Business Statistics

Tommaso Proietti

DEF - Università di Roma 'Tor Vergata'

Regression and Smoothing Splines

Introduction

Let us consider the regression model with a single input X:

$$Y = f(X) + \epsilon,$$

where $f(X) = \mathsf{E}(Y|X)$ is an unknown conditional mean function.

$f(X)$ is possibly nonlinear and non-additive.

In the linear regression framework we considered the global polynomial approximation

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 \cdots + \beta_p X^p.$$

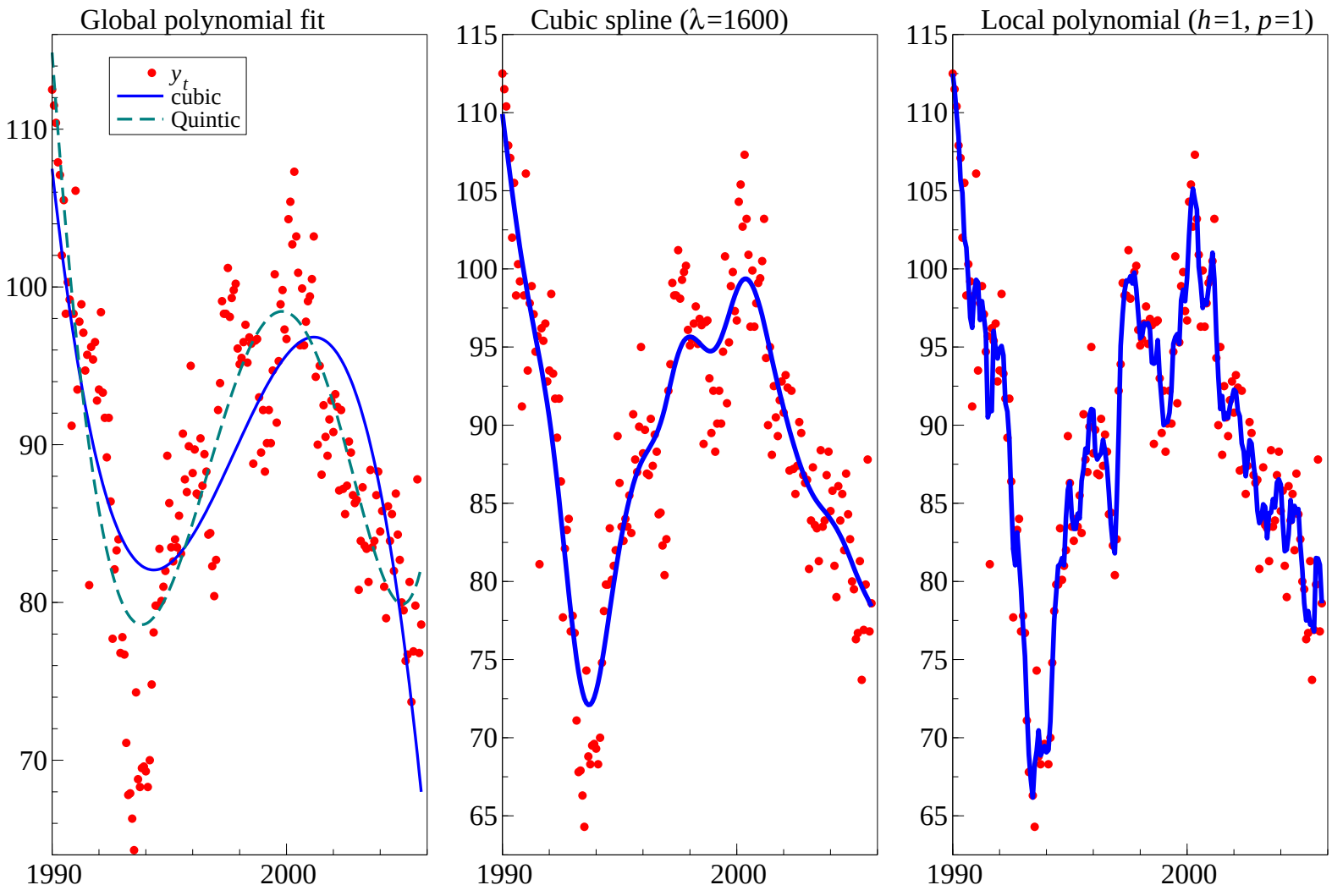
Here global means that the coefficients of the polynomial are constant across the sample span of X and it is not possible to control the influence of the individual observations on the fit.

Global polynomials are amenable to mathematical treatment, but are not very flexible: they can provide bad local approximations and behave rather weirdly at the extreme of the sample.

This point is illustrated by the first panel of figure 1, which plots the original series, representing the industrial production index for the Italian *Automotive* sector, and the estimate of the trend arising from fitting cubic and quintic polynomials of time.

In particular, it can be seen that a high order is needed to provide a reasonable fit (the cubic fit being very poor).

Figure: Industrial Production Index, Manufacture and Assembly of Motor Vehicles, seasonally adjusted, Italy, January 1990 - October 2005.



We will discuss the approximation

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

which retains linearity in the regression coefficients and uses a suitable set of transformation or functions of X, $h_m(X)$, called basis.

In the case of polynomial splines the idea is to add to a global polynomial of order p polynomial pieces at given points, called *knots*, so that the sections are joined together ensuring that certain continuity properties are fulfilled.

Given the set of points $\xi_1 < \ldots < \xi_k < \ldots \xi_K$, a polynomial spline function of degree p with K knots $\{\xi_k, k = 1, \ldots, K\}$ is a polynomial of degree p in each of the $k+1$ intervals $[\xi_k, \xi_{k+1})$, with $p-1$ continuous derivatives, whereas the p-th derivative has jumps at the knots.

The spline can be represented as follows:

$$f(X) = \beta_0 + \beta_1 X + \cdots + \beta_p X^p + \sum_{k=1}^{K} \beta_{p+k}(X - \xi_k)_+^p, \qquad (1)$$

where the set of functions

$$(X - \xi_k)_+^p = \begin{cases} (X - \xi_k)^p, & X - \xi_k \geq 0, \\ 0, & X - \xi_k < 0 \end{cases}$$

defines what is usually called the *truncated power basis* of degree p.

- According to (1) the spline is a linear combination of polynomial pieces; at each knot a new polynomial piece, starting off at zero, is added so that the derivatives at that point are continuous up to the order $p-1$.
- The truncated power representation has the advantage of representing the spline as a multivariate regression model.
- The piecewise nature of the spline "reflects the occurrence of structural change" (Poirer, 1973). The knot ξ_i is the location of a structural break. The change is "smooth", since certain continuity conditions are ensured.
- The coefficients β_k determines the size of the break.

Cubic splines and natural boundary conditions

The cubic spline model arises when $p = 3$:

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X - \xi_1)_+^3 + \cdots + \beta_{K+4} (X - \xi_K)_+^3.$$

The model has $K + 4$ parameters and the number of effective degrees of freedom is $df = 4 + K$.

A cubic spline (like any other high order spline) behaves too erratically at the boundary of the X support, i.e. the variance of the fit is very high.

Figure: Training sample (x_i, y_i) and true regression function $f(X) = [\exp(1.2X) + 1.5\sin(7X) - 1]/3$.

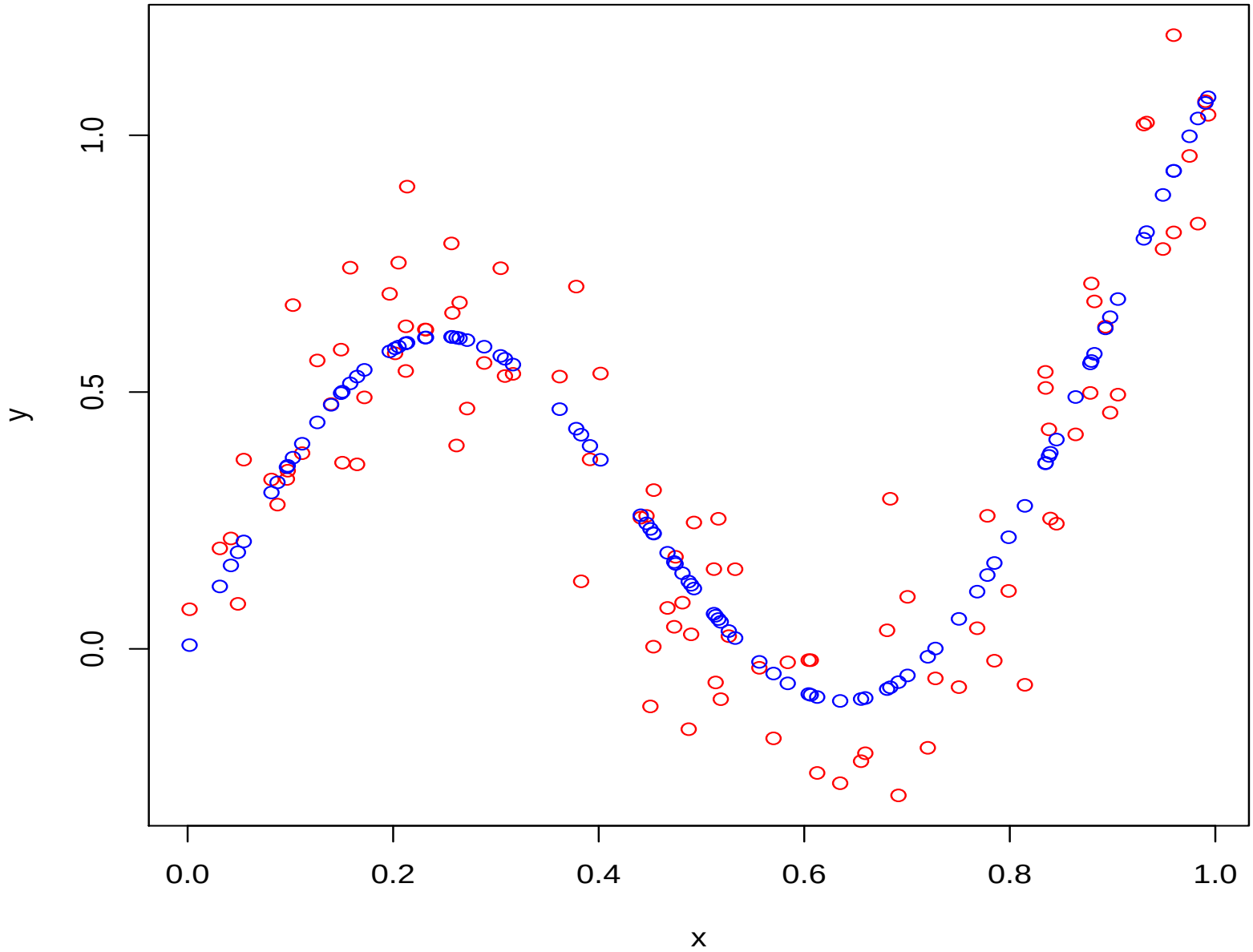


Figure: Comparison of cubic and linear spline fit with 4 internal equally spaced knots at 0.2, 0.4, 0.6, 0.8.

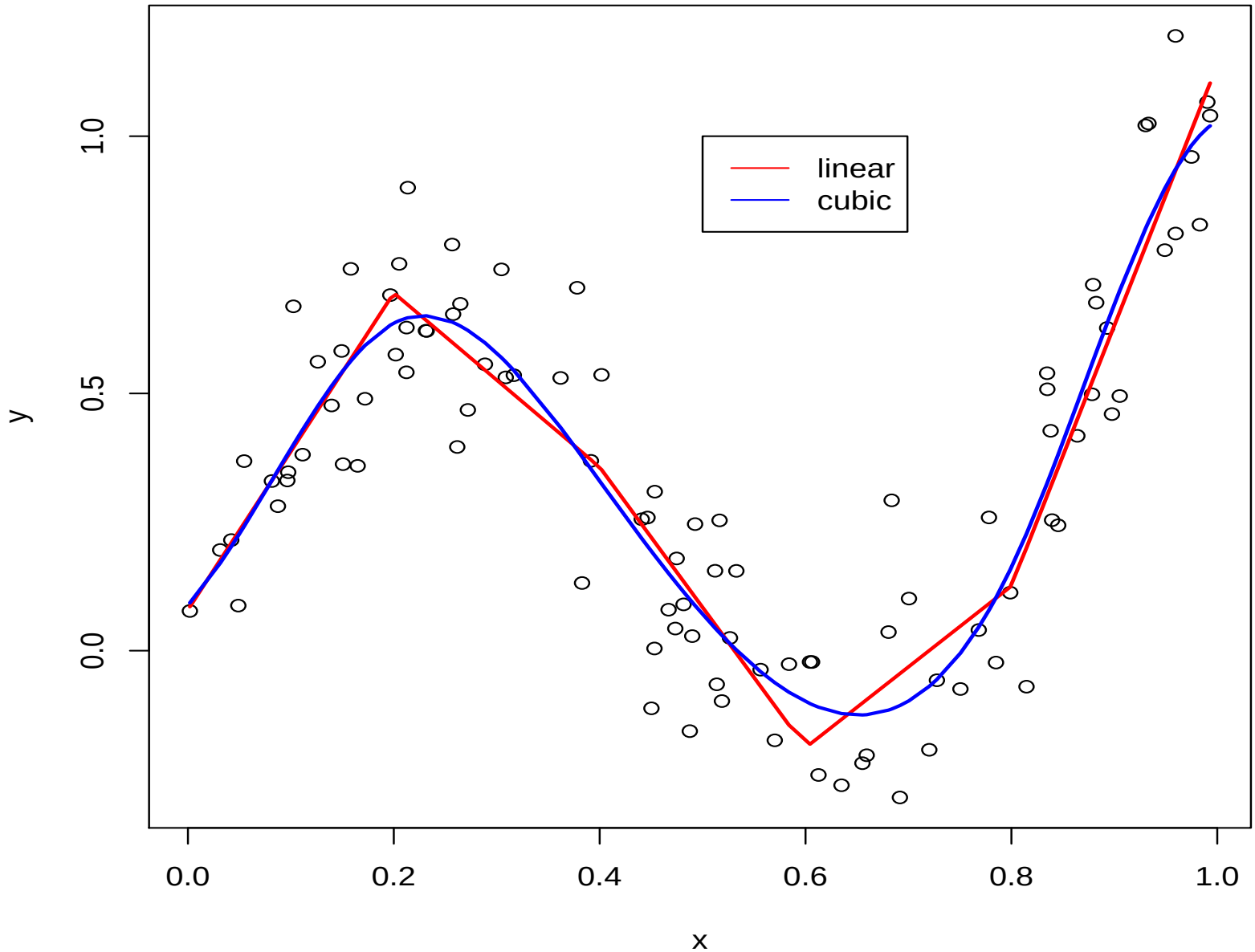
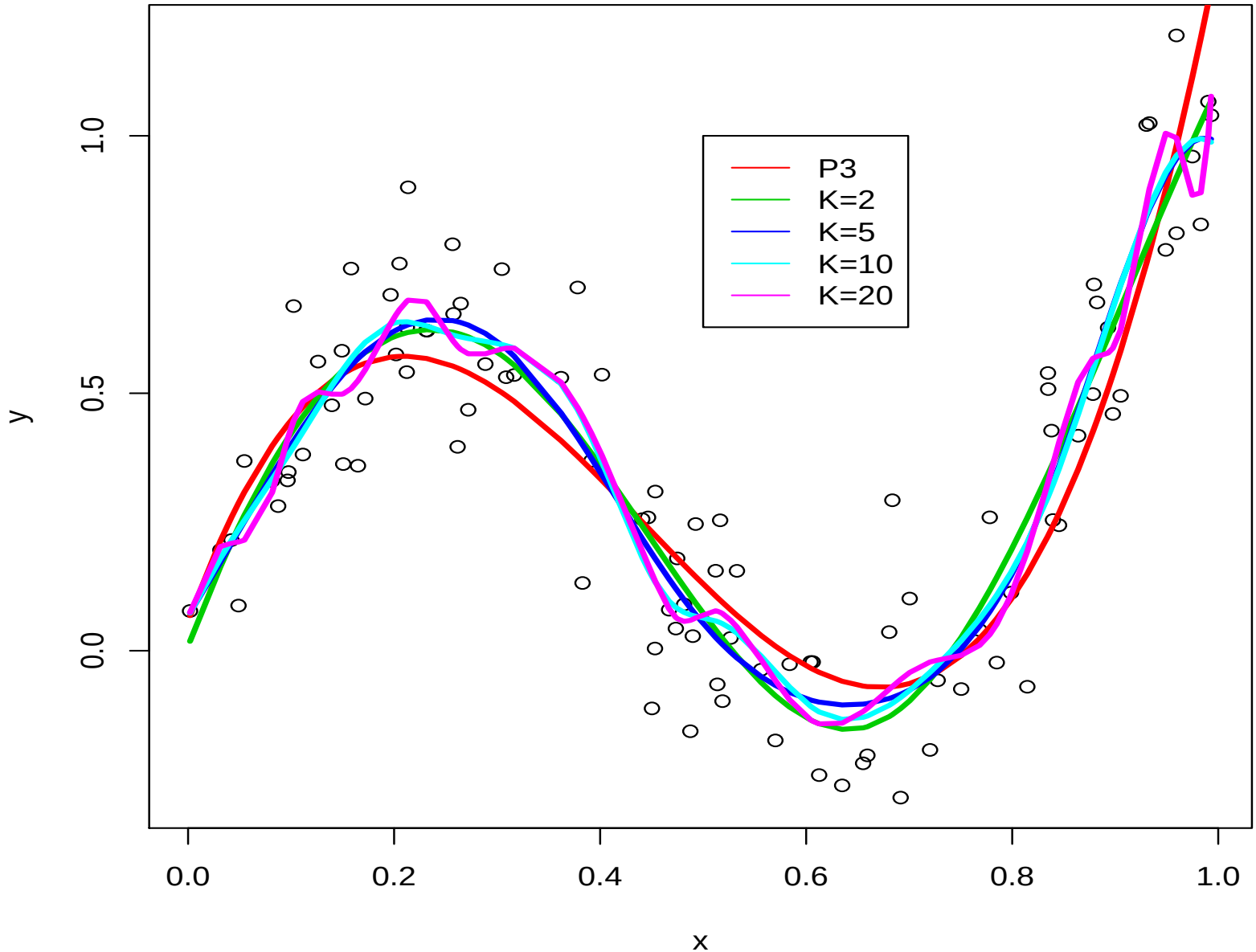


Figure: Cubic spline fit with different K and global cubic fit (knots are located automatically at quantiles of the X distribution.

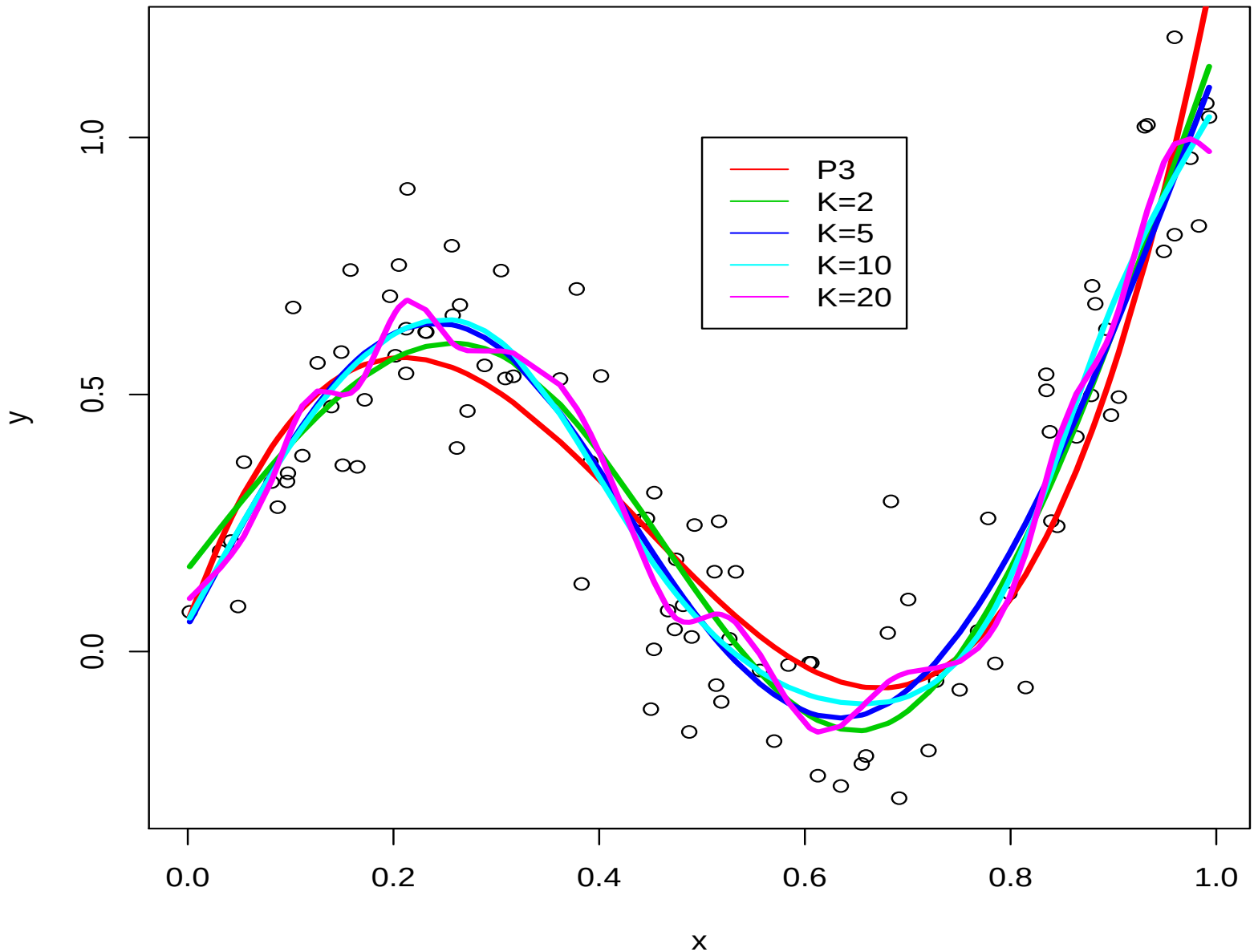


This is the reason why it is preferable to impose the so called *natural boundary conditions*, which constrain the spline be linear outside the boundary knots.

The natural boundary conditions require that the second and the third derivatives are zero for $x \leq x_{\min}$ and $x \geq x_{max}$. This amounts to imposing 4 restrictions, and this frees 4 df.

The complexity of the spline model is measured by the degrees of freedom, which is the trace of the hat matrix. This is equal to $p + 1 + K$ for polynomial splines and K for a natural cubic spline.

Figure: Natural cubic splines. Comparison of fit with different K and global cubic fit.



Model selection: how many knots?

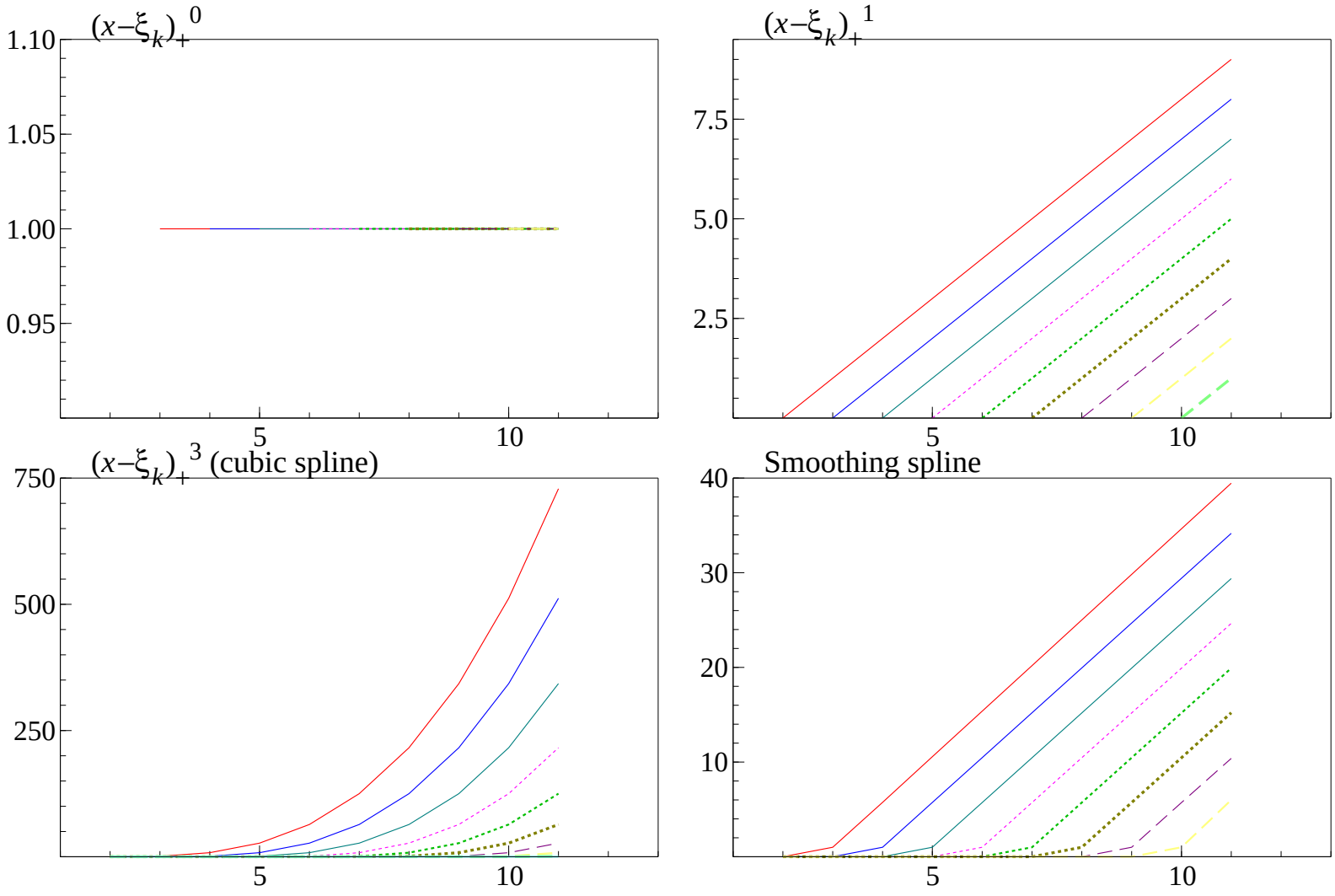
Model selection is carried out by the same methods considered for regression. It entails not only the selection of the number of knots, K, but also their location along the support of X.

The automatic option is to locate them at the $100k/(K+1), k = 1, \ldots, K$-th percentiles of the distribution of X. If there are K candidate knots, there are 2^K possible models to select. Stepwise selection has been proposed. An alternative is to use a regularization approach, i.e. smoothing splines.

Model complexity is measured by the degrees of freedom, e.g. $df = \text{trace}(\mathbf{H}) = p + 1 + K$ for a polynomial spline.

Note: the truncated power basis is easily interpretable; however, for computational efficiency the a linear transformation of this basis, the B-spline basis is used for estimation (the regressors are less collinear).

Figure: Truncated power basis for polynomial spline models.



Smoothing splines

A smoothing spline is a natural cubic spline with N knots placed at each observation $x_i, i = 1, \ldots, N$. Hence each new observation carries "news". Obviously, such a model is overparameterized as the number of parameters is N, i.e. there is one parameter for each observation.

Hence, we choose $f(X) = \sum_{i=1}^{N} \theta_i N_i(X)$, where $N_i(X)$ are the elements of the natural spline basis corresponding to the knots x_i's. The coefficients θ_i are estimated by minimising the following penalised residual sum of squares function:

$$PRESS(\lambda) = \min \left\{ \sum_{i=1}^{N} [y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 dx \right\}, \qquad (2)$$

where $\lambda \geq 0$ is the smoothness parameter, $f''(x)$ is second derivative of the function, and $\int [f''(x)]^2 dx$ is the curvature of the function.

- The parameter λ regulates the complexity of the model.
- For $\lambda = 0$, the spline fits the observations perfectly ($y_i = \hat{f}(x_i)$).
- For $\lambda \to \infty$ we obtain the linear fit $\hat{f}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (the linear function has zero second derivative and so $\int [f''(x)]^2 dx = 0$).

In vector notation

$$PRESS(\lambda) = (\mathbf{y} - \mathbf{N}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{N}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}' \boldsymbol{\Omega} \boldsymbol{\theta},$$

where $\mathbf{N}$ is the regression matrix of the natural spline and Ω has elements $\int N_h''(x) N_k''(x) dx$.

The solution is

$$\hat{\boldsymbol{\theta}} = (\mathbf{N}'\mathbf{N} + \lambda \boldsymbol{\Omega})^{-1} \mathbf{N}'\mathbf{y}$$

(notice the analogy with ridge regression).

The estimated regression function is $\hat{\mathbf{f}} = \mathbf{H}_\lambda \mathbf{y}$, where $\mathbf{H}_\lambda = \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}'$. The effective degrees of freedom of the spline fit is

$$df(\lambda) = \mathsf{tr}(\mathbf{H}_\lambda).$$

When $\lambda \to \infty$, $df(\lambda) \to 2$ (minimal complexity), whereas as $\lambda \to 0$ $df(0) \to N$ (maximum complexity).

The estimation of λ is carried out either by minimizing an information criterion or by crossvalidation. Denoting

$$RSS(\lambda) = \sum_{i=1}^{N} [y_i - f(x_i)]^2,$$

$$AIC(\lambda) = \ln[RSS(\lambda)] + 2df(\lambda)/N.$$

Crossvalidation

We seek the value of λ which minimises

$$CV(\lambda) = \sum_{i=1}^{N} \left(\frac{y_i - \hat{f}(x_i)}{1 - h_{i,\lambda}} \right)^2$$

where $h_{i,\lambda}$ is the i-th diagonal element of $\mathbf{H}_\lambda$, or

$$GCV(\lambda) = \frac{RSS(\lambda)}{[1 - N^{-1} df(\lambda)]^2}.$$

Figure: Simulated example: $N = 100$, $f(X) = \sin[12(X + 0.2)]/(X + 0.2)$, $\epsilon \sim \mathsf{N}(0, 1)$, $Y = f(X) + \epsilon$.

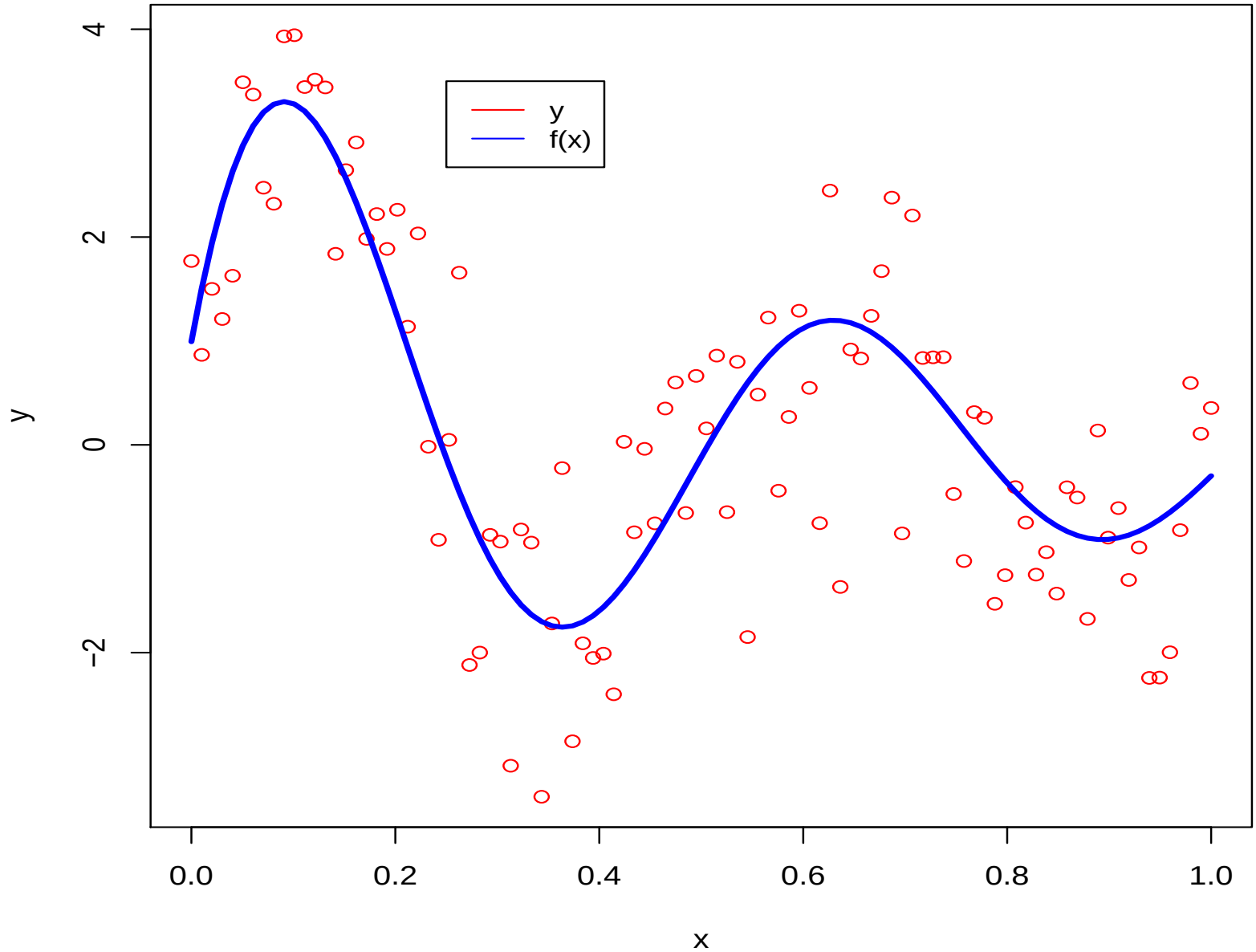
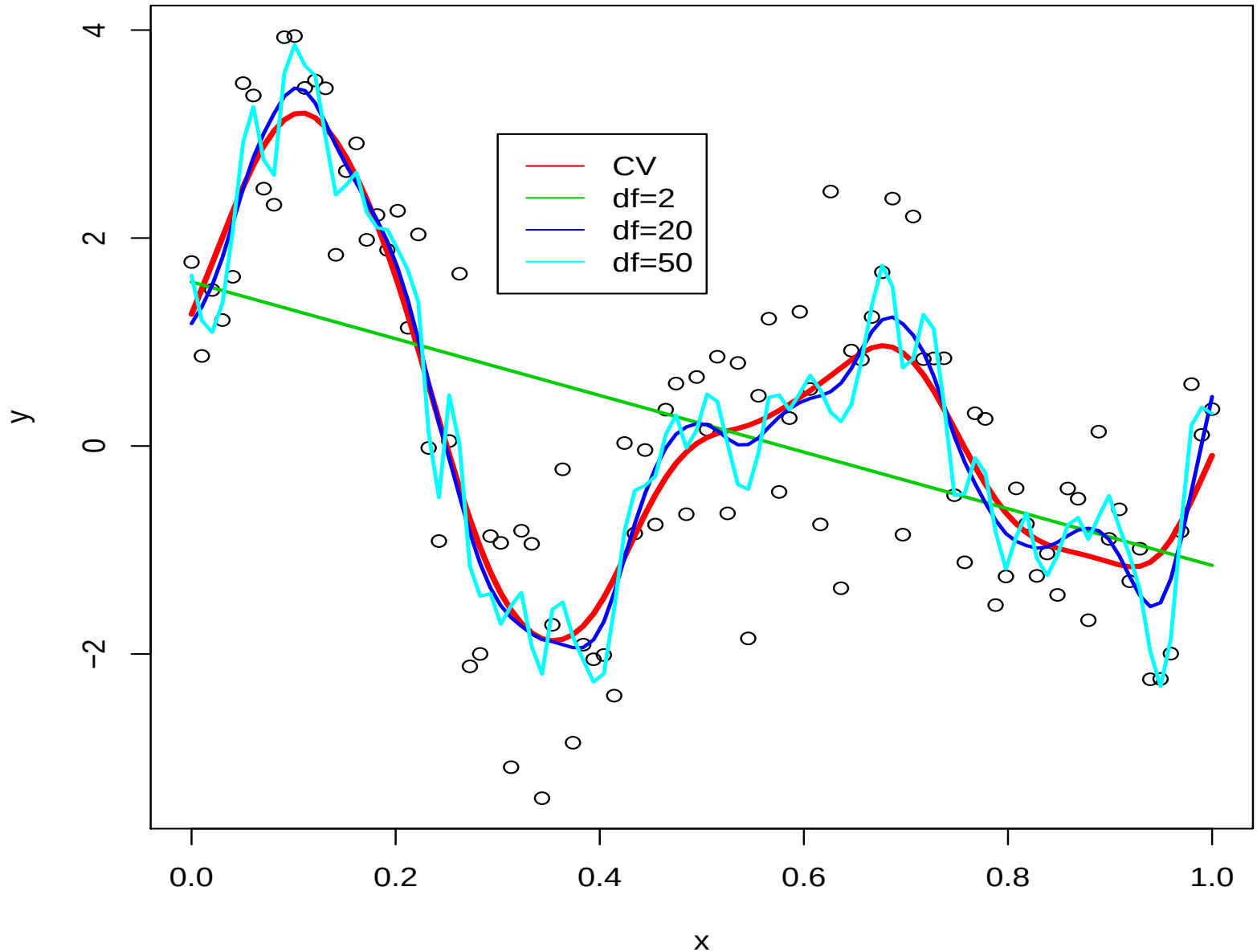


Figure: Simulated example: $N = 100$, $f(X) = \sin[12(X + 0.2)]/(X + 0.2)$, $\epsilon \sim \mathsf{N}(0, 1)$, $Y = f(X) + \epsilon$. Smoothing spline fit.



Nonparametric logistic regression

Polynomial splines can be easily adapted to logistic regression, where we set

$$\ln \frac{P(G=1|X=x)}{P(G=0|X=x)} = f(x).$$

As for spline smoothing, the function f can be estimated by minimizing the penalized log-likelihood, where the additional term penalizes the curvature of the function. Again, for large values of λ the logits are a linear function of X, whereas for small values a more complex fit is obtained.

Multiple predictors

Multidimensional splines are more difficult (curse of dimensionality). Restricted approaches, that impose additivity, like GAM (see later), are preferred.