Modelling Non-linear and Non-stationary Time Series

Chapter 2: Non-parametric methods

Henrik Madsen

Advanced Time Series Analysis

September 2016

Non-parametric estimation

- Free of parameters and model structure (apart from a smoothing constant).
- Useful if no prior information about the structure is available.
- The regression curve is a direct function of the data.
- Important methods: Kernel estimation, splines, and nearest neighbor.

Introduction

Assume that Y has PDF $f(y)$, then

$$f(y) = \lim_{h \to 0} \frac{1}{2h} \mathbb{P}\left(Y - h < Y \leq Y + h\right)$$

.

Assume further we have N observations. An estimator of $f(y)$ is:

$$\hat{f}(y) = \frac{1}{2hN} \left[\text{Number of observations in (y-h;y+h)}\right]$$

$$\hat{f}(x) = \frac{1}{2hN} \sum_{i=1}^{n} \chi_{]x-h,x+h]}(x_i)$$

where χ is the characteristic function:

$$\chi_{[x-h,x+h]}(x_i) = \begin{cases} 1 & \text{if } x_i \in]x-h, x+h], \\ 0 & \text{if } x_i \notin]x-h, x+h] \end{cases}$$

Introduction

Define

$$w(u) = \begin{cases} \frac{1}{2} & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

then

$$\hat{f}(y) = \frac{1}{N} \sum_{t=1}^{N} \frac{1}{h} w\left(\frac{y - Y_t}{h}\right)$$

The general kernel estimator:

$$\int k(u)\, \mathrm{d}u = 1$$

$$\hat{f}(y) = \frac{1}{Nh} \sum_{t=1}^{N} k\left(\frac{y - Y_t}{h}\right)$$

The kernel estimator

Assume $\{Y_t\}, \{Z_t\}$ are strongly (or strictly) stationary processes Madsen (2008).
The basic quantity estimated is then Robinson (1983):

$$\mathbb{E}\left[g(Z_t)|Y_t = y\right] f_{Y_t}(y) \qquad (*)$$

where $f_{Y_t}(y)$ is the PDF of Y_t. The solution to (*) is given by

$$[G(Z_t); y] = (N'h)^{-1} \sum_{t=1}^{N'} G(Z_t) k\left(\frac{y - Y_t}{h}\right)$$

The kernel estimator - examples

Example 1

$$\hat{f}_Y(y) = [1;y] = \frac{1}{Nh}\sum_{t=1}^{N} k\left(\frac{y-Y_t}{h}\right)$$

Example 2

$$\hat{\mathbb{E}}\left[G(Z_t)|Y_t = y\right] = \frac{[G(Z_t);y]}{[1;y]}$$
$$= \frac{\frac{1}{N'}\sum G(Z_t)k(\frac{y-Y_t}{h})}{\frac{1}{N}\sum k(\frac{y-Y_t}{h})}$$

Notice the special case

$$g(Z_t) = Y_{t+1}$$

leads to estimation of

$$\mathbb{E}\left[Y_{t+1}|Y_t = y\right]$$

The kernel estimator

Assume that the theoretical relation is

$$Y = g(X^{(1)}, \ldots, X^{(q)}) + \epsilon$$

where ϵ is white noise, and g is an unknown continuous function. Given N observations

$$O = \{(X_1^{(1)}, \ldots, X_1^{(q)}, Y_1), \ldots, (X_N^{(1)}, \ldots, X_N^{(q)}, Y_N)\}$$

the goal is now to estimate the function g.
If $q = 1$ the *kernel estimator* for g given the observations is

$$\hat{g}(x) = \frac{\frac{1}{N}\sum_{s=1}^{N} Y_s k\{h^{-1}(x - X_s)\}}{\frac{1}{N}\sum_{s=1}^{N} k\{h^{-1}(x - X_s)\}}$$

The kernel estimator

It is shown in Robinson (1983) that under reasonable assumptions about the stochastic process $\{X_t\}$ and the kernel, the kernel estimator converges in probability to the conditional expectation of Y:

$$\hat{g}(x) \xrightarrow{p} \mathbb{E}\left[Y|X=x\right]$$

More precisely:

$$N \to \infty, h \to 0, Nh \to \infty$$

Then at every point of continuity of $g(x)$:

$$\hat{g}(x) \xrightarrow{p} g(x)$$

Kernel functions

The function, k is the *kernel*, and h is called the *bandwidth* of the kernel.
A kernel is a real and bounded function which satisfies $\int k(u) \, \mathrm{d}u = 1$.
The kernel with bandwidth, h, is given by

$$K_h(u) = h^{-1} k(u/h)$$

Commonly used kernels are *Gaussian*, *Epanechnikov kernel*, rectangular, and triangular kernels.

Important kernels

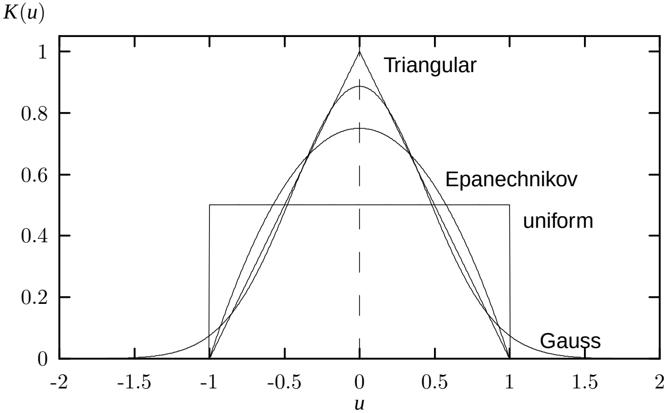
The *Epanechnikov kernel* with bandwidth h is given by :

$$K_h^{Epa}(u) = \frac{3}{4h}(1 - \frac{u^2}{h^2})I_{\{|u| \leq h\}}$$

The *Gaussian kernel* with bandwidth h is given by:

$$K_h^{Gau}(u) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{u^2}{2h^2}\right)$$

Kernel functions - illustration



Nearest Neighbor (NN) estimation

- Previously: h was fixed.
- Another (and often better) alternative is
  * k-NN (k Nearest Neighbor) (Note: here, k is a natural number - not the kernel).

Since it is reasonable to link the value of k to N we often consider a fraction (or percentage) of the data. Often the symbol, $\hbar$, is used in this case.

Multivariate kernels

If we allow $q > 1$ in (1) a kernel of q'th order have to be used, i.e. a real bounded function $k_q(u_1, \dots, u_q)$, where $\int k_q(\boldsymbol{u}) d\boldsymbol{u} = 1$.
By a generalization of the bandwidth h to the quadratic matrix $\boldsymbol{h}$ with the dimension $q \times q$ the more general kernel estimator for $g(X^{(1)}, \dots, X^{(q)})$ is

$$\hat{g}(\boldsymbol{x}) = \frac{\frac{1}{N} \sum_{s=1}^{N} Y_s k_q[(\boldsymbol{x} - \boldsymbol{X_s})\boldsymbol{h}^{-1}]}{\frac{1}{N} \sum_{s=1}^{N} k_q[(\boldsymbol{x} - \boldsymbol{X_s})\boldsymbol{h}^{-1}]}$$

where $\boldsymbol{x} = (x_1, \dots, x_q)$ and $\boldsymbol{X_s} = (X_s^{(1)}, \dots, X_s^{(q)})$. This is called the *Nadaraya-Watson estimator*.

Multivariate kernels - example

If $\boldsymbol{k}_q$ is parameterized using the normal distribution, $N_q(\mathbf{0}, \boldsymbol{I})$, it is seen that $|\boldsymbol{h}|^{-1} k_q\left((\boldsymbol{x} - \boldsymbol{z})\boldsymbol{h}^{-1}\right)$ is the value in the point, $\boldsymbol{x}$ of the q-dimensional normal density function with mean $\boldsymbol{z}$ and covariance $\boldsymbol{h} \cdot \boldsymbol{h}^T$.
Hence it is possible to take into account correlations in data.

$$\boldsymbol{Y} \sim N(\mathbf{0}, \boldsymbol{I}),\ \boldsymbol{Y} = (\boldsymbol{X} - \boldsymbol{z})\boldsymbol{h}^{-1} \Rightarrow$$

$$\boldsymbol{X} \sim N(\boldsymbol{z}, \boldsymbol{h}\boldsymbol{h}^T)$$

Product kernel

In practice product kernels are used, i.e.

$$k_q(\boldsymbol{x}\boldsymbol{h}^{-1}) = \prod_{i=1}^{q} k(x_i/h_i)$$

where k is a kernel of order 1.
Assuming that $h_i = h, \quad i = 1, \ldots, q$ the following generalization of the one dimensional case in (1) is obtained

$$\hat{g}(\boldsymbol{x}) = \frac{\frac{1}{N}\sum_{s=1}^{N} Y_s \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N}\sum_{s=1}^{N} \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}$$

In case of a Gaussian kernel, this corresponds to a diagonal covariance matrix.

Non-parametric LS

If we define the weight

$$w_s(\boldsymbol{x}) = \frac{\prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N}\sum_{s=1}^{N}\prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}$$

then it is seen that the estimator corresponds to the local average:

$$\hat{g}(\boldsymbol{x}) = \frac{1}{N}\sum_{s=1}^{N} w_s(\boldsymbol{x}) Y_s$$

Non-parametric LS - continued

The estimate is actually a *non-parametric least squares estimate* at the point $\boldsymbol{x}$. This is recognized from the fact that the solution to the least squares problem

$$\arg\min_{\theta} \frac{1}{N} \sum_{s=1}^{N} w_s(\boldsymbol{x})(Y_s - \theta)^2$$

is given by

$$\hat{\theta}_{LS}(\boldsymbol{x}) = \frac{\sum_{s=1}^{N} w_s(\boldsymbol{x}) Y_s}{\sum_{s=1}^{N} w_s(\boldsymbol{x})}$$

Non-parametric LS - continued

This shows that at each $\boldsymbol{x}$, $\hat{g}$ is essentially a weighted LS location estimate, i.e.

$$\hat{g}(\boldsymbol{x}) = \hat{\theta}_{LS}(\boldsymbol{x}) \frac{1}{N} \sum_{s=1}^{N} w_s(\boldsymbol{x}) = \hat{\theta}_{LS}(\boldsymbol{x})$$

The kernel estimate is equivalent to a local weighted least squares estimate.

Variance of the non-parametric estimates

To assess the variance of the curve estimate at point $\boldsymbol{x}$ Härdle (1990) proposes the point-wise estimator given by

$$\hat{\sigma}^2(\boldsymbol{x}) = \frac{1}{N}\sum_{s=1}^{N} w_s(\boldsymbol{x})(Y_s - \hat{g}(\boldsymbol{X}_s))^2$$

where the weights are given as shown previously by

$$w_s(\boldsymbol{x}) = \frac{\prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N}\sum_{s=1}^{N}\prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}$$

Remembering the WLS interpretation this estimate seems reasonable.

The bandwidth

In analogy with smoothing in spectrum analysis the bandwidth h determines the smoothness of $\hat{g}$:

- If h is large the variance is small but the bias is large.
- If h is small the variance is large but the bias is small.

In the limits:
As $h \to \infty$ it is seen that $\hat{g}(\boldsymbol{x}) = \bar{Y}$.
As $h \to 0$ it is seen that $\hat{g}(\boldsymbol{x}) = Y_i$ for $\boldsymbol{x} = (X_i^{(1)}, \dots, X_i^{(q)})$ and otherwise 0.

Selection of bandwidth - Cross Validation

The ultimate goal for the selection of h is to minimize (see Härdle (1990))

$$\text{MSE1}(h) = \frac{1}{N}\sum_{i=1}^{N}[\hat{g}(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)]^2\pi(\boldsymbol{X}_i)$$

Where $\pi(\dots)$ is a weighting function which screens off some of the extreme observations. However g is unknown.
An easy solution is the 'plug-in' method, where Y_i is used as an estimate of $g(\boldsymbol{X}_i)$. Hence, the criterion is

$$\text{MSE2}(h) = \frac{1}{N}\sum_{i=1}^{N}[\hat{g}(\boldsymbol{X}_i) - Y_i]^2\pi(\boldsymbol{X}_i)$$

but it is clear that $\widehat{\text{MSE2}}(h) \to 0$ for $h \to 0$.

The “Leave one out” method

The idea is to avoid that $(\boldsymbol{X}_i)$ is used in the estimate for Y_i. For every data $(\boldsymbol{X}_i, Y_i)$ we define an estimator $\hat{g}_{(i)}$ for Y_i based on all data except $(\boldsymbol{X}_i)$.
The N estimators $\hat{g}_{(1)}, \ldots, \hat{g}_{(N)}$ (called the “leave one out” estimators) are written

$$\hat{g}_{(j)}(\boldsymbol{x}) = \frac{\frac{1}{N-1}\sum_{s\neq j} Y_s \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}{\frac{1}{N-1}\sum_{s\neq j} \prod_{i=1}^{q} k\{h^{-1}(x_i - X_s^{(i)})\}}$$

The “Leave one out” method

Now the *Cross-Validation criterion* using the “leave one out” estimates $\hat{g}_{(1)}, \dots, \hat{g}_{(N)}$ is

$$CV(h) = \frac{1}{N} \sum_{i=1}^{N} [Y_i - \hat{g}_{(i)}(\boldsymbol{X}_i)]^2 \pi(\boldsymbol{X}_i)$$

and the optimal value of h is then found by minimizing $CV(h)$.
It can be shown that under weak assumptions the estimate of the bandwidth $\hat{h}$ that is obtained when minimizing the Cross-Validation criterion is asymptotic optimal, i.e. it minimizes (1) – see Härdle (1990).

Applications of kernel estimators

Assume that a realization $X_1, \ldots, X_N$ of a stochastic process $\{X_t\}$ is available. The goal is now to use the realization to estimate the functions $g(\cdot)$ and $h(\cdot)$ in the model

$$X_t = g(X_{t-1}, \ldots, X_{t-p}) + h(X_{t-1}, \ldots, X_{t-q})\epsilon_t$$

As in Tjöstheim (1994) we shall use the notation

$$\begin{aligned} M(x_{i_1}, \ldots, x_{i_d}) &= \mathbb{E}\left[X_t | X_{t-i_1} = x_{i_1}, \ldots, X_{t-i_d} = x_{i_d}\right] \\ V(x_{i_1}, \ldots, x_{i_d}) &= \mathbb{V}\left[X_t | X_{t-i_1} = x_{i_1}, \ldots, X_{t-i_d} = x_{i_d}\right] \end{aligned}$$

One solution is to use the kernel estimator (other possibilities are splines, nearest neighbor or neural network estimates).

Applications of kernel estimators - continued

Using a product kernel we obtain

$$\hat{M}(x_{i_1},\ldots,x_{i_d}) = \frac{\frac{1}{N-i_d}\sum_{s=i_d+1}^{N} X_s \prod_{j=1}^{d} k\{h^{-1}(x_{i_j}-X_{s-i_j})\}}{\frac{1}{N-i_d+i_1}\sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^{d} k\{h^{-1}(x_{i_j}-X_{s-i_j})\}}$$

Assuming that $E(X_t)=0$ the estimator for $V(\cdot)$ is

$$\hat{V}(x_{i_1},\ldots,x_{i_d}) = \frac{\frac{1}{N-i_d}\sum_{s=i_d+1}^{N} X_s^2 \prod_{j=1}^{d} k\{h^{-1}(x_{i_j}-X_{s-i_j})\}}{\frac{1}{N-i_d+i_1}\sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^{d} k\{h^{-1}(x_{i_j}-X_{s-i_j})\}}$$

Applications of kernel estimators - continued

If $\mathbb{E}\left[X_t\right] \neq 0$ it is clear that the above estimator is changed to

$$\hat{V}(x_{i_1}, \ldots, x_{i_d}) = \frac{\frac{1}{N-i_d} \sum_{s=i_d+1}^{N} X_s^2 \prod_{j=1}^{d} k\{h^{-1}(x_{i_j} - X_{s-i_j})\}}{\frac{1}{N-i_d+i_1} \sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^{d} k\{h^{-1}(x_{i_j} - X_{s-i_j})\}} - \hat{M}^2(x_{i_1}, \ldots, x_{i_d})$$

Identification of functional relationship

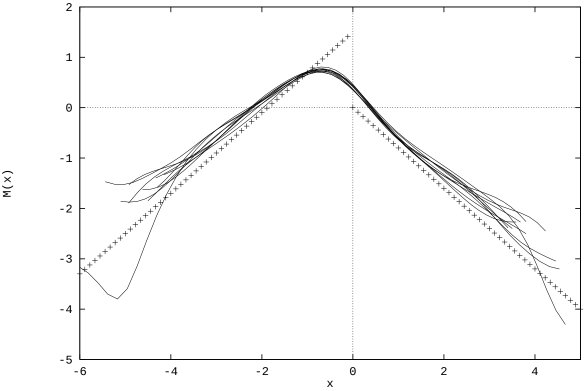
Estimation of conditional mean and variance
We have simulated 10 stochastic independent time series for both of the following non-linear models:

$$\begin{aligned} A: \quad X_t &= \begin{cases} -0.8X_{t-1} + \epsilon_t, & X_{t-1} \geq 0 \\ 0.8X_{t-1} + 1.5 + \epsilon_t, & X_{t-1} < 0 \end{cases} \\ B: \quad X_t &= \sqrt{1 + X_{t-1}^2}\epsilon_t \end{aligned}$$

What kind of models are model A and B?

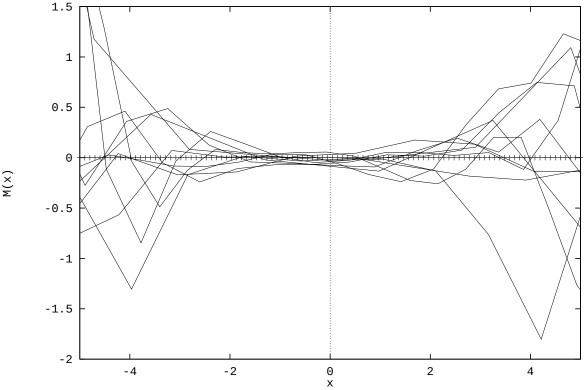
- Gaussian kernels with bandwidths, $h = 0.6$, are used.
- Simulation of 10 independent time series.

Identification of functional relationship - Estimated $M(x)$ for model A

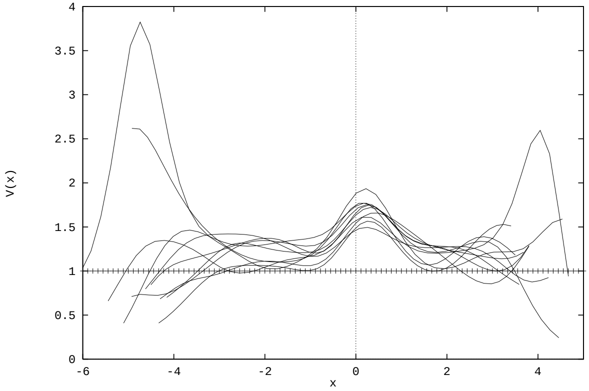


The theoretical conditional mean and variance are shown with '+'.

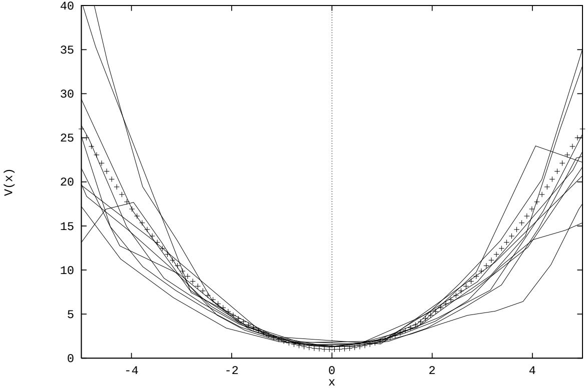
Identification of functional relationship - Estimated $M(x)$ for model B



Identification of functional relationship - Estimated $V(x)$ for model A



Applications of kernel estimators - Estimated $V(x)$ for model B



Example: non-parametric estimation of the PDF

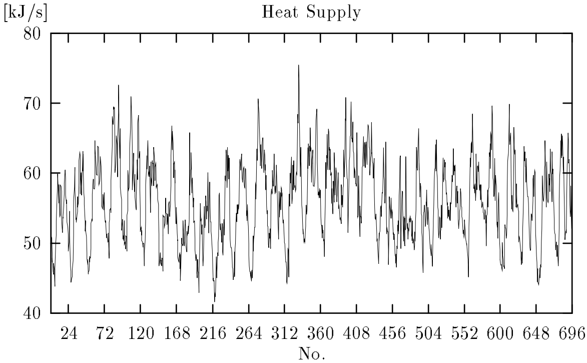
An estimate of the probability density function (PDF) is Robinson (1983) :

$$\hat{f}(x_{i_1}, \ldots, x_{i_d}) = \frac{1}{N - i_d + i_1} \sum_{s=i_d+1}^{N+i_1} \prod_{j=1}^{d} h^{-1} k\{h^{-1}(x_{i_j} - X_{s-i_j})\}$$

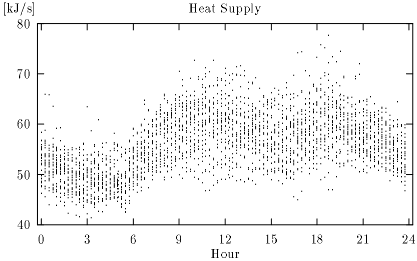
Example: Non-parametric estimation of non-stationarity

- Non-parametric methods can be applied to an identification of the structure of the existing relationships leading to proposals for a parametric model. An example of identifying the diurnal variation in a non-stationary time series is given in the following.
- Measurements of heat supply from 16 terrace houses in Kulladal, a suburb of Malmö in Sweden, are used to estimate the heat load as a function of the time of the day.

Example data
- Heat supply versus measurement number



Example data – heat supply versus time of day



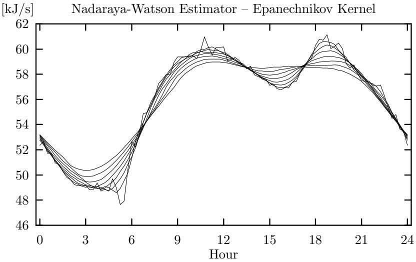
The observations are equidistantly sampled on the interval $[0, 24]$ with 0.25 hour between the sample points. The transport delay between the consumers and the measuring instruments is not more than a couple of minutes.

Dependence between data points

- Assumptions: The heat supply can be related to the time of day. There is no difference between the days for the considered period. Hence, the regression curve is a function only of the time of day, and it is this functional relationship, which will be considered.
- Thus, the assumption of independence between measurements is not fulfilled. The regression function contained in data is minor compared to a set of data with the same number of independent observations.

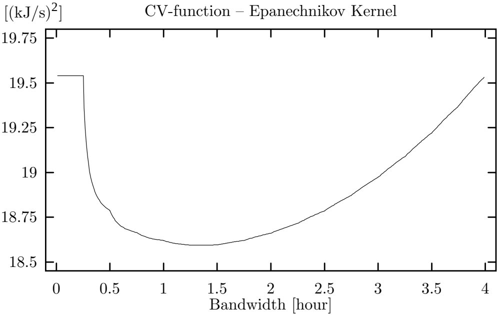
Smoothing with different bandwidths

Smoothing of the diurnal power load curve using an Epanechnikov kernel and bandwidths 0.125, 0.625,. . . , 3.625.



The characteristics of the non-smoothed averages gradually disappear, when the bandwidth is increased.

CV - approach for finding h



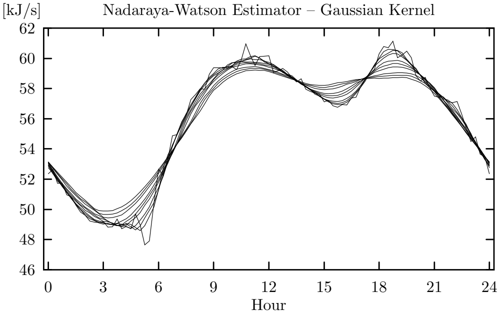
Obviously the minimum of the CV-function is obtained for a bandwidth close to 1.3. This value is most likely somewhat bigger than the visually chosen.

Cross-Validation - Comments

- The drop at 5.15 in the morning will clearly disappear in a curve estimate with this bandwidth.
- These results indicate the need of non-parametric curve estimates in which the bandwidth is not only a function of the density of the predictor variables, but for instance also of the gradient of the regression curve.

Cross-Validation – Gaussian kernels

Smoothing of the diurnal power load curve using a Gaussian kernel and bandwidths 0.025, 0.125,. . . , 0.925.

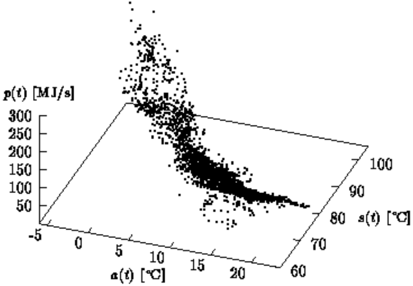


Cross-validation – Gaussian kernels

Clearly it is difficult visually to observe any difference between the curve estimates obtained with the Epanechnikov and Gaussian kernel. **A general conclusion :** The choice of the kernel (window) is not important as long as it is reasonably smooth. *It is the choice of the bandwidth that matters*.

Example: Non-parametric estimation of (static) dependence on external variables

The dependent variable is the heat load in a district heating system, and the independent variables are ambient and supply temperature. Data are collected in the district heating system in Esbjerg, Denmark, during the period August 14 to December 10, 1989.



Heat load $p(t)$ versus ambient air temperature $a(t)$ and supply temperature $s(t)$.

Applying the Epanechnikov kernel

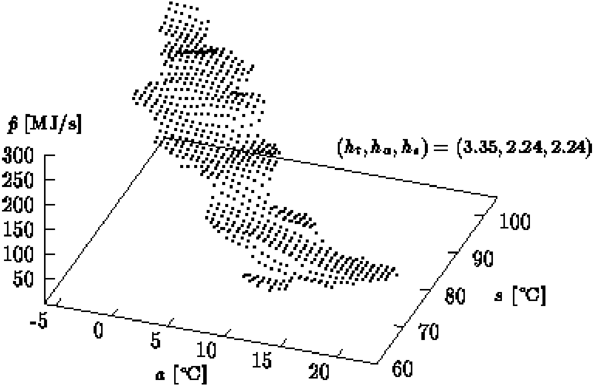
For the estimation of the regression surface the kernel method is applied using the Epanechnikov kernel (1). A product kernel is used, i.e. the power load estimate, p, at time t, at ambient air temperature a and for supply temperature s is estimated as

$$\hat{p}_{h_t,h_a,h_s}(t,a,s) = \frac{\sum_{i=1009}^{3843} p_i K_{h_t}(t-t_i) K_{h_a}(a-a_i) K_{h_s}(s-s_i)}{\sum_{i=1009}^{3843} K_{h_t}(t-t_i) K_{h_a}(a-a_i) K_{h_s}(s-s_i)}$$

where $K_h(u) = h^{-1}k(u/h)$. The bandwidths were chosen using a combination of Cross-Validation and visual inspection.

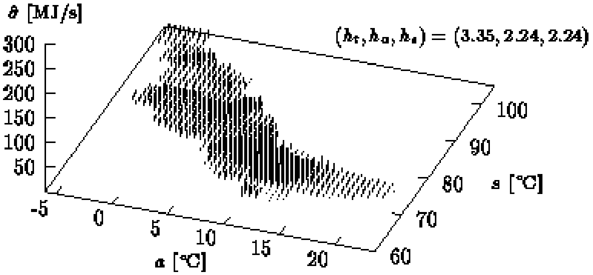
The Epanechnikov kernel estimate

Kernel estimate of the dependence on time of day, t_d, ambient air temperature, a and supply temperature, s, shown at $t_d = 7$.



The Epanechnikov kernel estimate – Estimate of the standard deviation

Point-wise estimate of standard deviation of the surface estimates.



Härdle, W. (1990), *Applied nonparametric regression*, Cambridge University Press, New York.

Madsen, H. (2008), *Time Series Analysis*, 1. edn, Chapman & Hall/CRC.

Robinson, P. (1983), ‘Nonparametric estimators for time series’, *Journal of time series analysis* **4**(3), 185–207.

Tjöstheim, D. (1994), ‘Non-linear time series: A selective review’, *Scandinavian Journal of Statistics* **21**, 97–130.