Approaches to Function-free Local Smoothing

with Application to Speed-flow Diagrams

Jochen Einbeck

joint work with Gerhard Tutz, University of Munich

10th of February 2005

- Local principal curves
- Multi-valued nonparametric regression

What is "Smoothing"?

Typical definition:

"Given observations from an explanatory variable X and a response variable Y, construct a function, a "smoother", which at point x estimates the average value of Y given that $X = x$.

Also called nonparametric regression."

(Holmström et.al, 2003)

Nonparametric regression

Given n data points $(X_i, Y_i), i = 1, \ldots, n$, one seeks a smooth function $m : \mathbb{R} \longrightarrow \mathbb{R}$ relating X and Y in a proper way, which can be generally expressed in the form

$$m(x) = \Phi(Y|X = x).$$

Possible choices of the operator $\Phi(\cdot)$ are obtained as solutions of the minimization problem

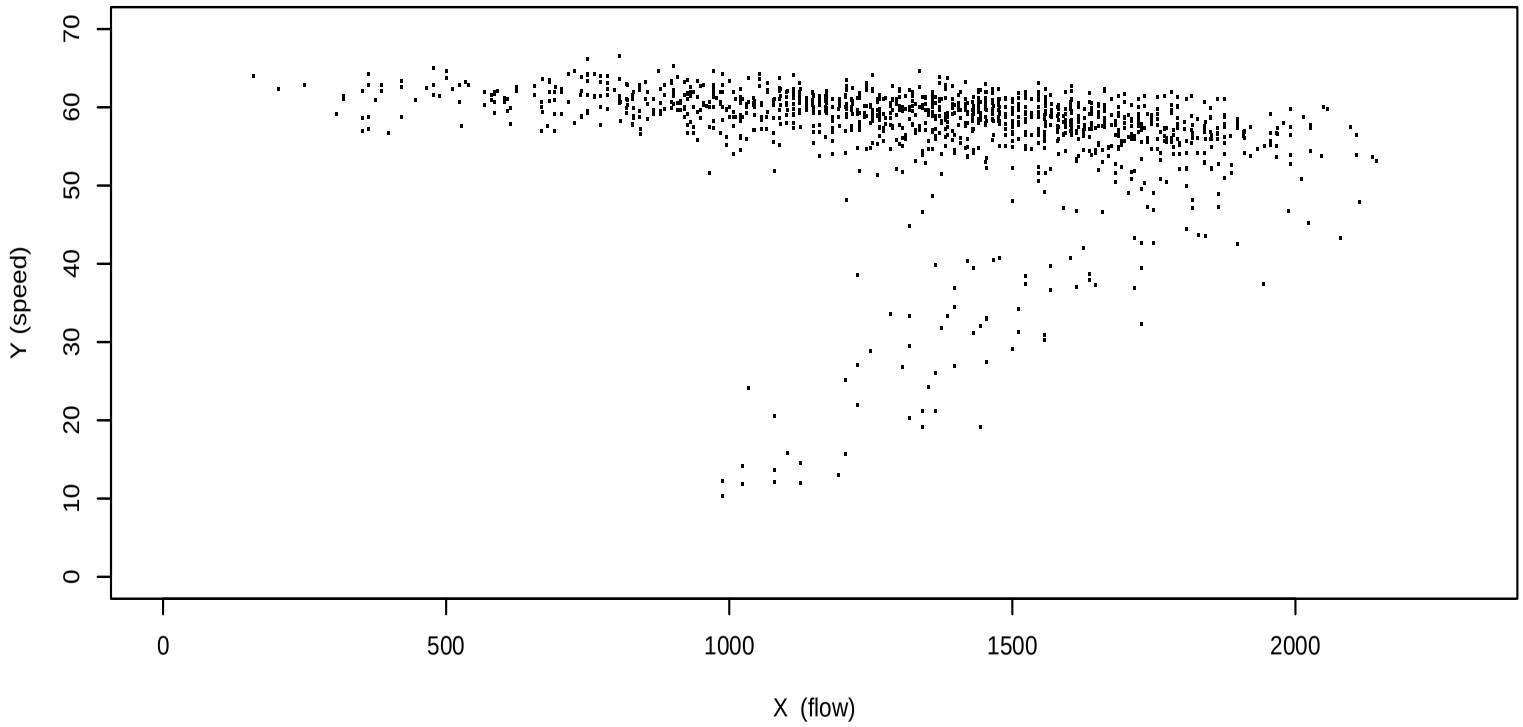
$$m_l(x) = \arg\min_a E(l(Y - a)|X = x),$$

yielding

$l(z)$	z^2	$\|z\|$	$-\delta(z)$
$\Phi(\cdot)$	$E(\cdot)$	$Median(\cdot)$	$Mode(\cdot)$

Limits of ordinary nonparametric regression

Speed-Flow diagram:

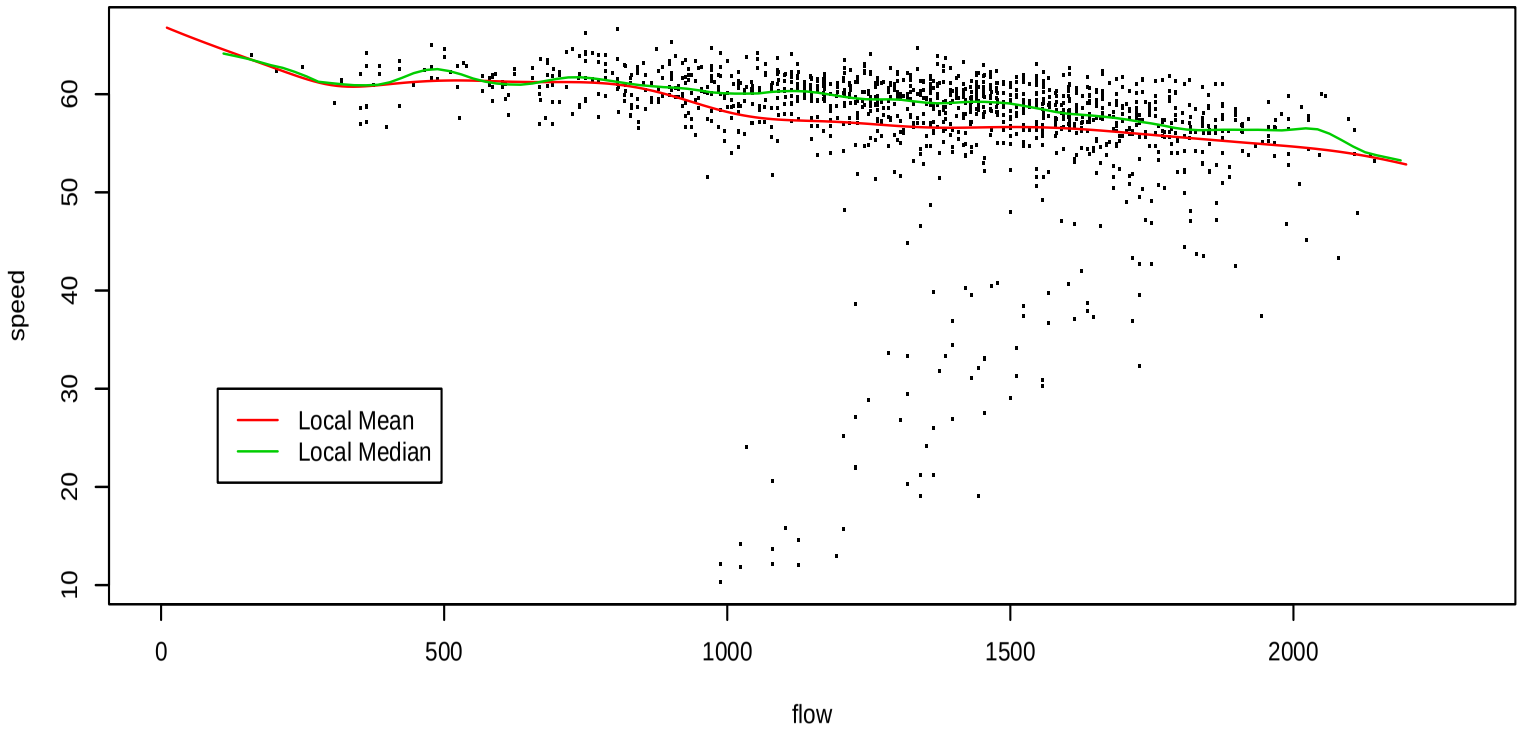


X: traffic flow in cars/hour, Y: speed in Miles/hour

recorded on a californian "freeway".

Attempt: Ordinary nonparametric regression

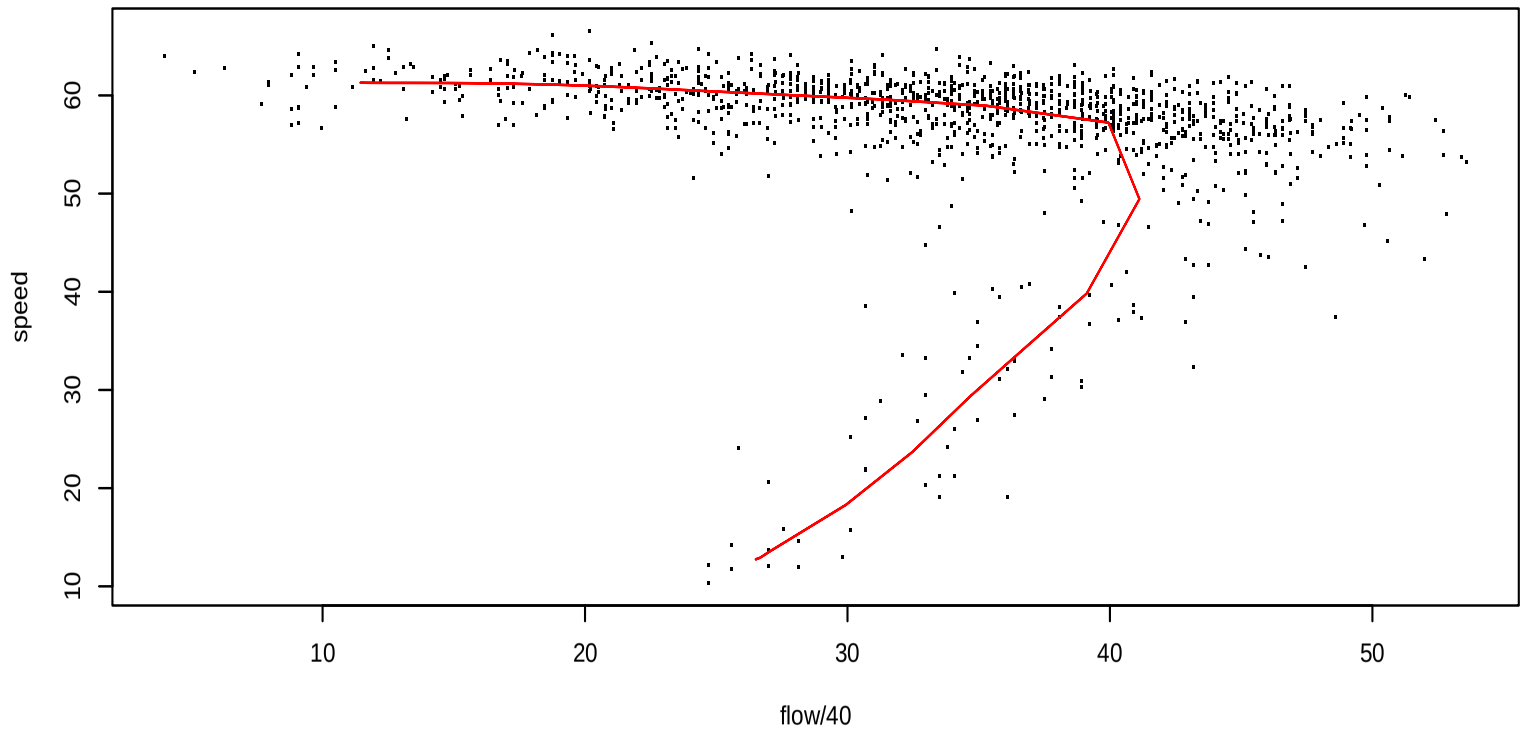
$$Y = m(X) + \epsilon\,;\ \text{with a function}\ m : \mathbb{R} \longrightarrow \mathbb{R}$$



$\longrightarrow$ Obviously some information is discarded!

1st alternative Approach: Principal curves

$$\begin{pmatrix} X \\ Y \end{pmatrix} = f(t) + \epsilon\,;\; f : \mathbb{R} \longrightarrow \mathbb{R}^2$$

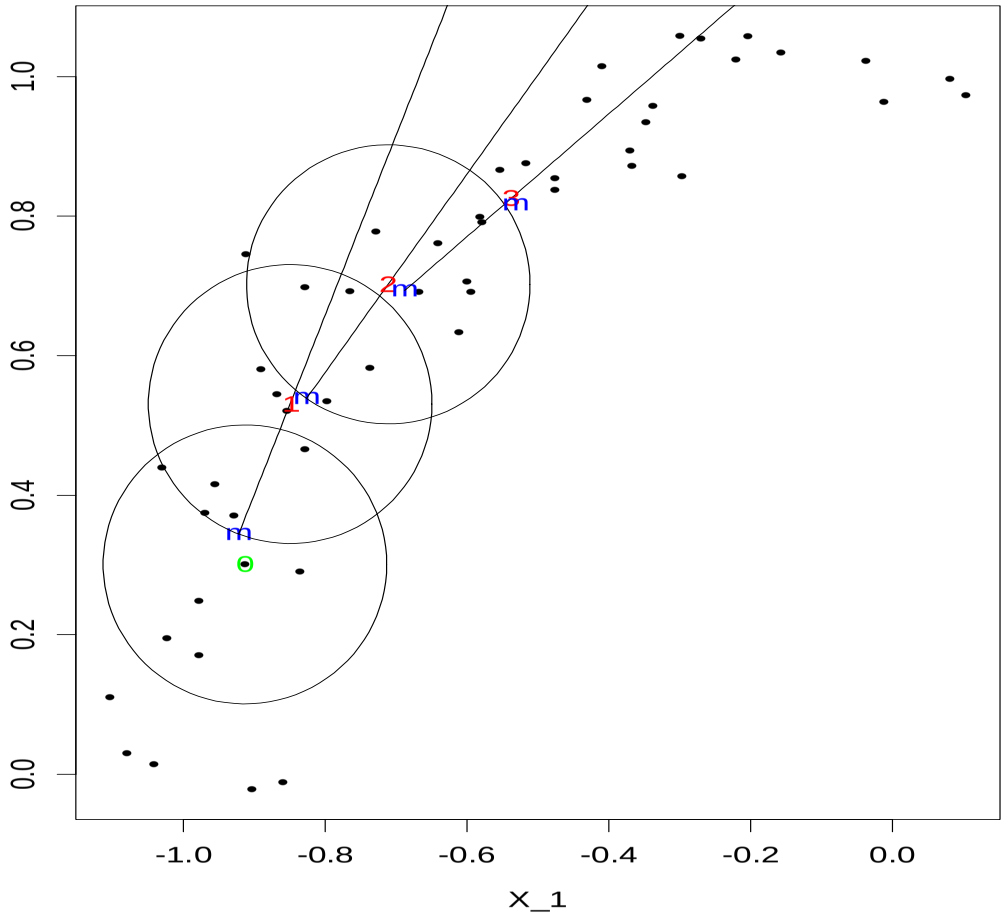


Intuitive Definition:

Principal curves are smooth functions passing through the middle of the data cloud.

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: Local principal curves

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

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. Perform a PCA locally at x, i.e. compute the 1^{st} eigenvector γ^x

   of the local covariance matrix $\Sigma^x = (\sigma^x_{jk})_{(1 \le j,k \le d)}$, 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 in direction of the first principal component to $x := \mu^x + t_0 \gamma^x_1$.

5. Repeat steps 2. to 4. until the μ^x remain constant.

A central tool of the LPC algorithm is the Mean Shift,

$$M(x) = \mu(x) - x. \tag{1}$$

i.e. the movement from the current point to the local center of mass. For a kernel density estimate $\hat{f}_K(x)$, Comaniciu & Meer (2002) show that

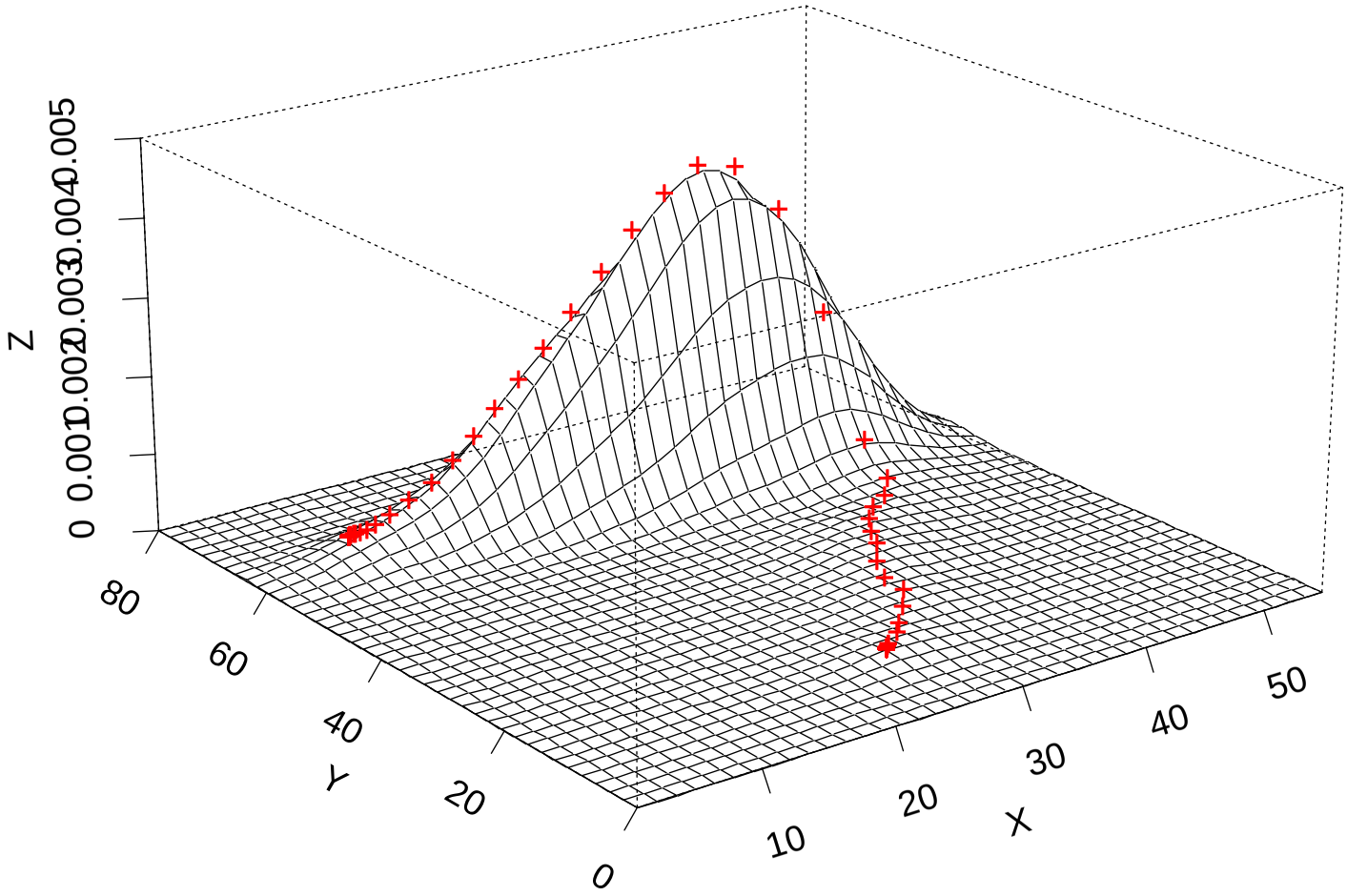
(A) $M(x) \sim \nabla \hat{f}_K(x)$

(B) and the sequence

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

converges to a neighboring critical point of $\hat{f}_K(\cdot)$.

Conclusion: The LPC algorithm approximates the crest line of a kernel density estimate.

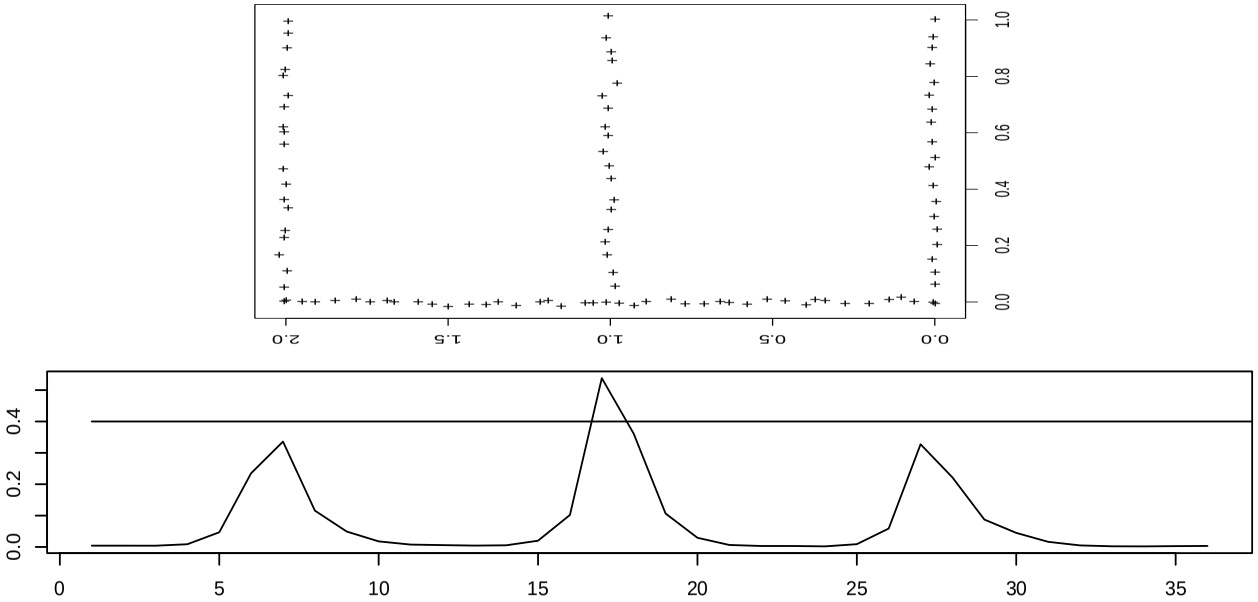


Kernel density estimate of speed-flow data and LPC (+).

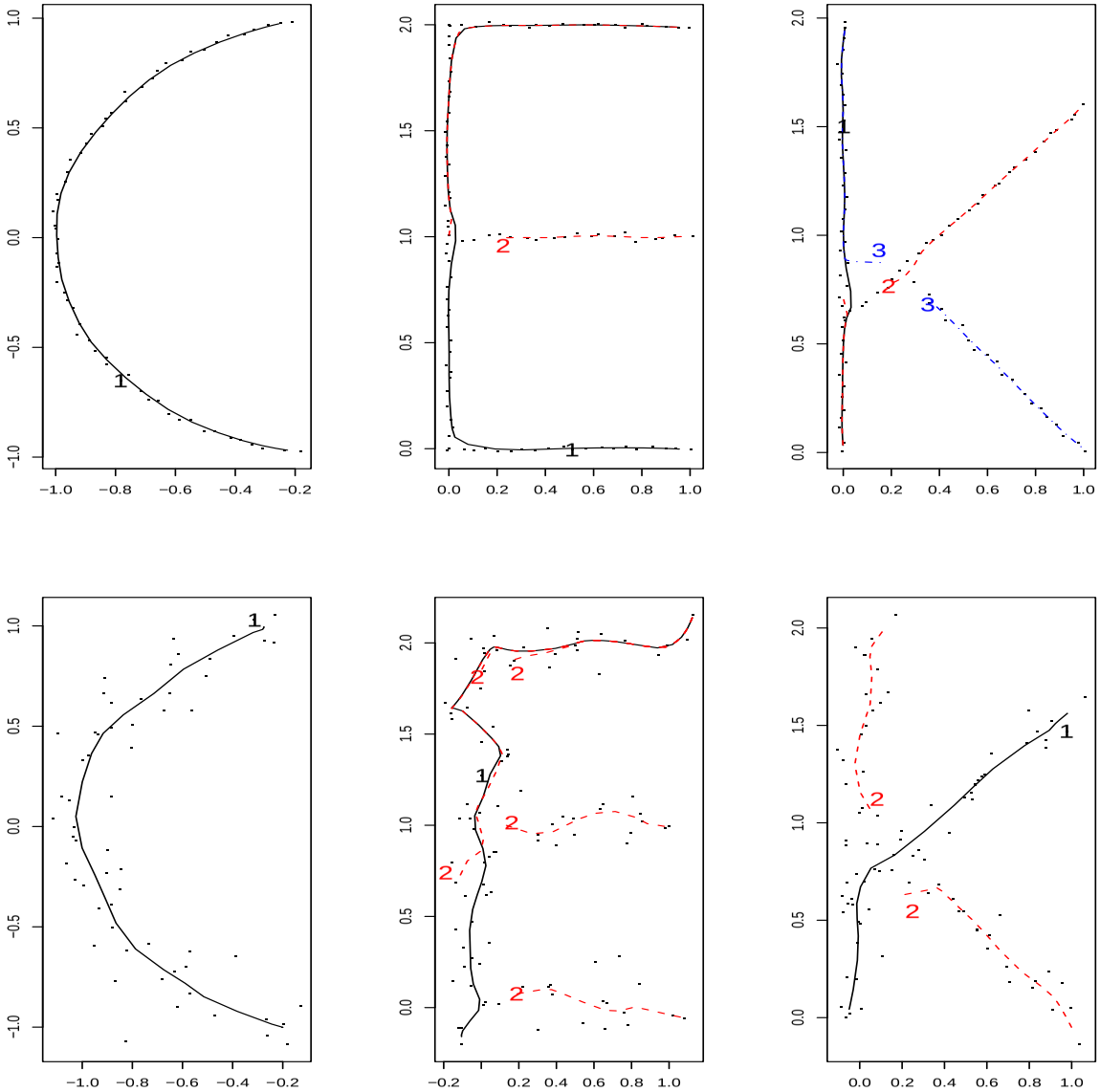
LPC's of higher order

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.

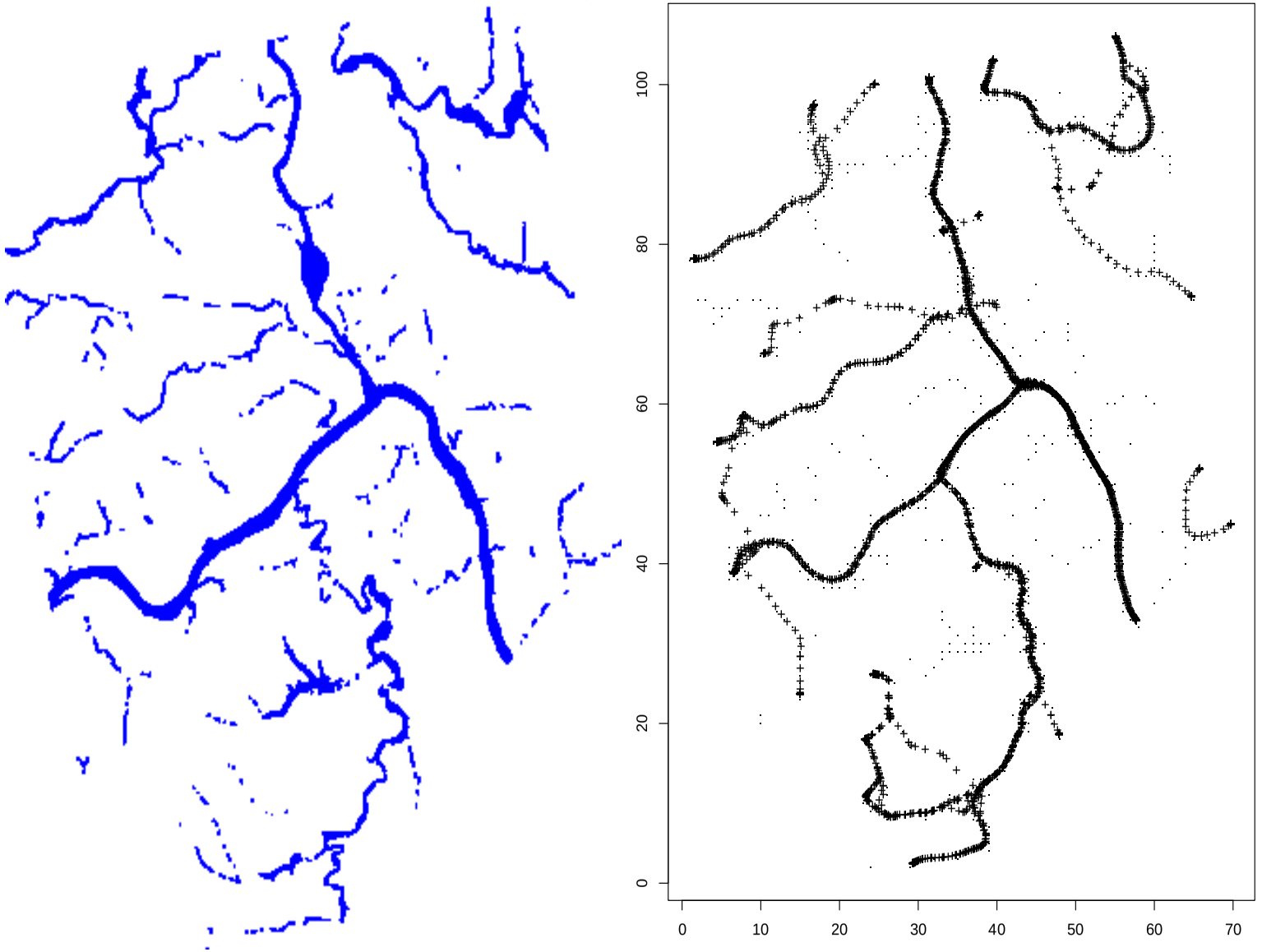
Example: Simulated "E" and flow diagram of the quantity λ_2^x/λ_1^x .



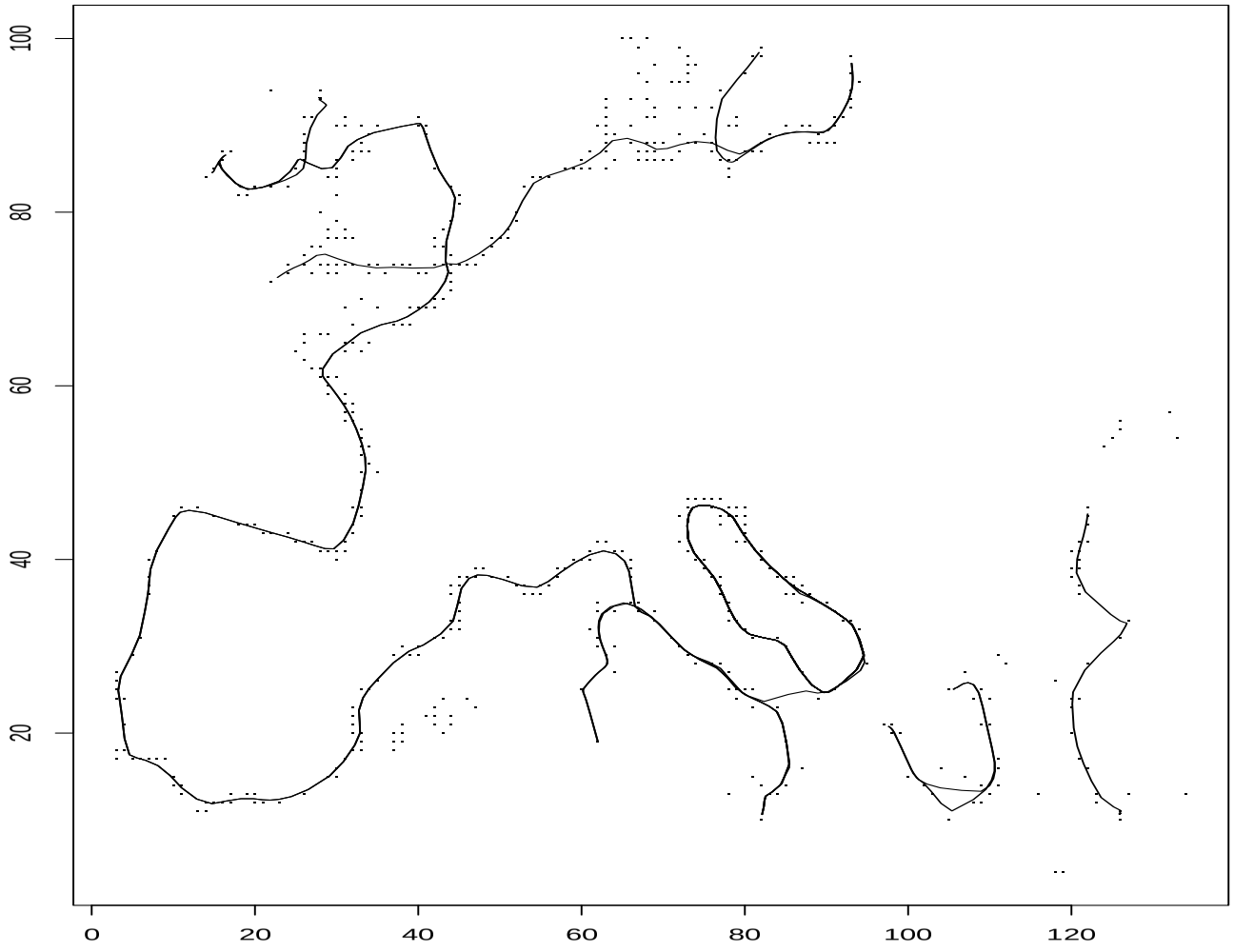
LPC's through simulated letters (C,E,K)



A further example: Floodplains in Pennsylvania



A further example: Coastal resorts in Europe

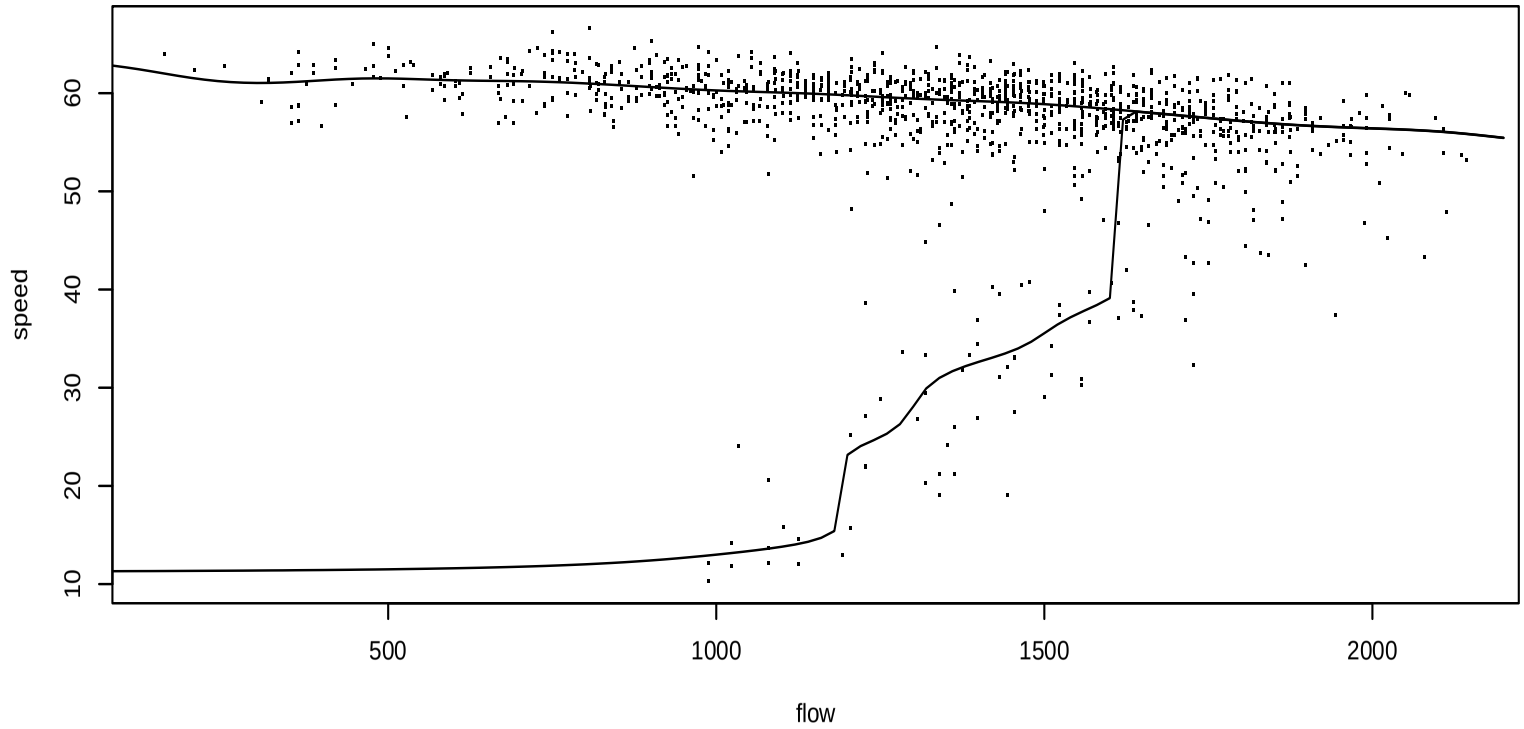


Concluding remarks

- LPC's work well in a variety of data situations, and show (in particular for complicated data structures) a better performance than the "global" algorithms from Hastie & Stuetzle (1989) und Kegl et al. (2000) .
- However, the theoretical grounding is somewhat waggly.
- Bandwidth selection works by means of a Coverage measure.
- (Local) Principal Curves are not suitable for prediction of Y from a given $X = x$.

2nd alternative approach: Multi-valued nonparametric regression

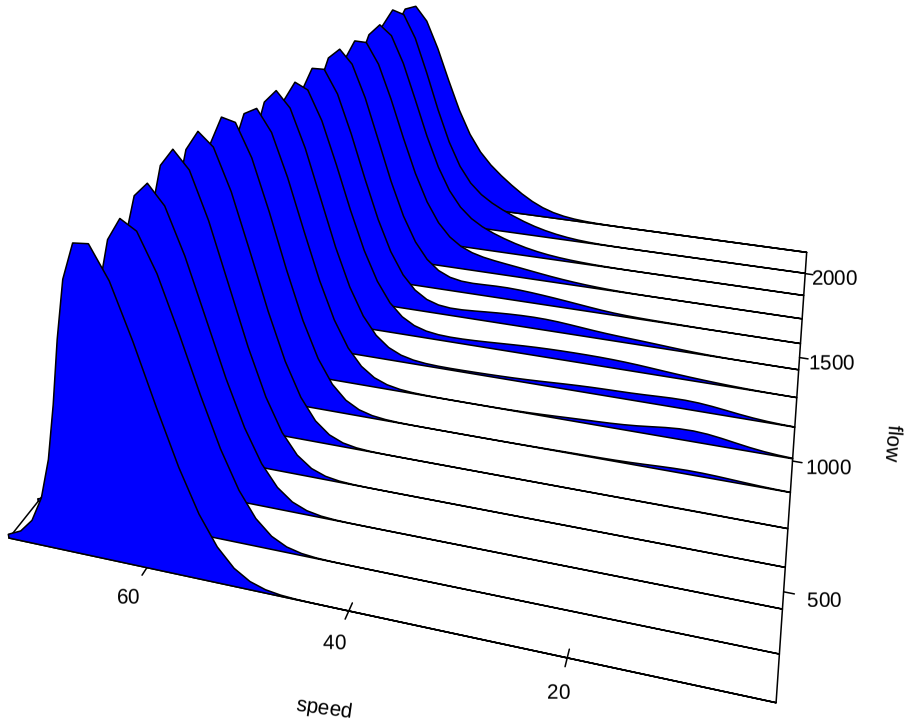
$$Y = r(X) + \epsilon, \text{ with a multifunction } r : \mathbb{R} \longrightarrow \mathbb{R}$$



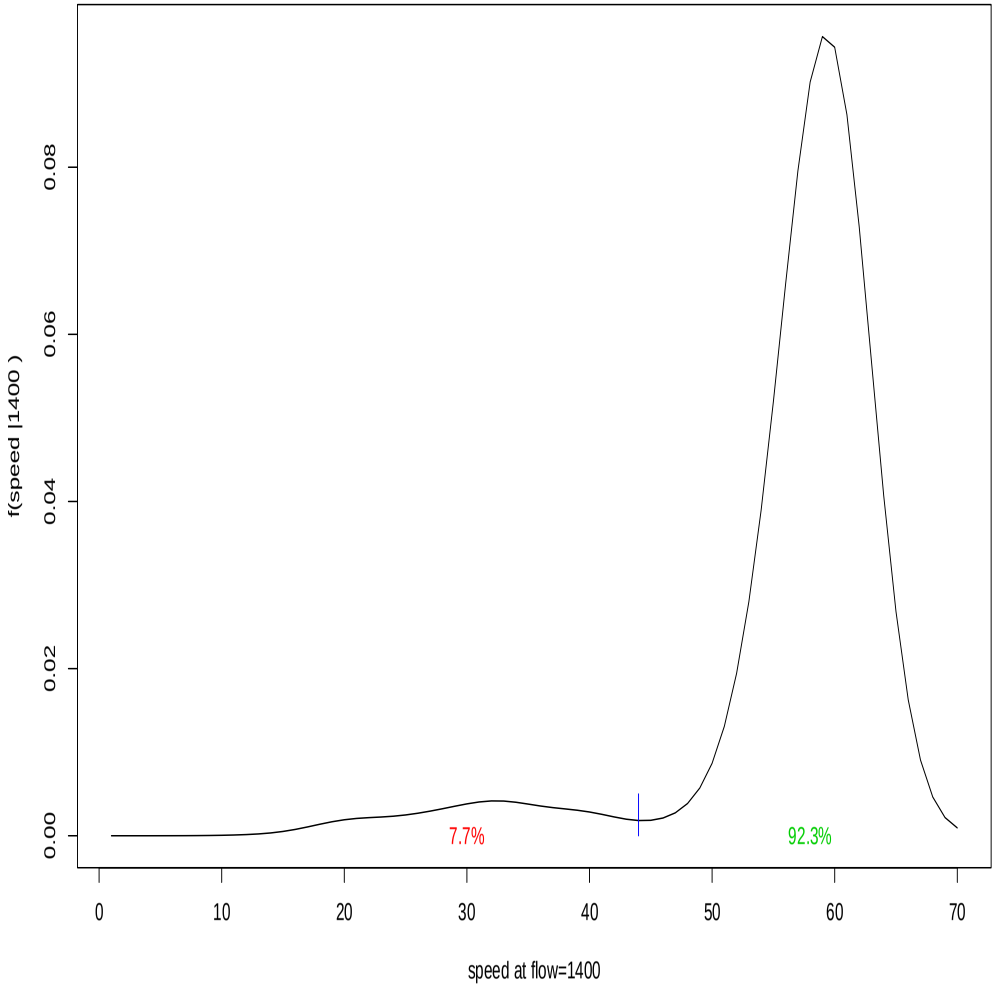
The data cloud is assumed to consist of several (smooth) branches.

For every $X = x$, more than one predicted value is possible.

Basic idea: Consider the conditional densities.

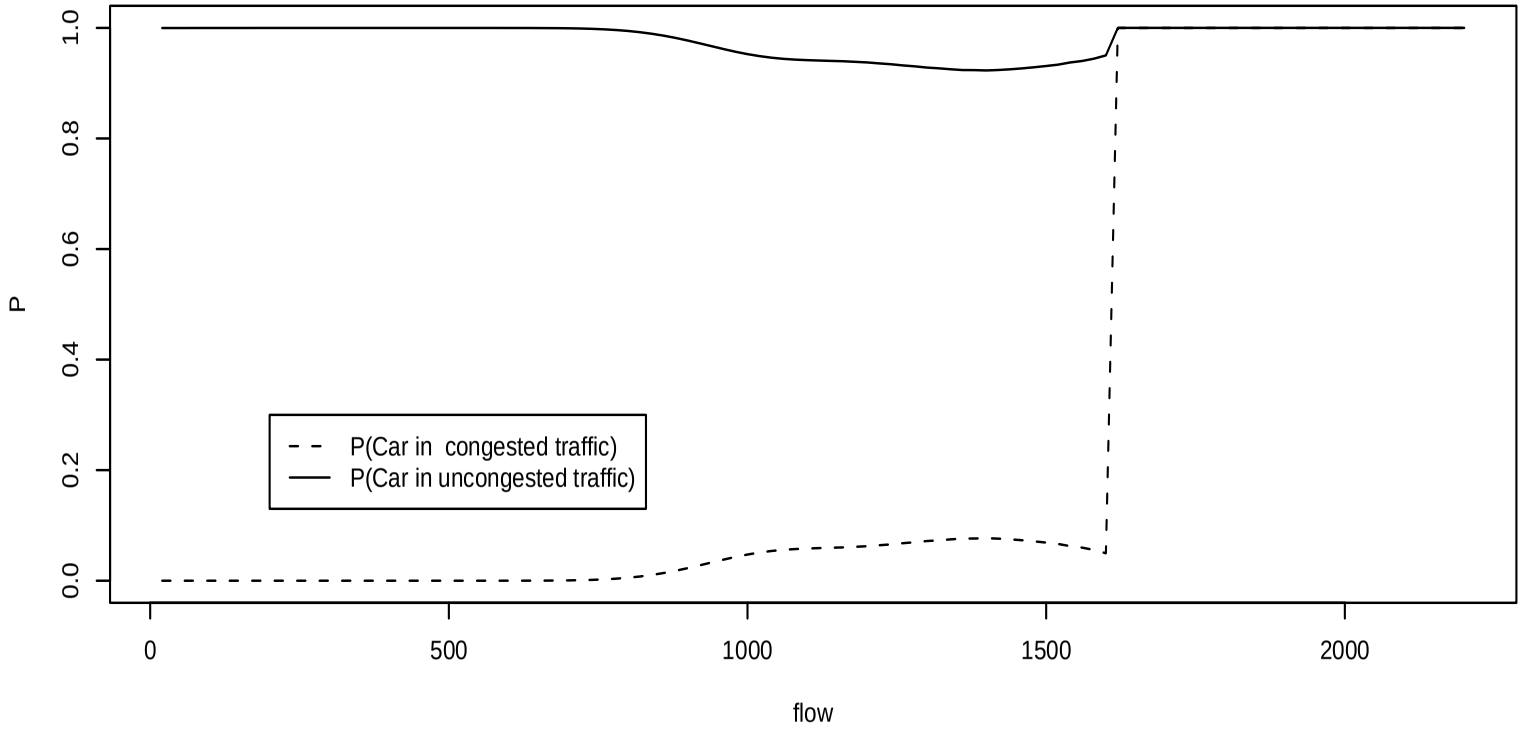


For instance, conditional density at a flow $= 1400$.



- 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.

Relevance assessment



The relevance of the branches varies smoothly for varying x, as long as as the branches can be separated.

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

$$y = \frac{\sum_{i=1}^{n} K_1\left(\frac{x-X_i}{h_1}\right) G\left(\frac{y-Y_i}{h_2}\right) Y_i}{\sum_{i=1}^{n} K_1\left(\frac{x-X_i}{h_1}\right) G\left(\frac{y-Y_i}{h_2}\right)}. \tag{2}$$

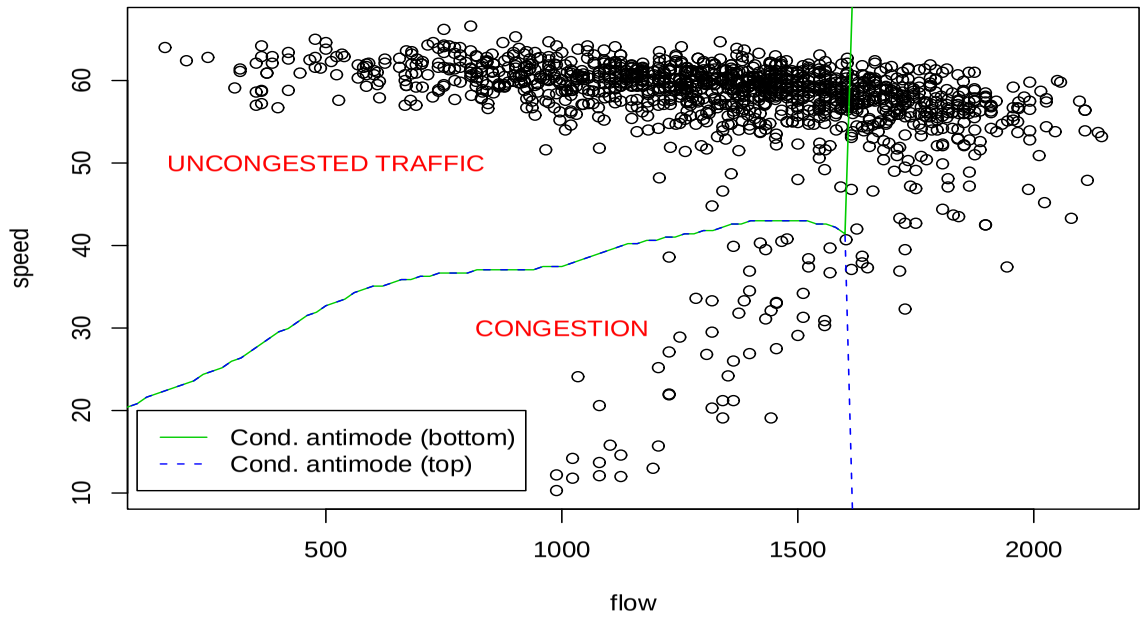
with $G(\cdot) = -k'((\cdot)^2)$.

Remarks:

- Applying (2) recursively, this can be seen as a conditional mean shift! Unlike for LPC's, this iteration has to be run until convergence.
- The right side of (2) corresponds to the "Sigma-Filter" used in digital image smoothing. Thus, the sigma filter is a one-step approximation to the conditional mode.

Outlook: 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.

Summary:

- The conditional mode is more useful for the analysis of multimodal data as the conditional mean or median.
- The mode is robust to outliers and is edge-preserving.
- Maxima of the conditional density can be calculated fast and easily via a conditional mean shift procedure.

To do:

- Bandwidth selection (for conditional densities already available: Fan et al., 1996)
- Bias, Variance? Asymptotics? (for cond. modes: Berlinet et al., 1998)
- Relation to mixture models?

Literature

- Principal Curves

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

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

⋮

Einbeck, J, Tutz, G. & Evers, L. (2003): Local Principal Curves. *SFB 386 Discussion Paper No. 320, accepted in Statistics and Computing*

Einbeck, J, Tutz, G. & Evers, L. (2005): Exploring Multivariate Data Structures with Local Principal Curves. *to appear in Proceedings GFKL 2004*

- Multi-valued nonparametric regression

  Einbeck, J, & Tutz, G.(2004): Modelling beyond regression functions: An application of multimodal regression to speed-flow data , *SFB 386 Discussion Paper No. 395.*

  ⋮ ?

  "Related Literature":

  Berlinet, Gannoun & Matzner-Løber (1998): Normalité aymptotique d'estimateurs convergents du mode conditionnel. *Canadian Journal of Statistics, 26, 365–380*.

  Fan, Yao & Tong (1996): Estimation of conditional densities and sensivity measures in nonlinear dynamical systems. *Biometrika, 83, 189–206*