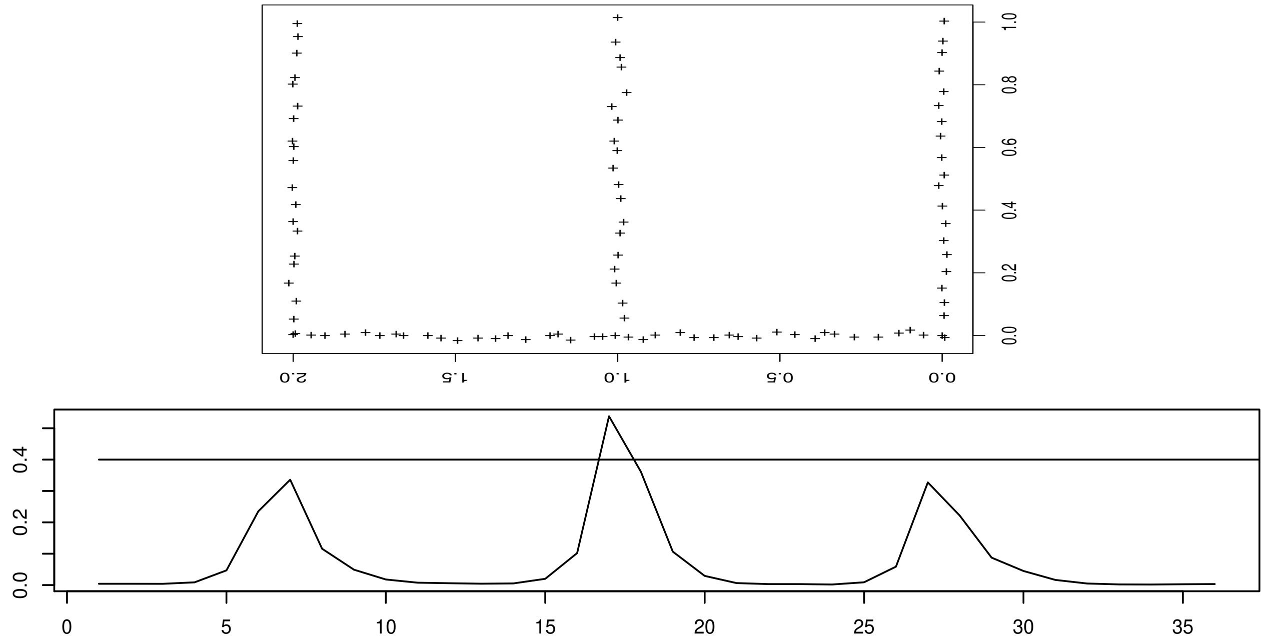
We conclude:

- Calculation of the local center of mass leads us to a point nearby a density mode.
- The LPC follows the crest of a kernel density estimate.

Higher-order-LPC's

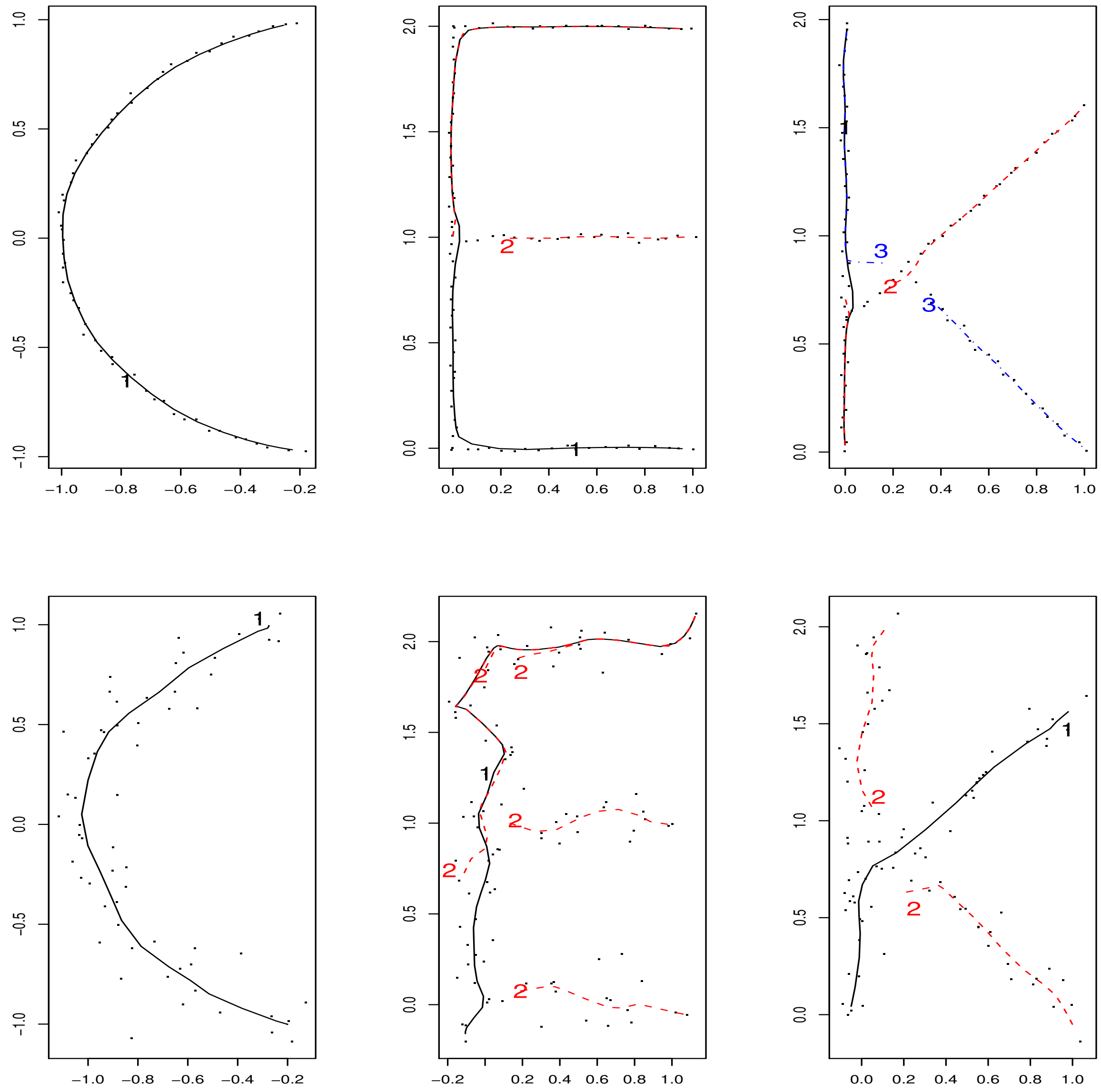
Consider the second local eigenvalue λ_2^x, i.e. the second largest eigenvalue of Σ^x: If this value is large at a certain point of the original LPC, a new LPC is launched in direction of the second local eigenvector γ_2^x. Every bifurcation raises the depth of the LPC tree.

Example



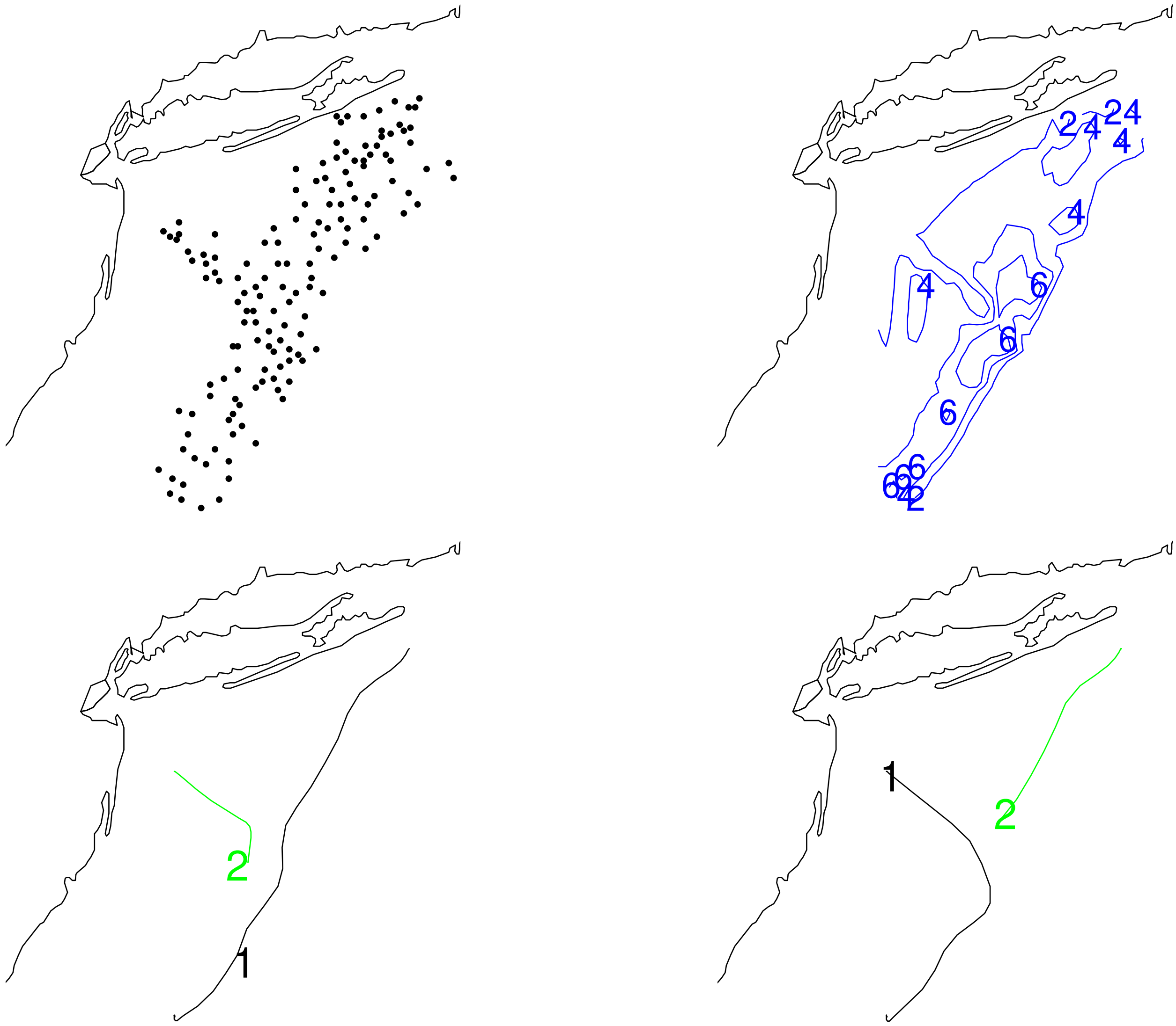
Simulated E and flow diagram of relation $\lambda_2^x / \lambda_1^x$.

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:	Contour plot of water depth
Bottom left, right:	Two results of the LPC algorithm
1, 2:	LPC's with depth 1 or 2, resp.

Parameter selection

- Step length t_0: Set $t_0 = h$.

- Bandwidth h:

Given a principal curve m consisting of a set P_m of points, the coverage is defined by

$$C_m(\tau) = \frac{\#\{x \in \mathbf{X} | \exists p \in P_{m(\tau)} \text{ mit } ||x - p|| \leq \tau\}}{n}.$$

This a monotonely increasing function of τ. For a principal curve $m(\tau)$, calculated with a bandwidth τ, we define further the self-coverage as

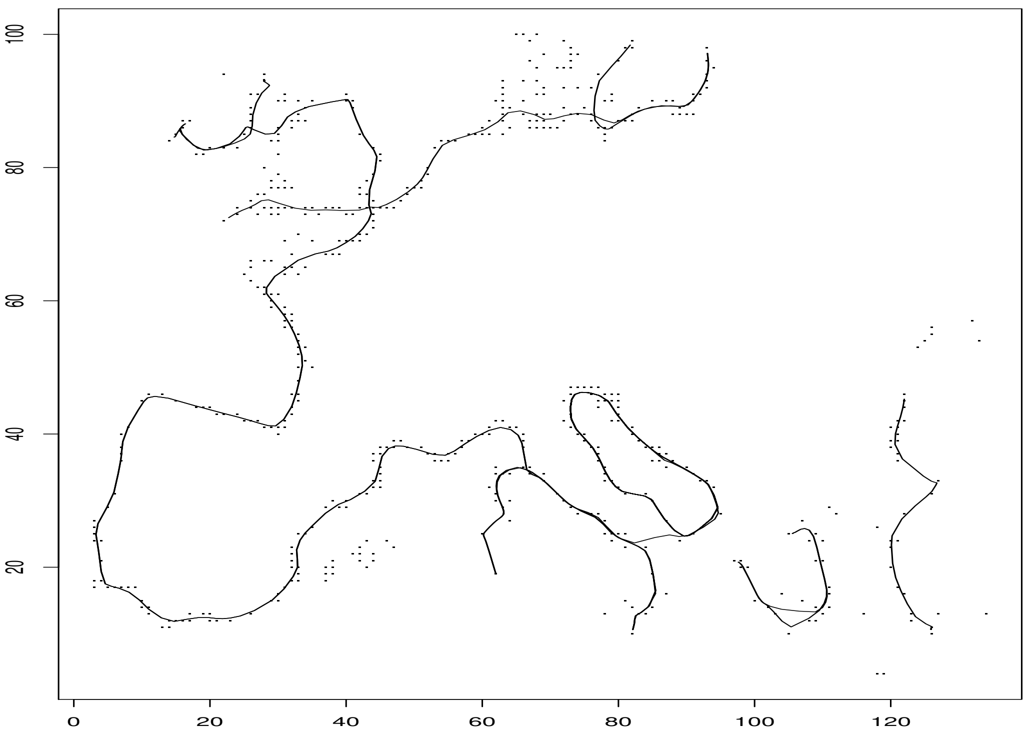
$$C_{m(\tau)}(\tau) = \frac{\#\{x \in \mathbf{X} | \exists p \in P_{m(\tau)} \text{ mit } ||x - p|| \leq \tau\}}{n}$$

and select

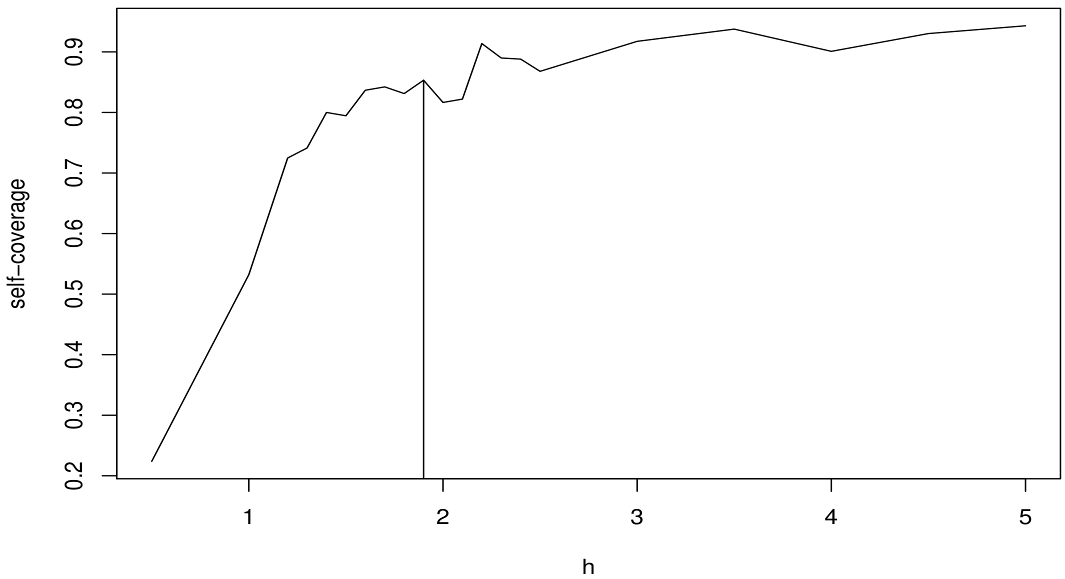
$$h = \text{first (distinct) local maximum of} \quad C_{m(\tau)}(\tau).$$

Example: European Coastal Resorts

LPC (solid line) through european coastal resorts (·):



Bandwidth selection via self-coverage:



Alternative Approaches

Local ("bottom-up"):

Delicado (2001) works with *principal directions*, i.e. vectors orthogonal to the hyperplane that locally minimize the variance of the data points projected on it.

Disadvantage: Cannot be solved by PCA.

Global ("top-down"):

Hastie & Stützle (1989) via self-consistency,

Tibshirani (1992) via a probabilistic approach,

Kégl, Krzyzak, Linder & Zeger (2000)
Polygonal line algorithm.
Application on recognition of hand-written letters in Kégl, B., and Krzyzak, A., (2002)

References

COMANICIU & MEER (2002): Mean Shift: A robust appproach towards feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell. 24, 603-619*.

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

HASTIE & STUETZLE, L. (1989): Principal Curves. *JASA, 84, 502-516*.

KÉGL, B., and KRZYZAK, A., (2002): Piecewise linear skeletonization using principal curves. *IEEE Transactions Patt. Anal. Mach. Intell., 24, 59-74*.

KÉGL, B., KRZYZAK, A., LINDER, T. and ZEGER, K. (2000): Learning and design of principal curves, *IEEE Transactions Patt. Anal. Mach. Intell., 22, 281-297*.

TIBSHIRANI, R. (1992): Principal Curves Revsited. *Statistics and Computing, 2 , 183-190*.