Exponential Smoothing and Kernel Regression

Marie Forbelská

Masaryk University Brno, Department of Mathematics and Statistics

Podlesí, 11.9.–13. 9. 2012

Content

Characteristics of Time Series

Time Series

- A time series is a sequence of random variables

$$\{Y_t, t = 0, \pm 1, \pm 2, \ldots\}$$

Stationarity

- We introduce weak stationarity which require that time series exhibit certain time-invariant behavior.

Dependence

- The dependence in the data marks the fundamental difference between time series analysis and classical statistical analysis.
- Different measures are employed to describe the dependence at different levels to suit various practical needs.

Autocovariance and Autocorelation

Autocovariance

- Autocovariance $C_Y(t,s)$ of a random process $\{Y_t, t \in \mathbb{Z}\}$ is defined as the covariance of Y_t and Y_s:
$$C_Y(t,s) = E(Y_t - EY_t)(Y_s - EY_s)$$
- In particular, when $t = s$, we have
$$C_Y(t,t) = E(Y_t - EY_t)^2 = DY_t$$

Autocorelation

- Autocorrelation coefficient is defined as
$$R_Y(t,s) = \frac{C_Y(t,s)}{\sqrt{DY_t DY_s}}$$

Weak Stationarity

Definition

- A time series $\{Y_t, t \in \mathbb{Z}\}$ is (weak) stationary if $EY_t < \infty$ for each t, and
  - (i) $EY_t = \mu$ is a constant, independent of t, and
  - (ii) $C_Y(t, t+k)$ is independent of t for each k.

Notation

- If $\{Y_t, t \in \mathbb{Z}\}$ is (weak) stationary denote by

$$\begin{aligned} \gamma_Y(k) &= C_Y(t, t+k) \\ \rho_Y(k) &= R_Y(t, t+k) \end{aligned} \quad \text{for all } t.$$

Local Regression

Local regression or loess is used to model a relation between a predictor variable and response variable. It is an approach to fitting curves and surfaces to data by smoothing.

Local nature of the method

It is called local since the fit at a generic point x_0 is the value of a parametric function fitted only to those observations that are close to x_0.

To keep things simple we will consider the fixed design model. We assume a model of the form

$$Y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $m(x)$ is an unknown function and ε_i an error term, representing random errors in the observations or variability from sources not included in the x_i. We assume the errors ε_i are IID with mean 0 and finite variance $D\varepsilon_i = \sigma^2$.

Local Polynomial Approximation

We make no global assumptions about the function $m(x)$ but assume that locally it can be well approximated with a member of a simple class of parametric function, e.g. a constant or straight line.

Taylor's theorem

says that any continuous function can be approximated with polynomial. If the $(p+1)$th derivative of $m(x)$ at the point x_0 exists, we can approximate $m(x)$ locally by a polynomial of order p:

$$m(x) = \underbrace{m(x_0)}_{\beta_0(x_0)} + \underbrace{m'(x_0)}_{\beta_1(x_0)}(x - x_0) + \ldots + \underbrace{\frac{m^{(p)}(x_0)}{p!}}_{\beta_p(x_0)}(x - x_0)^p,$$

for x in a neighbourhood of x_0.

Local Polynomial Regression

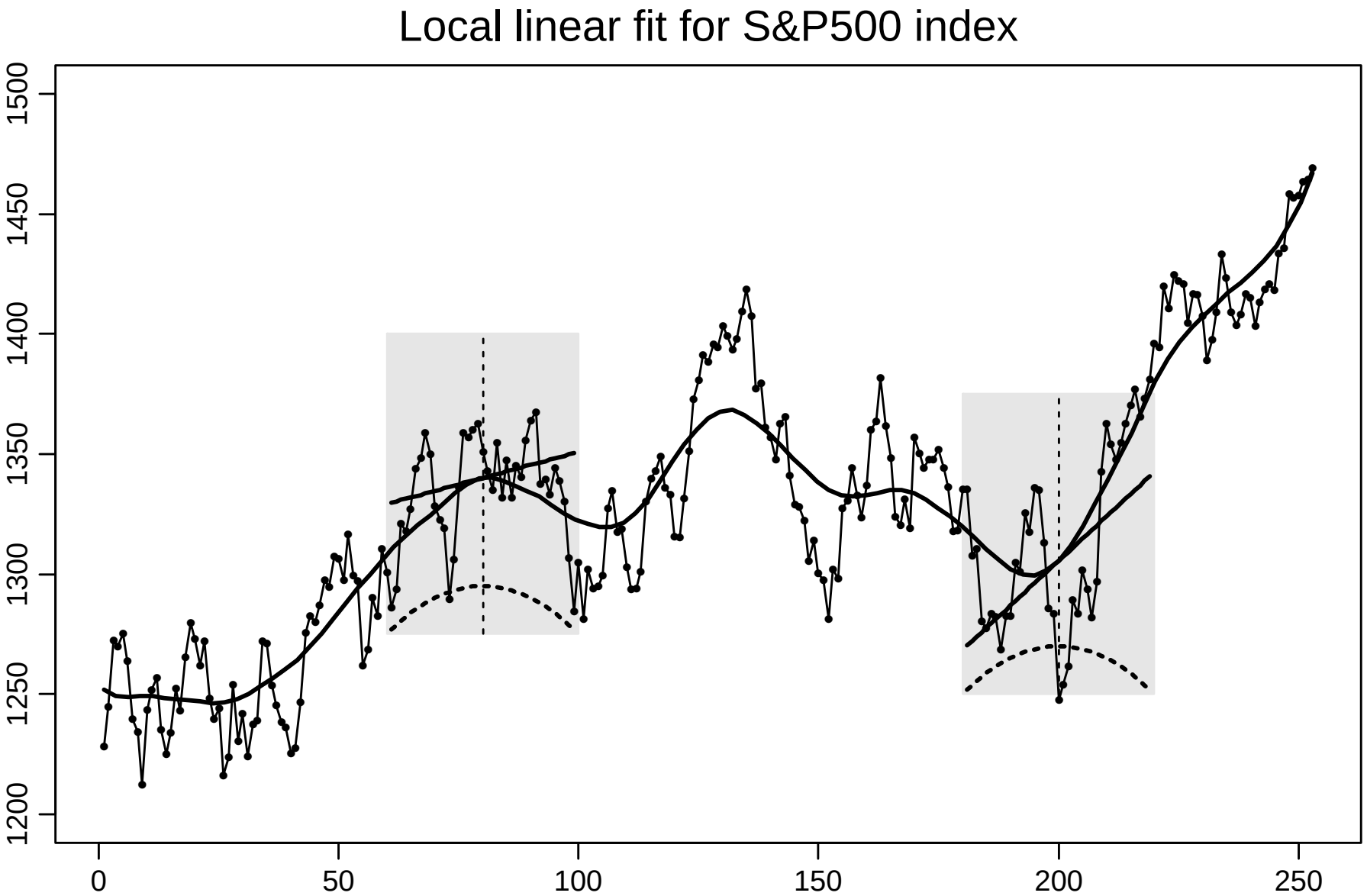
Almost always, we will want to incorporate a weight function, $w(x)$, that gives greater weight to the x_i in the neighbourhood that are close to generic point x_0 and lesser weight to those that are further.
The criterion of estimation depends on the assumption made about the distribution of the Y_i.
For example, if we suppose that the Y_i are approximately Gaussian with constant variance then it makes sense to base estimation on least-squares.

Least-squares criterion

$$\sum_{i=1}^{n} w\left(\frac{x_i - x_0}{h}\right)\left[Y_i - \sum_{j=0}^{p}\beta_j(x_0)(x_i - x_0)^j\right]^2$$

where the span or window size h (also called bandwidth) controls the "smoothness".

Local Linear Regression



Matrix Notation

It is more convenient to write the above least squares problem in matrix notation.

- Denote by $\mathbf{W}_h(x_0)$ diagonal matrix

$$\mathbf{W}_h(x_0) = diag\left\{w\left(\frac{x_1-x_0}{h}\right), \ldots, w\left(\frac{x_n-x_0}{h}\right)\right\}$$

- Let

$$\mathbf{X}_p(x_0) = \begin{pmatrix} 1 & x_1 - x_0 & \cdots & (x_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_0 & \cdots & (x_n - x_0)^p \end{pmatrix}$$

- $\boldsymbol{\beta}(x_0) = (\beta_0(x_0), \ldots, \beta_p(x_0))^{\mathrm{T}}$
- $\mathbf{e}_j = (0, \ldots, 0, \underbrace{1}_{j\text{th element}}, 0, \ldots, 0)^{\mathrm{T}}$.

Locally Weighted Least Squares

Then, for the locally weighted least squares problem we can write

LWLS criterion

$$\widehat{\beta(x_0)} = \arg\min_{\beta(x_0)} [\mathbf{Y} - \mathbf{X}_p(x_0)\beta(x_0)]^{\mathrm{T}} \mathbf{W}_h(x_0) [\mathbf{Y} - \mathbf{X}_p(x_0)\beta(x_0)]$$

Weighted least squares theory provides the solution

LWLS estimates

$$\begin{aligned} \widehat{\beta(x_0)} &= [\mathbf{X}_p(x_0)^{\mathrm{T}} \mathbf{W}_h(x_0) \mathbf{X}_p(x_0)]^{-1} \mathbf{X}_p(x_0)^{\mathrm{T}} \mathbf{W}_h(x_0) \mathbf{Y} \\ \widehat{m(x_0)} &= \mathbf{e}_1^{\mathrm{T}} \widehat{\beta(x_0)} \end{aligned}$$

if matrix $[\mathbf{X}_p(x_0)^{\mathrm{T}} \mathbf{W}_h(x_0) \mathbf{X}_p(x_0)]^{-1}$ is regular.

Modelling the Data

The use local regression in practice, we must choose

- the weight function $w(x)$,
- the bandwidth h,
- the parametric family of $m(x)$,
- and the fitting criterion.

The first three choices depend on assumptions we make about the behaviour of $m(x)$. The fourth choice depends on the assumptions we make about other aspects of the distribution of the Y_i.

Kernel Regression and Local Least Squares Regression

- Estimator

$$\widehat{m(x_0)} = \mathbf{e}_1^{\mathrm{T}} \widehat{\boldsymbol{\beta}(x_0)} = \mathbf{e}_1^{\mathrm{T}} \left[\mathbf{X}_p(x_0)^{\mathrm{T}} \mathbf{W}_h(x_0) \mathbf{X}_p(x_0)\right]^{-1} \mathbf{X}_p(x_0)^{\mathrm{T}} \mathbf{W}_h(x_0) \mathbf{Y}$$

  is just one member of a hierarchical class of local least squares kernel estimators since one may choose to fit locally polynomials of arbitrary order.

- This class includes the Nadaraya-Watson kernel estimator

$$\widehat{m(x_0)} = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)}$$

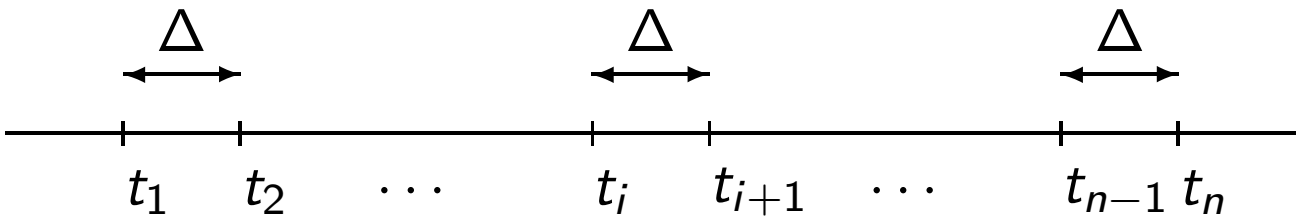
  which corresponds to local constant fits, i.e $p = 0$.

Exponential Smoothing – Introduction

- The formulation of exponential smoothing forecasting methods arose in the 1950's from the original work of Brown (1959, 1962) and Holt (1960) who were working on creating forecasting models for inventory control systems.
- Exponential smoothing can be viewed as a special type of local polynomial regression procedure where the fitting at a particular location uses only data to the left of that location.
- In fact the simple exponential smoothing (also called EWMA – Exponentially Weighted Moving Average) is virtually identical to the Nadaray-Watson kernel estimator with a kernel function that is zero in its positive arguments, something which we refer to as a "half–kernel".

Fixed design for time series

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ be a time series observed at equally-spaced time points $t_1, \ldots, t_n$. We consider the problem of using these data to forecast Y_{n+1} at time t_{n+1}.



Denote by $\Delta = t_{i+1} - t_i$. Then $\begin{aligned} t_i &= t_1 + (i-1)\Delta \\ i &= \frac{t_i - t_1}{\Delta} + 1 \end{aligned}$

Without loss of generality, we can therefore assume that $t_i = i$.

Polynomial approximation for fixed time design

Again assuming Taylor expansion of regression function $m(x)$

$$m(x) = \underbrace{m(x_0)}_{a_0(x_0)} + \underbrace{m'(x_0)}_{a_1(x_0)}(x - x_0) + \ldots + \underbrace{\frac{m^{(p)}(x_0)}{p!}}_{\frac{a_p(x_0)}{p!}}(x - x_0)^p$$

Denote by $\tau = x - x_0 = t - t_0$. In this case, time series model

$$Y_{t_0+\tau} = \sum_{k=0}^{p} \frac{a_k(t_0)}{k!}\tau^k + \varepsilon_{t_0+\tau}$$

and classical exponential smoothing of order p is based on minimization

Locally Weighted Least Squares

$$\sum_{j=0}^{\infty}(1-\alpha)^\tau \left[Y_{t_0-\tau} - \sum_{k=0}^{p} \frac{a_k(t_0)}{k!}(-\tau)^k\right]^2 \quad \text{for} \quad 0 < \alpha < 1.$$

Simple Exponential Smoothing

It has been shown that if the model is a constant ($p = 0$)

EWMA (Exponential weighted moving average) model

$$Y_{t_0+\tau} = a_0(t_0) + \varepsilon_{t_0+\tau},$$

and the smoothing process is based on minimization of the weighted least squares

$$\sum_{\tau=0}^{\infty}(1-\alpha)^{\tau}\left[Y_{t_0-\tau} - a_0(t_0)\right]^2$$

then fitted value at point t is given by means the recurrence formula

Solution

$$\widehat{Y}_t = \widehat{a_0(t)} = \sum_{\tau=0}^{\infty}\alpha(1-\alpha)^{\tau}Y_{t-\tau}$$

One Step Ahead Prediction

The attraction of exponential weighting is that estimates can be updated by a simple recursion, that is

Recursion

$$\widehat{Y}_t = \alpha Y_t + (1-\alpha)\widehat{Y}_{t-1}.$$

Exponential smoothing can also be expressed in terms of the one step ahead prediction, so $\widehat{Y}_t$ is replaced by $\widehat{Y}_{t+1|t}$.
Thus the recursion can be written

Recursion by means prediction

$$\widehat{Y}_{t+1|t} = \alpha Y_t + (1-\alpha)\widehat{Y}_{t|t-1}$$

Higher Order Exponential Smoothing

Simple exponential smoothing can be represented by an operator as follows

First Order Operator

$$S_t(Y) = \alpha Y_t + (1-\alpha) S_{t-1}(Y)$$

For higher degrees of smoothing, we define the pth order operator by

Higher Order Operator

$$S_t^p(Y) = S[S_t^{p-1}(Y)] = \alpha S_t^{p-1}(Y) + (1-\alpha) S_{t-1}^p(Y)$$

with "no smoothing" as the identity operator

$$S_t^0(Y) = I_t(Y) = Y_t$$

In this case follow

Fundamental Theorem

$$\widehat{a_0(t)} = [I - (I-S)^{p+1}]_t(Y).$$

Examples

Double Exponential Smoothing ($p = 1$)

$$\begin{aligned}
\widehat{Y}_{t+\tau} &= \widehat{a_0(t)} + \widehat{a_1(t)}\tau \\
\widehat{a_0(t)} &= 2S_t(Y) - S_t^2(Y) \\
\widehat{a_1(t)} &= \tfrac{\alpha}{1-\alpha}[S_t(Y) - S_t^2(Y)]
\end{aligned}$$

Triple Exponential Smoothing ($p = 2$)

$$\begin{aligned}
\widehat{Y}_{t+\tau} &= \widehat{a_0(t)} + \widehat{a_1(t)}\tau + \tfrac{1}{2}\widehat{a_2(t)}\tau^2 \\
\widehat{a_0(t)} &= 3S_t(Y) - 3S_t^2(Y) + S_t^3(Y) \\
\widehat{a_1(t)} &= \tfrac{1}{2}\alpha(1-\alpha)^2[(6-5\alpha)S_t(Y) - 2(5-4\alpha)S_t^2(Y) + (4-3\alpha)S_t^3(Y)] \\
\widehat{a_2(t)} &= \tfrac{\alpha^2}{(1-\alpha)^2}[S_t(Y) - 2S_t^2(Y) + S_t^3(Y)]
\end{aligned}$$

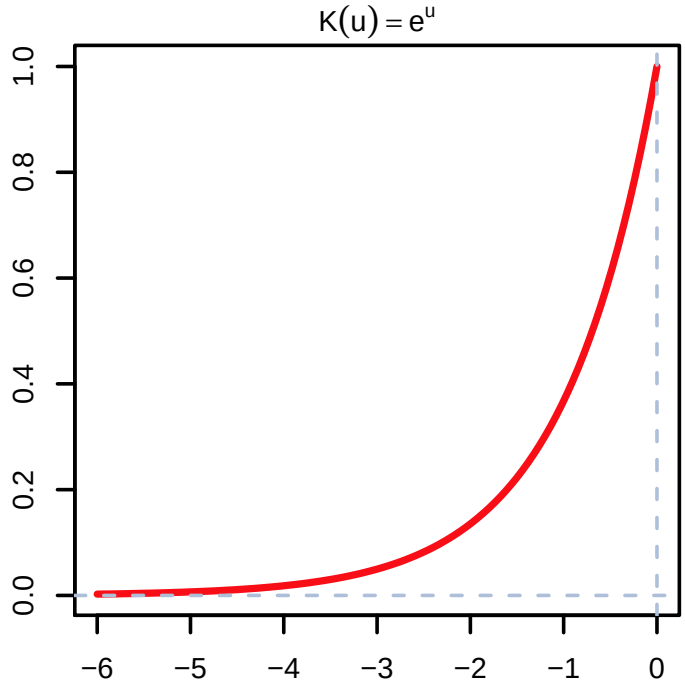
Simple Exponential Smoothing and Kernel Regression

If we define

$$h = -\frac{t_n - t_1}{(n-1)log(1-\alpha)} = -\frac{\Delta}{log(1-\alpha)}$$

and

$$K_e(u) = e^u I_{\{u \leq 0\}}$$



then it is easily shown that

$$\widehat{Y}_{n+1} = \frac{\sum_{k=1}^{n} K_e\left(\frac{t_k - t_{n+1}}{h}\right) Y_{n-k}}{\sum_{k=1}^{n} K_e\left(\frac{t_k - t_{n+1}}{h}\right)}.$$

This shows that the EWMA is equivalent to a Nadaraya-Watson, or zero-degree local polynomial, kernel estimate at point t_{n+1}.

Asymptotic theory (introduction)

Let

LWLS (Locally Weighted Least Squares) Estimate of the $m(x)$

$$\widehat{m(x)} = \mathbf{e}_1^{\mathrm{T}} \widehat{\boldsymbol{\beta}(x)} = \mathbf{e}_1^{\mathrm{T}} \left[\mathbf{X}_p(x)^{\mathrm{T}} \mathbf{W}_h(x) \mathbf{X}_p(x)\right]^{-1} \mathbf{X}_p(x)^{\mathrm{T}} \mathbf{W}_h(x) \mathbf{Y}$$

for a general kernel K with the properties $K(x) = 0$, $x > 0$, and $\int K = 1$. We shall refer to such a kernel as a half-kernel. Since K is a half-kernel, the estimate is always based on data to the left of x or at x itself. Define the half-kernel

$$K_p(u) = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_p & \mu_{p+1} & \cdots & \mu_{2p} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ u \\ \vdots \\ u^p \end{pmatrix} K(u)$$

with $\mu_j = \int u^j K(u) du$.

Assumptions

(a) $m^{(p+1)}$ is continuous and square integrable on $(0,1)$.

(b) K is square integrable and has compact support on the interval $[-\tau, 0]$ or $\tau > 0$ such that $K(0) > 0$. Also, K is $p+1$ times differentiable on its support and $K^{(p+1)}$ is Lipschity continuous.

(c) Data are available in the interval $[-h\tau, 0]$ and are used in the construction of $\widehat{m(x)}$. This condition insures that here are no left-hand boundary effects.

(d) The errors ε_t are obtained by application of a causal linear filter to independent and identically distributed random variables with mean 0 and all moments finite.

(e) The autocovariance function γ of ε_t satisfies $0 < \sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty$.

(f) The minimizer of $\sum_{t=1}^{n} (Y_t - \widehat{Y}_t)^2$ is searched on the interval $H_n = [an^{-1/(2p+3)}, bn^{-1/(2p+3)}]$ for each n, for some $b > a > 0$.

Bias and variance of the $\widehat{m(x)}$

Under assumptions given above, and assuming that

$$h = h_n \to 0 \qquad \text{and} \qquad nh \to \infty \quad \text{as} \quad n \to \infty,$$

for all

$$x \in [0, 1]$$

Bias

$$E\left\{\widehat{m(x)} - m(x)\right\} = \frac{\int u^{p+1} K_p(u) du}{(p+1)!} m^{(p+1)}(x) h^{p+1} + o\left(h^{p+1}\right)$$

Variance

$$var\left\{\widehat{m(x)}\right\} = \int K_p(u)^2 du \left\{\sum_{k=-\infty}^{\infty} \gamma(k)\right\} (nh)^{-1} + o\left\{(nh)^{-1}\right\}$$

Measures of the global error

Average of the Squared Residual

$$ASR(h) = \frac{1}{n}\sum_{t=1}^{n}\left\{\widehat{m(x_t)} - Y_t\right\}^2$$

Average Squared Error

$$ASE(h) = \frac{1}{n}\sum_{t=1}^{n}\left\{\widehat{m(x_t)} - m(x_t)\right\}^2$$

In nonparametric regression it is convenient to work with

$$MASE(h) = E\left\{ASE(h)\right\}$$

The optimal bandwidths

The optimal bandwidths (under the respective measures) are

$$\widehat{h}_{MASE} = \left\{ \frac{V_p}{(2p+2)B_p^2} \right\}^{1/(2p+3)} n^{-1/(2p+3)}$$

$$\widehat{h}_{ASR} = \left\{ \frac{V_p - 2K_p(0)\sum_{k=1}^{\infty}\gamma(k)}{(2p+2)B_p^2} \right\}^{1/(2p+3)} n^{-1/(2p+3)}$$

where

$$\begin{aligned} V_P &= \int K_p(u)^2 du \left\{ \sum_{k=-\infty}^{\infty} \gamma(k) \right\} \\ B_p^2 &= \left\{ \int u^{p+1} K_p(u) du/(p+1)! \right\}^2 \int m^{(p+1)}(x)^2 dx \end{aligned}$$

ASR and Crossvalidation

- For p=0 note that, because K is a half-kernel,

$$\widehat{Y}_\tau = \widehat{Y}_{\tau|\tau-1} = \frac{\sum_{k=1}^{\tau-1} K_e\left(\frac{t_k - t_\tau}{h}\right) Y_\tau}{\sum_{k=1}^{\tau-1} K_e\left(\frac{t_k - t_\tau}{h}\right)} = \frac{\sum_{k\neq\tau} K_e\left(\frac{t_k - t_\tau}{h}\right) Y_\tau}{\sum_{k\neq\tau} K_e\left(\frac{t_k - t_\tau}{h}\right)}$$

  so $\widehat{h}_{ASR}$ minimizes $\sum_{t=1}^{n}\{Y_t - \widehat{m_{-t}(x_t)}\}^2$ where $\widehat{m_{-t}(x)}$ is the same as $\widehat{m(x)}$, but based on the data with (x_t, Y_t) omitted.
- The same result can be easily shown to hold for general p.
- Therefore, $\widehat{h}_{ASR}$ is the same as cross-validation with a half-kernel.

Estimating the Correlation Function

- Usually the correlation function is unknown and must be estimated from the data.
- A simple approach:
  - compute the low-order sample autocorrelations of the residuals
  - and fit an autoregressive–moving average (ARMA) model.

References

- Altman, N. S. (1990): Kernel Smoothing of Data With Correlated Errors. *Journal of the American Statistical Association*. Vol. **85**, No. 411, pp. 749–759.
- Brabanter, K., Brabanter, J., Suykens, J. A. K. (2011): Kernel Regression in the Presence of Correlated Errors. *Journal of Machine Learning Research* 12, pp. 1955–1976.
- Brown, R. G., Meyer, R. F., D'Esopo, D. A. (1961): The Fundamental Theorem of Exponential Smoothing. *Operations Research*, Vol. 9, No. 5, pp. 673–687.
- Gijbels, I., Pope, A., Wand, M. P. (1999): Understanding exponential smoothing via kernel regression. *Journal of the Royal Statistical Society*. B, **61**, Part 1, pp. 39–50.