Measuring goodness-of-fit in nonparametric unsupervised learning problems

Jochen Einbeck

Department of Mathematical Sciences, Durham University

jochen.einbeck@durham.ac.uk

Piraeus, 18th of August 2010

Outline

- Supervised and Unsupervised Learning
- Principal curves
- Measuring goodness-of-fit via *Coverage*
- Bandwidth selection via *Self-coverage*
- Mode detection and Clustering
- Discussion

Statistical Learning

- Supervised Learning
  - Data $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{p+1}$, i=1, ..., n.
  - Aim: Recover a continuous or discrete mapping $\boldsymbol{x}_i \mapsto m(\boldsymbol{x}_i)$, yielding fitted values $\hat{y}_i = \hat{m}(\boldsymbol{x}_i)$ ("Regression" or "Classification", respectively).
  - Estimation: Make y_i and $\hat{m}(\boldsymbol{x}_i)$ "as close as possible" (For instance, least squares $\sum_{i=1}^{n}[y_i - \hat{m}(\boldsymbol{x}_i)]^2$).
  - The y_i play the role of a "teacher" $\Longrightarrow$ Supervised Learning.

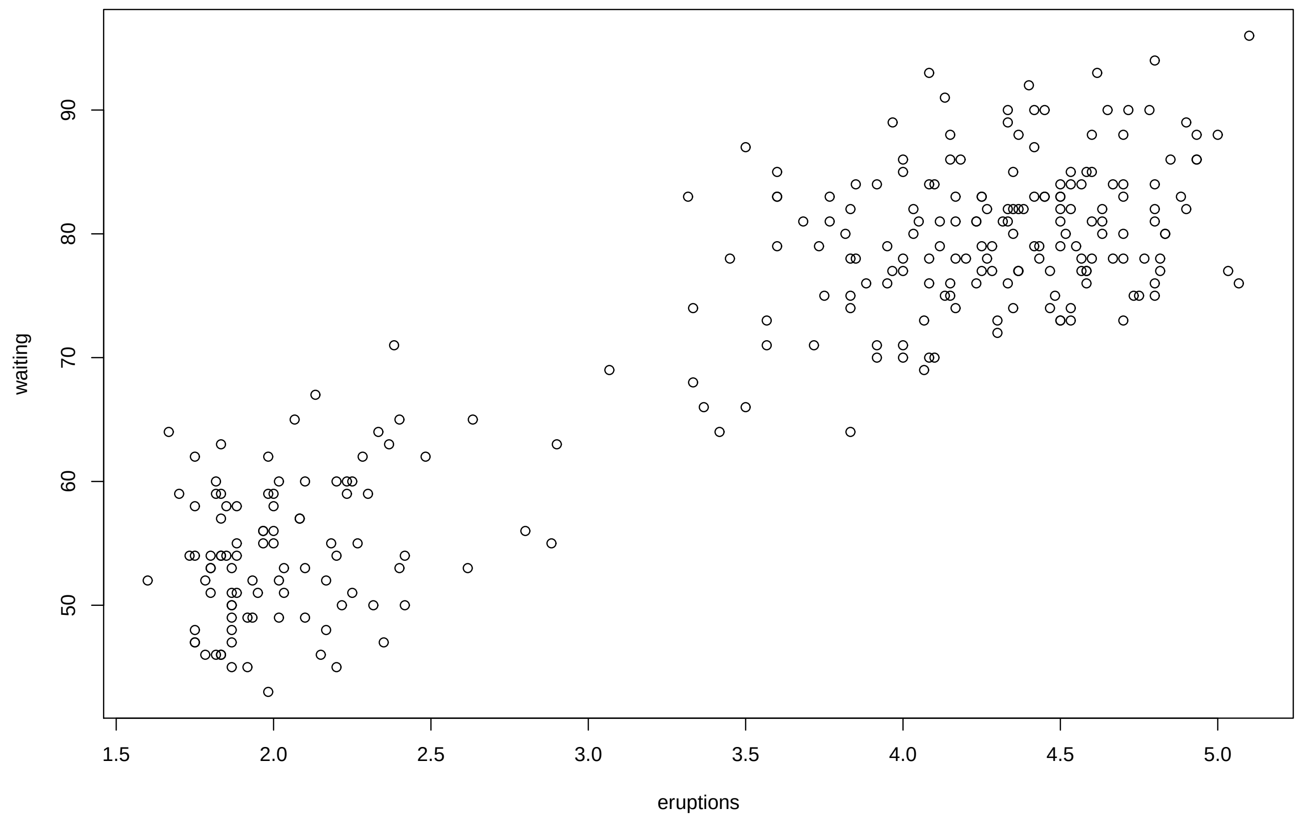
Statistical Learning

- Supervised Learning
  - Data $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{p+1}$, i=1, ..., n.
  - Aim: Recover a continuous or discrete mapping $\boldsymbol{x}_i \mapsto m(\boldsymbol{x}_i)$, yielding fitted values $\hat{y}_i = \hat{m}(\boldsymbol{x}_i)$ ("Regression" or "Classification", respectively).
  - Estimation: Make y_i and $\hat{m}(\boldsymbol{x}_i)$ "as close as possible" (For instance, least squares $\sum_{i=1}^{n}[y_i - \hat{m}(\boldsymbol{x}_i)]^2$).
  - The y_i play the role of a "teacher" $\Longrightarrow$ Supervised Learning.
- Unsupervised Learning
  - Data $(\boldsymbol{x}_i) \in \mathbb{R}^p$, i=1, ..., n. No response!
  - Aim: Learn "something" about the inner structure of the data cloud (density, linear summary, clusters, best fitting manifold).
  - No "teacher" available $\Longrightarrow$ Unsupervised Learning.

Example: Old Faithful geyser data

$n = 272$ measurements from the Old Faithful geyser in Yellowstone National Park, Wyoming, USA:

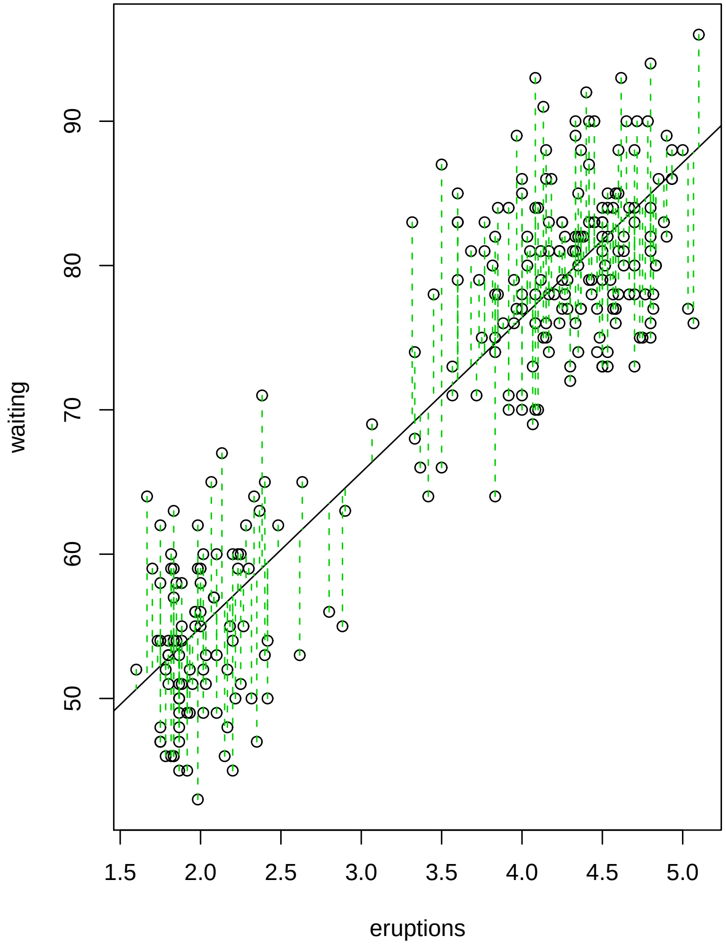
- the `waiting` time between eruptions;
- the duration of the `eruptions`.

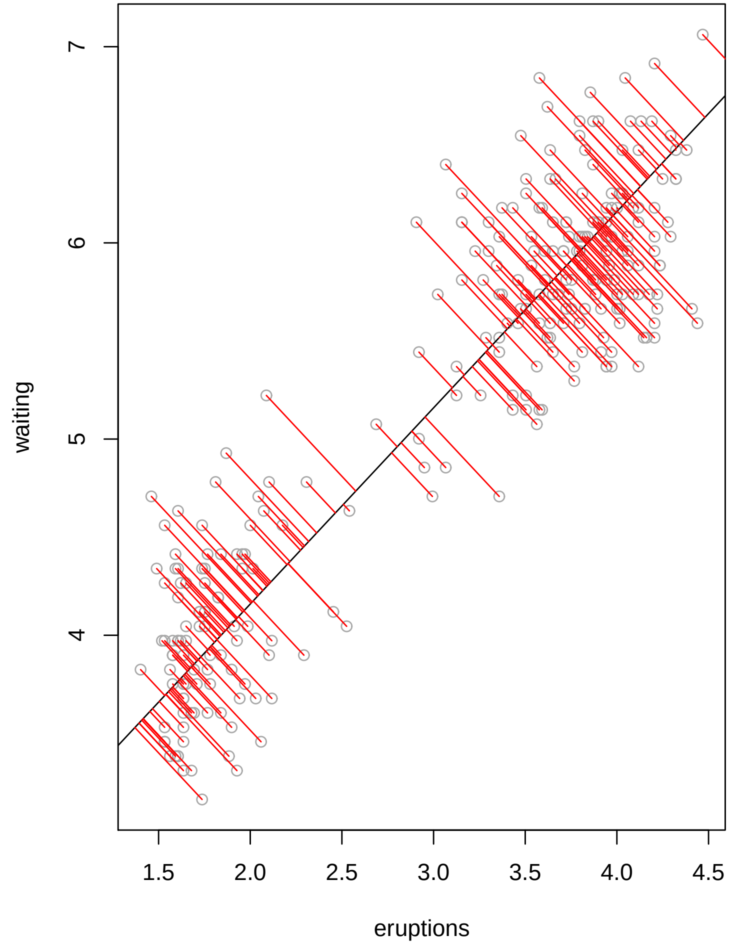


Parametric estimation

Linear regression
Supervised Learning

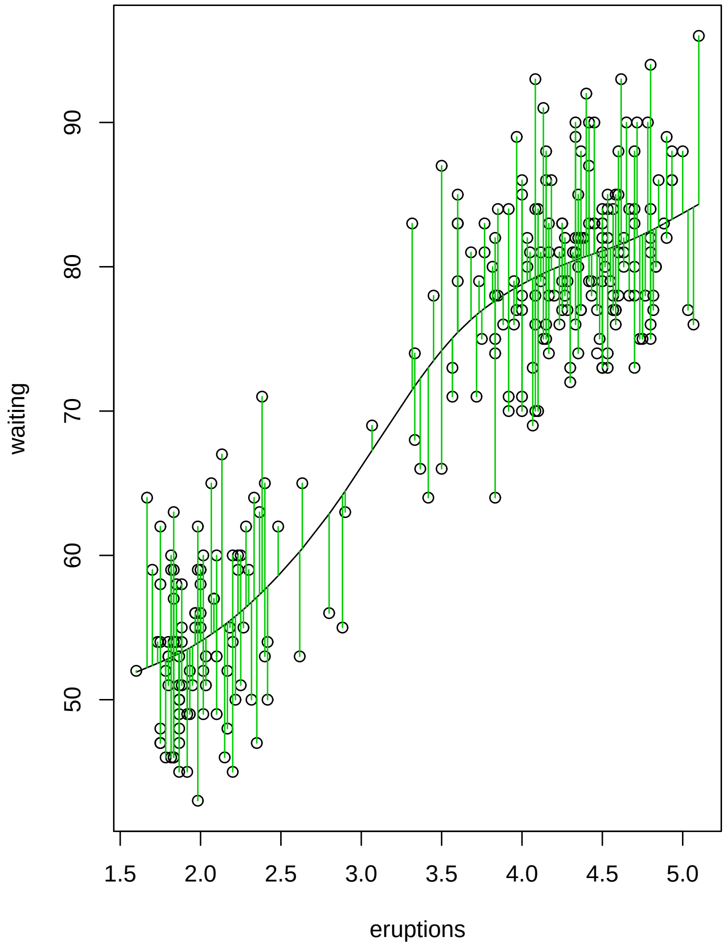
PCA
Unsupervised Learning



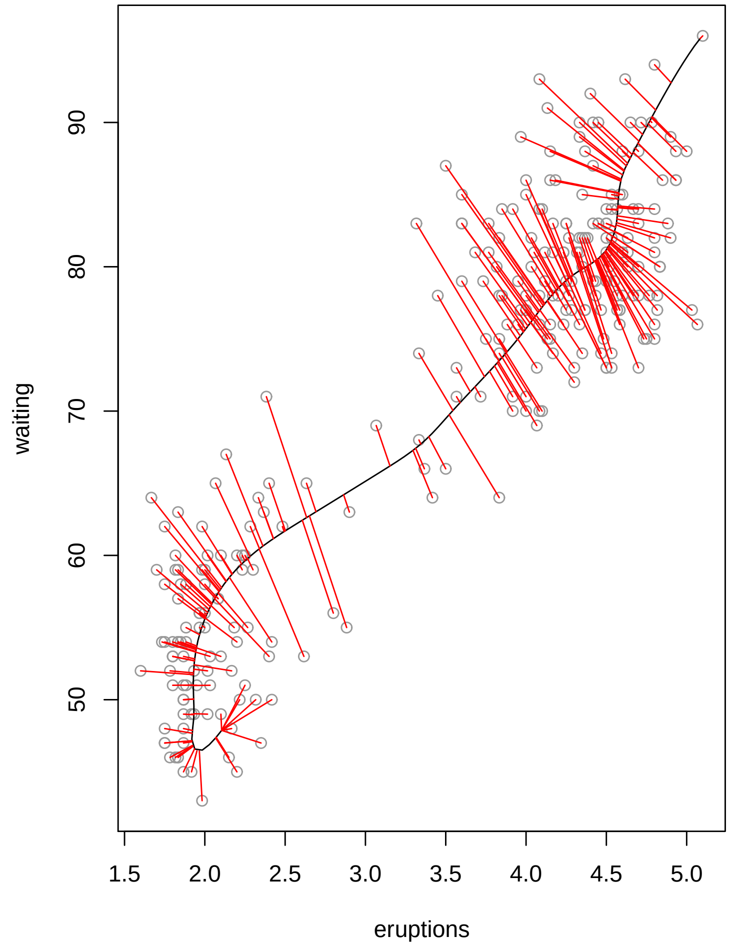


Nonparametric estimation

Nonparametric regression
Supervised Learning



Principal curve
Unsupervised Learning



Principal curves

- Descriptively, a principal curve is a smooth curve through the "middle" of a data cloud $\boldsymbol{X}$.
- A principal curve is symmetric w.r.t. interchanging the coordinate axes.
- As such, a principal curve is a representant of a "nonparametric unsupervised learning technique".
- Today exist a variety of different notions of principal curves, roughly dividable in two categories:

'Top-down' algorithms start with a globally fitted initial line (e.g. the 1st PC) and bend this line or concatenate other lines to it until some convergence criterion is met.

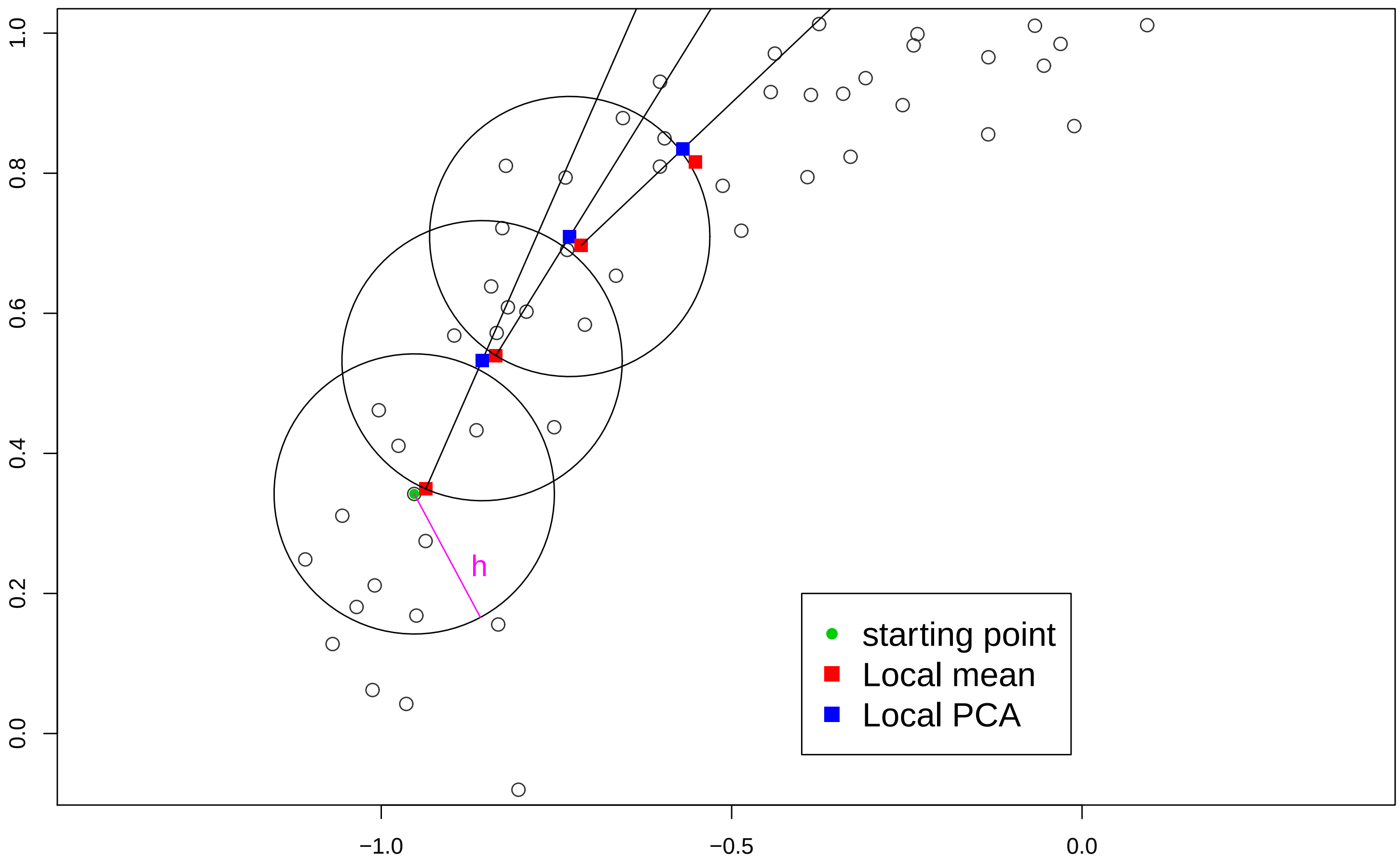
- Hastie & Stuetzle 1989 (HS),...

'Bottom-up' algorithms estimate the principal curve locally moving step by step through the data cloud.

- Einbeck, Tutz & Evers 2005 (LPC), ...

Local principal curves (LPC)

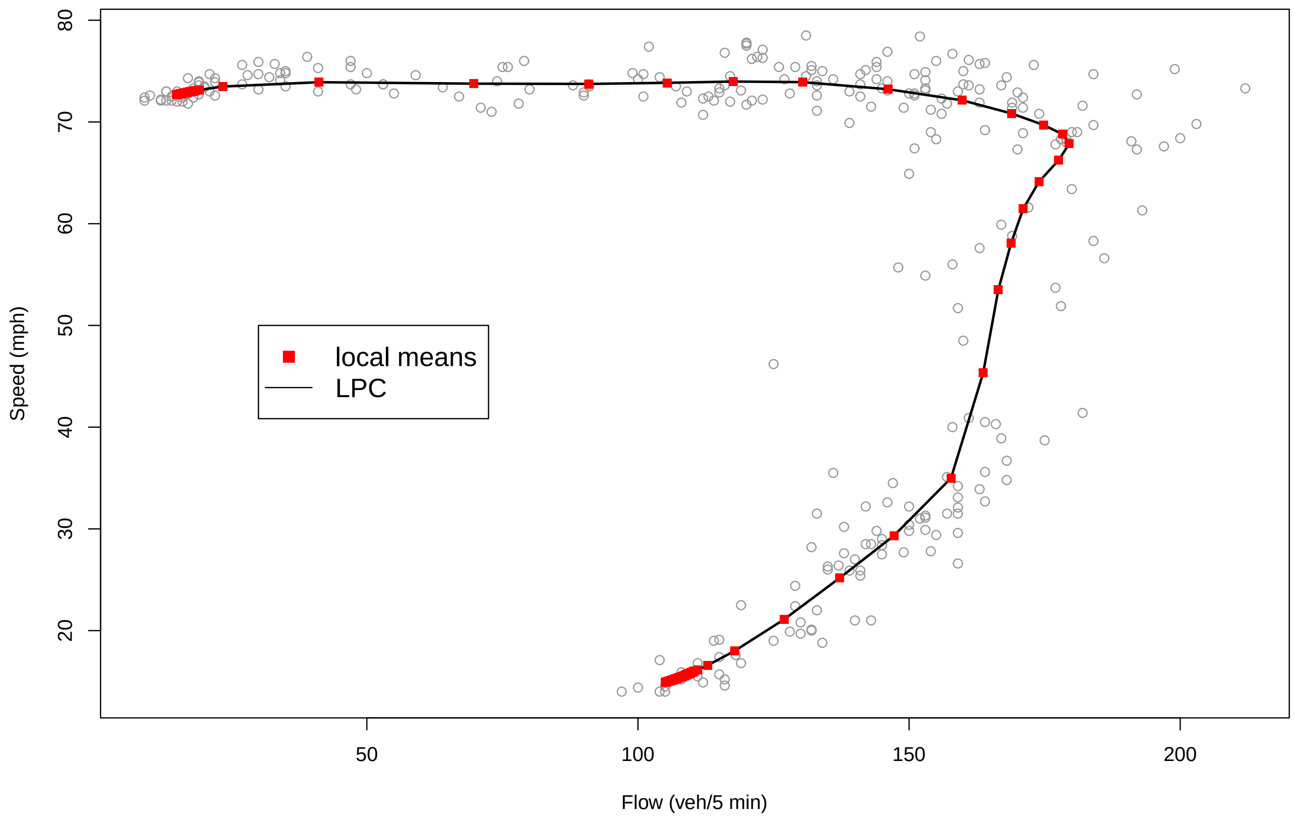
- Idea: Calculate alternately a local mean and a first local principal component, each within a certain radius ("bandwidth") h.



- The LPC is the series of local means.

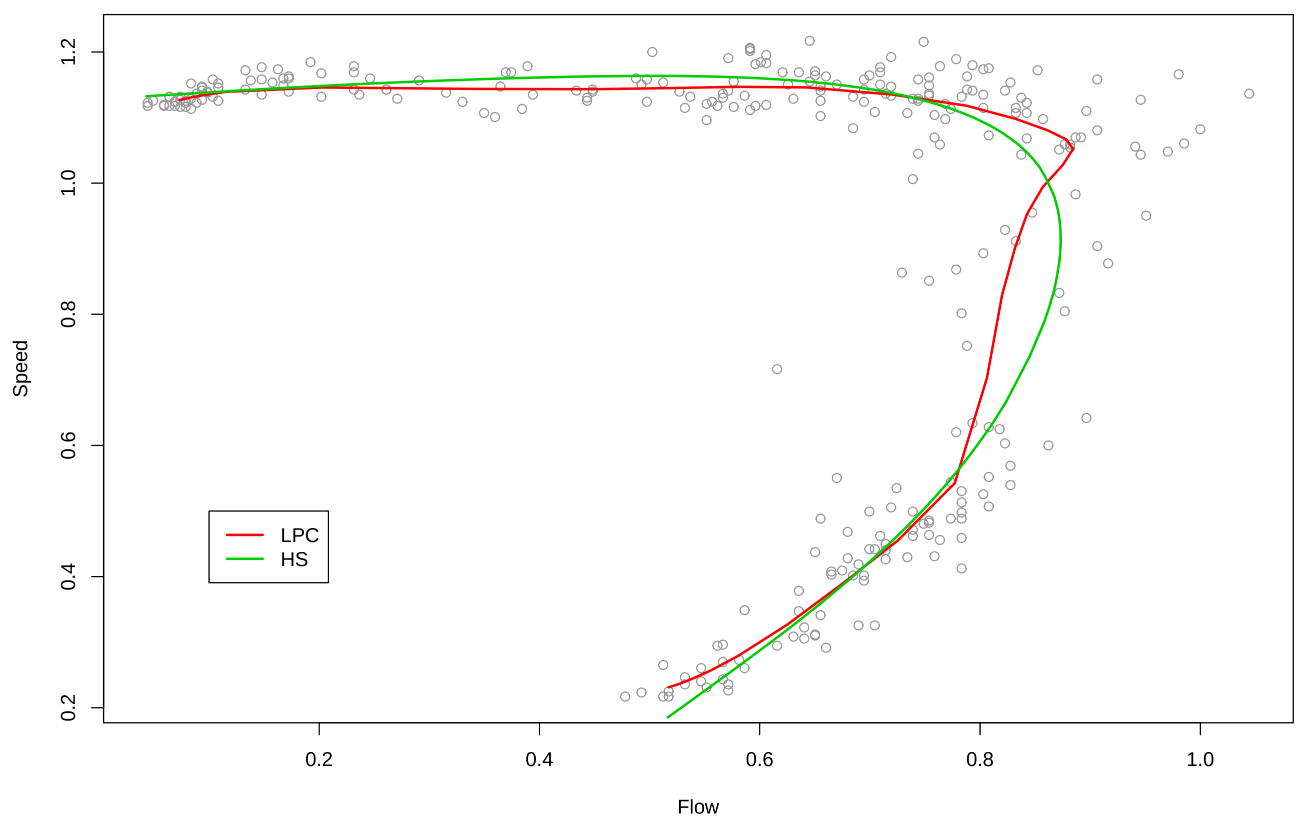
Second Example: Speed-Flow data

- $n = 288$ measurements of traffic speed and vehicle flow on a Californian Freeway, with local principal curve.



Speed-Flow data (cont.)

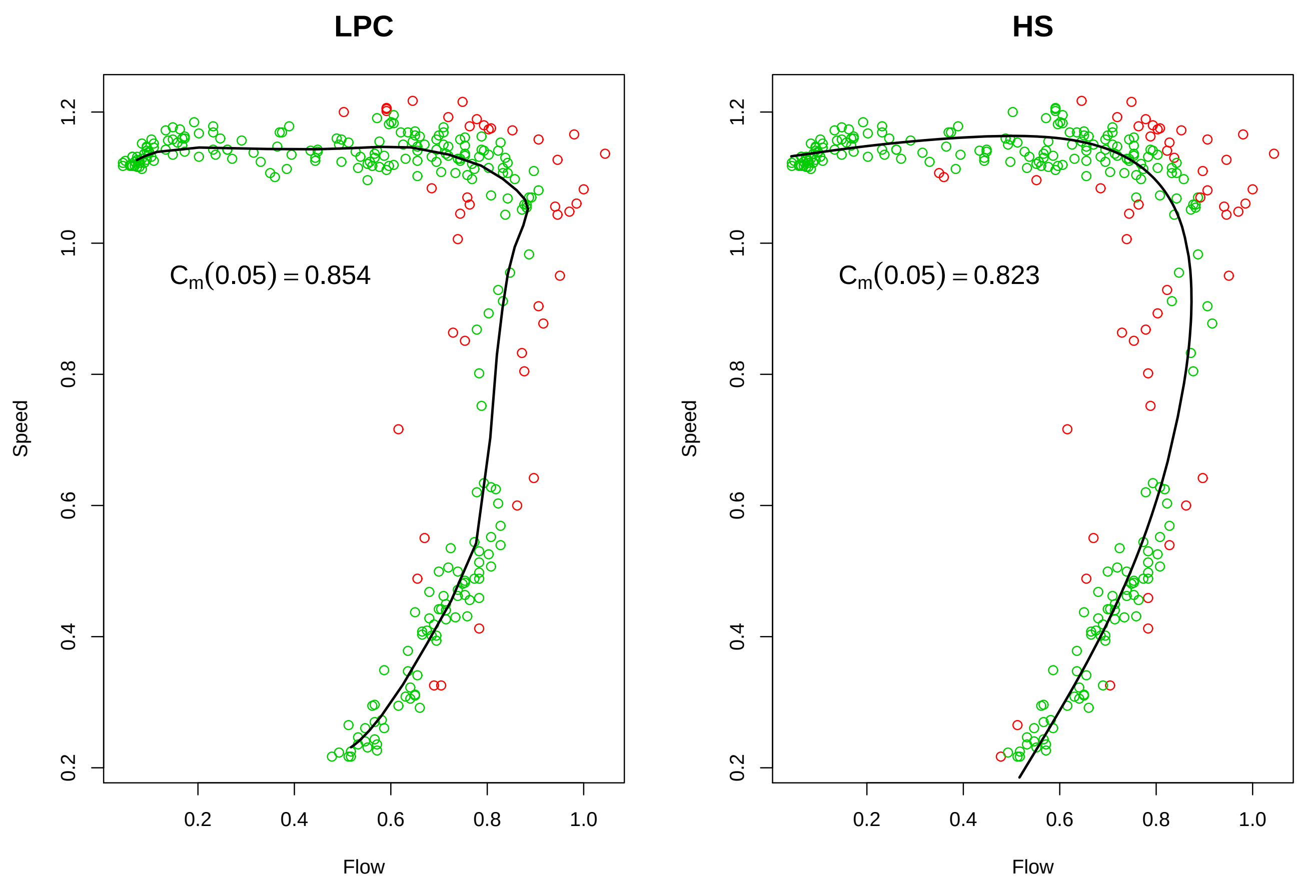
- Compare with HS curve (variables now standardized):



- How can we measure which curve fits better?

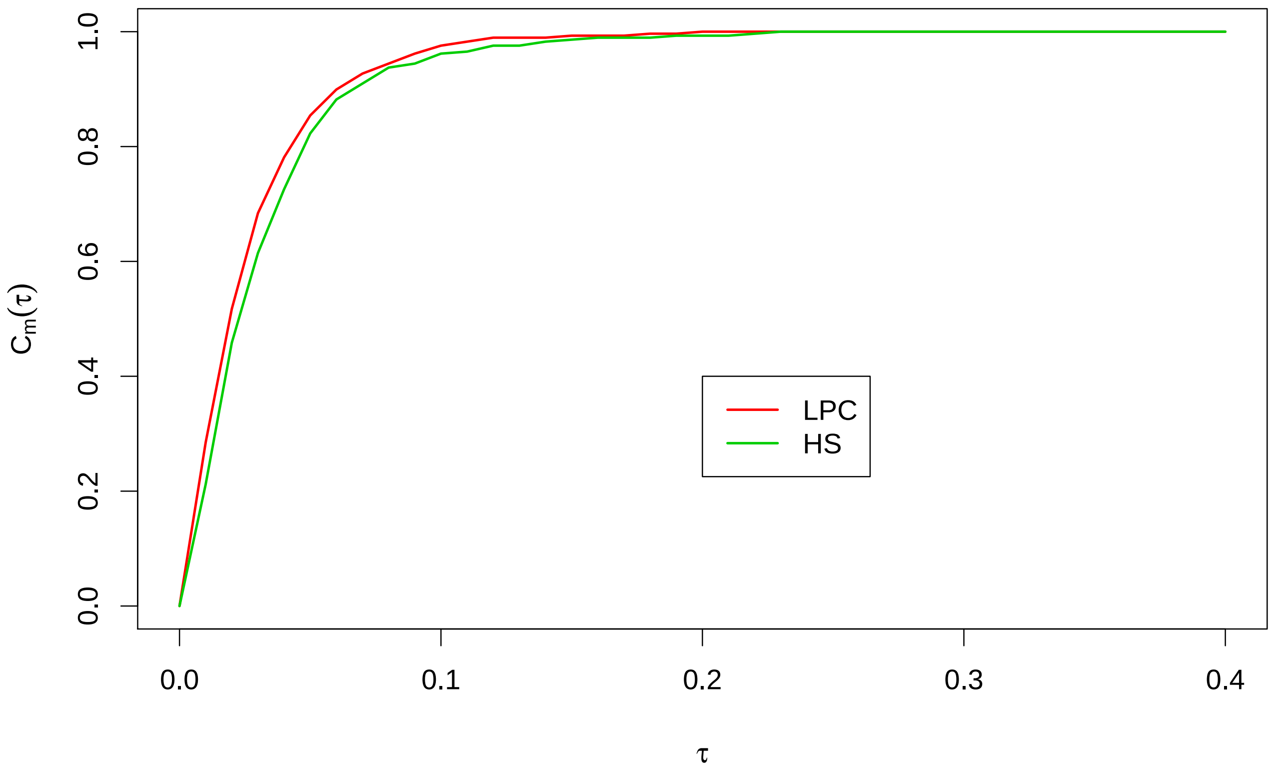
Coverage

- The coverage $C_{\boldsymbol{m}}(\tau)$ of a principal curve $\boldsymbol{m}$ is the proportion of all data points lying within a tube around $\boldsymbol{m}$ with radius τ.
- Compute $C_{\boldsymbol{m}}(0.05)$ for the two principal curves fitted before:



Coverage (cont.)

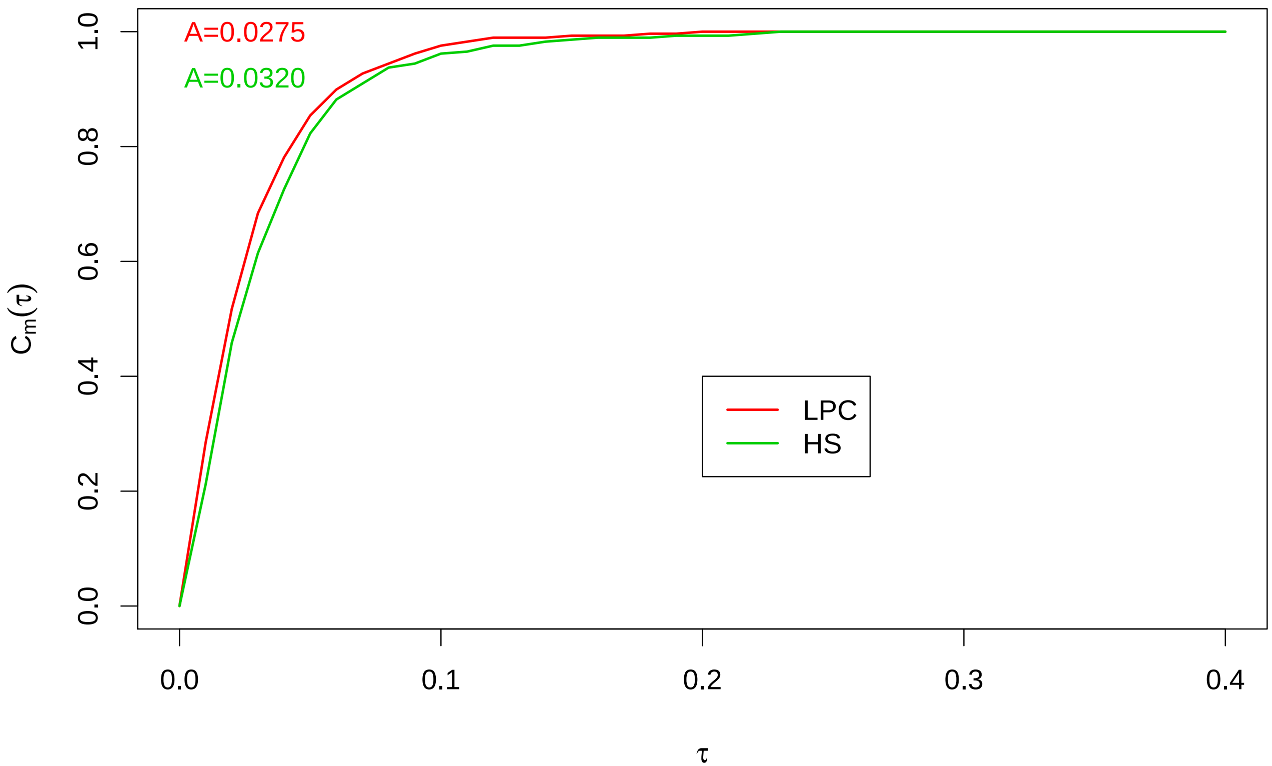
- Of course, this measure depends on the tube width τ, but we can compute the coverage curve over all τ.



- A "good" coverage curve will be concave and rise quickly.
- Compute left top area, say A, between $\tau = 0$, $C_{\boldsymbol{m}}(\tau) = 1$, and the curve.

Coverage (cont.)

- Of course, this measure depends on the tube width τ, but we can compute the coverage curve over all τ.



- A "good" coverage curve will be concave and rise quickly.
- Compute left top area, say A, between $\tau = 0$, $C_{\boldsymbol{m}}(\tau) = 1$, and the curve.
- Small advantage for LPC!

Interpretation

- Theoretically, this area has an appealing interpretation. Denote $||\boldsymbol{\epsilon}_i|| = ||\boldsymbol{x}_i - \boldsymbol{m}||$ the norm of the "residuals", i.e. the shortest distance between a point $\boldsymbol{x}_i$ and the principal curve $\boldsymbol{m}$.
- Note that

$$C_{\boldsymbol{m}}(\tau) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{||\boldsymbol{\epsilon}_i||\leq\tau\}}|| \equiv F_n(\tau)$$

which is the empirical distribution function of the residuals. Then

$$A = \int_0^\infty (1 - F_n(\tau))d\tau = \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty 1_{\{||\boldsymbol{\epsilon}_i||>\tau\}}d\tau = \frac{1}{n}\sum_{i=1}^{n}||\boldsymbol{\epsilon}_i||$$

is just the mean length of the residuals!

R_C

- Next, we set this area A in proportion to the corresponding area A_{PC} which would be obtained when fitting a linear principal component line (the parametric benchmark). Computing "1 minus this ratio" yields the coverage coefficient, R_C

$$R_C \equiv 1 - \frac{A}{A_{\text{PC}}} = 1 - \frac{\sum_{i=1}^{n} ||\boldsymbol{\epsilon}_i||}{\sum_{i=1}^{n} ||\boldsymbol{\epsilon}_i^{(\text{PC})}||} = \frac{\sum_{i=1}^{n} \left(||\boldsymbol{\epsilon}_i^{(\text{PC})}|| - ||\boldsymbol{\epsilon}_i|| \right)}{\sum_{i=1}^{n} ||\boldsymbol{\epsilon}_i^{(\text{PC})}||}$$

- Hence, R_C can be interpreted as the **mean reduction in residual length**.
- For instance, $R_C = 0.8$ means that the mean residual length has been reduced by 80% when using a principal curve instead of a principal component.

R_C (cont.)

- R_C has values in $(-\infty, 1]$, with
  - 1 corresponding to the best possible fit,
  - 0 corresponding to a 'bad' fit of the same quality as PCA,
  - negative values corresponding to a fit being worse than PCA.
- Similar in spirit to coefficient of determination (R^2).
- For instance, for the two principal curves fitted to the traffic data, one has:

  LPC $R_C = 0.8692$

  HS $R_C = 0.8485$
- Both curves give a good fit; LPC sightly better.

Bandwidth selection

- Coverage and R_C are goodness-of-fit criteria. Using these for bandwidth selection would clearly lead to overfitting.

Bandwidth selection

- Coverage and R_C are goodness-of-fit criteria. Using these for bandwidth selection would clearly lead to overfitting.
- However, intuitively, if a certain bandwidth h leads to a "good" principal curve, then a tube with *the same* radius h around this curve should warrant a high coverage.

Bandwidth selection

- Coverage and R_C are goodness-of-fit criteria. Using these for bandwidth selection would clearly lead to overfitting.
- However, intuitively, if a certain bandwidth h leads to a "good" principal curve, then a tube with *the same* radius h around this curve should warrant a high coverage.
- This leads to the idea of self-coverage: Use the same bandwidth **for the curve fitting and for the coverage estimation**:

$$S(\tau) = C_{\boldsymbol{m}(\tau)}(\tau)$$

where $\boldsymbol{m}(\tau)$ is a local principal curve estimated using bandwidth $h = \tau$.

Bandwidth selection

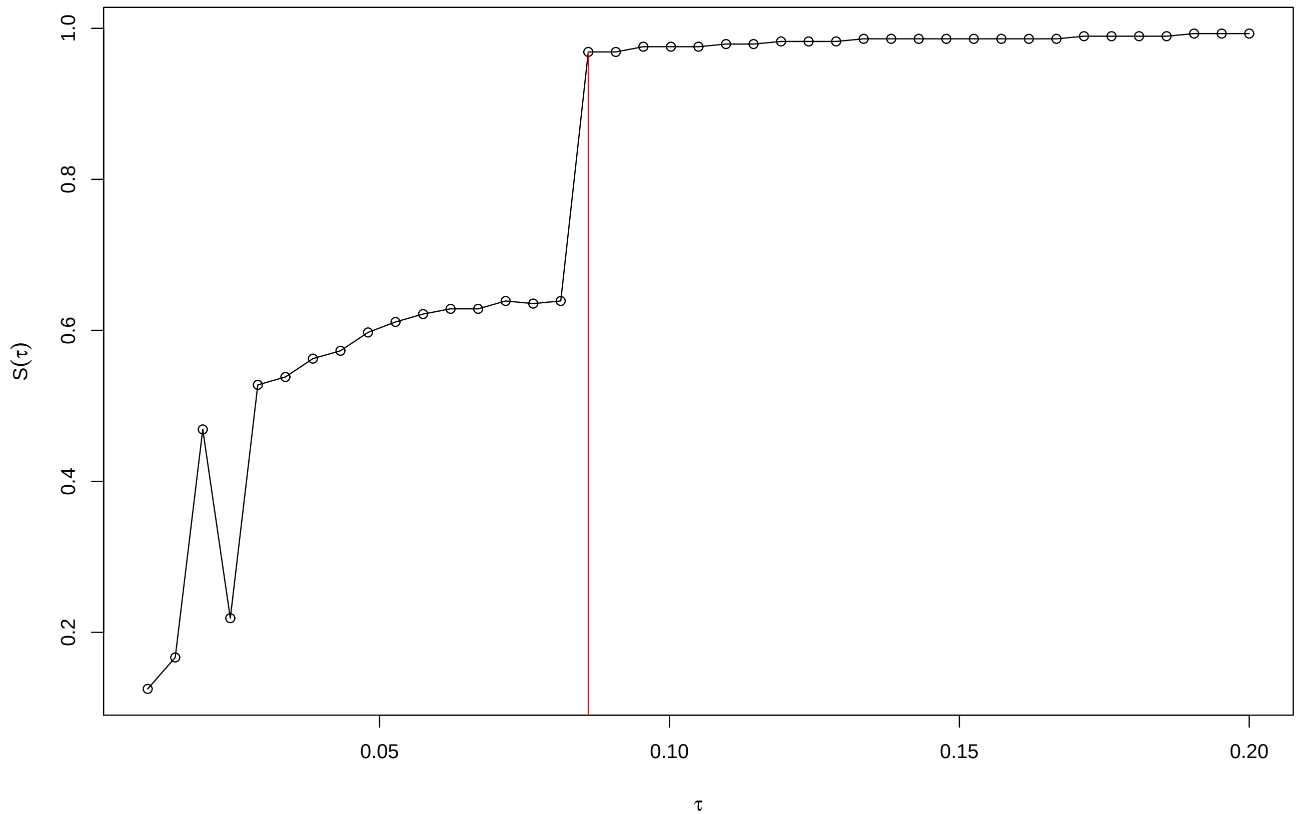
- Coverage and R_C are goodness-of-fit criteria. Using these for bandwidth selection would clearly lead to overfitting.
- However, intuitively, if a certain bandwidth h leads to a "good" principal curve, then a tube with *the same* radius h around this curve should warrant a high coverage.
- This leads to the idea of self-coverage: Use the same bandwidth **for the curve fitting and for the coverage estimation**:

$$S(\tau) = C_{\boldsymbol{m}(\tau)}(\tau)$$

  where $\boldsymbol{m}(\tau)$ is a local principal curve estimated using bandwidth $h = \tau$.
- Unlike $C_{\boldsymbol{m}}(\tau)$, the curve $S(\tau)$ is not necessarily monotone, but has usually local maxima or jumps which correspond to good bandwidths.

Self-coverage curve

- We compute the self-coverage curve for the Californian speed-flow diagram:



- Selected bandwidth: $h = 0.086$
  - The resulting curve has $R_C = 0.8745$.

Generalization

- These ideas generalize to other unsupervised learning problems.
- Examples include density mode detection and clustering.
- The essential device is the computation of the local mean ("mean shift"):

$$\hat{\boldsymbol{\mu}}(\boldsymbol{x}) = \frac{\sum K_h(\boldsymbol{x}_i - \boldsymbol{x})\boldsymbol{x}_i}{\sum K_h(\boldsymbol{x}_i - \boldsymbol{x})}$$

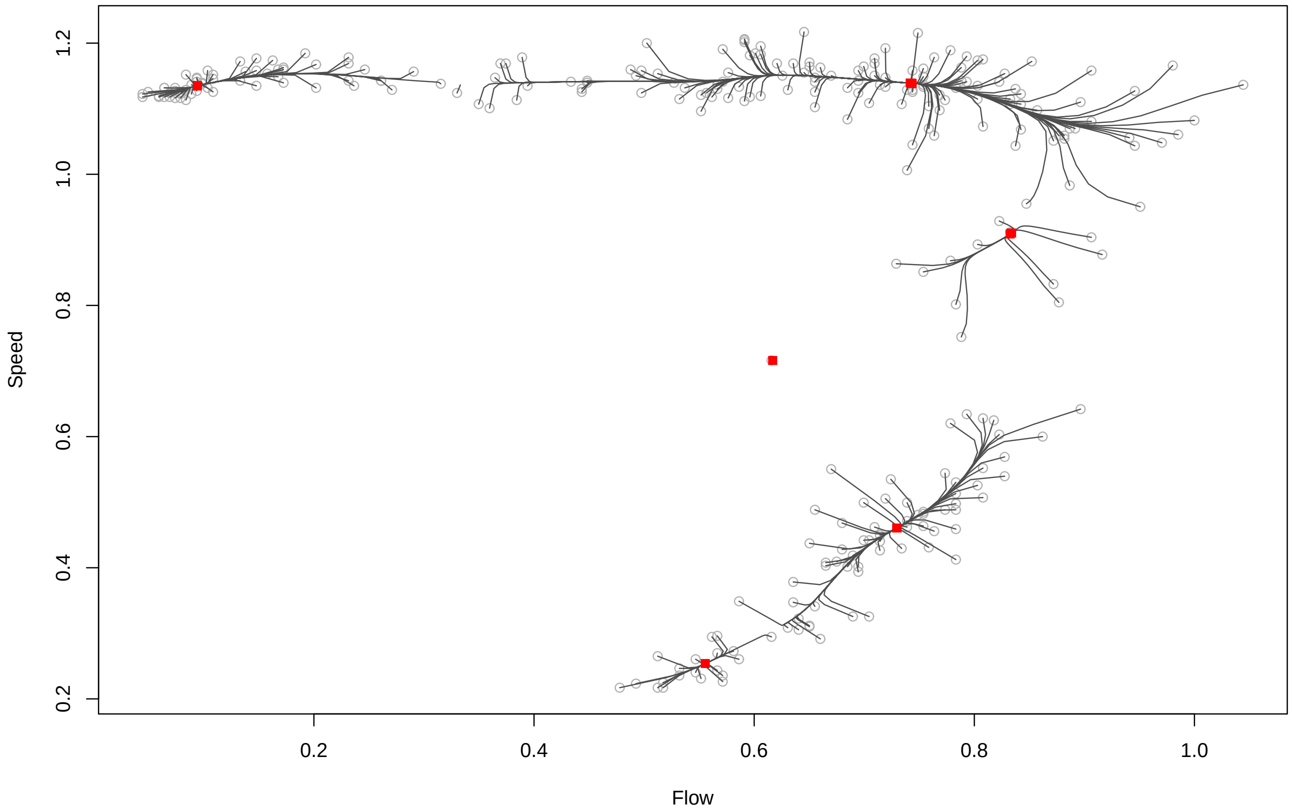
  with

$$K_h(\boldsymbol{x}_i - \boldsymbol{x}) = \frac{1}{h^d} K\left(\frac{||\boldsymbol{x}_i - \boldsymbol{x}||}{h}\right)$$

- Iterating the mean shift, i.e. $\boldsymbol{x}^{(j+1)} = \hat{\boldsymbol{\mu}}(\boldsymbol{x}^{(j)})$, leads to a local mode of the kernel density estimate $\hat{f}_h$ of the true density f. (Comaniciu & Meer, 2002).

Mean-shift based mode detection

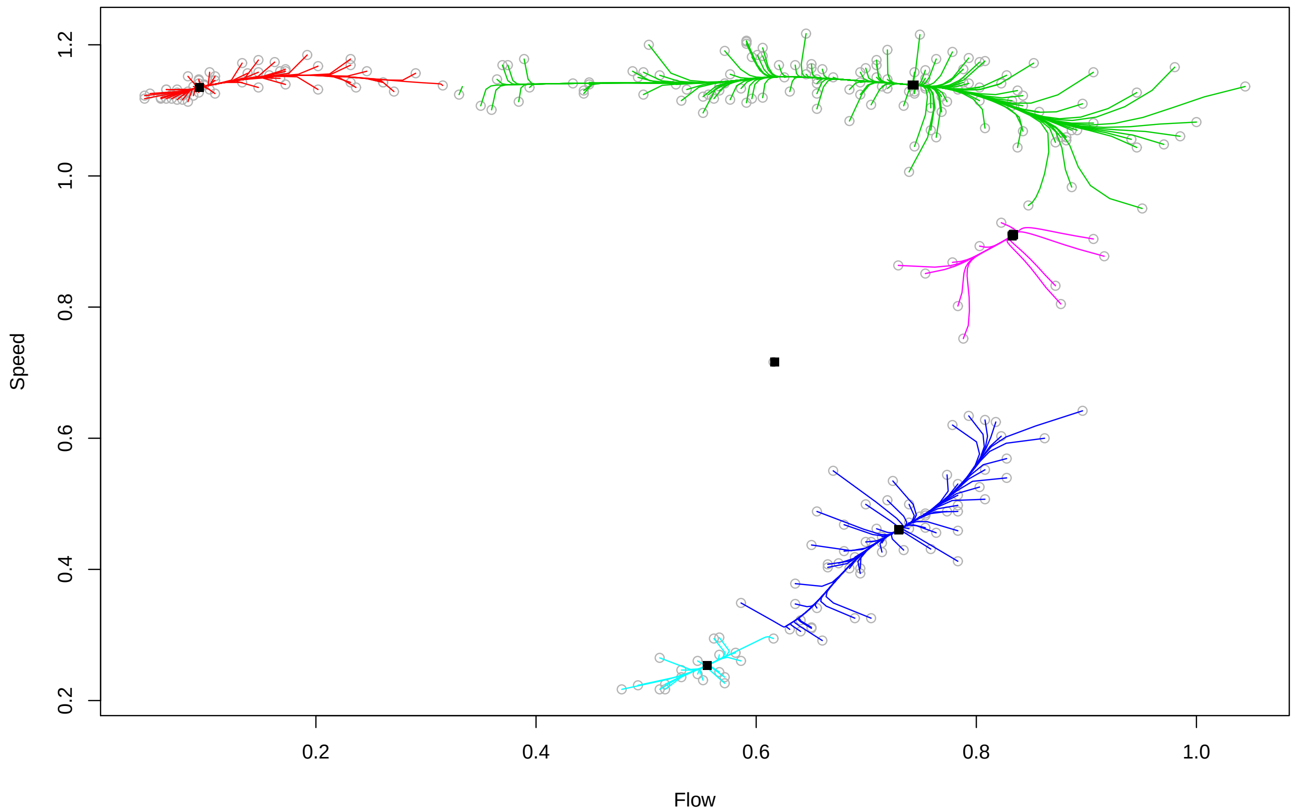
- Starting from each data point $\boldsymbol{x}_i$, iterate the mean shift until convergence:



- for $h = 0.05$, six distinct modes are detected.

Mean-shift based clustering

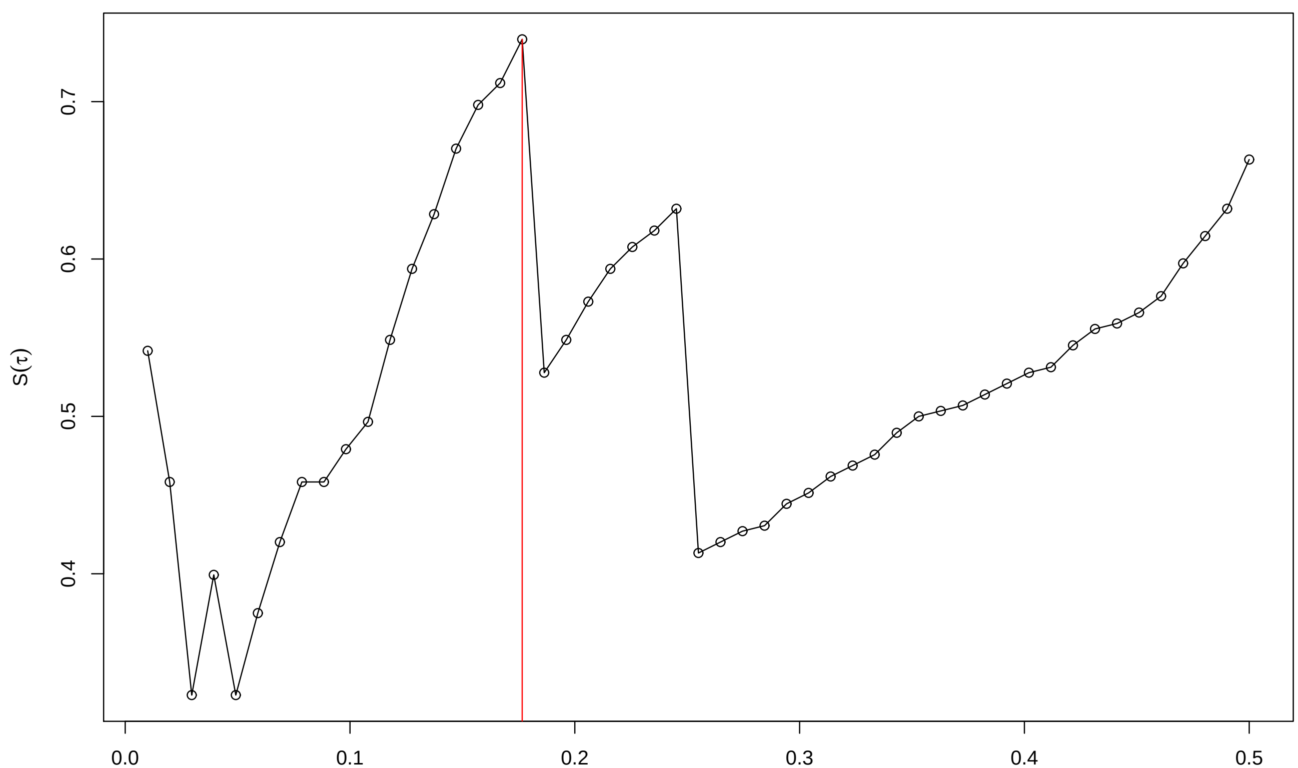
- By assigning each data point to the mode to which it converged, this turns into a clustering technique:



- for $h = 0.05$, six distinct clusters are detected.

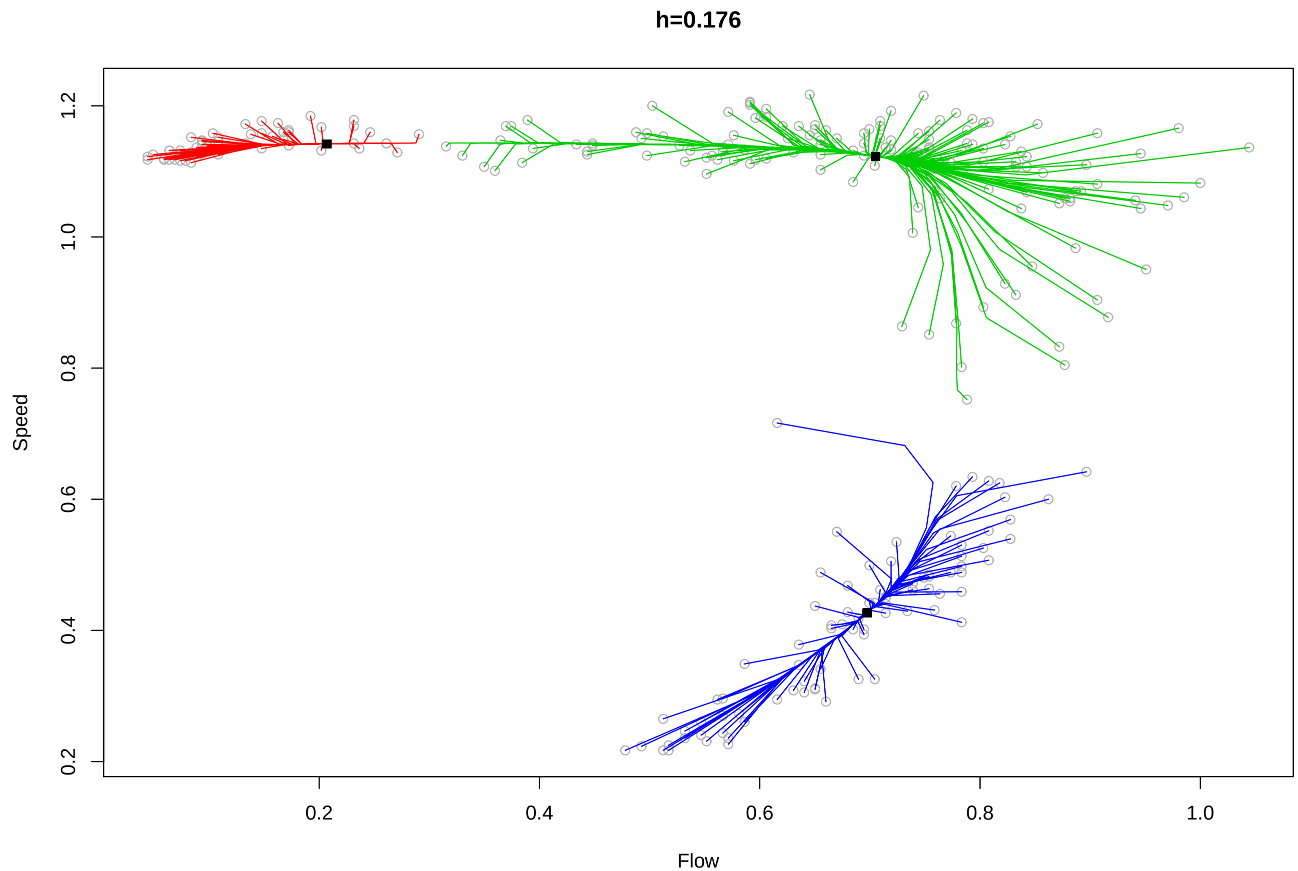
Bandwidth selection

- In contrast to other clustering techniques (such as k-means), mean shift clustering does not require pre-specification of the number of clusters, k.
- However, one needs to specify a bandwidth h instead.
- Self-coverage is calculated as before: The proportion of points in a circle of radius τ, where $h = \tau$ is used for the mean shift clustering.



Bandwidth selection

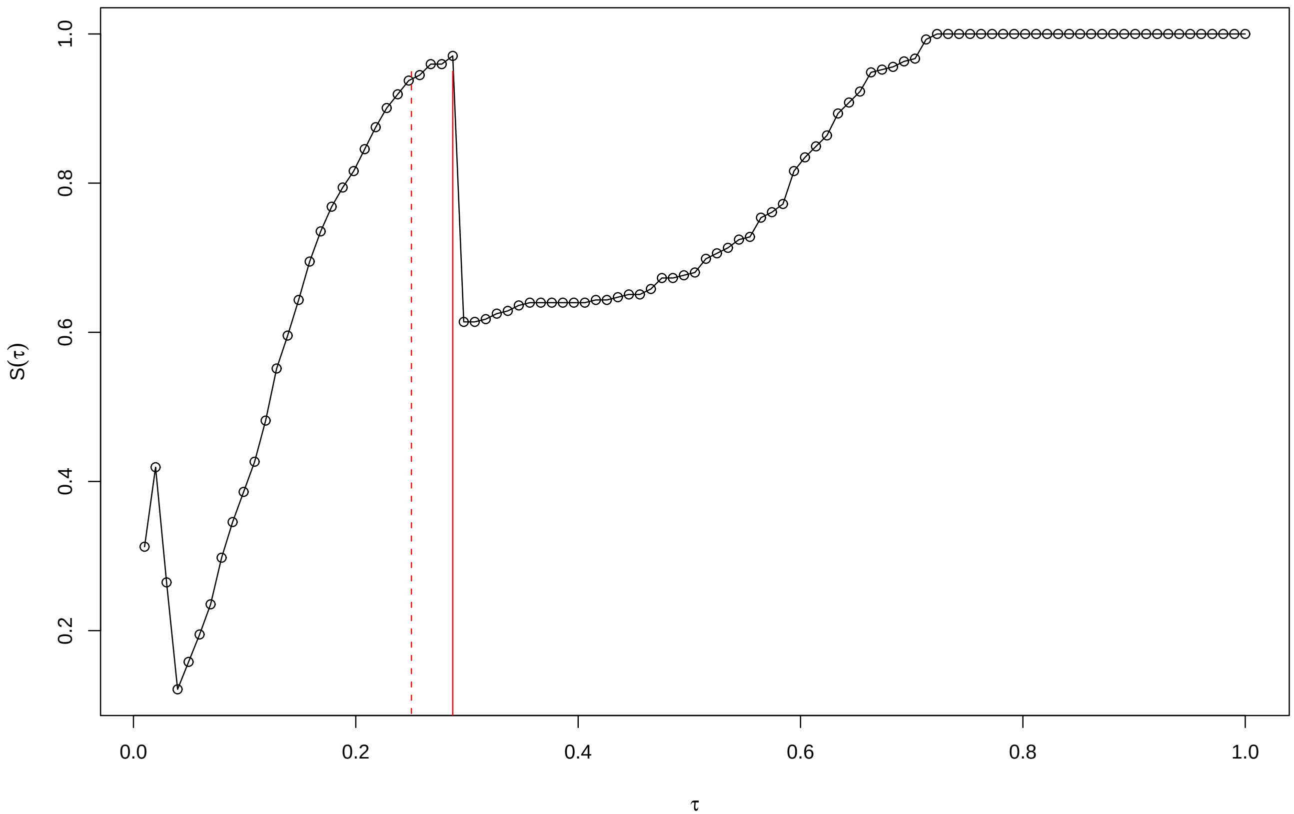
- Mean shift clustering using bandwidth selected via self-coverage:



- $h = 0.176$ corresponds to $k = 3$ clusters.

Old Faithful data

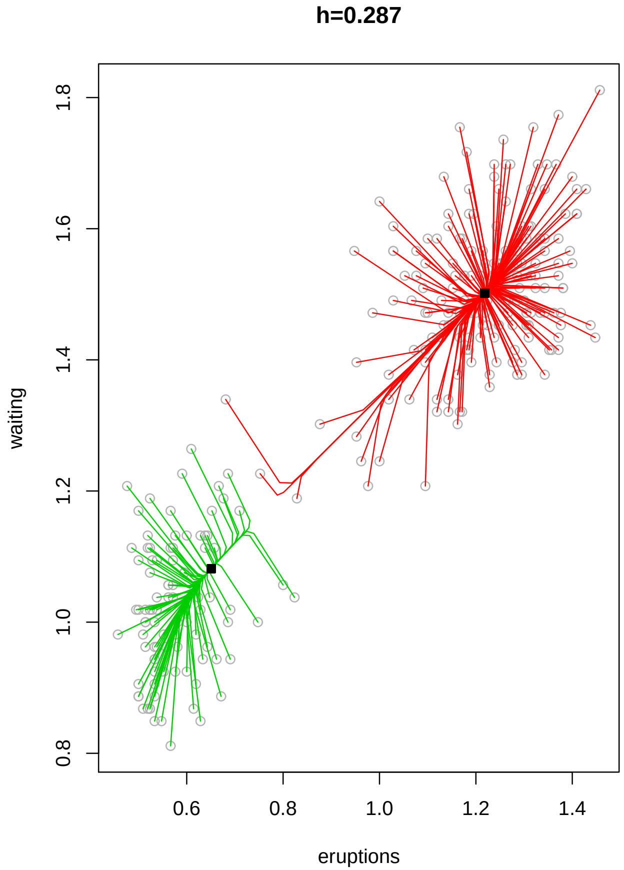
- Self-coverage curve for Old Faithful data:

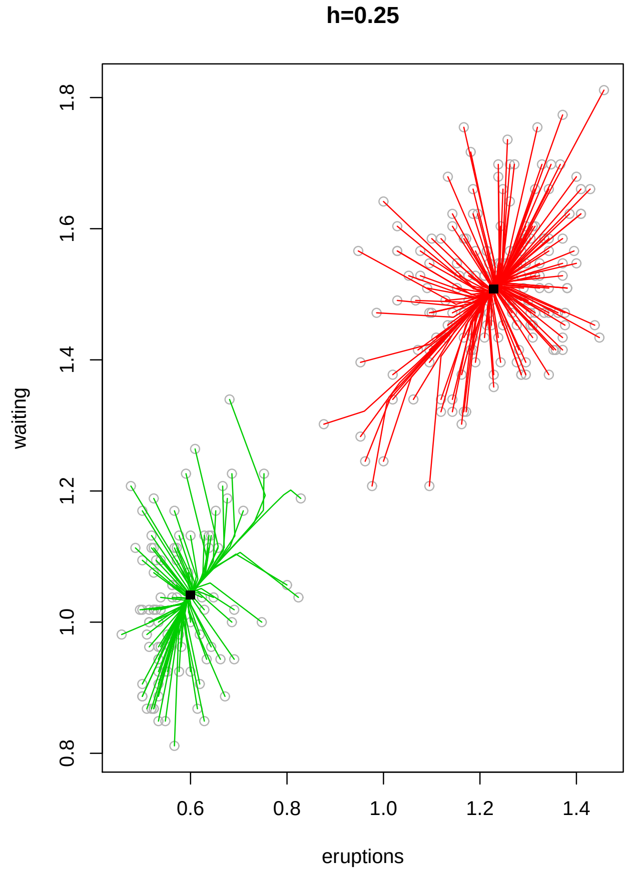


- peaks at $h = 0.287$.

Old Faithful data (cont.)

- Don't be greedy.....





$R_C = 0.4577$ $R_C = 0.5065$

Discussion

- Checking for **goodness-of-fit** should be separated from **model selection** (here bandwidth selection). This is not different than in the regression context (supervised learning): The value R^2 is a goodness-of-fit criterion, and should not be used for model selection!
- The goodness-of-fit of principal curves or clustering methods can be assessed qualitatively (through a coverage curve) or quantitatively (through the relative mean reduction in residual length, R_C).
- For bandwidth selection in this context, a self-coverage measure works well.

References

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.

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

Einbeck & Evers (2010): **LPCM** (Local principal curves and manifolds). R package version 0.39-1, `http://www.maths.dur.ac.uk/~dma0je/software.html`.