The Fitting of Multifunctions:
An approach to Nonparametric Multimodal Regression

Jochen Einbeck
National University of Ireland, Galway

joint work with Gerhard Tutz, University of Munich

GfKI, 10th of March 2005
What is “Nonparametric Regression”?

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 $X = x$. ”

(Holmström et.al, 2003)
Nonparametric regression

Given \( n \) data points \((X_i, Y_i), i = 1, \ldots, n\), one seeks a smooth function \( m : \mathbb{R} \rightarrow \mathbb{R} \) 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

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

recorded on a californian “freeway”.
Attempt: Ordinary nonparametric regression

One assumes a asymmetric relationship between $X$ and $Y$:

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

→ Obviously some information is discarded!
Possible solution: Principal curves

\[
\begin{pmatrix} X \\ Y \end{pmatrix} = f(t) + \epsilon; \quad f : \mathbb{R} \rightarrow \mathbb{R}^2 \quad \text{(X and Y symmetric!)}
\]

- Principal curves are smooth functions passing through the middle of the data cloud.
- They uncover nicely the underlying branched structure.
- However, they are not suitable for prediction of \( Y \) from a given \( X = x \).
New approach: Multi-valued nonparametric regression

As for ordinary regression, we assume an asymmetric relationship

\[ Y = M(X) + \epsilon, \]

but now employing a multifunction (multi-valued ‘function’) \( M : \mathbb{R} \to \mathbb{R} \).

- The data cloud is thought of to consist of several smooth branches (regimes).
- For every \( X = x \), more than one predicted value is possible.
- The estimator for \( M \) at \( x \) is a random set \( \hat{M}(x) = \{ \hat{m}_1(x), \ldots, \hat{m}_k(x) \} \), with \( k \)
  being the number branches.
Basic idea: Consider the conditional densities.
For instance, conditional density at a flow $= 1400$.

- For estimation of $M(x)$, 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.
The estimated branches correspond to the uncongested and congested regimes.
The relevance of the branches varies smoothly for varying $x$, as long as the branches can be separated.
Estimation of conditional modes

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

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

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

\[ \frac{\partial \hat{f}(y \mid x)}{\partial y} = \frac{2c_k}{h_2^3} \sum_{i=1}^{n} K_1 \left( \frac{x - X_i}{h_1} \right) k' \left( \left( \frac{y - Y_i}{h_2} \right)^2 \right) (y - Y_i) \bigg| = 0 \]
and obtains

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

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

Remarks:

- The right side of (1) can be seen as a conditional mean shift! Thus, applying (1) iteratively, the sequence \((y_j)\) converges to a conditional mode of \(Y|X = x\) (Comaniciu & Meer, 2002).

- The right side of (1) corresponds also 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)

• Automatic detection of the number of branches.
Literature

- Principal Curves


- Multi-valued nonparametric regression

• “Related Literature”:


