TP: Kernel methods and dimension reduction for regression

Stéphane Girard & Jérôme Saracco

Inria Grenoble Rhône-Alpes & Inria Bordeaux Sud-Ouest

Solutions can be found in the file TPkernel+SIR(solutions).R

1. One-dimensional kernel regression: Influence of the bandwidth

2. Multidimensional kernel regression with dimension reduction

Simulations

- Generate 100 pairs (X_i, Y_i) from the model $Y = f(X) + \varepsilon$, $X \sim U[0,1]$, $\varepsilon \sim N(0, 1/9)$ with (i) $f(x) = f_1(x) = 2 + 3x$ and (ii) $f(x) = f_2(x) = \sin(4x)$.
- Plot the data, and superimpose the true link function.

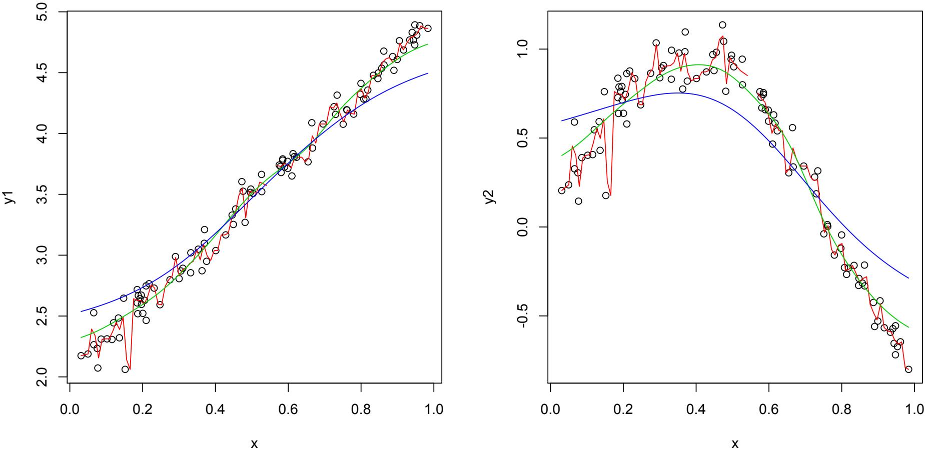
Estimation of the link function

On each of the previous two simulated models, estimate the link function using

- the linear model (lm command),
- the kernel estimator (ksmooth command) with the Gaussian kernel and bandwidth $h \in \{0.01, 0.2, 0.5\}$.

and superimpose the estimators to the previous graphes.

Estimation of the link function: results



Cross-validation (1/2)

Implement the cross-validation procedure for selecting the bandwidth

- For $h \in \{h_{\min}, \dots, h_{\max}\}$ (with *nbh* trials)
- For $j \in \{1, \dots, n\}$
- Compute the estimator at point X_j on the training set excluding X_j with bandwidth h.

$$\hat{f}_{-j}(X_j) = \sum_{i \neq j} K\left(\frac{X_j - X_i}{h}\right) Y_i \Bigg/ \sum_{i \neq j} K\left(\frac{X_j - X_i}{h}\right)$$

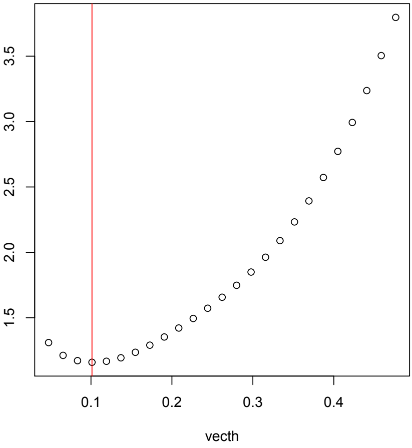
- Compute the associated prediction error: $\hat{\varepsilon}_j^2 = (Y_j - \hat{f}_{-j}(X_j))^2$
- Choose h such that $\sum_{j=1}^{n} \hat{\varepsilon}_j^2$ is the smallest.

Cross-validation (2/2)

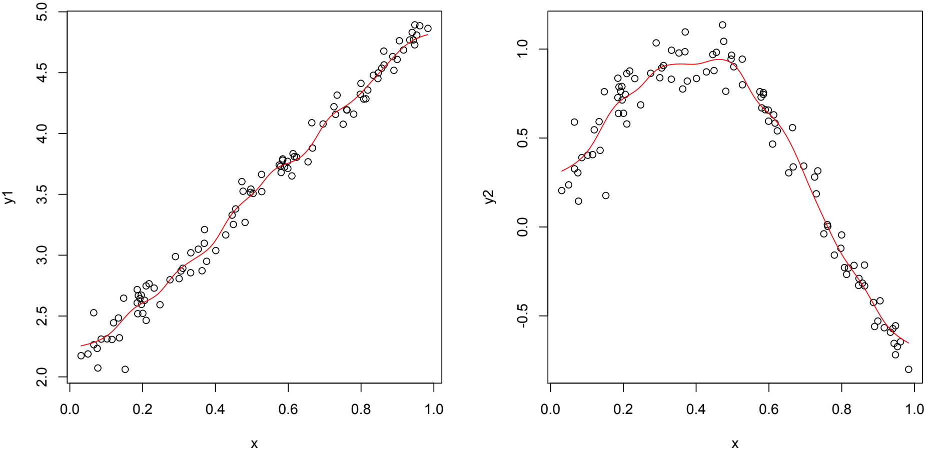
On each of the previous two simulated models,

- plot the cross-validation criteria,
- compute the “optimal bandwidth”, *i.e* minimizing the cross-validation criteria,
- superimpose the kernel estimator computed the “optimal bandwidth” to the simulate data.

Cross-validation criteria



Kernel estimators with the “optimal bandwidth”



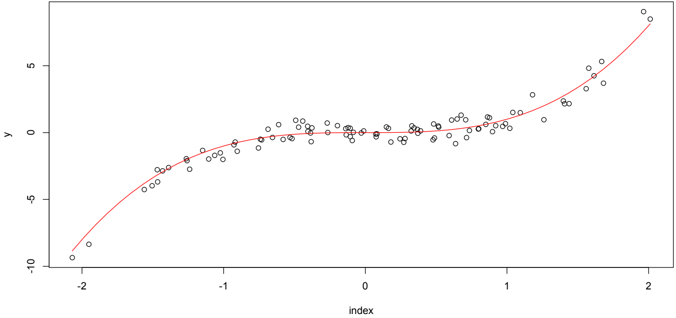
1. One-dimensional kernel regression: Influence of the bandwidth

2. Multidimensional kernel regression with dimension reduction

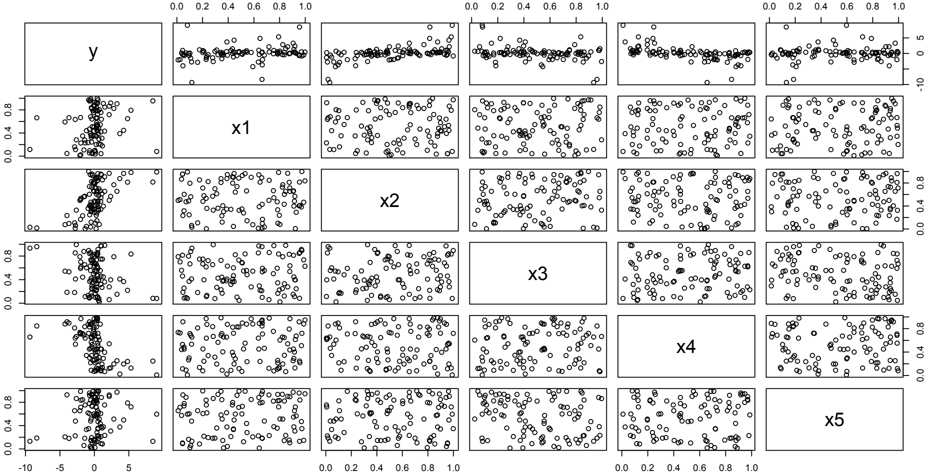
Simulations

- Generate 100 pairs (X_i, Y_i) from the model $Y = f(\beta' X) + \varepsilon$, $X \sim U[0,1]^5$, $\varepsilon \sim N(0, 1/2)$ with $f(x) = x^3$ and $\beta = (1, 2, -1, -2, 0)'$.
- Plot the pairs $(\beta' X_i, Y_i)$, $i = 1, \dots, 100$ and superimpose the link function.

Plot of the pairs $(\beta' X_i, Y_i)$ and true link function



Remark

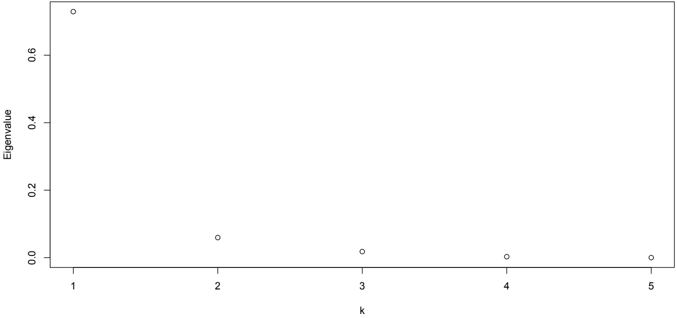


SIR

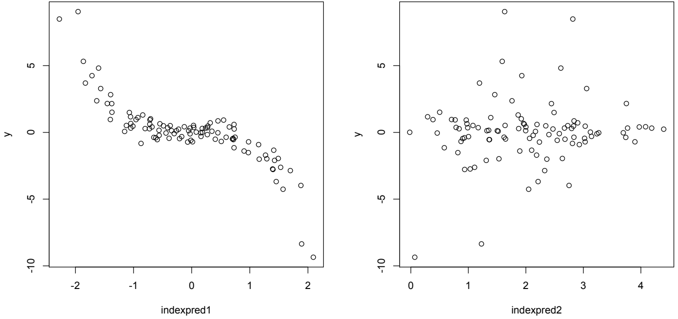
Use the function edr from the edrGraphicalTools package in order to:

- Plot the eigenvalues screeplot and select the dimension of the EDR subspace.
- Plot the pairs $(\hat{b}_1' X_i, Y_i)$, $i = 1, \ldots, 100$ where $\hat{b}_1$ is the first EDR direction. Compare to the plot of $(\hat{b}_2' X_i, Y_i)$, $i = 1, \ldots, 100$ where $\hat{b}_2$ is the second EDR direction.
- Visualize the pairs $(\hat{b}_1' X_i, \beta' X_i)$, $i = 1, \ldots, 100$.

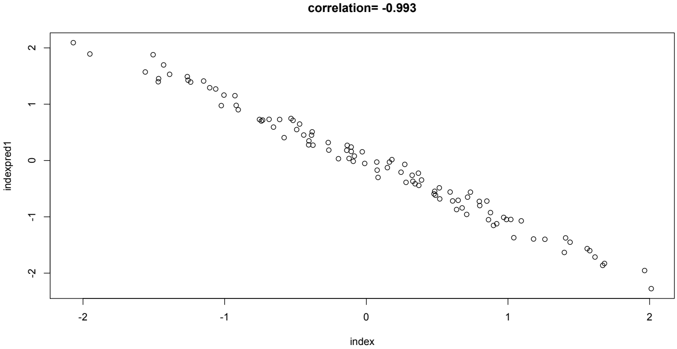
Eigenvalues screeplot



Plots of $(\hat{b}_1' X_i, Y_i)$ and $(\hat{b}_2' X_i, Y_i)$



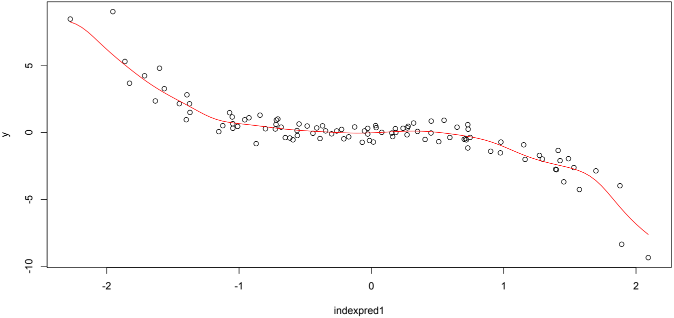
Plot of $(\hat{b}_1' X_i, \beta' X_i)$



Kernel regression on the estimated index

- Use a one-dimensional kernel estimator (with bandwidth selected by cross-validation) to estimate the link function between $\hat{b}_1' X_i$ and Y_i, $i = 1, \ldots, 100$.
- Plot the pairs $(\hat{b}_1' X_i, Y_i)$, $i = 1, \ldots, 100$ and superimpose the estimated link function.

Plot of the pairs $(\hat{b}_1' X_i, Y_i)$ and estimated link function


y
indexpred1