Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Matúš Maciak

Aditivní regresní modely s regresními spliny

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Doc. Petr Volf, CSc.

Studijní program: matematika

Obor: pravděpodobnost, matematická statistika a ekonometrie

Charles University in Prague

Faculty of Mathematics and Physics

DIPLOMA THESIS

Matúš Maciak

Additive Regression Models with Regression Splines

Department of Probability and Mathematical Statistics

Supervisor: Doc. Petr Volf, CSc.

Program of study: mathematics

Branch of study: Probability, Mathematical Statistics and Econometry

Děkuji panu docentovi Petrovi Volfovi, že ochotně vedl moji diplomovou práci a profesorům Karlovy University v Praze, kteří mě zasvětili do tajů matematiky a statistiky. Chtěl bych také poděkovat svému bratru Palovi za vytvoření příjemné atmosféry pro práci a Petrovi Dolinovi, který byl ochotný ujmout se úlohy korektora textu. V neposlední řadě děkuji všem, kteří jakýmkoliv způsobem přispěli ke vzniku této práce.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 10.04.2006 Matúš Maciak

Contents

Název práce: **Aditivní regresní modely s regresními spliny**
Autor: **Matúš Maciak**
Katedra: **Katedra pravděpodobnosti a matematické statistiky**
Vedoucí diplomové práce: **Doc. RNDr. Petr Volf, CSc.**
E-mail vedoucího: **volf@utia.cas.cz**

Abstrakt: Tato práce pojednává o problematice odhadování mnohorozměrné regresní funkce založeném na neparametrickém přístupu. Zabývá se problémy spojenými s rostoucím počtem dimenzí a způsoby usilujícmi o jejich eliminaci. Vlastní odhad neznámé mnohorozměrné regresní funkce je založen na metodách splinů, přičemž práce srovnává několik základních strategií (vyhlazovací spliny, penalizační spliny, regresní spliny, B spliny a též P spliny). Hlavní část práce se věnuje nízkodimenzionální redukci modelu, zejména aditívní formě regresního modelu. Důkaz o konzistenci splinového odhadu a rychlosti konvergence je odvozen pro případ aditívních regresních modelů pro Euklidovskou L^2 normu (Ch. Stone) a také pro L^∞ normu. Uvedené jsou i sofistikovanější postupy a to hlavně metoda PPR a MARS algoritmus.

Klíčová slova: mnohorozměrná regrese, aditívní regresní modely, spliny, optimální konvergence regresních odhadů, PPR, MARS, prokletí dimenzionality, nzkodimenzionální redukce

Title: **Additive Regression Models with Regression Splines**
Author: **Matúš Maciak**
Department: **Department of Probability and Mathematical Statistics**
Supervisor: **Doc. RNDr. Petr Volf, CSc.**
Supervisor's E-mail address: **volf@utia.cas.cz**

Abstract: This thesis deals with a problem of estimating a multivariate regression function based on a nonparametric approach. A curse of dimensionality problem and different methods how to eliminate it are mentioned too. The estimate of the regression function is based on a spline approach where different spline strategies are presented (smoothing splines, splines with penalties, regression splines, B splines, P splines). The main part of diploma thesis concerns to a low dimensional reduction principle and an additive form of a regression model. A proof of consistence of spline estimates is given for L^2 norm (Ch. Stone) and also for L^∞ norm. Finally some sophisticated methods are mentioned, especially Projection Pursuit Regression and MARS Algorithm (Multivariate Adaptive Regression Splines).

Keywords: multivariate regression, additive regression models, splines, optimal rate of convergence for regression estimates, Projection Pursuit Regression, MARS, curse of dimension, low dimensional reduction principe

Est modus in rebus...

(Horatius)

Dedication:

This diploma thesis is dedicated to my parents as a thanksgiving for everything they have done for me so far.

Chapter 1

Introduction

Regression analysis:

An idea of a regression analysis was used by Francis Galton[1] for the first time in the 19^{th} century. He studied the dependence of average heights between parents and their sons. Since that time the concept of regression analysis has become very popular and a useful tool for mathematical statistics. Standard methods have been improved and new computational approaches for regression analysis have been proposed.

Regression analysis is usually posed in statistics as an optimization problem where we are attempting to find a solution to some problem while the error is at a minimum. The aim of regression analysis is to estimate the conditional expectation of a random variable Y given $\mathbf{X} = (X_1, \ldots X_J)$ on the basis of a random sample $\{(\mathbf{X}_i, Y_i),\ i = 1, \ldots N\}$. The components of $\mathbf{X}$ are called the predictor variables and the random variable Y is called the response. The dependence between variables is usually described by the given regression function. We can take many different assumptions on modelling of dependence between the variable $\mathbf{X}$ and Y. On the other hand sometimes only minimum conditions are necessary to get an ideal approximation of exploring dependence. With respect to different assumptions we take on a regression function we distinguish different approaches to the problem solution.

Let $(\mathbf{X}, Y)$ be a pair of random variables, each ranging over some space, and let f be an unknown regression function that depends on a joint distribution of variables $\mathbf{X}$ and Y. Suppose that the joint distribution of the variables $\mathbf{X}$ and Y is also unknown and consider a problem of estimating the regression function

$$f(\mathbf{x}) = \mathsf{E}\left[Y|\mathbf{X} = \mathbf{x}\right], \tag{1.1}$$

based on a random sample $\{(\mathbf{X}_i, Y_i),\ i = 1, \ldots, N\}$ from the joint distribution. According to assumptions we take on the function f, we choose an estimation technique which will be used to estimate the regression function f.

[1] Francis Galton (1822-1911) was an English explorer and anthropologist. He was the first who used a notion "*regression*". He is therefore considered as a pioneer of statistical correlation and regression.

In general we distinguish between **parametric** and **nonparametric** approaches:

- **Parametric regression estimates:**
  The parametric approach starts with an assumption of an a priori model for the function f that contains only finite number of unknown parameters. Those parameters are estimated and their estimates define the estimate of the regression function f (e.g. maximum likelihood, likelihood ratio, least squares method, minimum variance estimation and others).

- **Nonparametric regression estimates:**
  At the opposite site of the parametric approach nonparametric methods take the function f as an unrestricted function. Sometimes f is a subject to some smoothness assumptions. The estimate of f is driven directly from the data points and do not depends on parameters in global character.

In this thesis I will focus on nonparametric estimation techniques[2], especially the thesis highlights the different spline estimates in univariate and multivariate regression estimation problem.

Problem formulation:

Let $(\mathbf{X}, Y)$ be a pair of random variables such that $\mathbf{X} = (X_1, \ldots, X_J)$. Suppose that $X_j \in [0,1]$, for $1 \leq j \leq J$ and let Y is a real valued variable with mean $\mu = \mathsf{E}Y$, and a finite second moment $\mathsf{E}Y^2 < \infty$. Suppose also that the distribution of the variables $\mathbf{X}$ and Y is unknown.

Let f be the regression function of Y on $\mathbf{X}$, so that $f(\mathbf{x}) = \mathsf{E}\left[\ Y|\mathbf{X} = \mathbf{x}\ \right]$ for $\mathbf{x} \in C$ where $C := [0,1]^J$. Let $\mu = \mathsf{E}Y = \mathsf{E}f(\mathbf{X})$. Suppose from now on that f is an additive function which means f can be written in a form

$$f(x_1, \ldots, x_J) = \mu + \sum_{j=1}^{J} f_j(x_j), \tag{1.2}$$

where $\mathsf{E}f_j(X_j) = 0$, for $1 \leq j \leq J$. We assume that f is a smooth function with smooth derivatives up to some degree and f is bounded on C. Functions f_j for $1 \leq j \leq J$ are called functional components. Under the specific conditions[3] the functional components f_j are uniquely determined up to sets of measure zero and there is at most one continuous version of each such function. Even if the regression function f is not genuinely additive, an additive approximation to f may be sufficiently accurate. It will be shown that such additive function is an

[2] Nonparametric regression approaches are briefly mentioned in the next chapter.

[3] Conditions which are required to hold the statements will be mentioned in later chapters.

adequate approximation of the real regression function. Let f_j^* for $1 \leq j \leq J$, be chosen subject to the constraints $\mathsf{E} f_j^*(X_j) = 0$ for $1 \leq j \leq J$ to minimize $\mathsf{E}[f^*(\mathbf{X}) - f(\mathbf{X})]^2$, where function f^* is written as an additive combination

$$f^*(x_1, \ldots, x_J) = \mu + \sum_{j=1}^{J} f_j^*(x_j). \tag{1.3}$$

The functional components f_j^* are also uniquely determined up to sets of measure zero under the same conditions as in (1.2). In a case that function f is a genuinely additive then of course $f^* = f$.

Splines estimates:

Let $(\mathbf{X}_1, Y_1)$, $(\mathbf{X}_2, Y_2)$, $(\mathbf{X}_3, Y_3)$, $\ldots$, $(\mathbf{X}_N, Y_N)$ denote an independent random sample where each pair $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, N$ from this sample has the same distribution as $(\mathbf{X}, Y)$. Let $\mathbf{X}_i$ denotes a J-dimensional random vector $(X_{i1}, \ldots, X_{iJ})$.

We will consider different additive spline estimates $\widehat{f_N}$ of the regression function f or f^* respectively, which are based on the random sample $(\mathbf{X}_1, Y_1)$, $(\mathbf{X}_2, Y_2)$, $(\mathbf{X}_3, Y_3)$, $\ldots$, $(\mathbf{X}_N, Y_N)$. Spline estimates $\widehat{f_N}$ of the true underlying regression function f or its approximation f^* are thought as nonparametric techniques. Therefore we will discuss nonparametric regression models. Although, splines are an evolution of classical parametric interference indeed and they bridge the gap between parametric and nonparametric estimation methods. A variety of nonparametric regression models will be discussed in relation to three aspects. It is the flexibility, the dimensionality, and the interpretability of a regression model.

- **Flexibility:**
  It is an ability of the model to provide accurate fits in a wide variety of situations. Inaccuracy here leading to **bias** in estimation. The nonparametric regression methods try to provide a flexible model which could be useful for many situations. Too many assumptions decrease the model flexibility.

- **Dimensionality:**
  It can be thought of in terms of the **variance** in estimation. Problems occur in high dimensions. It is known as the **curse of dimensionality**[4]. It means that the amount of data required to avoid an unacceptably large variance increases rapidly with increasing dimensionality. Therefore, there is an inevitable trade-off needed between flexibility and dimensionality. In practice it is known as **Bias-Variance** trade-off.

[4] The curse of dimensionality problem will be closely discussed in chapter 4. The problem of Bias-Variance trade off will be also briefly mentioned there.

- **Interpretability:**
  Interpretability is an ability to simply explain the dependence of the variable Y on variables $\mathbf{X}$ given by the regression model. Usually we try to achieve easy interpretable models. However, it depends on the model dimensionality, on the form of a regression function and many other aspects. In real regression problem it is especially easy to interpret low dimensional[5] models (number of dimensions up to two - easy graphical interpretation) or models with a simple functional form (e.g. additive models).

It is necessary to know what kind of model do we want to fit. Sometimes there is not enough data to fit high-dimensional models so lower-dimensional models are automatically taking in advance. Sometimes even if there is enough data to fit a high-dimensional model and if the presence of interactions[6] is even detected, lower-dimensional models may be preferable because of greater interpretability. It is an important objective to chose the right model to achieve an optimal flexibility and an ideal trade-off between bias and variance. Optimal selection criteria will be mention later as well.

In the next chapter a brief overview on nonparametric regression problem will be given. A complex view on splines in statistics and different methods used to fit regression models will be given in chapter 3. Chapter 4 focuses on some problems arising from multivariate regression (curse of dimensionality, bias-variance trade-off). Generalization of spline estimation methods into more dimensions is formulated in chapter 5. For additive spline estimates it will be shown (chapter 6) that under some smoothness assumptions on functional components f_j, for $1 \leq j \leq J$, these estimates achieve the same rate of convergence for a general J as they do for $J = 1$. Such estimates are consistent. In chapters 7 and 8 the model selection criteria are discussed. The right positions of knots used to fit a model (chapter 7) and the optimal choice of smoothing parameter (chapter 8). More sophisticated regression strategies as Projection Pursuit Regression and Multivariate Adaptive Regression Splines algorithm are gone through in chapters 9 and 10.

This diploma thesis draws especially from the articles of Charles J. Stone (Stone [23] and Stone [24]) which highlight the optimal rate of convergence for nonparametric regression models.

[5] It is convenient to think of an arbitrary function of d real variables as being "d-dimensional." Consider a nonparametric model in which function f is defined explicitly in terms of other functions (e.g. $f_1, \ldots, f_J$), at least one of which is d-dimensional (another functions are lower dimensional). Such a model will also be thought of as being d-dimensional. More dimensional function can be thought as an interaction between variables.

[6] Once we include interactions into the model we necessarily increase the model dimensionality which makes worse the interpretability.

Chapter 2

Nonparametric Regression

The most common regression model used in parametric approach in statistics is to fit a simple parametric function defined by a set of parameters which are usually estimated by the least-squares method. The estimate of a regression function f takes a form

$$\hat{f}(\mathbf{x}) = g\left(\mathbf{x} \mid \{\hat{\pi}_t\}_{t=1}^{T}\right), \tag{2.1}$$

where parameters $\{\hat{\pi}_t\}_{t=1}^{T}$ are real valued and are given by the following equation

$$\{\hat{\pi}_t\}_{t=1}^{T} = \arg \min_{\{\pi_t\}_{t=1}^{T}} \sum_{i=1}^{N} \left[y_i - g\left(\mathbf{x}_i \mid \{\pi_t\}_{t=1}^{T}\right)\right]^2. \tag{2.2}$$

Such parametric regression model has only a limited flexibility and is likely to produce an accurate approximation only if the form of the true underlying regression function $f(\mathbf{x})$ is close to the prespecified parametric form[1] (2.1).

To overcome such limitation of the flexibility in parametric regression we try to employ nonparametric strategies. These methods are nonparametric in global character but can be parametric in local character which means that the behavior of a function can be determined by sets of parameters in small subregions. The object of nonparametric regression is to estimate the regression function f directly, rather than to estimate single parameters. Unlike parametric regression the regression function which is estimated by nonparametric approaches can be totally unrestricted (sometimes we need some smoothness assumption). Most methods of nonparametric regression implicitly assume that the regression function f is a smooth, continuous function[2]. The final shape of the regression surface is determined by data-driven techniques used to fit the model. There are three related paradigms used in nonparametric regression - piecewise and local paramet-

[1] A limitation is generally given by a set of parameters which define the general form of the function estimate. The advantage is that parametric models are easy to interpret.

[2] Continuity conditions can be also imposed on low order derivatives of a regression function.

ric methods and roughness penalty methods. Statistical methods used in nonparametric regression analysis are based on those three paradigms. The idea of a piecewise parametric fitting is to approximate the function f by several simple parametric functions (usually polynomials) where each one is defined over a different subregion of the domain. The roughness penalty methods are used to control the trade-off between bias and variance of the estimate.

Nonparametric models:

Let $(\mathbf{X}, Y)$ be a pair of random variables such that $\mathbf{X} = (X_1, \dots, X_J)$ and Y is a real valued variable. Let $(\mathbf{X}_1, Y_1)$, $(\mathbf{X}_2, Y_2) \dots (\mathbf{X}_N, Y_N)$, $N > 0$ be independent pairs of random variables each having the same distribution as $(\mathbf{X}, Y)$. The general nonparametric regression model is written in a similar manner as the parametric model

$$Y_i = f(X_1, X_2, \dots, X_J) + \epsilon_i, \qquad i = 1, 2, \dots, N, \tag{2.3}$$

where the random error ϵ_i is an independent random value with a distribution $\mathsf{N}(0, \sigma^2)$, $0 < \sigma \leq B \in \mathbb{R}$ and the function f is let unspecified. The aim of the nonparametric regression is to give a function $\hat{f}$ which is a reasonable approximation of the true underlying function f.

To measure a rationality of the approximation we define a **lack of accuracy**[3] as the integral error (2.4) or the expected error (2.5)

$$I(f, \hat{f}) = \int w(\mathbf{x}) \rho(f(\mathbf{x}), \hat{f}(\mathbf{x})) d\mathbf{x}, \tag{2.4}$$

$$E(f, \hat{f}) = \frac{1}{N} \sum_{i=1}^{N} w(\mathbf{x}_i) \rho(f(\mathbf{x}_i), \hat{f}(\mathbf{x}_i)), \tag{2.5}$$

where $\rho(.,.)$ is a measure of distance (euclidean, maximum) and $w(\mathbf{x})$ is an optional weight function. The integral error characterizes the average accuracy over the entire domain whereas the expected error reflects average accuracy only on the design points. There are many different methods[4] for nonparametric regression (Kernel estimation, M-smoothers, spline estimates, etc.) which often use different types of simple, local models in different sections of the data to build up an overall model of the data. This makes nonparametric regression a good alternative to nonlinear regression for modelling situations in which a theoretical model is not known, or is difficult to fit. Nonparametric regression models can generally be used for the same types of applications, estimation, prediction, calibration, and optimization, that traditional regression models are used for.

[3] The idea of using the Lack of Accuracy as defined above to measure the rationality of the final fit was proposed J.H. Friedman in [13].

[4] I will discuss spline estimation techniques and some adaptive methods related to splines.

Chapter 3

Splines in Statistics

3.1 Classical Spline Theory

Spline[1] approximation techniques are probably the most successful approximation methods for practical applications so far discovered. They have found a great use as approximating functions in mathematics and statistics especially because of some of their excellent properties. In the most general setting a mathematical spline can be thought as the solution to a constrained optimization problem.

Definition 1 (**Piecewise polynomial spline**)

A simple spline function is defined as a piecewise polynomial of $\mathbf{n}^{\mathbf{th}}$ ***degree, where*** $\mathbf{n} \in \mathbb{N}$***. The single pieces join in the points, so called knots and they fulfill continuity conditions for the function itself and the first*** $\mathbf{n-1}$ ***derivatives. Thus a spline function of degree*** $\mathbf{n}$ ***is a continuous function with*** $\mathbf{n-1}$ ***continuous derivatives.***

While splines are not parametric in functional form, in most cases they may be written as a linear combination of basis functions that usually have a polynomial representation. Locally they are defined by parameters. Thus there is certainly a parametric flavor. However, the set of admissible functions that may be splines has the cardinality of $\mathbb{R}^{\mathbb{R}}$ so there is an extremely rich class of all admissible functions. This is the reason why they are treated as “overparametric” or nonparametric estimates techniques. A feature of being piecewise polynomial causes that splines behavior in one region may be totally unrelated to their behavior in another region. Polynomials and most other functions have just the opposite property. Their behavior in a small region determines their behavior everywhere

[1] The original mechanical spline was a thin, flexible piece of wood (used by draftsmen) curved to desired shape and tacked down at selected points. Using of splines in regression analysis was presented in works of different authors (Svante Wold [26], Edward Wegman and Ian Wright [25], Patricia L. Smith [20] and R. L. Eubank [11]).

else. Another advantage of spline functions as piecewise polynomials with continuity conditions is that they can represent any variation of variable y with x arbitrarily well over wide intervals of a variable x. Furthermore, due to the local properties of the spline function, they are excellent tools for differentiation and integration of empirical data. Since the splines are everywhere represented by simple polynomials, they are computationally easy to handle and their integrals and derivatives are also spline function of degree higher or lower respectively.

Probably the most common choice of the spline order n and also the recommended choice by different authors (S. Wold [26], E. Wegman and I. Wright [25]) is $n = 3$. Splines of 3^{rd} degree represent nice and smooth curves in the physical world[2]. The same feature is satisfied for any $n \geq 3$ but then it is much more difficult to handle the problem. Anyway, splines of degree $n = 3$ are computationally simple and they have sufficient flexibility for most purposes.

All those properties make spline functions excellent for use in mathematical statistics and curve fitting problem. They are used to interpolate a curve through specific points in a plane (interpolation splines) or to fit a smooth curve through the data with random components (smoothing splines). Interpolation splines will be briefly mentioned in the next section however, they are not so used in statistics because they can not work with noisy data. Smoothing splines will be discussed in various modifications.

3.2 Interpolation Splines in Statistics

The simplest task regarding to splines is to interpolate data in a plane. The interpolation problem[3] means to fit a smooth curve though specific points in the plane (e.g. to fit a curve through the points $\{(x_i, y_i), i = 1, \dots N\}$). Piecewise polynomials (all of degree n) joint in the points called **knots** obeying the continuity conditions for the function itself and its first $n-1$ derivatives. To interpolate a curve through the points $\{(x_i, y_i), i = 1, \dots N\}$ a mesh $\Delta = \{\xi_i;\ 1 \leq i \leq K;\ x_1 = \xi_1 < \xi_2 < \cdots < \xi_K = x_N\}$ has to be chosen. Points ξ_i for $i \leq i \leq K$ are knots. For computational reasons the knots usually correspond to the original points $\{x_i, i = 1, \dots N\}$. The most frequently used splines are cubic splines ($n = 3$). Such interpolation curve is smooth and it obeys the continuity conditions also for the first and the second derivative.

[2] The cubic spline ($n = 3$) is a nice and smooth curve in a physical world because the human eye is skilled at picking up second and lower order discontinuities, but not higher. So it seems to be nice and smooth for human.

[3] I will focus on the one dimensional problem regarding splines estimation problem. Multivariate problems will be discussed later for smoothing splines.

Let the mesh Δ coincides with $\{x_i, i = 1, \dots N\}$ and suppose that $s_i(x)$ is a polynomial interpolation of (x_i, y_i) and (x_{i+1}, y_{i+1}) defined as:

$$s_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, \quad x \in (x_i, x_{i+1}],$$

where coefficients a_i, b_i, c_i, d_i, for $i = 1, \dots, N - 1$, are specified for each interval $(x_i, x_{i+1}]$ separately. Then the whole interpolating spline $s_\Delta(x)$ of 3^{rd} degree with the mesh Δ can be written as

$$s_\Delta(x) = \sum_{i=1}^{N-1} \mathbb{I}_{\{x \in (x_i, x_{i+1}]\}} \left[a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i\right]. \tag{3.1}$$

Coefficients a_i, b_i, c_i, d_i can be easily computed by simple system of linear equations. Let $h_i = x_{i+1} - x_i$ for $i = 1, \dots, N - 1$ and define M_i as the second derivative of $s_\Delta(x)$, where $M_i = s''_\Delta(x_i)$, for $i = 1, \dots, N$.

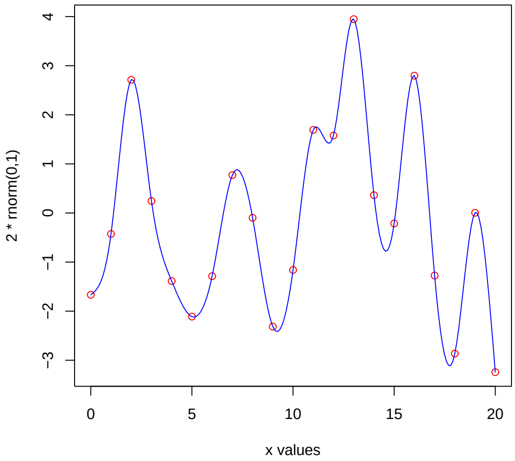


Figure 3.1: Interpolating spline of 3^{rd} degree with the mesh $\Delta = \{x_i, i = 1, \dots N\}$

By taking various derivatives of the interpolating spline and evaluating them at the knot points (M_i values), coefficients can be expressed as

$$b_i = \frac{M_i}{2},$$

$$a_i = \frac{(M_{i+1} - M_i)}{6h_i},$$

$$c_i = \frac{(y_{i+1} - y_i)}{h_i} - \frac{2(h_i M_i + h_i M_{i+1})}{6},$$

$$d_i = y_i, \;\; i = 2, 3, \dots, n - 1.$$

Fitting problem reduces to find values of M_i. With a special requirement that $M_0 = M_N = 0$ and using the continuity of the first derivative of the spline problem can be solved by finding a solution to the following equations:

$$h_{i-1}M_{i-1} + \frac{2h_{i-1}}{M_i} + h_i M_{i+1} = 6\left(\tfrac{y_{i+1} - y_i}{h_i} - \tfrac{y_i - y_{i-1}}{h_{i-1}}\right) \tag{3.2}$$

for $i = 2, 3, \dots, N - 1$. It can be solved by Gaussian elimination for example.

3.3 Smoothing Splines in Statistics

Interpolating splines as mentioned before were predicated on nonnoisy data. Therefore it was desirable to create a new type of splines (called **smoothing splines**) that could pass near in some sense to the data but not be constrained to interpolate exactly. Smoothing splines represent an important alternative to kernel regression. It was even shown that in a certain sense, spline smoothing corresponds approximately to smoothing by a kernel method with bandwidth depending on the local density of design points (B. W. Silverman [19]). Unfortunately, smoothing splines do not generally admit closed forms that make them easy to present or interpret.

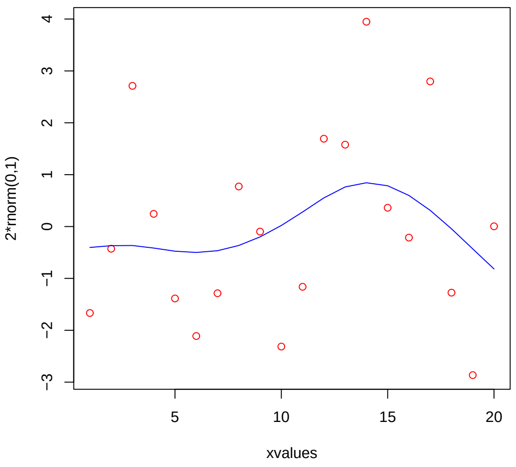


Figure 3.2: Smoothing spline of 3^{rd} degree with 10 equidistant knots and the smoothing parameter $\lambda = 0.06$. (same data as in fig. 3.1)

There are two different approaches to spline fitting methods corresponding to different points of view in dealing with the “noise” in the data.

The most frequently used method is parallel to the least squares curve-fitting procedure by minimizing a criterion that depends on a least-squares-like term plus a term penalizing roughness. This method is called **Penalized Least Squares** (S. Wold [26]).

In a case when we are able to set fairly 100% confidence limit for each data point **100 Percent Confidence Intervals** method is recommended.

3.3.1 Penalized Least Squares

It is appropriate to use the Penalized Least Squares method when error shocks have an infinite or semi-infinite support. Suppose that the x values of the data lie in a finite interval. Without any loss of generality, we can assume the interval is [0,1] and that we have $0 \leq x_1 \leq \cdots \leq x_N \leq 1$. The fitted spline is then the solution to the optimization problem

$$Minimize \quad \sum_{i=1}^{N} \left(f(x_i) - y_i\right)^2 + \lambda \int_0^1 (d^n f(x))^2 dx, \tag{3.3}$$

subject to functions $f \in W_n$ and a smoothing parameter $\lambda > 0$. The symbol W_n denote the set of functions f on interval $[0,1]$ such that $d^j f$ is absolutely continuous for $j \leq n-1$ and $d^n f$ is in L_2. Symbol d denotes a differentiation operator and L_2 is the set of measurable square integrable functions on $[0,1]$. The first term in (3.3) measures the closeness to the data while the second term penalizes the curvature in the function. The most common smoothing splines are cubic smoothing splines. They are solutions to the following optimization problem:

$$Minimize \quad \sum_{i=1}^{N} \big(f(x_i) - y_i\big)^2 + \lambda \int_0^1 (f''(x))^2 dx, \tag{3.4}$$

subject to all functions $f \in W_3$ with two continuous derivatives. The smoothing parameter λ in (3.3) and (3.4) controls the amount of smoothing. Large values of the parameter λ produce smoother curves while smaller values produce more wiggly curves. At the one extreme, as $\lambda \to \infty$, the penalty term dominates, forcing $d^n f(x) = 0$ everywhere, and thus the solution is the least-squares line, which removes not only the noise but also the signal. At the other extreme, as $\lambda \to 0$, the penalty term becomes unimportant and the solution tends to an interpolating n-differentiable function. The correct choice of the smoothing parameter λ is a sophisticated problem. Different methods are used to estimate parameter λ. They will be discussed in detail later.

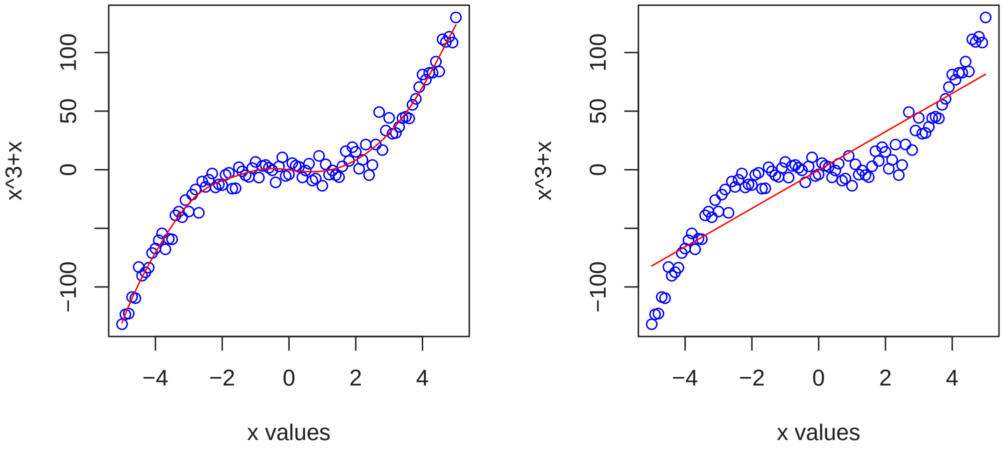


Figure 3.3: Smoothing spline of degree $n = 3$ with 10 equidistant knots and the smoothing parameter $\lambda = 0.1$ in the first case, and $\lambda \to \infty$ in the second case.

3.3.2 100 Percent Confidence Interval

An alternative approach to fit a smooth spline is 100 Percent Confidence Interval method. This method can be used if 100% confidence intervals can be set for the values $y_i,\ i = 1, \ldots, N$ (e.g. if data are taken from a calibrated instrument). Suppose that $[a_i, b_i]$ is a 100% confidence interval for ordinate at x_i, where $a_i < b_i$. Then fitting problem reduces to

$$Minimize \quad \int_0^1 (d^{\,n} f(x))^2 dx, \tag{3.5}$$

subject to $f \in W_n$ and $a_i \leq f(x) \leq b_i$. The solution to this problem is a piecewise polynomial spline of degree $2n - 1$ with knots at those data points where the constraints are active.

3.4 Regression Splines in Statistics

Regression Splines method is quite similar to smoothing splines with penalties. It can be even said that in some certain ways those two methods are even the same (see Eubank [11]). The regression splines were progressively derived from smoothing splines and the main difference is that regression splines are closely related to regression model. Moreover regression splines allow for arbitrary chosen knot points not even from the data points.

Necessary parameters:

Regression splines are also piecewise polynomial of degree n, where single pieces join smoothly and fulfill the continuity conditions for the function itself and the first $n - 1$ derivatives. They could be thought as a generalization of smoothing splines, where we omit the penalty term and we minimize the problem subject to basis coefficients.

1. The most important parameter is the degree of the regression spline. The most reasonable and the most frequently used choice is $n = 3$ which defines cubic regression splines.

2. Another parameters - the number of knots, where polynomial pieces joint together satisfying the continuity conditions up to order $n - 1$.

3. Once the number of knots is known their exact positions have to be determined - setting a mesh $\Delta = \{\xi_i,\ i = 1, \ldots, K\}$.

4. The last parameters which are required are free (basis) coefficient of the spline function. In a case of regression splines the basis coefficient are estimated.

Suppose that $u_+ = u$ if $u > 0$ and let $u_+ = 0$ otherwise. Then the general form of a polynomial spline estimate for a regression function f is following

$$s_\Delta(x_i) = \sum_{j=0}^{n} \beta_{0j} x_i^j + \sum_{k=1}^{K} \beta_{kn}(x_i - \xi_k)_+^n. \tag{3.6}$$

Spline estimate s_Δ satisfies the continuity conditions up to and including order $n-1$. Coefficients β_{0j} for $j = 0, \ldots, n$ and β_{kn} for $n = 1, \ldots, K$ are free coefficients and they can be determined by the ordinary least squares estimate method. There are $n+1+K$ coefficients which habe to be estimated. Function s_Δ as defined in (3.6) evidently has the required properties:

- Function s_Δ is a polynomial function of degree n in any subinterval $[\xi_k, \xi_{k+1})$.
- Function s_Δ has $n-1$ continuous derivatives.
- Function s_Δ has an n^{th} derivative which is a step function with jumps at K points $\xi_1, \ldots, \xi_K$.

The problem of estimating the regression function f is then the minimizing problem where we compute $s_\Delta(x)$ based on a least squares approach

$$Minimize \qquad \sum_{i=1}^{N} (s_\Delta(x_i) - y_i)^2, \tag{3.7}$$

subject to coefficients $\beta_{00}, \beta_{01}, \ldots, \beta_{0n}, \beta_{1n}, \beta_{2n}, \ldots, \beta_{Kn} \in \mathbb{R}$, where $s_\Delta(x)$ is defined by (3.6) and functions $x^0, \ldots, x^n, (x-\xi_1)_+^n, \ldots, (x-\xi_K)_+^n$ can be thought as spline basis functions.

In regression splines there is usually no smoothing parameter available[4] which would control the amount of smoothness in a regression fit. Therefore the role of knots positions is much more important than at smoothing splines. It is the only way how to fairly influence the smoothness of the regression fit. The amount of smoothness is regulated by the number and the positions of knots.

3.5 B-Splines in Statistics

Although (3.6) has a nice algebraic appeal, it is not the recommended form for estimating the regression function based on a regression splines method. For a real problem it is usually more convenient to define regression splines in terms of **B-splines**[5]. Using of B-splines brings some nice properties and it is also simpler because of computational and interpretation reasons.

[4] Some authors have proposed methods whose use a combination of the regression splines method and a penalty term (so called Regression Splines with Penalties).

[5] The idea of using B-splines in regression curve fitting problem is mentioned in de Boor [3] and Dierckx [8]

3.5.1 Classical B-splines

B-splines are defined in the similar way as regression splines as piecewise polynomials of n^{th} degree with $n-1$ continuous derivatives. They are used to define a spline basis which is necessary to estimate the unknown regression function. B-splines became favorite because of some of their nice properties.

- Each B-spline function consists of $n+1$ polynomial pieces, where number n is a degree of B-spline (the degree of each spline basis function).
- Single polynomial pieces join at n inner knots obeying the continuity conditions.
- At the joining points (knots) all derivatives up to order $n-1$ are continuous.
- The B-spline function is positive on a domain spanned by $n+2$ knots, everywhere else it is zero by definition.
- B-spline is overlap by $2n$ another polynomials (except boundary B-splines).
- At a given x in a domain, there are $n+1$ B-splines which are nonzero in x.

Let $\Delta = \xi_1, \xi_2, \dots, \xi_K$ is a given mesh of K knots. Then the classical B-spline estimate of n^{th} degree is defined as a linear combination

$$s_\Delta(x) = \sum_{k=1}^{K+n+1} \vartheta_k \cdot B_{kn}(x), \tag{3.8}$$

where $\vartheta_k \in \mathbb{R}$ for $k = 1, \dots, K+n+1$ and $B_{k,n}$ are singe B-splines of n^{th} degree whose are defined by means of divided differences as

$$B_{kn}(x) = (-1)^{n+1} \cdot \mathbb{I}_{[x \, \leq \, \xi_k]} \cdot \sum_{t=k-n-1}^{k} \Big((x-\xi_t)_+^n / \textstyle\prod_{\substack{l=k-n-1 \\ l \neq t}}^{k} (\xi_t - \xi_l) \Big). \tag{3.9}$$

We consider a mesh $\Delta = \{\xi_1, \xi_2, \dots, \xi_K\}$ but in a definition of B-splines in (3.9) we use also some additional knots. For example if we take $n = 3$, we need $K+8$ knots. So we have to define formally $2(n+1)$ additional knots. They are defined by the next equations:

$$\xi_t = \begin{cases} \xi_1 - (1-t) \cdot (\xi_1 - \min x) & for \ t \leq 0 \\ \xi_K + (t-K) \cdot (\max x - \xi_K) & for \ t \geq K+1 \end{cases} \tag{3.10}$$

B-splines are very attractive as base functions for nonparametric regression. It means single B-splines are used to set up a spline basis of functions which will be applied to estimate a regression function f. The main feature of such a basis is that any given basis spline function $B_{kn}(x), \ k = 1, \dots, K+n+1$ is nonzero only

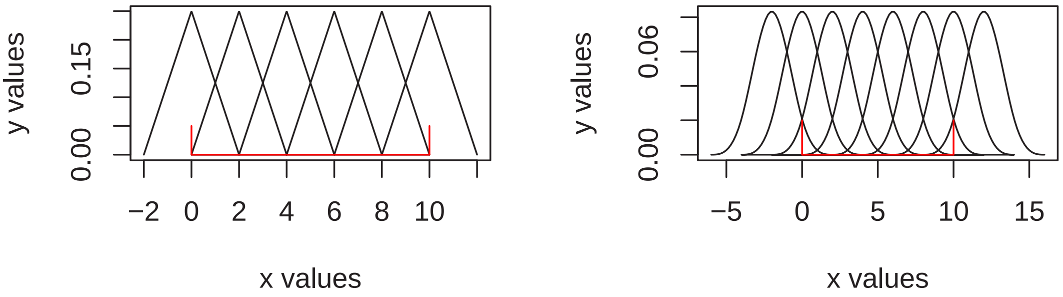


Figure 3.4: B-spline basis of 1^{st} degree (in the left part) over the interval [0,10] with four equidistant inner knots and B-spline basis of 3^{rd} degree (in the right part) over the same interval with the same inner knots.

over a span of $n+2$ distinct knots. Two different B-spline bases for the same interval with the same number of equidistant inner knots but different degree of spline basis are illustrated in the figure 3.4.

Let the $B_{kn}(x),\ k=1,\dots,K+n+1$ be the B-splines basis functions. Then the problem of estimating a regression function f based on B-splines is formulated as a following minimization problem:

$$Minimize \qquad \sum_{i=1}^{N}(s_\Delta(x_i)-y_i)^2 \tag{3.11}$$

where the spline estimate $s_\Delta(x)$ of a regression function f is defined as a simple linear combination

$$s_\Delta(x)=\sum_{k=1}^{K+n+1}\vartheta_k\cdot B_{kn}(x), \tag{3.12}$$

where $\{B_{kn}(x),\ k=1,2,\dots,K+n+1\}$ is a B-spline basis (each B-spline of n^{th} degree) and $\vartheta_k\in\mathbb{R}$ for $k=1,2,\dots,K+n+1$, are the free coefficients of a linear combination (3.12) leading to minimization of (3.11).

Hence the fitting of a spline function is a linear problem once the set of knots is specified. I treat only equidistant set of knots but B-splines work also for any arbitrary knots selection[6].

[6] Once we specify non-equidistant set of knots we have to remember that all additional knots derived from (3.10) will be **equidistant**.

3.5.2 B-splines with Penalties

Consider a problem of estimating a regression function f based on B-splines approach. Once we specify a mesh $\Delta = \{\xi_1, \xi_2, \ldots, \xi_K\}$, where K is relatively large the fitted curve will show more variation than we expect or than is justified by the data. Similar to regression splines method there is no smoothing parameter λ to control the amount of smoothness. To make the fitted curve more flexible a new method was presented by O'Sullivan [18] by including a penalizing term in the minimizing problem (3.11). The objective function for the spline minimizing problem with a penalty term can be formulated as

$$Minimize \quad \sum_{i=1}^{N} (s_\Delta(x_i) - y_i)^2 + \lambda \cdot \int_{\xi_0}^{\xi_{K+1}} (s''_\Delta(x))^2 dx, \tag{3.13}$$

subject to parameter λ and coefficients $\vartheta_k \in \mathbb{R}$ for $k = 1, 2, \ldots, K + n + 1$ where $s_\Delta(x)$ is defined by (3.12) as a linear combination of B-spline basis functions. The integral of the square of the second derivative of a fitted function $s_\Delta(x)$ in (3.13) can be thought as smoothness penalty. There is nothing special about the second derivative in the penalty term so in fact lower and or higher orders of derivative may be used as well.

In (3.13) we use the second derivative of $s_\Delta(x)$. A simple formula for derivatives of B-splines in the linear combination (3.12) was given by De Boor [3]

$$s'_\Delta(x) = \sum_{k=1}^{K+n+1} \vartheta_k \cdot B'_{kn}(x) \tag{3.14}$$

where

$$\sum_{k=1}^{K+n+1} \vartheta_k \cdot B'_{kn}(x) = \sum_{k=1}^{K+n+1} \vartheta_k \cdot B'_{k\ n-1}(x) - \sum_{k=1}^{K+n} \vartheta_{k+1} \cdot B'_{k+1\ n-1}(x).$$

More general version of B-splines method with penalties can also be used. When we base the penalty term on a higher order of finite differences of the coefficients ϑ_k we can formulate the minimizing problem as

$$Minimize \quad \sum_{i=1}^{N} (s_\Delta(x_i) - y_i)^2 + \lambda \cdot \sum_{k=p+1}^{K+n+1} (\Delta^p \vartheta_k)^2, \tag{3.15}$$

subject to smoothing parameter λ and coefficients ϑ_k where we define $\Delta^p \vartheta_k$ in a recursive way as $\Delta^p \vartheta_k = \Delta^{p-1}\vartheta_k - \Delta^{p-1}\vartheta_{k-1}$. B-splines are usually known as **P-splines**[7]. Their the most important properties are inherited from classical B-splines. Smoothing parameter λ is taken by model selection criteria.

[7] New method (B-Splines method with penalties) has been proposed by Paul H. Eilers and Brian D. Marx [10]. B-splines with penalties are known as **P-splines**.

Chapter 4

Problems Arising with Multidimensional Data

In physical situations there usually arises a need to interpolate data using multiple predictors. In many cases the interpolated surface required for application is two or even more dimensional. It is theoretically simple to generalize one dimensional methods into multivariate approaches and to use them to fit a multivariate data. But there is a problem to do that in real data. A common problem that arises when fitting multivariate data is that smoothing methods are usually limited by the fact that estimating a J-variate function with no constraints on its structure, apart from smoothness, requires data sets of impractical size for larger values of J. To be more precise, number of required data to fit J dimensional plane increases exponentially with increasing number of dimensions J. This problem is referred to as the **curse of dimensionality**.

4.1 Curse of Dimensionality

The term "curse of dimensionality" (Bellman [1], Bishop [2]) generally refers to the difficulties involved in fitting models, estimating parameters, or optimizing a function in many dimensions. As the dimensionality of the input data space (i.e., the number of predictors) increases, it becomes exponentially more difficult to find global optima for the parameter space, i.e., to fit models.

Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ be independent and identically distributed (i.i.d) random variables with $\mathbf{X}$ uniformly distributed in the hypercube $[0,1]^J$. If $\mathbf{X}_i,\ i = 1, \ldots, N$ take values in a high-dimensional space (it means that $J >> 2$) it is not possible to densely cover the space of $\mathbf{X}$ (hypercube $[0,1]^J$) with finitely many sample points, even if the sample size N is very large. I will illustrate this with a simple example.

We will generate N random uniformly distributed variables in the hypercube $[0,1]^J$ for different choices of number of dimensions $(J = 1, J = 2, J = 3, J = 5, J = 10, J = 20)$. We will observe the average minimum distance between variables $\mathbf{X}_i,\ i = 1, \ldots, N$ and the random variable $\mathbf{X}$. We will take two different norms to measure the distance between two variables.

Supremum-norm:

$$\| \mathbf{x} \|_{sup} \quad = \max_{j=1,\ldots,J} |x_j| \tag{4.1}$$

Euclidean-norm:

$$\| \mathbf{x} \|_{euc}^2 \quad = \sum_{j=1,\ldots,J} x_j^2 \tag{4.2}$$

The results are shown in next tables. The size N of the random sample was chosen from 100 observations to 100000 observations and number of dimensions was chosen in a range from one to twenty. For each combination of N and the dimension J a calculation was repeated 20 times. The average minimum distance between variables $\mathbf{X}$ and $\mathbf{X}_i$ was written down into the tables:

Supremum-norm

	$J = 1$	$J = 2$	$J = 3$	$J = 5$	$J = 10$	$d = 20$
$n = 100$	0.003838	0.054951	0.094651	0.232593	0.366015	0.571504
$n = 1000$	0.000506	0.015051	0.053464	0.129761	0.273968	0.440982
$n = 10000$	0.000044	0.004691	0.021613	0.044339	0.213186	0.402223
$n = 100000$	0.000006	0.001178	0.009108	0.030709	0.159703	0.353620

Euclidean-norm

	$J = 1$	$J = 2$	$J = 3$	$J = 5$	$J = 10$	$d = 20$
$n = 100$	0.003838	0.060434	0.118966	0.328987	0.660090	1.264582
$n = 1000$	0.000506	0.017274	0.063800	0.191530	0.498007	1.000749
$n = 10000$	0.000041	0.005546	0.027003	0.060081	0.376753	0.909363
$n = 100000$	0.000005	0.001376	0.011672	0.052891	0.289131	0.795231

As we see in the tables when the number of dimensions is increased just a little then the minimum distance between variables rapidly increases and problem becomes difficult to solve. To avoid that, an unacceptably large sample is necessary to populate a high dimensional space.

We can even give an approximate lower bounds for the Supremum-norm and the Euclidean-norm as well. We will show there are numbers $B_{sup}(N, J)$ and $B_{euc}(N, J)$ such that

$$\mathsf{E}[\min_{i=1,\ldots,N} \| \mathbf{X} - \mathbf{X}_i \|_{sup}\ ] \geq B_{sup}(N, J), \tag{4.3}$$

and

$$\mathsf{E}[\min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{euc}\ ] \geq B_{euc}(N,J). \qquad (4.4)$$

At first we will show an existence of the bound $B_{sup}(N,J)$.
Set $Y = \min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{sup}$. Then the random variable Y is a non-negative random variable so for $\mathsf{E}Y$ we can write

$$\mathsf{E}Y = \int_0^\infty \mathsf{P}\left[\ \min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{sup}\ \ >\ t\ \right] dt =$$
$$= \int_0^\infty 1-\mathsf{P}\left[\ \min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{sup}\ \ \leq\ t\ \right] dt.$$

If we use a property of uniformly distributed random variables $\mathbf{X}_i$ and $\mathbf{X}$ we can bound the probability

$$\mathsf{P}\left[\ \min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{sup}\ \ \leq\ t\right]\ \leq\ N\cdot\mathsf{P}\left[\ \parallel \mathbf{X}-\mathbf{X}_1 \parallel_{sup}\ \ \leq\ t\right]\ =$$
$$=\ N\cdot\mathsf{P}\left[\ \max_{j=1,\dots,J}\left|X_j-X_{1j}\right|\leq t\right]\ \leq\ N\cdot\{\mathsf{P}\left[\left|X_1-X_{11}\right|\leq t\right]\}^J\ \leq\ N\cdot(2t)^J,$$

where $N\cdot(2t)^J \leq 1$. Therefore $t\in[0,1/2n^{1/J}]$ and we can express the lower bound for expected minimum distance in supremum norm.

$$\mathsf{E}Y \geq \int_0^{\frac{1}{2N^{1/J}}} \left(1-N\cdot(2t)^J\right) dt\ =\ \left[t-N\cdot 2^J\cdot\tfrac{t^{J+1}}{J+1}\right]_{t=0}^{\frac{1}{2N^{1/J}}}\ =\ \frac{1}{2\cdot N^{1/J}}\cdot\frac{J}{1+J}$$

By analogy we can also express the lower bound $B_{euc}(N,J)$ for the expected average minimum distance in Euclidean norm. Set $Z = \min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{euc}$. Then we can write

$$\mathsf{E}Z = \int_0^\infty \mathsf{P}\left[\ \min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{euc}\ \ >\ t\ \right] dt =$$
$$= \int_0^\infty 1-\mathsf{P}\left[\ \min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{euc}\ \ \leq\ t\ \right] dt.$$

Using the uniformly distributed variables we can bound the probability as

$$\mathsf{P}\left[\ \min_{i=1,\dots,N} \parallel \mathbf{X}-\mathbf{X}_i \parallel_{euc}\ \ \leq\ t\right]\ \leq\ N.\mathsf{P}\left[\ \parallel \mathbf{X}-\mathbf{X}_1 \parallel_{euc}\ \ \leq\ t\right]\ =$$
$$=\ N\cdot\mathsf{P}\left[\textstyle\sum_{j=1,\dots,J}(X_j-X_{1j})^2\ \leq\ t^2\right]\ \leq\ N\cdot\{\mathsf{P}\left[(X_1-X_{11})^2\leq\tfrac{t^2}{J}\right]\}^J\ =$$
$$=\ N\cdot\{\mathsf{P}\left[\left|X_1-X_{11}\right|\leq\tfrac{t}{\sqrt{J}}\right]\}^J\ \leq\ N\cdot\left(2t/\sqrt{J}\right)^J.$$

where $N \cdot \left(2t/\sqrt{J}\right)^J \leq 1$. Now we can express the lower bound for the expected average minimum distance between two uniformly distributed variables in Euclidean norm.

$$\mathsf{E}Z \geq \int_0^{\sqrt{J}/2N^{1/J}} (1 - N(2t/\sqrt{J})^J)\, dt = \left[t - \tfrac{N2^J}{J^{J/2}} \cdot \tfrac{t^{J+1}}{J+1}\right]_0^{\frac{\sqrt{J}}{2N^{1/J}}} \; = \; \frac{\sqrt{J}}{2N^{1/J}} \cdot \frac{J}{J+1}$$

The lower bounds for the expected average minimum distance between two uniformly distributed variables in d-dimensional hypercube are:

$$B_{sup}(N, J) = \frac{1}{2} \cdot \frac{1}{N^{1/J}} \cdot \frac{J}{J+1}, \qquad B_{euc}(N, J) = \frac{1}{2} \cdot \frac{\sqrt{J}}{N^{1/J}} \cdot \frac{J}{J+1} \tag{4.5}$$

The only common way to overcome the "curse of dimensionality" problem is to incorporate additional assumptions about the regression function beside the sample. Now I have to remark that problem mentioned above is not longer valid if the components of a variable $\mathbf{X}$ are not independent (e.g. there is no problem to fit a regression curve in multidimensional cube if all values $\mathbf{X}$ lie on a line in a hypercube $[0, 1]^J$).

4.2 Bias-Variance Trade-Off

I have already mentioned a relation between flexibility and dimensionality in a regression fit. These two aspects work in opposite way. In many cases the improvement of flexibility leads to the increase of dimensionality. This problem is in statistics referred to as a "Bias-Variance Trade-Off" (see Laszlo Gyorfi and coll. [15]).

Consider a regression function f such that $f(\mathbf{x}) = \mathsf{E}[Y|\mathbf{X} = \mathbf{x}]$. Let $\widehat{f}_N$ is an estimate of f based on a random sample. Than we can write the expected squared error (ESE) of $\widehat{f}_N$ at $\mathbf{x}$ as:

$$\mathsf{ESE} = \mathsf{E}\left[|\widehat{f}_N(\mathbf{x}) - f(\mathbf{x})|^2\right] \tag{4.6}$$

We can also express the expected squared error ESE as a sum of two components

$$\mathsf{ESE} = \mathsf{E}\left[|\widehat{f}_N(\mathbf{x}) - f(\mathbf{x})|^2\right] = \mathsf{E}\left[|\widehat{f}_N(\mathbf{x}) - \mathsf{E}(\widehat{f}_N(\mathbf{x}))|^2\right] + \mathsf{E}\left[|\mathsf{E}(\widehat{f}_N(\mathbf{x})) - f(\mathbf{x}))|^2\right],$$

where the first component is the variance of $\widehat{f}_N(\mathbf{x})$ and the second component is athe squared $bias(\widehat{f}_N(\mathbf{x}))$. Therefore it is called "Bias-Variance Trade-Off".

There is still an important task to choose an optimal decomposition of bias and variance. If we reduce variance too much bias increases. On the other hand if we excessively reduce the bias then variance increases as well.

Chapter 5

Multivariate Spline Regression

We have discussed only one dimensional regression problem so far and a problem of estimating one variable regression function f based on a random sample $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$ where X_i and Y_i are real valued variables. The nonparametric estimate of the univariate regression function f was based on a spline approach. However, in real regression problem we are usually forced to use multiple predictors, which means we have to expand the regression problem to more dimensions. As a consequence of such extension the regression surface becomes more dimensional. Consider a problem of estimating a regression function $f(\mathbf{x}) = f(x_1, \ldots, x_J)$ where $J > 1$ and $f(\mathbf{x}) = \mathsf{E}[Y|\mathbf{X} = \mathbf{x}]$ based on a multivariate random sample $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_N, Y_N)\}$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})$. Now we want to construct a function $\hat{f}$ which is a reasonable approximation to an unknown multivariate regression function f where $\hat{f}$ is based on spline approach[1].

The direct extension of univariate piecewise polynomial semi-parametric spline modelling (smoothing splines, smoothing splines with penalties, regression splines, B-splines, or B-splines with penalties) is straightforward in principle but too difficult in practice. These difficulties are related as mentioned to the curse-of-dimensionality problem. There are two separate common methods used to bypass the curse of dimensionality problem and to estimate the!multivariate regression function f based on spline approach. The!first one is to use generalized spline approaches, so called low dimensional expansions which means to use a set of lower dimensional functions (usually one or two dimensional) to estimate the true underlying function. The other way is to employ more sophisticated strategies so called adaptive spline methods[2] which try to overcome some limitations associated with high dimensions.

[1] Different nonparametric multivariate methods could be thought as a generalization of univariate splines methods (smoothing splines, regression splines, B-splines, P-splines).

[2] Adaptive methods for estimation of a multivariate regression function f will be discussed in chapter 9 and chapter 10.

In the case of generalized spline approach the subregions in high dimensions are constructed as tensor products of $K+1$ indexknotsintervals defined by K knots over the J variables. The corresponding global basis is then the tensor product over the $K+n+1$ basis functions associated with each one variable. This gives rise to $(K+n+1)^J$ coefficients to be estimated from the data points. Even with a coarse grid[3] a very large data sample is required.

The ability of the nonparametric methods to adequately approximate functions in low dimensions, has motivated approximations that take the form of an expansion in low dimensional functions

$$\widehat{f}(\mathbf{x}) = \sum_{t=1}^{T} \widehat{g}_t(\mathbf{z}_t), \tag{5.1}$$

where each $\mathbf{z}_t$ is a comprised of a small preselected subset of $(x_1, \dots, x_J)$. Any variable $x_1, \dots, x_J$ can be supplied in more than just one subset $\mathbf{z}_t$. Such an expansion means that one J-dimensional function is estimated by more low dimensional functions. The functions $\{\widehat{g}_t(\mathbf{z}_t)\}_1^T$ are taken to minimize

$$Minimize \qquad \sum_{i=1}^{N} \left[\textstyle\sum_{t=1}^{T} g_t(\mathbf{z}_{it}) - Y_i\right]^2 + \sum_{t=1}^{T} \lambda_t \rho(g_t), \tag{5.2}$$

subject to all admissible functions g_t. The second term in (5.2) is an optional roughness penalty[4]. Parameter $\lambda = (\lambda_1, \dots, \lambda_T)$ is a multiple smoothing parameter. The optimal choice is provided by some of the model selection criteria[5].

The most intensively used low dimensional expansion is the additive regression model which takes a form (1.2). Because estimation strategies as already mentioned work best for one-dimensional functions it is also the most useful low dimensional expansion model. The underlying regression function is estimated as a sum of one dimensional functions (functional components). Moreover, the number of functional components in the sum is the same as the number of variables that is why additive models have such easy interpretation and popularity. Such additive model overcomes curse of dimension problem and is a simple extension of standard spline methodology for multivariate regression problem. The another advantage of such a model is a property that many true underlying functions can be approximated fairly well by a function in additive form.

[3] The effect of a coarse grid is causes by a small number of knots K and the high number of dimensions. (effect of the curse of dimension)

[4] The roughness penalty term is an optional with an arbitrary choice of penalty functional form ρ. The most common choice of ρ is an integral of n^{th} derivative of the function g_t over the particular domain of $\mathbf{z}$.

[5] Model selection criteria are in detail discussed in chapter 8

In general we can write the additive estimate of the multivariate regression function $f(\mathbf{x})$, for $\mathbf{x} = (x_1, \dots, x_J)$ in a form

$$\widehat{f}(\mathbf{x}) = \sum_{j=1}^{J} \widehat{f}_j(x_j), \tag{5.3}$$

where $\widehat{f}_j$ are estimates of the appropriate functional components in the additive model based on univariate spline approach.

Let the response Y is centered which means $\sum_{i=1}^{N} Y_i = 0$. Then once a regression function takes an additive form (1.2) or the true underlying regression function is estimated as an additive function the regression problem is formulated as a minimizing problem

$$Minimize \quad \sum_{i=1}^{N} \Big(\textstyle\sum_{j=1}^{J} f_j(X_{ij}) - Y_i \Big)^2, \tag{5.4}$$

subject to all acceptable functions $f_j, j = 1, \dots, J$ from sets of all admissible functions for single variables. Since we formulated a regression problem based on a spline approach the classes of admissible functions are composed by splines of the same degree over the span of each variable x_j.

More general we can omit the condition of centered response Y. Then the additive estimate of the regression function[6] f is written in a form

$$\hat{f}(\mathbf{x}) = \hat{f}(x_1, \dots, x_J) = \hat{\mu} + \sum_{j=1}^{J} \widehat{f}_j(x_j), \tag{5.5}$$

where $\hat{\mu} = \overline{Y}_N = N^{-1} \sum_i Y_i$, and spline estimates are determined by the least squares minimizing problem

$$Minimize \quad \sum_{i=1}^{N} \Big[Y_i - \overline{Y}_N - \textstyle\sum_{j=1}^{J} f_j(X_{ij}) \Big]^2 + \sum_{j=1}^{J} \lambda_j \int_0^1 f_j^{(n)}(x_j) dx_j, \tag{5.6}$$

subject to all admissible functions f_j, $j \in \{1, 2, \dots, J\}$ and a multivariate smoothing parameter $\lambda = (\lambda_1, \dots, \lambda_J)$. Regarding to the assumption $\mathsf{E} f_j(X_j) = 0$ we have to consider that $\sum_{i=1}^{N} f_j(X_{ij}) = 0$ for $j = 1, 2, \dots, J$. The second term in (5.6) is an optional roughness penalty which is taken as a sum of individual smoothness penalties for each functional component f_j with an univariate smoothing parameter[7] $\lambda_j \in \mathbb{R}$.

[6] The true underlying regression function does not even have to be in the additive form.

[7] Once a smoothing parameter is different for each variable in (5.6) we discuss a multivariate smoothing parameter. If the weight of smoothness penalty is the same for each variable, we talk about single smoothing parameter $\lambda \in \mathbb{R}$.

The optimal choice of the smoothing parameters is selected in respect to the optimal model which can be established by the model selection criteria. The smoothness penalty in (5.6) is taken as an integral of n^{th} derivative of each functional component where n is a degree of splines used to estimate components f_j. However, the order of derivative is optional and the most common choice is the second derivative where we penalize the roughness. The lower and upper bounds used in integral appear from the assumption $X_j \in [0, 1]$ for each $j \in \{1, \dots, J\}$ from the chapter 1.
A question of knots positions and their number is similar to the univariate problem. The set of knots define subintervals[8] on interval $[0, 1]$ for each single variable x_j and single spline estimates of the function components f_j are constructed in the same way as spline estimates in the case of one-dimensional regression problem.

Sometimes it might happened that the estimate of a multivariate regression surface made by one-dimensional expansion is not sufficient and there arises a need to estimate a dependence not only on single variables but also using some interactions. An implementation of interaction between two single variables is not as simply as it appears. In the sense of definition of a multivariate regression model once we include interaction into the model it becomes more dimensional. Such model with double interactions takes a form

$$f(\mathbf{x}) = f(x_1, \dots, x_J) = \mu + \sum_{i=1}^{N} f_j(x_j) + \sum_{i=1}^{N} \sum_{l<i} f_{il}(x_i, x_l),$$

where functional components f_{il} are **two** dimensional function. This strategy is not very convenient because at first we use a one dimensional expansion to overcome curse of dimensionality next we increase the dimension and the curse of dimension problem appears again. Although, in the next chapter a convergence of spline approach will be shown also for a case of including interaction terms in a model.

More sophisticated methods which intend to limit some restrictions and attempt to approximate regression functions in high dimensions are based on adaptive computation. An adaptive computation is one that dynamically adjusts its strategy to take into account the behavior of the particular problem to be solved. An adaptive computation is not limited by the amount of data as much as additive models therefore we can consider also some interaction terms without a problem of exponential increase of necessary parameters. Such methods will be briefly mentioned in chapter 9 (Multivariate Adaptive Regression Splines) and chapter 10 ((Projection Pursuit regression).

[8] As we assume each $x_j \in [0, 1]$ therefore we can use the same knots to define subintervals for each variable separately. Then we construct the regions as tensor products over all J dimension.

Chapter 6

Rates of Convergence for Additive Models

In this chapter it will be shown that the additive spline estimates $\widehat{f_N}$ of the additive regression function f^* as mentioned above are consistent for $N \to \infty$ and the rate of convergence does not depend on the number of dimensions J (the optimal rate of convergence for additive spline estimates is the same for general J as for $J = 1$). Stone [21] derived a proof of consistence for a rich class of nonparametric regression estimates (nearest neighbor estimates, kernel estimates, partition estimates, local polynomial estimates). Later he derived (Stone [22] and Stone [23]) the optimal rate of convergence for nonparametric regression estimates.

Let $\| \cdot \|_q$ denotes the L^q norm for a collection of functions on $C = [0,1]^J$ where $q \in (0, \infty]$. Let $\| g \|_q = \sup_{\mathbf{x} \in [0,1]^J} |g(\mathbf{x})|$, for $q = \infty$ and $\| g \|_q = (\int_C |g(\mathbf{x})|^q d\mathbf{x})^{1/q}$ for $q \in (0, \infty)$. Let $\{b_N\}$ be a sequence of positive constants. The sequence $\{b_N\}$ is called an optimal rate of convergence if is it a lower rate of convergence and an achievable rate of convergence at the same time. The sequence $\{b_N\}$ is called the lower rate of convergence if

$$\lim_{c \to 0} \liminf_{N \to \infty} \sup_{f \in \kappa} \mathsf{P}(\| \widehat{T}_N - T(f) \|_q \; > c \cdot b_N) = 1, \tag{6.1}$$

and the sequence $\{b_N\}$ is called the achievable rate of convergence if

$$\lim_{c \to \infty} \limsup_{N \to \infty} \sup_{f \in \kappa} \mathsf{P}(\| \widehat{T}_N - T(f) \|_q \; > c \cdot b_N) = 0, \tag{6.2}$$

where $\widehat{T}_N$ is a sequence of estimators of $T(f)$ based on a random sample of size N and $T(f)$ represents a derivative[1] of the true underlying regression function $f \in \kappa$ of the m^{th} order where κ is the collection of k-times continuously differentiable functions on $[0,1]^J$ and $m \leq k$.

[1] Let $\alpha = (\alpha_1, \ldots, \alpha_J)$ represents a J-tuple of nonnegative integers where $m = \sum_j \alpha_j$. Then the derivative of f of the m^{th} order is written as a functional T where $T(f) = \frac{\partial^{\alpha_1 + \cdots + \alpha_J}}{\partial x_1^{\alpha_1} \ldots \partial x_J^{\alpha_J}} f$.

Theorem 6.1 (Rates of Convergence for Nonparametric Estimates) *Let conditions* [2] *hold. Let $\beta \in (0,1]$ and set $p = k + \beta$. Let $0 < q \leq \infty$ and set $r = \frac{p-m}{2p+J}$. Then the optimal global rate of convergence for nonparametric estimates $\widehat{T}_N$ of the derivative $T(f)$ is*

$$\{N^{-r}\} \quad \textit{for} \;\; q \in (0, \infty),$$

and

$$\{(N^{-1} \cdot \ln N)^r\} \quad \textit{for} \;\; q = \infty.$$

A proof was given in Stone [23]. Some notes can also be found in Stone [22]. The optimal global rate of convergence given by Theorem 6.1 is strictly depended on the number of dimensions J. Generally, the increasing number of predictor variables used in a regression model causes a lower rate of convergence (according to the fact $r = \frac{p-m}{2p+J}$).

Theorem 6.1 can be successfully generalized if the usual L^q norm in (6.1) and (6.2) is replaced by $\mid \widehat{T}_N(\mathbf{x}_0) - T(f)(\mathbf{x}_0) \mid$ where $\mathbf{x}_0 \in [0,1]^J$ is fixed. Stone [22] showed that n^{-r} is also the optimal rate of convergence for $\mid \widehat{T}_N(\mathbf{x}_0) - T(f)(\mathbf{x}_0) \mid$. Such a rate of convergence can be thought as a local rate of convergence.

In the case of additive regression models a special dimensional reduction principle can be established not only for the function f (written in the form (1.2)) but even for the optimal global rate of convergence for the estimates based on the random sample of size N where the optimal global rate of convergence does not depend on the number of predictor variables used in the final fit. Problems can arise if the additive estimate is given to a non-additive regression function. However, the additive approximation to the true underlying regression function can be successfully obtained be minimizing the least squares term. In such a case the dimensional reduction principle is established for a sequence of estimators of the additive approximation f^* which has the form (1.3).

6.1 Additive Approximation - Decomposition

Let $f(\mathbf{x}) = f(x_1, \ldots, x_J)$ be a function on $[0,1]^J$ which is not genuinely additive. Then the function f can be easily decomposed into "main effects" (if interactions are required they can be obtained be the same decomposition).

[2] Conditions required to hold the statement are in detail listed in Stone [23]. Generally three conditions are required. We assume that the logarithm of $h(y|\mathbf{x}, f(\mathbf{x}))$ which is a conditional density of Y is locally bounded by another function $M(y|\mathbf{x}, f(\mathbf{x}))$ with a bounded conditional mean. The second condition required that $\mathsf{E}[\exp(s|y - f(\mathbf{x})|)|\mathbf{X} = \mathbf{x}]$ is bounded for some $s > 0$ and the third condition required a densely packed space $[0,1]^d$ with increasing N.

Condition 1
Let the distribution of $\mathbf{X}$ *is absolutely continuous and let its density* g *is bounded away from zero and infinity (*$\exists b > 0, \quad \exists B > 0, \quad b \leq g \leq B$ *on* $C = [0,1]^J$*).*

Let $\mathsf{E}Y^2 = \mathsf{E}(f(\mathbf{x}))^2 < \infty$. Let the variables $X_j, \ldots, X_J$ are independent. Then the additive decomposition[3] of the function f which minimize the least squares term can be written in the form

$$f^*(\mathbf{x}) = \mathsf{E}(f(\mathbf{x})) + \sum_{j=1}^{J} \left[\mathsf{E}[f(\mathbf{x})|X_j = x_j] - \mathsf{E}(f(\mathbf{x}))\right], \tag{6.3}$$

where $f_j^* = \mathsf{E}[f|X_j] - \mathsf{E}(f)$ is a function of the variable x_j, for $j = 1, \ldots, J$ and it also satisfies the condition that $\mathsf{E}f_j^*(X_j) = 0$ because

$$\mathsf{E}\left[\mathsf{E}[f(\mathbf{x})|X_j = x_j]\right] - \mathsf{E}(f(\mathbf{x})) = 0.$$

Lemma 1
Let the random variable $\sum_j h_j(X_j)$ *has a finite second moment where* h_j *are functions on* $[0,1]$*. Set* $\delta_1 = \sqrt{(1 - b/B)}$ *and let* $SD(\cdot)$ *denotes the standard deviation. Then each* $h_j(X_j)$ *has a finite second moment and it holds that*

$$SD(\textstyle\sum_j h_j(X_j)) \geq ((1-\delta_1)/2)^{(J-1)/2} \cdot (SD(h_1(X_1)) + \cdots + SD(h_J(X_J))).$$

Under the Condition 1 and Lemma 1 and considering the fact that the conditional mean is uniquely determined up to set of measure zero the additive decomposition defined by (6.3) is uniquely determined up to sets of measure zero.

Condition 2
Let the regression function f *is bounded on* $C = [0,1]^J$ *and let the conditional variance* $\mathsf{Var}(Y|\mathbf{X} = \cdot)$ *is also bounded on* C*.*

Let $\mathcal{H}$ be a collection of functions on $[0,1]$ whose k^{th} derivative exists and satisfies the Hölder condition with exponent β

$$\forall_{h \in \mathcal{H}} \forall_{x_1, x_2 \in [0,1]} : \qquad \|h^{(k)}(x_1) - h^{(k)}(x_2)\| \leq M \|x_1 - x_2\|^{\beta},$$

where $M \in (0, \infty)$. The following smoothness assuption has to be imposed on the functional components of the additive approximation[4] f^* to hold the statements of the main results.

Condition 3
Let $f_j^*, \ j = 1, \ldots, J$ *be the functional components of the additive approximation* f^* *and let* $f_j^* \in \mathcal{H}$ *for* $j = 1, \ldots, J$*.*

[3] The additive decomposition was proposed in Efron [9]. Similar decomposition was proposed by Hegland and Pestov [16] where the functional components were defined as $f_j^* = [f - \mathsf{E}(f|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_J)] \cdot \mathsf{E}[f|x_j]$ which is a function of the variable x_j and by using The Fubini Theorem and properties of the conditional mean it also satisfies $\mathsf{E}f_j^* = 0$.

[4] From now on without any loss of generality the additive approximation f^* will be assumed rather then the additive function f. If f is additive function then of course $f^* = f$.

6.2 Convergence for Additive Estimates

Let $\parallel \varphi \parallel$ denote the L^2 norm of a function φ defined on $C = [0,1]^J$, defined by $\parallel \varphi \parallel^2 = \mathsf{E}\varphi^2(\mathbf{X}) = \int_C \varphi^2(\mathbf{x})g(\mathbf{x})d\mathbf{x}$. For $1 \leq j \leq J$ let $\parallel h \parallel_j^2$ denotes the L^2 norm of a function h on $[0,1]$, defined by $\parallel h \parallel_j^2 = \mathsf{E}h^2(X_j) = \int_0^1 h^2(x_j)g_j(x_j)dx_j$. The existence of the marginal density g_j of the variable X_j follows easily from the condition 1. From the fact that the density g is bounded away from zero and infinity on C it also follows that g_j are bounded away form zero and infinity on $[0,1]$. Set $\gamma = 1/(2p+1)$.

Let N_N denote a positive integer and let $I_{N\nu}$, for $1 \leq \nu \leq N_N$, denote the subintervals of $[0,1]$ defined by

$$I_{N\nu} = [\tfrac{(\nu-1)}{N_N}, \tfrac{\nu}{N_N}) \quad \text{for} \quad 1 \leq \nu \leq N_N - 1, \quad \text{and} \quad I_{NN_N} = [1 - \tfrac{1}{N_N}, 1].$$

Let $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_N, Y_N)$ denote independent random pairs, each having the same distribution as $(\mathbf{X}, Y)$ and consider an additive spline estimate in the form (5.5) (piecewise polynomial on intervals $I_{N\nu}$, $\nu = 1, 2, \ldots, N_N$) of the regression function f^* based on the random sample of size N. The dimensional reduction principle for the rate of convergence of additive spline estimates is established by the following theorem of Stone [24].

Theorem 6.2 (Rate of Convergence for Additive Estimates) *Suppose that Conditions 1, 2, 3 hold and let $N_N \sim N^\gamma$. Recall that $r = \frac{p-m}{2p+1}$, where $0 \leq m \leq k$. Then*

$$\mathsf{E}(\parallel \widehat{f}_{Nj}^{(m)} - (f_j^*)^{(m)} \parallel_j^2 |X_1, \ldots, X_N) = O_{pr}(N^{-2r}), \quad \text{for } 1 \leq j \leq J, \tag{6.4}$$

$$\mathsf{E}(\parallel \widehat{f}_{Nj} - f_j^* \parallel_j^2 |X_1, \ldots, X_N) = O_{pr}(N^{-2r}), \quad \text{for } r = p/(2p+1), \tag{6.5}$$

$$\mathsf{E}((\overline{Y}_N - \mu)^2 |X_1, \ldots, X_N) = O_{pr}(N^{-2r}), \quad \text{for } r = p/(2p+1), \tag{6.6}$$

$$\mathsf{E}(\parallel \widehat{f}_N - f^* \parallel^2 |X_1, \ldots, X_N) = O_{pr}(N^{-2r}), \quad \text{for } r = p/(2p+1). \tag{6.7}$$

The rates of convergence in Theorem 6.2 do not depend on the dimension J of the random vector $\mathbf{X}$. It was shown in Stone [23] that these rates of convergence are optimal for $J = 1$ (it satisfies (6.1) and (6.2)).

The last statement (6.7) follows easily form (6.5) and (6.6) by using the definition of the norm $\parallel \cdot \parallel$, the linearity of integral, the Fubini Theorem and the fact that $\mathsf{E}f_j^*(X_j) = 0$. Recall that functions $f^*, \widehat{f}_N$ are additive which means $\frac{\partial^2 f^*}{\partial x_{j_1} \partial x_{j_2}} = \frac{\partial^2 \widehat{f}_N}{\partial x_{j_1} \partial x_{j_2}} = 0$ if $j_1 \neq j_2$. The only reasonable derivatives are partial derivatives for the **same** variable $\frac{\partial^m f^*}{\partial x_j^m} = (f_j^*)^{(m)}$ and by analogy for $\widehat{f}_N$. In such a case the rate of convergence is defined by the statement (6.4) in Theorem 6.2. If we define the derivative of the m^{th} order of the function f^* or $\widehat{f}_N$ respectively, as a linear combination of partial derivatives we can write

$$(f^*)^{(m)} = \frac{\partial^m f^*}{\partial x_1^m} + \cdots + \frac{\partial^m f^*}{\partial x_J^m} = (f_1^*)^{(m)} + \cdots + (f_J^*)^{(m)}. \tag{6.8}$$

Under the Condition 3 the functional components $(f_j^*)^{(m)}$ are $(k-m)$-times continuously differentiable on $[0,1]$ thus from Theorem 6.2 $\{N^{-r}\}$ is also the optimal global rate of convergence for $\| \widehat{f}_N^{(m)} - (f^*)^{(m)} \|$ where $r = \frac{p-m}{2p+1}$.

6.3 Proof of Theorem 6.2

Consider an additive spline estimate $\widehat{f}_N$ of the regression function f^* where we minimize the least squares term $\sum_i (f_N(\mathbf{x}) - Y_i)^2$ subject to functional components $f_{Nj} \in \mathcal{H}$ such that a restriction of f_{Nj} to $I_{N\nu}$ is a polynomial of degree $l \leq k$ and f_{Nj} is $(l-1)$-times continuously differentiable on $[0,1]$ and $\sum_i f_{Nj}(X_{ij}) = 0$. Firstly it will be shown that such an estimate is uniquely determined up to set of measure zero.

Lemma 2
Let $N_N \sim N^\gamma$ and let $\delta_2 \in (\delta_1, 1)$, where δ_1 is defined in Lemma 1. Let $SD_N(H_j(h))$ denotes the standard deviation corresponding to the empirical distribution of $\mathbf{X}_i$ of a function on $[0,1]^J$ such that $H_j(h)(\mathbf{x}) = h(x_j)$, where $h \in \mathcal{H}$ is a function on $[0,1]$. Then except on an event whose probability tends to zero with $N \to \infty$, the following statement holds for arbitrary choices of spline functions $h_1, \ldots, h_J \in \mathcal{H}$:

$$SD_N(\textstyle\sum_j H_j(h_j)) \geq ((1-\delta_1)/2)^{(J-1)/2} \textstyle\sum_j SD_N(H_j(h_j))$$

The proof of Lemma 2 follows from Lemma 1 and it is obtained in Stone [24]. Consider two additive spline estimates $\widehat{f}_N^{[1]}, \widehat{f}_N^{[2]}$ of the same regression function such that

$$\mathsf{P}\left[|\widehat{f}_N^{[1]}(\mathbf{X}) - \widehat{f}_N^{[2]}(\mathbf{X})| = 0 \right] \longrightarrow 1$$

If we define functions $H_j(h_j)$ from Lemma 2 as $H_j(h_j) = \widehat{f}_{Nj}^{[1]} - \widehat{f}_{Nj}^{[2]}$ then it follows from Lemma 2 that under the Condition 1 the functional components are uniquely determined except on an event whose probability tends to zero.

If the constrained minimization problem has a unique solution then the functional components of the additive estimate of f^* can by written as

$$\widehat{f}_{Nj}(x_j) = \sum_{i=1}^{N} W_{Nij}(x_j) Y_i, \tag{6.9}$$

where $W_{Nij}(\cdot)$ are weight functions on $[0,1]$, for $1 \leq i \leq N$ and $1 \leq j \leq J$.

From Lemma 2 it also follows that the weight functions are uniquely determined W_{Nij} except on an event whose probability tends to zero. Moreover

$$\sup_{x_j \in [0,1]} |W_{Nj}(x_j)|^2 = \sup_{x_j \in [0,1]} \left[\sum_{i=1}^{N} W_{Nij}^2(x_j) \right] = O_{pr}(N^{-1} \cdot N_N) \tag{6.10}$$

Let $\overline{\mu} = \frac{1}{N} \sum_i f(\mathbf{X}_i)$ and consider an additive function $\overline{f}_N(\mathbf{x}) = \overline{\mu} + \sum_j \overline{f}_{Nj}(x_j)$ where the function $\overline{f}_N$ minimize the least square term

$$\sum_{i=1}^{N} (\overline{f}_N(\mathbf{X}_i) - f(\mathbf{X}_i))^2.$$

Now, reformulate the problem: $\widehat{f}_{Nj}^{(m)} - (f_j^*)^{(m)} = (\widehat{f}_{Nj}^{(m)} - \overline{f}_{Nj}^{(m)}) + (\overline{f}_{Nj}^{(m)} - (f_j^*)^{(m)})$. To give a proof of Theorem 6.2 it will be shown that all following statements hold. The results of Theorem 6.2 than follow easily from the definition of the norm $\| \cdot \|$ and the triangle inequality. Recall that $r = (p-m)/(2p+1)$.

$$\| \widehat{f}_{Nj} - \overline{f}_{Nj} \|_j^2 = O_{pr}(N^{-2r}), \quad \text{for} \quad m = 0 \tag{6.11}$$

$$\| \overline{f}_{Nj} - f_j^* \|_j^2 = O_{pr}(N^{-2r}), \quad \text{for} \quad m = 0 \tag{6.12}$$

$$\| \widehat{f}_{Nj}^{(m)} - \overline{f}_{Nj}^{(m)} \|_j^2 = O_{pr}(N^{-2r}), \tag{6.13}$$

$$\| \overline{f}_{Nj}^{(m)} - (f_j^*)^{(m)} \|_j^2 = O_{pr}(N^{-2r}), \tag{6.14}$$

The first equation follows from (6.9) and (6.10). Let $\widehat{f}_{Nj}(x_j) = \sum_i W_{Nij}(x_j) Y_i$ and $\overline{f}_{Nj}(x_j) = \sum_i W_{Nij}(x_j) f(\mathbf{X}_i)$. Consider $N \in \mathbb{N}$ fixed. Then

$$\begin{aligned}
\| \widehat{f}_{Nj} - \overline{f}_{Nj} \|_j^2 &= \left\| \sum_i W_{Nij}(\cdot) Y_i - \sum_i W_{Nij}(\cdot) f(\mathbf{X}_i) \right\|_j^2 \\
&= \left\| \sum_i W_{Nij}(\cdot)(Y_i - f(\mathbf{X}_i)) \right\|_j^2 \\
&\leq \sum_i \left\| W_{Nij}(\cdot) \max_{1 \leq i \leq N} |Y_i - f(\mathbf{X}_i)| \ \right\|_j^2 \\
&\leq K^2 \cdot \sum_{i=1}^{N} \int_{[0,1]} W_{Nij}^2(x_j) g_j(x_j) dx_j \qquad (6.15) \\
&\leq K^2 \int_{[0,1]} \sup_{x_j \in [0,1]} \left(\sum_i W_{Nij}^2(x_j) \right) g_j(x_j) dx_j \\
&= K^2 \cdot \sup_{x_j \in [0,1]} \left(\sum_i W_{Nij}^2(x_j) \right),
\end{aligned}$$

which follows from the triangle inequality for L^2 norm and The Lebesgue Theorem and the fact that the conditional variance is bounded on C. Thus for $N \to \infty$ the desired conclusion holds and $\| \widehat{f}_{Nj} - \overline{f}_{Nj} \|_j^2 \;\; = O_{pr}(N^{-2r})$.

Lemma 3
Suppose that Conditions 1, 2 and 3 hold and let $N_N \sim N^\gamma$. Then

$$\| \overline{f}_{Nj} - f_j^* \|_{Nj}^2 \;\; = \;\; \frac{1}{N}\sum_{i=1}^{N}(\overline{f}_{Nj}(X_{ij}) - f_j^*(X_{ij}))^2 = O_{pr}(N^{-2r}) \qquad (6.16)$$

where the norm $\| \cdot \|_{Nj}$ denotes the L^2 norm of a function on $[0,1]$ with respect to the empirical distribution.

Lemma 4
Let $N_N \sim N^\gamma$. Then there is a number $M_1 \in (0,\infty)$ such that, except on an event whose probability tends to zero with $N \to \infty$, the next statement holds for an arbitrary choice of $h \in \mathcal{H}$ and a spline function $s \in \mathcal{H}$:

$$\| s - h \|_j^2 \;\; \leq \;\; M_1 \left[N_N^{-2p} + \| s - h \|_{Nj}^2 \right] \qquad (6.17)$$

Lemma 5
Let $N_N \sim N^\gamma$. Then there is a number $M_2 \in (0,\infty)$ such that, for an arbitrary choice of $h \in \mathcal{H}$ and a spline function $s \in \mathcal{H}$:

$$\| s^{(m)} - h^{(m)} \|_j^2 \;\; \leq \;\; M_2 \left[N_N^{-2(p-m)} \; + \; N_N^{2m} \| s - h \|_j^2 \right] \qquad (6.18)$$

The proofs of Lemmas 3, 4 and 5 are given in Stone [24]. To prove the statement (6.12) we use Lemma 4:

$$\begin{aligned} \| \overline{f}_{Nj} - f_j^* \|_j^2 &\leq M_1 \left[N_N^{-2p} \; + \; \| \overline{f}_{Nj} - f_j^* \|_{Nj}^2 \right] \\ &= M_1 N_N^{-2p} + M_1 \| \overline{f}_{Nj} - f_j^* \|_{Nj}^2 \\ &= O_{pr}(N^{-2r}), \end{aligned} \qquad (6.19)$$

which follows from Lemma 3 and the fact $N_N \sim N^\gamma$. Similarly, we proof (6.13):

$$\begin{aligned} \| \widehat{f}_{Nj}^{(m)} - \overline{f}_{Nj}^{(m)} \|_j^2 &\leq M_2 \left[N_N^{-2(p-m)} \; + \; N_N^{2m} \;\; \| \widehat{f}_{Nj} - \overline{f}_{Nj} \|_j^2 \right] \\ &= M_2 N_N^{-2(p-m)} + M_2 N_N^{2m} \| \widehat{f}_{Nj} - \overline{f}_{Nj} \|_j^2 \\ &= O_{pr}(N^{\frac{-2(p-m)}{2p+1}}) + O_{pr}(N^{\frac{-2(p-m)}{2p+1}}) = O_{pr}(N^{-2r}), \end{aligned} \qquad (6.20)$$

which follows from Lemma 5 and (6.15). The statement (6.14) can be shown by analogy from Lemma 5, Lemma 4 and finally from Lemma 3:

$$\begin{aligned} \| \overline{f}_{Nj}^{(m)} - (f_j^*)^{(m)} \|_j^2 &\leq M_2' \left[N_N^{-2(p-m)} \; + \; N_N^{2m} \;\; \| \overline{f}_{Nj} - f_j^* \|_j^2 \right] \\ &\leq M_2' N_N^{-2(p-m)} + M_2' N_N^{2m} \;\; \| \overline{f}_{Nj} - f_j^* \|_j^2 \\ &\leq O_{pr}(N^{-2r}) + M_1' M_2' N_N^{2m} \;\; \| \overline{f}_{Nj} - f_j^* \|_{Nj}^2 \\ &= O_{pr}(N^{-2r}) + O_{pr}(N^{-2r}) = O_{pr}(N^{-2r}), \end{aligned} \qquad (6.21)$$

for $r = \frac{p-m}{2p+1}$. Thus the statements (6.4) and (6.5) follows easily from (6.15) and (6.19) – (6.21). Finally, to complete the proof of Theorem 6.2 the validity of (6.6) has to be shown. We use an inequality for a norm $\| \cdot \|$ in a Hilbert space $\| v \|^2 \leq (\max_i \| v_i \|) \cdot (\sum_i \| v_j \|)$, where $v = \sum_i v_i$. Then

$$\left\|(\overline{Y}_N - \mu)\right\|^2 = (\overline{Y}_N - \mu)^2 \leq \frac{1}{N}\left(\max_{1\leq i \leq N} |Y_i - \mu|\right) \cdot \sum_{i=1}^{N} \frac{|Y_i - \mu|}{N}. \tag{6.22}$$

Thus for $N \to \infty$ the desired conclusion holds.

6.4 Rate of Convergence for Supremum Norm

Now we want to generalize the previous results and to set the optimal rate of convergence for the additive estimates in the Supremum L^∞ norm. Suppose that the response Y has a zero mean $\mu = 0$.

Theorem 6.3 (Rate of Convergence for Supremum Norm)
Let Conditions 1, 2 and 3 hold and let $N_N \sim N^\gamma$ Suppose that $\mu = 0$. Then

$$\|\widehat{f}_N - f^*\|_\infty = \sup_{\mathbf{x}\in[0,1]^J} |\widehat{f}_N(\mathbf{x}) - f^*(\mathbf{x})| = O_{pr}(N^{-r} \cdot \log^r N), \tag{6.23}$$

$$\|\widehat{f}_{Nj} - f_j^*\|_{\infty,j} = \sup_{x_j\in[0,1]} |\widehat{f}_{Nj}(x_j) - f_j^*(x_j)| = O_{pr}(N^{-r} \cdot \log^r N). \tag{6.24}$$

The first statement follows easily from (6.24). The validity of the second statement can be showed through Theorem 6.1 for a special case of $J = 1$, since we separate the multivariate regression problem into J univariate regression problems. Since the functional components f_j^* are defined by the additive decomposition (6.3) as

$$f_j^*(x_j) = \mathsf{E}[f(\mathbf{x})|X_j = x_j] - \mathsf{E}(f(\mathbf{x})) \quad \text{for} \;\; j = 1, \dots, J, \tag{6.25}$$

one can estimate the functional components f_j^* as simple univariate regression problems based on random samples (X_{ij}, Y_{ij}^*) where $\widehat{f}_{Nj}$ is a solution to

$$Minimize \quad \sum_{i=1}^{N} (Y_{ij}^* - f_{Nj}(X_{ij}))^2, \quad \text{where} \quad Y_{ij}^* = \overline{Y}_{Nij} - \overline{Y}_N,$$

subject to all spline functions $f_{Nj} \in \mathcal{H}$ (the same constraints than the multivariate regression case), where $\overline{Y}_{Nij} = N^{1-J} \sum_{\iota,\ \iota_j = i} \tilde{Y}_\iota$, where[5] $\iota = (\iota_1, \dots, \iota_J)$ and $1 \leq \iota_j \leq N$. Then all necessary conditions hold and the statement (6.24) follows from Theorem 6.1 for $J = 1$.

[5] The number $\overline{Y}_{Nij}$ can be thought as an empirical conditional mean and $\tilde{Y}_\iota$ is defined by an alternative experiment (see Stone [24]) $(\tilde{\mathbf{X}}_\iota, \tilde{Y}_\iota)$ having N^J cases (something like a permutation of vector components of $\mathbf{X}_1, \dots, \mathbf{X}_N$), where $\tilde{\mathbf{X}}_\iota \in [0,1]^J$ is defined as $\tilde{X}_{\iota j} = X_{\iota_j j}$.

Chapter 7

Positioning of the Knots

In the previous chapters some estimation techniques for a regression function $f(\mathbf{x}) = \mathsf{E}[\, Y|\mathbf{X} = \mathbf{x}\,]$ were mentioned. The regression estimate of the true underlying function f (univariate or multivariate) is defined as a minimization problem where we minimize some penalty term (e.g. least-squares term) through a class of all admissible functions. In the case of spline approaches we take the class of admissible functions to be a set of piecewise polynomial functions (splines). However, splines, B-splines or P-splines eventually are uniquely determined by the number $n \in \mathbb{N}$ which is a degree[1] of all piecewise polynomials which compose a spline and by the number of knots and their exact positions. The number of knots K used to fit a curve and their positions (a mesh $\Delta = \{\xi_1, \ldots, \xi_K\}$) have to be determined otherwise one is not able to minimize the regression fitting problem in the right way. Therefore the number K and the mesh $\Delta = \{\xi_1, \ldots, \xi_K\}$ can be thought as parameters or semi-parameters[2] of the spline regression problem. To reap the full benefits of the spline approach, the right choice of the number K and the positions of the knots (the mesh Δ) is a necessary and usually also a difficult problem. Once we specify the number and the positions of the knots the fitting problem usually becomes a simple linear problem.

In general one can not separate decisions made on the number of knots and their positions. It is a complex and related task. By the number of knots one can control the smoothness of the final fit and by the positioning of the knots one can control the shape of the regression curve (surface). Many different instructions for knots selection have been proposed by different authors but there is hardly one good universal method used in real regression problems.

[1] In our case the degree of spline estimates is $n = 3$ because splines of 3^{rd} degree are smooth enough. Of course the problem of selection the right number K and the mesh Δ remains the same for any choice of spline degree $n \in \mathbb{N}$. In general the degree of the spline function depends on what is a realistic assessment of the number of derivatives available in the regression function.

[2] It is possible to set a number and positions of knots by minimizing some criterion (Cross-Validation or Generalized Cross-Validation) which measures the goodness of the final fit.

7.1 Positions of Knots for Interpolation Spline

Once we use interpolation splines for fitting problem there is nothing to worry about considering the number of knots and their positions. For computational reasons the mesh $\Delta = \{\xi_1, \ldots, \xi_K\}$ usually coincides with individual observations $X_i,\ i = 1, \ldots, N$, therefore the number K and the positions of knots are defined by the number of observations and the values of single data points. Moreover, if knots are placed in each observed data point each point will be treated equivalently which is an important property for working with interpolation splines.

7.2 Positions of Knots for Smoothing Spline

Another problems arise when one wants to use smoothing splines for fitting some regression function. The idea of using all observations as knot points is not available because it leads to high overfitting of the regression problem and the large number of observations can cause an efficiency increase. Therefore, some another methods have to be used to define the set of knots, a mesh Δ. One can take a mesh as a subset of $\{X_1, \ldots, X_N\}$ (in the case of smoothing splines) or a new set of knots can be determined and it does not even have to coincide with single observations $X_i,\ i = 1, \ldots, N$ (in the case of regression splines).

Generally, the problem of positioning of knots is much more important than the selection of the number of knots K. The knots placement can be crucial for both, the shape and the quality of the spline fit. Therefore, it is important to choose the positions of knots as well as possible. Some known strategies for the knots positioning will be mentioned in the next sections.

7.2.1 Equidistant Knots

Probably the simplest method for a determination of knots positions is the method of Equidistant knots. It means we place K uniformly spaced knots over a span of the variable X. The accurate choice of the number K is not so important. However, one has to remember that a large number of knots leads to overfitting a problem and a small number of knots will cause an inaccurate estimate of the regression function f. The recommended choice of K is at about $K \sim \frac{N}{5}$ (see S. Wold [26]).

This method is not very popular in real data fitting problem. The spline estimates based on uniformly spaced knots are often too loose. It is clear also from the figure 7.1 that more intuitive placing of knots over the span of the variable X brings much better estimation of the behavior of data points than the estimate based on the uniformly spaced knots.

7.2.2 Intuitive Placement of Knots

An opposite method to uniformly spaced knots is an intuitive placing of knots. Such a strategy is based on an a priori information regarding to the shape of the regression function f and is an implementation of rules of thumb mentioned later. An intuitive placing of knots is closely associated with the degree of splines used to fit a curve. Another placements of knots are necessary for linear spline fitting problem and different knots have to be defined for cubic splines. The knots should be chosen to correspond to the overall behavior of the data[3].

The difference between two curves fitted using uniformly spaced knots and intuitive selected knots can be seen in figure 7.1.

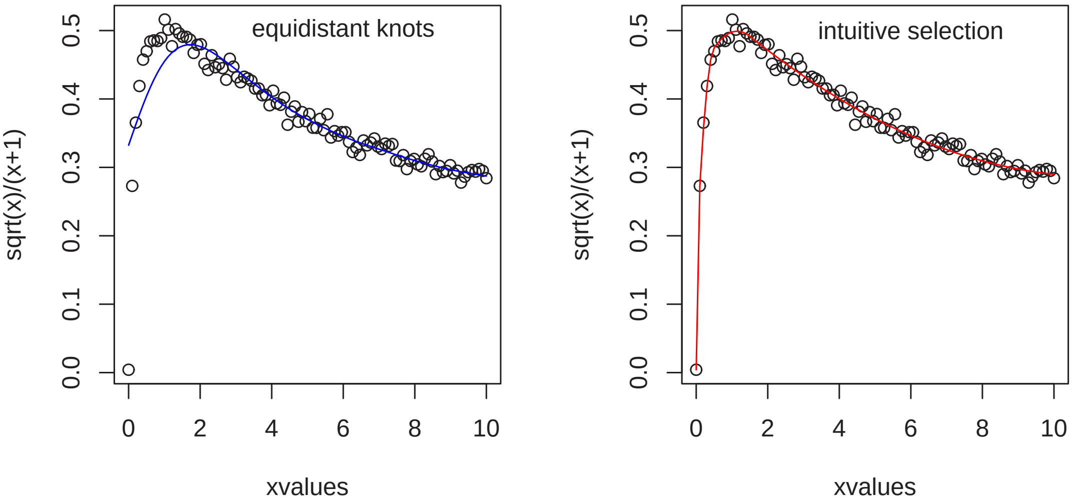


Figure 7.1: The smoothing curve on the left side was fitted by smoothing spline of 3^{rd} degree with equidistant knots while the curve on the right part of the figure was fitted by using more intuitive placing of the knots (simulated data with number of observations $N = 100$).

The intuitive knots placing strategy works quite well with real data anyway, sometimes some simulations are used to recondition the positions of knots and to find the best positions for the knots. Moreover, some regression problems are formulated as minimizing problems where a penalty term is even minimized subject to all possible positions of knots.

[3] The behavior of data means for example the number of single observations, positions of minima and maxima points and positions of flex points, etc.

7.2.3 Rules of Thumb

The Rules of Thumb[4] strategy for knots positioning is based on some intuitions and some practical experiences from the real regression problems solved by different authors. The notion "rules of thumb" has came from Svante Wold [26]. The Rules of Thumb strategy works well when the number of observations is larger than about $30 - 40$ and more. However, with fewer observations they are a little bit useless because one can not get an adequate information about a real behavior of data points. In real problem the Rules of Thumb usually serve as a starting points for the design of the simulations. The real knots positions are then derived from a subsequent simulation study.

The following rules are used with cubic splines:

1. To make an afford to have as few knot points as possible, ensuring that there are at least $4 - 5$ data points between two neighboring knots.
2. To have mostly one extremum data point in each interval defined by two nearest knots, where the minimum, maximum respectively is centered in the interval.
3. To have mostly one inflection point between two knots, where the inflection point is situated close to the knot.

If the number of observations is sufficiently large or if the X values are roughly equidistant[5] the Rules of Thumb strategy is not recommended.

7.2.4 Some Another Strategies for Knots Placement

More objective knot selection strategies can be accomplished by optimizing the estimation criterion of interest with respect to both the knot set $\Delta = \{\xi_1, \ldots, \xi_K\}$ and the class of all admissible functions.

Another method for knots placing is based on the sum of squared residuals between two neighboring knots. It means a start mesh Δ_1 is chosen[6] at first and then the sum of squared residuals is calculated for each interval $(\xi_i, \xi_{i+1}]$, $i = 1, \ldots, K - 1$. The knot positions are thereafter systematically varied until the sum of squared residuals gets to the minimum. Such a mesh Δ is then used as a final mesh to fit a regression spline estimate.

[4] The Rules of Thumb are different for any choice of spline degree n. The Rules of Thumb as presented are designated only for the spline degree $n = 3$. For another choice of the spline degree n they can be easily generalized.

[5] The notion "roughly equidistant" means that the random variable $X_i - X_{i-1}$ remains almost the same for more than 50% of all data points.

[6] The inceptive set of knots can be taken as a mesh of uniformly spaced knots over the span of the variable X.

Chapter 8

Model Selection Criteria

Once we consider a regression situation where the information available on the problem under the consideration does not favor a specific model then one has to choose a good one from those models being tentatively proposed. In the least squares methods for fitting smoothing splines or B-splines respectively with the smoothness penalty term like in (3.3), (3.4), (3.13) or (3.15) the amount of smoothness can be easily influenced by a different choice of a smoothing parameter λ, where $\lambda \in [0, \infty)$. However, different values of the parameter λ define different models. These models belong to the class of all admissible models which are good approximates in some sense. Therefore, we need to define some way how to choose an "optimal" value for the smoothing parameter λ which means to take one model from the class of admissible models to be the final model. The right choice of the smoothing parameter is of a substantial importance[1] for the fitting problem and there are many different methods to choose it. The optimal choice of the smoothing parameter determines the final appearance of the regression surface. Model selection criteria will be mentioned in the next sections.

The most popular model selection criteria are based on the **cross-validation**[2] CV and the **generalized cross-validation** GCV. Both criteria can be thought as estimates of the mean squared error of prediction MSEP of the fit where

$$\mathsf{MSEP} = \frac{1}{N} \sum_{i=1}^{N} \mathsf{E} \left[\widehat{f}(x_i) - y_i \right]^2 , \tag{8.1}$$

where $\widehat{f}$ is an estimate of the original regression function f. Because we use a class of admissible functions defined by different values of λ we modify (8.1) to be

[1] It was mentioned earlier that different choices of the smoothing parameter λ allow to easily control the Bias-Variance trade-off in the final fit.

[2] Model selection criteria based on the cross-validation was proposed by S. Wold and G. Wahba [27]. The selection criteria based on generalized version of CV were introduced by P. Graven and G. Wahba [6].

a function of λ, for different values $\lambda \in [0,\infty)$ as follows:

$$\mathsf{MSEP}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} \mathsf{E}\left[\widehat{f}_{\lambda}(x_i) - y_i^*\right]^2, \tag{8.2}$$

where the function $\widehat{f}_{\lambda}$ is a spline estimate of the true regression function f for the appropriate value of the parameter λ and the variables y_i^* are new observations[3] taken at x_i as $y_i^* = f(x_i)$ plus an independent centered random error.

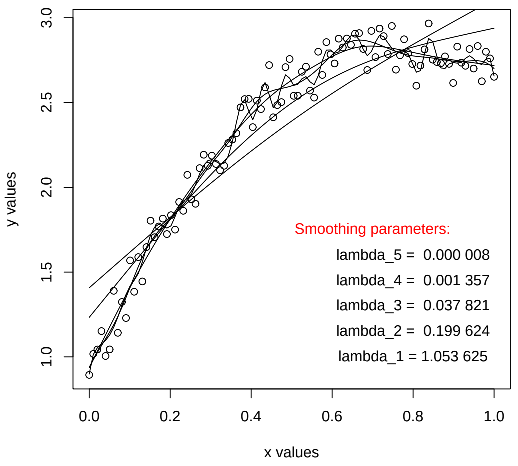


Figure 8.1: Fitted curves for a regression problem with five different choices of a smoothing parameter λ.

Model selection procedure could by ideally thought as a minimizing problem where we minimize the $\mathsf{MSEP}(\lambda)$ subject to all allowable values of a parameter λ. Unfortunately the $\mathsf{MSEP}(\lambda)$ is unknown therefore, one has to use another method to take the optimal smoothing parameter. We want to estimate the $\mathsf{MSEP}(\lambda)$ and then we will take λ which minimize the estimate of the $\mathsf{MSEP}(\lambda)$. We can write[4]

$$\mathsf{MSEP}(\lambda) = \mathsf{MSE}(\lambda) + \sigma^2,$$

where σ^2 is an error variance and MSE is the mean squared error. Such estimates of MSEP are cross-validation and generalized cross-validation. We can also think of some another estimates of $\mathsf{MSEP}(\lambda)$ but methods based on CV and GCV have a great advantage - they do not require an estimation of the error variance which is a positive quality. The most popular MSEP estimates used in practical applications is the **cross-validation criterion** .

Cross-validation criterion and Generalized Cross-validation criterium work by leaving points (x_i, y_i) out one at a time and estimating the smooth at x_i based only on the remaining $N-1$ data points. The main difference between CV and GCV is in the weights which are used to compute the residuals. The optimal smoothing parameter λ is chosen to minimize the estimate of $\mathsf{MSEP}(\lambda)$.

[3] The single observations y_i^*, $i = 1, \ldots, N$ can be thought as bootstraped values (Model-based resampling bootstrap for a regression model).

[4] Mean squared error of prediction $\mathsf{MSEP}(\lambda)$ can be easily write using the fact that the cross product term is zero as a sum of mean squared error $\mathsf{MSE}(\lambda)$ and the error variance σ^2 where the $\mathsf{MSE}(\lambda)$ function of λ is defined as $\mathsf{MSE}(\lambda) = \frac{1}{N}\cdot\sum_{i=1}^{N}\mathsf{E}[\widehat{f}_{\lambda}(x_i) - f(x)]^2$

8.1 Cross-validation

Cross validation is a model evaluation method which gives better insight on model than residuals. The problem with residual evaluations is that they do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on "new" data.

The cross-validation function criterion is defined as follows:

$$\mathsf{CV}(\lambda) = \frac{1}{N} \cdot \sum_{i=1}^{N} \Big[y_i - {}_{(-i)}\widehat{f}_\lambda(x_i) \Big]^2, \tag{8.3}$$

where the function ${}_{(-i)}\widehat{f}_\lambda(\mathbf{x})$ is the spline estimate for the regression function f computed by leaving out the i^{th} data point. The optimal smoothing parameter λ is taken to minimize $\mathsf{CV}(\lambda)$ function.

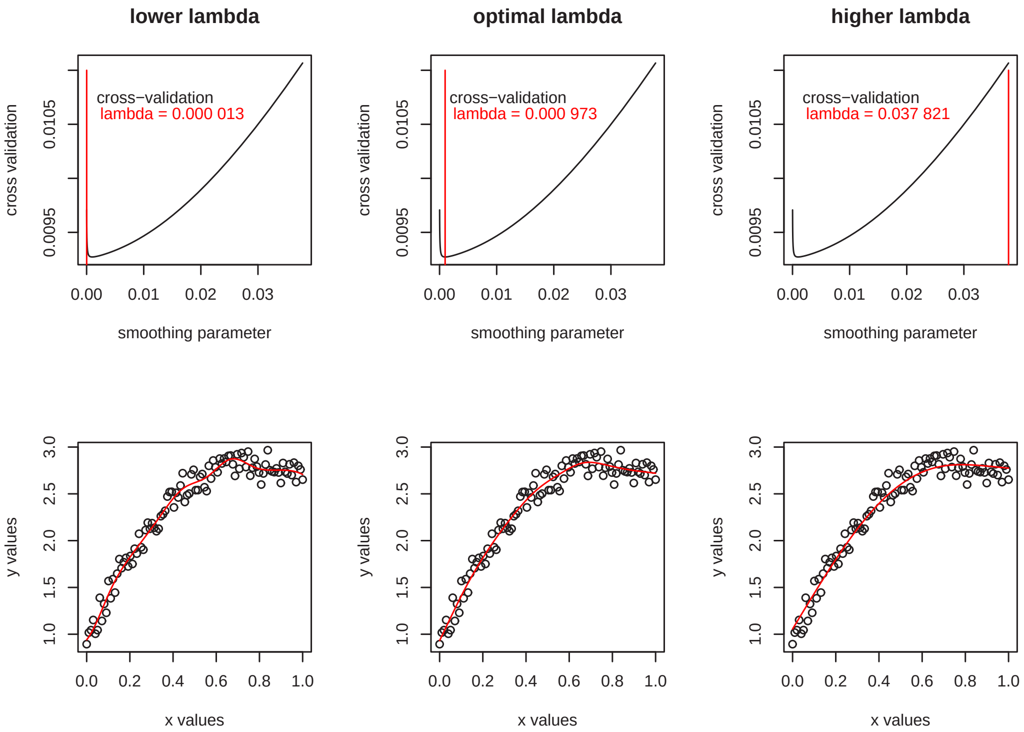


Figure 8.2: Figure with three different choices of the smoothing parameter λ. The optimal choice of the parameter λ was provided using **cross-validation** criterion. In the first row $\mathsf{CV}(\lambda)$ function is drawn and the appropriate choice of λ. In the second row a fitted curve is drawn for the appropriate choice of the smoothing parameter λ.

To overcome some difficulties regarding to CV criterion especially in nonlinear regression a modification is used where y_i is substituted by its estimate instead of omitting it. It is called the **full cross-validation** criterion

$$\mathsf{FCV}(\lambda) = \frac{1}{N} \cdot \sum_{i=1}^{N} \left[y_i - \tilde{f}_\lambda(x_i) \right]^2, \tag{8.4}$$

where $\tilde{f}_\lambda(x_i)$ is the least squares prediction at x_i with substituting y_i by $\widehat{f}_\lambda(x_i)$. By assuming that ${}_{(-i)}\widehat{f}_\lambda(x_i) \approx \widehat{f}_\lambda(x_i)$ we can write[5] the fact $\mathsf{E}[\mathsf{CV}(\lambda)] \approx \mathsf{MSEP}(\lambda)$.

8.2 Generalized Cross-validation

Craven and Wahba [6] have proposed a generalized version of CV method called the **generalized cross-validation** GCV. The GCV weights the ordinary residuals $(y_i - \widehat{f}_\lambda(x_i))$ with a weight function which depends on the parameter λ. The most common choice of the weight function is the average of the weights used for ordinary CV criterion. Such a selection of the weights is usually easier to compute. We can write the $\mathsf{GCV}(\lambda)$ function as

$$\mathsf{GCV}(\lambda) = \frac{1}{N} \cdot \sum_{i=1}^{N} w_i(\lambda) \cdot \left[y_i - {}_{(-i)}\widehat{f}_\lambda(x_i) \right]^2, \tag{8.5}$$

where $w_i(\lambda)$ is a weight function which depends on λ and the function ${}_{(-i)}\widehat{f}_\lambda(\mathbf{x})$ is an estimate computed by leaving out the i^{th} data point. The optimal smoothing parameter λ is chosen to minimize $\mathsf{GCV}(\lambda)$ function.

Let $\hat{\mathbf{y}}(\lambda) = (\hat{y}_1(\lambda), \ldots, \hat{y}_N(\lambda))^{\mathbf{T}}$, where $\hat{y}_i(\lambda) = \widehat{f}_\lambda(x_i)$ for $i = 1, 2, \ldots, N$. Then we can write $\hat{\mathbf{y}}(\lambda)$ as a uniquely determined projection of $\mathbf{y} = (y_1, \ldots, y_N)^{\mathbf{T}}$

$$\hat{\mathbf{y}}(\lambda) = \mathbf{H}(\lambda)\mathbf{y},$$

where $\mathbf{H}(\lambda)$ is the uniquely determined $N \times N$ hat matrix. Once we have the projection matrix $\mathbf{H}(\lambda)$ we can rewrite (8.3) as weighted ordinary residuals

$$\mathsf{CV}(\lambda) = \frac{1}{N} \cdot (\mathbf{y} - \hat{\mathbf{y}}(\lambda))\mathbf{H}_d(\lambda)(\mathbf{y} - \hat{\mathbf{y}}(\lambda))^{\mathbf{T}}, \tag{8.6}$$

where the matrix $\mathbf{H}_d(\lambda)$ is a diagonal matrix with components defined by the equation $h_d^{(ii)}(\lambda) = 1/(1 - h_{ii}(\lambda))^2$, where $\mathbf{H}(\lambda) = (h_{ij}(\lambda))_{i,j=1}^{N}$. For $\mathsf{GCV}(\lambda)$ function we use the average of the weights used for CV function defined by (8.6).

[5] Such an approximation is an application of the fact $\mathsf{MSEP}(\lambda) = \mathsf{MSE}(\lambda) + \sigma^2$ and the fact that $\mathsf{E}[y_i - {}_{(-i)}\widehat{f}_\lambda(x_i)]^2 = \mathsf{E}[y_i - f(x_i) + f(x_i) - {}_{(-i)}\widehat{f}_\lambda(x_i)]^2 = \mathsf{E}[f(x_i) - {}_{(-i)}\widehat{f}_\lambda(x_i)]^2 + \sigma^2$

The generalized cross-validation function $\mathsf{GCV}(\lambda)$ has the following form

$$\mathsf{GCV}(\lambda) = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{(y_i - \widehat{f}_\lambda(x_i))^2}{(1 - \frac{1}{N} \cdot tr(\mathbf{H}(\lambda)))^2}, \tag{8.7}$$

which means that identical weights $w_i(\lambda) = (1 + \frac{1}{N} \cdot tr(\mathbf{H}(\lambda)))^{-2}$ has to be taken. Analogous to ordinary cross-validation criterion one can modify $\mathsf{GCV}(\lambda)$ function to the **generalized full cross-validation** as

$$\mathsf{GFCV}(\lambda) = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{(y_i - \widehat{f}_\lambda(x_i))^2}{(1 + \frac{1}{N} \cdot tr(\mathbf{H}(\lambda)))^2}. \tag{8.8}$$

It was shown that FCV and GFCV are in some sense even better estimates[6] of MSEP. In particulary, the absolute values of the biases of FCV and GFCV are smaller than those for CV and GCV.

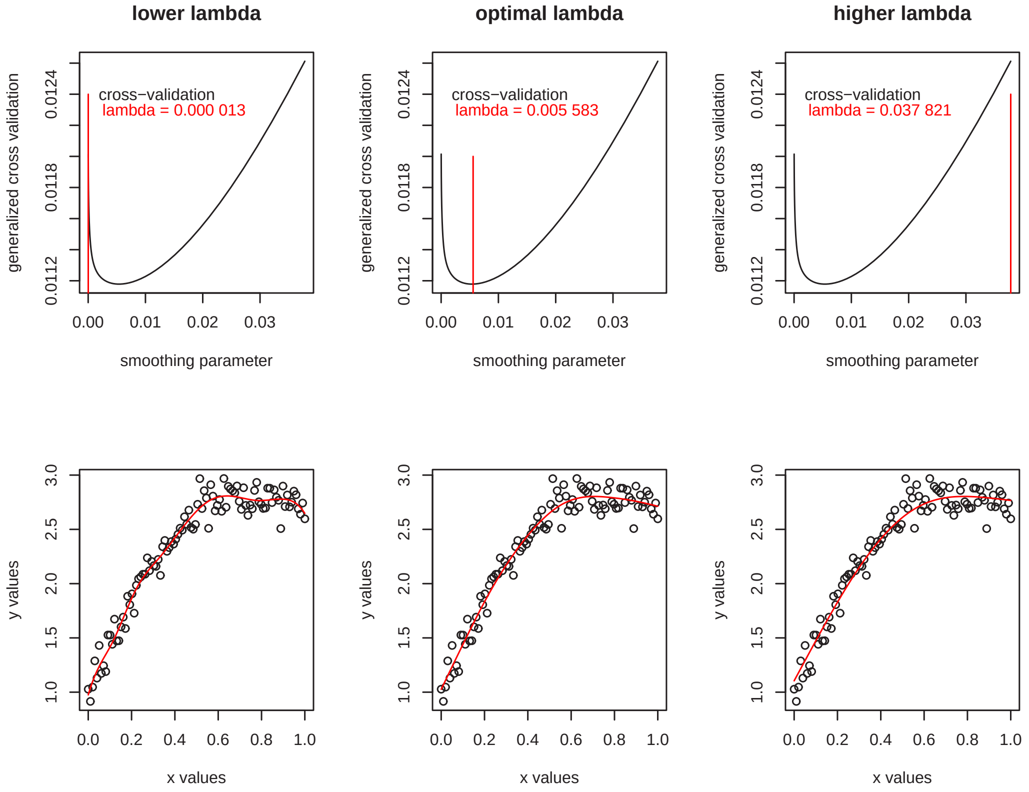


Figure 8.3: Three different choices of the smoothing parameter λ. The optimal choice of the parameter λ was taken to minimize the **generalized cross-validation** criterion.

[6] Under the assumption of normally distributed errors the next results hold: $\mathsf{Var}(\mathsf{GFCV}) < \mathsf{Var}(\mathsf{GCV})$ and also $\mathsf{Var}(\mathsf{FCV}) < \mathsf{Var}(\mathsf{CV})$

8.3 Another Model Selection Criteria

For the sake of completeness some another alternative criteria are mentioned. The can be also used to select an optimal model. They may also be interpreted as estimates of the mean squared error of prediction (MSEP). Indeed, they are functions of the residual sum of squares $\mathsf{RSS}(\lambda) = \frac{1}{N} \parallel y - \widehat{f}_\lambda(x) \parallel^2$.

Akaike Information Criterion: $$\mathsf{AIC}(\lambda) = \log(\mathsf{RSS}(\lambda)) + \frac{2 \cdot tr(\mathbf{H})}{N}$$

Bayesian Information Criterion: $$\mathsf{BIC}(\lambda) = \log(\mathsf{RSS}(\lambda)) + \frac{\log(N) \cdot tr(\mathbf{H})}{N}$$

Mallows Information Criterion: $$\mathsf{C}_p(\lambda) = \mathsf{RSS}(\lambda) + 2 \cdot \frac{tr(\mathbf{H})\sigma^2}{N}$$

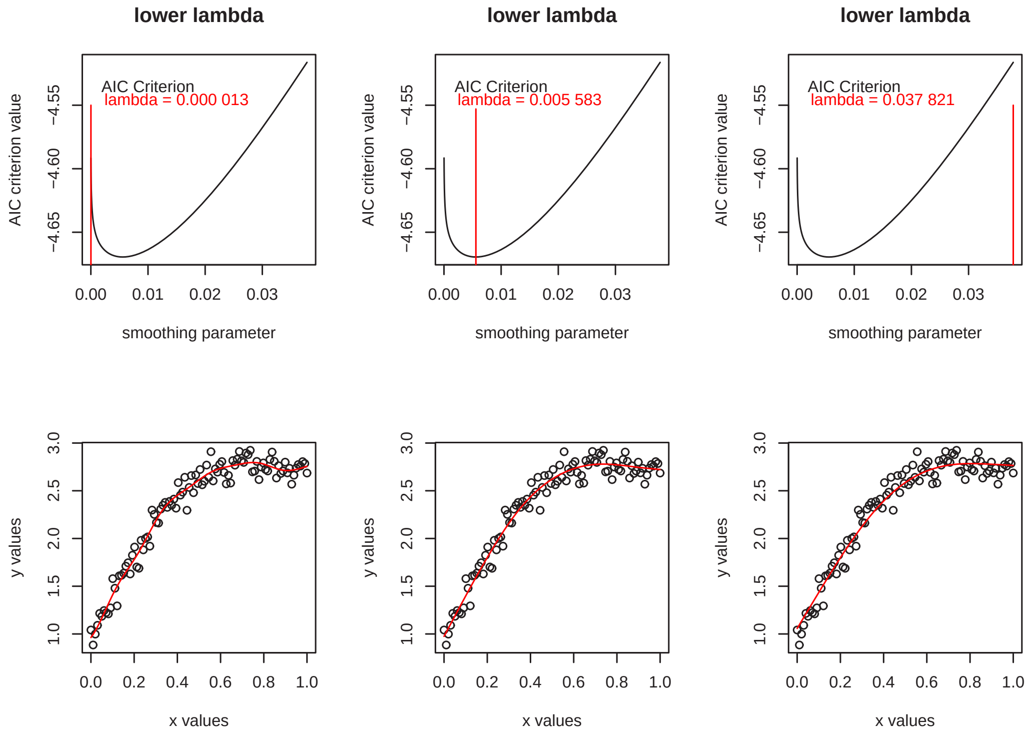


Figure 8.4: Three different choices of the smoothing parameter λ for the final fit. The optimal choice of the parameter λ was provided using **Akaike information criterion**. The optimum value (plots in the middle) was chosen to minimize the Akaike criterion function $\mathsf{AIC}(\lambda)$.

[6] Data used it figures (8.2), (8.3) and (8.4) were simulated and they are the same for all three different methods of selecting the optimal parameter λ. The value of $\lambda \in [0, \infty)$ was taken in discrete way with a step $\Delta_s = 0.005$ therefore, it might possible that optimal λ_0 is a little bit different but not more than ±0.005.

Chapter 9

MARS - Adaptive Regression Splines Approach

More sophisticated way to multivariate nonparametric spline regression is the Multivariate Additive Regression Spline (so called MARS) approach. The MARS Algorithm was proposed by J. Friedman [13] in 1991 and it presents an alternative way for dealing with multivariate regression and a multivariate function. The goal of this procedure is to overcome some limitations associated with the multivariate regression as curse of dimensionality or insufficient interpretability. Multivariate Adaptive Regression Spline approach can be thought as a generalization of a recursive partitioning regression[1]. Therefore recursive partitioning method will be mentioned firstly despite it is not a spline approach indeed.

9.1 Recursive Partitioning Regression

The Partitioning Regression is a recursive procedure which is based on splitting of the entire domain of the predictor variable $\mathbf{X} \in \mathbb{R}^J$. At each stage of the partitioning algorithm all existing subregions are split into two daughter subregions where the split is optimized over all covariates $x_1, \ldots, x_J$ and a split boundary. Such a procedure generates hyper-rectangular axis orientated subregions D_v. In general the recursive partitioning regression model takes the form

$$\hat{f}(\mathbf{x}) = g_v\left(\mathbf{x} | \{\hat{\pi}_{vt}\}_{t=1}^{T}\right), \quad \mathbf{x} \in D_v \subseteq \mathbb{R}^J, \tag{9.1}$$

where D_v are rectangular disjoint subregions of the entire domain of the variable $\mathbf{X} \in \mathbb{R}^J$ and $\{\hat{\pi}_{vt}\}_{t=1}^{T}$ are parameters which define the parametric form of the function estimate $\hat{f}$ for each subregion D_v separately.

[1] Recursive partitioning regression strategy has been proposed by Morgan and Sonquist [17]. More details on RPR can be found in works of Brieman and Meisel [5] or Friedman [12].

The most frequently used choice (e.g. proposed by Morgan and Sonquist [17]) of a function g_v is a simple constant function for each disjoint subregion D_v. The regression surface is than estimated as a piecewise constant without continuity conditions between two regions which could be a problem especially if the true underlying regression function f is continuous.

Once the estimate of the regression function f is defined as a piecewise constant over disjoint subregions D_v one can formulate the partitioning regression problem as a problem to derive a good set of basis functions (which means to define disjoint subregions) and adjust the coefficients to best fit the data.

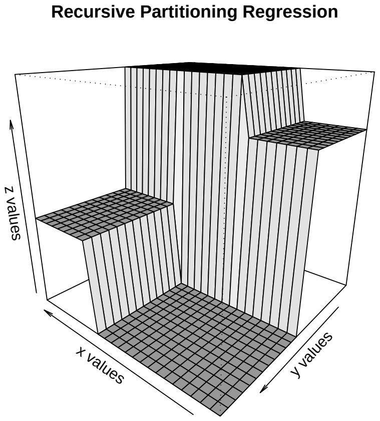


Figure 9.1: Multivariate regression surface fitted by the Recursive Partitioning procedure for simulated 2-dimensional data.

The estimate of the function f given by RPR algorithm takes a form

$$\hat{f}(\mathbf{x}) = \sum_{v=1}^{V} \hat{\pi}_v B_v(\mathbf{x}), \tag{9.2}$$

where $\{B_v(\mathbf{x})\}_{v=1}^{V}$ are basis functions which take forms of simple indicator functions

$$B_v(\mathbf{x}) = \mathbb{I}_{[\mathbf{x} \in D_v]},$$

and $\{\pi_v\}_{v=1}^{V}$ are coefficients which values are optimized to give the best fit to the data.

Recursive Partitioning Regression[2] is designed to be very good at finding a local low dimensional structure in functions that show a high dimensional global dependence. It means that the multivariate regression function which usually depends on too many variables in global character can be dependent only on few variables in local character. These few variables may be different in different subregions D_v. This method is consistent and it has a powerful graphic representation as a decision tree which increases interpretability.

However, many elementary functions are awkward for this method and it is difficult to discover when the fitted piecewise constant model approximates a standard smooth function. Another problem is a discontinuity of the estimate given by RPR even if the true underlying regression function f is continuous.

[2] A complex view on Recursive Partitioning Regression Method was given e.g. by L. Breiman, J. Friedman, R. Olshen and C. Stone [4]. More details on Recursive Partitioning Regression especially partitioning algorithms can be found in J. Friedman [13].

9.2 Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (MARS) is a nonparametric regression procedure that makes no assumption about the underlying functional relationship between dependent and independent variables. Instead, MARS constructs this relation from a set of coefficients and basis functions that are entirely "driven" from the regression data. In a sense, the method is based on the "divide and conquer" strategy, which partitions the input space into regions, each with its own regression equation. This makes MARS particularly suitable for problems with higher input dimensions (i.e., with more than 2 variables), where the curse of dimensionality would likely create problems for other techniques.

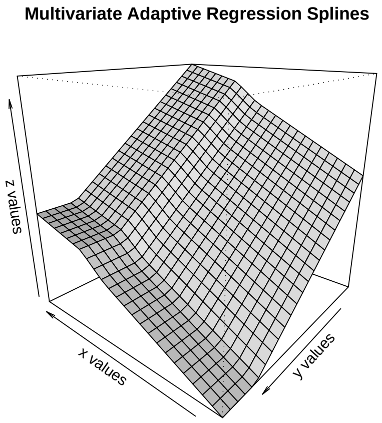


Figure 9.2: Multivariate ($J = 2$) Regression surface with the same simulated data as in a case of partitioning regression fitted by MARS Algorithm.

The same process is done by the Recursive Partitioning Regression algorithm but unlike recursive partitioning MARS approach was established to produce continuous models with continuous derivatives[3].

Actually, MARS algorithm[4] can be thought as a simple generalization of Partitioning Regression Algorithm (See J. Friedman [13] for details) where we replace a discontinuous step function in the algorithm by a continuous function with continuous derivatives. Usually we take a truncated two sided function of the form

$$K_n^{\pm}(x-t) = \left[\pm(x-t)\right]_+^n , \qquad (9.3)$$

where $K_n^{\pm}(x-t)$ is a univariate continuous function and n is an arbitrary integer value[5] and t can be thought as a knot or split location. By including such functions into the MARS algorithm we ensure continuous conditions for spline estimates and their derivatives up to degree $n-1$. The estimate computed by the Recursive Partitioning Regression can be thought as a spline of zero degree.

[3] The number of continuous derivatives depends on the degree of spline basis used to fit the data. If basis of n^{th} degree is used then all derivatives up to order $n-1$ will be continuous.

[4] MARS algorithm has been successfully generalized by S. Bakin, M. Hegland and M. Osborne, using B-splines instead of truncated power basis (CompStat 1999).

[5] The choice of the number n is influenced by a requirement on number of continuous derivatives - it is the order of the spline approximation. The most common choice is $n = 3$.

The idea of MARS fitting algorithm is to use basis functions as tensor products of univariate spline functions. Once we use two sided truncated functions as defined by (9.3) to create multivariate basis functions we can write the basis functions in the following form

$$B_v^{(n)}(\mathbf{x}) = \prod_{k=1}^{K_v} \left[s_{kv} \cdot (x_{j(kv)} - t_{kv})\right]_+^n . \tag{9.4}$$

The number K_v in (9.4) is the quantity of recursive splits in the MARS algorithm that gave rise to spline functions $B_v^{(n)}(\mathbf{x})$. The term s_{kv} takes values (± 1) and a concrete value depends on splitting phase k and the order v of the spline basis function $B_v^{(n)}(\mathbf{x})$. The value $x_{j(kv)}$ is the j^{th} component of the vector $\mathbf{x}$ and the number t_{kv} is a real valued and defines the split boundary for $x_{j(kv)}$.

The result of applying Multivariate Adaptive Regression Splines algorithm is a general model in the form

$$\hat{f}(\mathbf{x}) = \hat{\pi}_0 + \sum_{v=1}^{V} \hat{\pi}_v \cdot B_v^{(n)}(\mathbf{x}), \tag{9.5}$$

where π_0 is the coefficient of the constant basis and the sum is over all basis functions $B_v^{(n)}(\mathbf{x})$ constructed by (9.4).

MARS algorithm as just mentioned is often discussed as Multivariate Additive Regression Splines algorithm[6]. MARS algorithm gives an approximation to a function in additive form indeed or the estimate of the true underlying function is in additive form. If all basis functions that involve identical predictor variable will be put together into one sum MARS model as defined by (9.5) can be rewritten into the form

$$\hat{f}(\mathbf{x}) = \hat{\pi}_0 + \sum_{K_v=1} \sum_{j \in V_v} \hat{\pi}_v B_v^{(n)}(x_j) + \sum_{K_v=2} \sum_{j,k \in V_v} \hat{\pi}_v B_v^{(n)}(x_j, x_k) + \ldots , \tag{9.6}$$

where the first sum is over all basis functions that involve only a single variable and the second is over single basis functions which include only x_j. In the second term the first sum is over all basis functions that involve two variables - two variable interactions and the second sum is over all basis which include variables x_j and x_k. By the similar way one can include three variable interactions, etc.

Once a notation (9.6) is used then the interpretation of the MARS model becomes facilitated. Such a representation identifies the particular variables the level of interactions and the other variables that participate in them.

[6] Multivariate Additive Regression Splines or the additive form of the estimated regression function given by MARS algorithm is known as MARS ANOVA Decomposition (Friedman [13]).

Chapter 10

Projection Pursuit Regression

Projection pursuit regression can be viewed as a low dimensional expansion method where the low-dimensional arguments are not prespecified but instead are adjusted to best fist the data. This method models the regression surface as a sum of general smooth functions of linear combinations of the predictor variables in an iterative manner. It constructs a model of the regression surface based on projections of the data usually into the one dimension. For our purposes we will assume that general smooth functions used in Projection Pursuit Regression are estimated as splines, B-splines[1] respectively.

Let $(\mathbf{X}, Y)$ be a pair of random variables such that $\mathbf{X} = (X_1, \ldots, X_J)$, where $X_j \in [0,1]$ and let $Y \in \mathbb{R}$. Consider a random sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, N$, where all $(\mathbf{X}_i, Y_i)$ have the same distribution as $(\mathbf{X}, Y)$. Let f is a regression function of Y on $\mathbf{X}$ such that $f(\mathbf{x}) = \mathsf{E}[Y|\mathbf{X} = \mathbf{x}]$. Then the approximation of a regression function f based on the Projection Pursuit approach assumes the following form:

$$\widehat{f}(\mathbf{x}) = \sum_{v=1}^{V} g_v\left(\mathbf{b}_v^{\mathbf{T}}\mathbf{x}\right) = \sum_{v=1}^{V} g_v\left(\textstyle\sum_{j=1}^{J} b_{vj}x_j,\right) \tag{10.1}$$

where functions g_v are smooth functions (according to the assumption, g_v are spline functions, B-splines respectively) and $\mathbf{b}_v^{\mathbf{T}}\mathbf{x}$ denotes the inner product (it is a projection of J-dimensional vector $\mathbf{x}$ into $\mathbb{R}^1$). It was shown (Diaconis and Shahshahani [7]) that any smooth function of J variables can be represented well enough by equation (10.1) for $V \in \mathbb{N}$ large enough. The effectiveness of the approach lies in the fact that even for relatively small values of $V \in \mathbb{N}$ many classes of functions can be closely fit by approximations given by Projection Pursuit Regression method. The main advantage of the projection pursuit regression method is an easy interpretation of interactions in a model. Standard additive

[1] In statistics also another methods are used to estimate smooth functions which appear in Projection Pursuit Regression. (e.g. nearest neighbor, kernel estimate, local averaging, etc.)

models as mentioned above approximate the regression surface as a sum of functions of the individual predictors. In such a definition of the additive model it is difficult to include interactions (interactions can be included as separated functions of multiple predictors).

10.1 PPR Algorithm

The projection pursuit algorithm was proposed by Jerome H. Friedman and Werner Stuetzle [14]. The algorithm works in an iterative manner which means one formulates a hierarchy of models of increasing complexity[2]. At each step of the algorithm the model of the subsequent level of the hierarchy that best fits the data is selected. The aim is to find a particular model that when estimated from the data, best approximates the regression surface. Algorithm attempts to overcome the limitations of the recursive partitioning and some problems arising with splitting of regions and reducing the sample.

1. Step (Projection Pursuit Algorithm)

We assume that the response variable Y is centered ($\sum_i Y_i = 0$). The algorithm is a recursive method hence, we assume we already have determined the first $\nu - 1$ terms, which means we have determined vectors $\mathbf{b}_v$ and smooth functions g_v for $v = 1, \ldots, \nu - 1$, in the approximation defined by (10.1). Let $r_i^{\nu-1}$ are partial residuals of the regression fit defined by (10.2) after the first $\nu - 1$ cycles.

$$r_i^{\nu-1} = Y_i - \sum_{v=1}^{\nu-1} g_v(\mathbf{b}_v^{\mathbf{T}}\mathbf{X}_i) \tag{10.2}$$

2. Step (Projection Pursuit Algorithm)

Let $\mathbf{b} \in \mathbf{R}^J$ be any unit vector and let $c_i = \mathbf{b}^{\mathbf{T}}\mathbf{X}_i$ is a projection of J-dimensional vectors $\mathbf{X}_i$ into $\mathbb{R}^1$. Consider a simple regression problem of estimating a regression function $g(c_i) = \mathsf{E}[\ r_i^{\nu-1} \mid c_i\ ]$. Once a smooth estimate of the regression function g is given one can calculate the sum of the squared residuals related to function g and the sum is minimized subject to all possible vectors $\mathbf{b} \in \mathbb{R}^J$.

$$Minimize \quad \sum_{i=1}^{N}(r_i^{\nu-1} - g(\mathbf{b}^{\mathbf{T}}\mathbf{X}_i))^2 \tag{10.3}$$

[2] The complexity of a regression model can be thought as a number of degrees of freedom used to fit the final model.

Once we posed a condition on regression functions g_v to be a spline estimate (B-splines respectively) we have to consider a partial regression problem mentioned in the second step as a spline regression problem for every cycle in the algorithm.

3. Step (Projection Pursuit Algorithm)

When we find the vector $\mathbf{b} \in \mathbb{R}^J$ which minimize (10.3) we define the vector $\mathbf{b}_\nu$ and the function g_ν as follows:

$$\mathbf{b}_\nu := \mathbf{b} \qquad g_\nu := g \tag{10.4}$$

The previous process is iterated until the improvement defined by (10.3) becomes acceptable small.

The algorithm directly follows the successive refinement concept. The models at the ν^{th} level of the hierarchy are sums of ν smooth functions (splines or B-splines respectively) of arbitrary linear combinations of the predictors.

10.2 PPR Properties

Although, simple in concept, Projection Pursuit Regression method overcomes many limitations of other nonparametric regression approaches. PPR does not require a specification of a metric in the predictor space. Unlike recursive partitioning, PPR does not split the sample, thereby allowing, when necessary, more complex models.

I have already mentioned a simplicity of including interactions into the fit estimated by Projection Pursuit Regression method. For example, consider a simple interaction between two variables: $Z = X_1X_2$. A standard additive models can not represent this multiplicative dependence by a simple inclusion into the fit.

Once we use Projection Pursuit Regression method to estimate a regression function f we can express the interaction Z in a term (10.1) for a special choice of vectors $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^J$ such that $\mathbf{b}_1 = (1/2, 1/2, 0, \ldots, 0)$ and $\mathbf{b}_2 = (1/2, -1/2, 0, \ldots, 0)$. For the optimal choice of smooth functions $g_1(y) = y^2$ and $g_2(y) = -y^2$ we obtain a dependance of the response variable Y on the interaction Z. By an analogous way any similar interaction can be included in the final fit.

To be objective it has to be mentioned that Projection Pursuit Regression method brings also some problems when fitting a multidimensional regression surface. There exist some simple functions that require relatively too large V for good approximation given by PPR. Therefore sometimes it is not suitable to use Projection Pursuit Regression because of wicked interpretation.

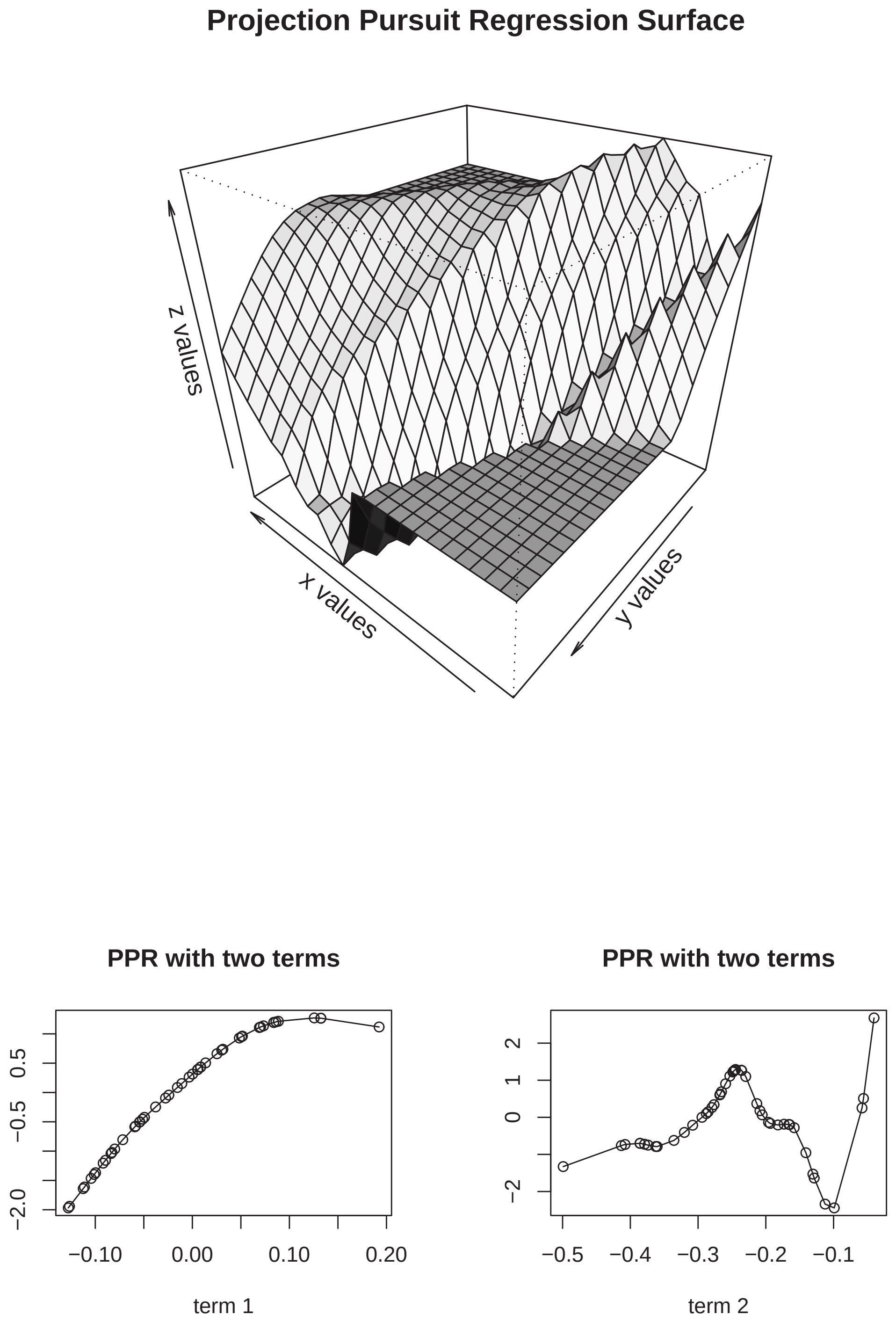


Figure 10.1: Projection Pursuit Regression model fitted from simulated 2-dimensional data where the regression surface was given by PPR algorithm as a sum of two smooth functions (estimated by smoothing spline method with penalty - two bottom plots) of one dimensional projection, where the projection vectors are $\mathbf{b}_1 = (0.3416, -0.9398)$ and $\mathbf{b}_2 = (-0.6320, 0.7749)$.

Chapter 11

Application in Real Regression Problem

A simple example of using multivariate nonparametric regression techniques based on spline approaches is given in this section. Examples are not supposed to serve as a manual for nonparametric regression it is just a brief overview of using different methods in multivariate spline regression. Data used for this example were obtain from the Statistical Database of University of Massachusetts Amherst[1]. We want to search for the dependence of the average Life Expectancy in forty largest countries in the world on two predictor variables (the number of people per TV and the number of people per physician in each country).

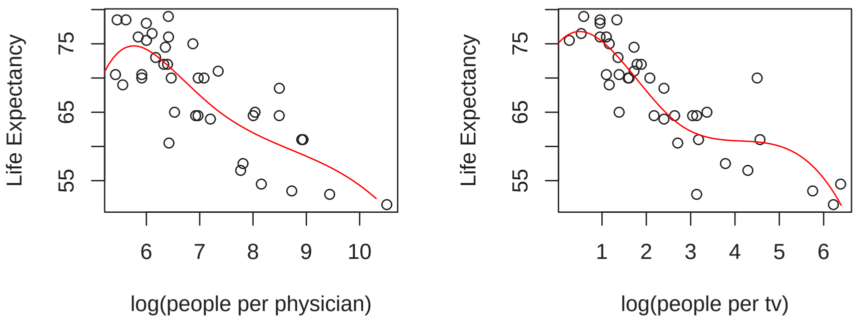
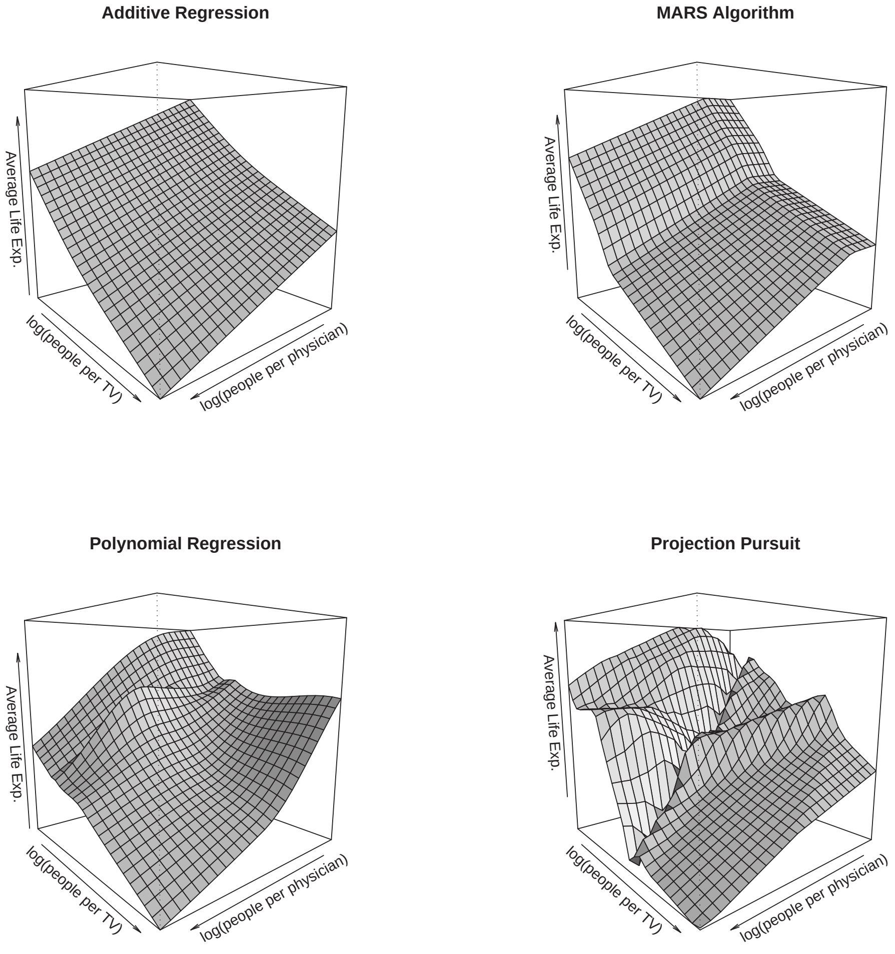


Figure 11.1: The dependence of Life Expectancy on two predictors (people per physician and people per TV). The regression curves were fitted by nonparametric B-splines of the 3^{rd} degree for both cases. For both predictors the logarithm transformation has been used.

[1] Data I used for this example were designed for multivariate nonparametric regression. More notes can by found in Database archive of University of Massachusetts USA on the following web site *http://www-unix.oit.umass.edu/ statdata/statdata.* I used low dimensional data because of better graphical interpretation (3D Plots).

From the Figure 11.1 we can measure the dependence on singe variables. However, if one wants to find a dependence on both variables at the same time one has to employ multivariate regression techniques. The next regression surfaces were fitted by four multivariate regression methods based on spline approach.


Additive Regression

MARS Algorithm

Polynomial Regression

Projection Pursuit

Figure 11.2: Multivariate regression surfaces fitted by four different regression methods.

From the figure 11.2 we can make a decision about interpretability of proposed regression models. While additive regression method and MARS Algorithm give pretty easy interpretable regression surfaces (in general one can say that with an increasing number of TVs (it could be thought as an increasing standard of living style) and an increasing number of physicians (which could be thought as an increasing medical care in a country) the average life expectancy increases) polynomial regression and PPR Algorithm give much more precise regression surface but a cost of that is a wicked interpretability (almost impossible in a case of PPR Algorithm). The lowest variance in estimation in the additive surface is supply by the highest bias in estimation. On the other hand Projection Pursuit Regression can give the best fit to the data (the lowest bias in estimation) but the cost of that is the highest variance in estimation (Bias - Variance Trade Off).

Additive models are usually preferred because of their easy interpretation. Even if the high dimensional regression surfaces fitted be the additive regression method one can always extracts functional components which allows for an easy interpretation of the dependence on single predictor variables.

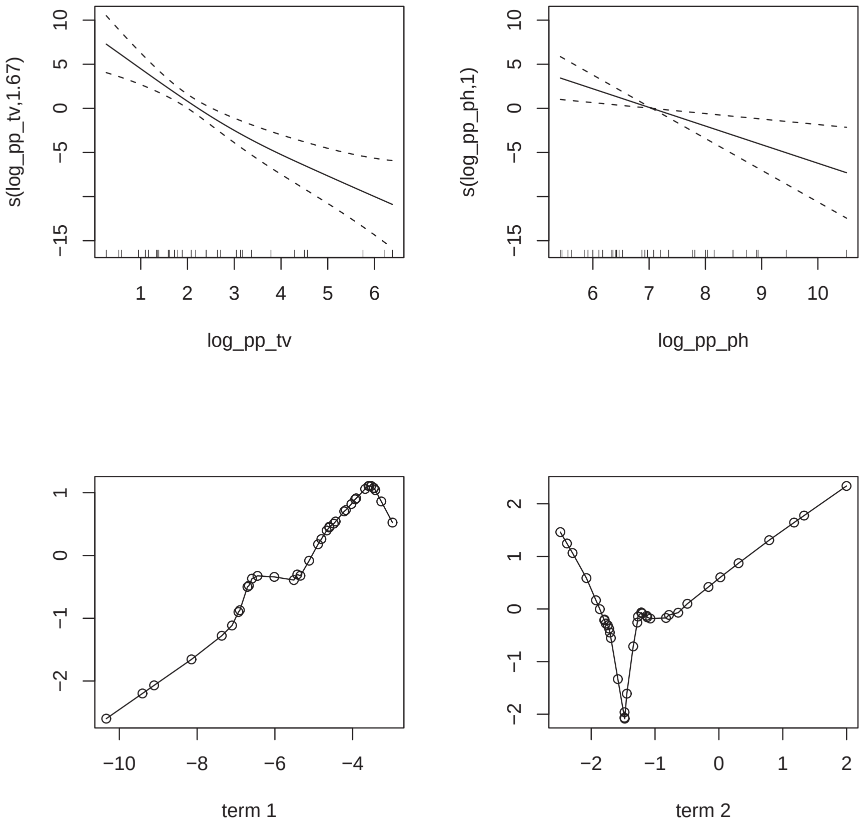


Figure 11.3: The first two plots give a dependence of the average life expectancy on single predictor variables as estimated by additive regression and 95% confidence area for both functional components. The dependence given by Additive Regression is almost linear. The last two plots give a dependence of the average live expectancy on both predictor variables (at the same time) projected into the one dimensional space where the projection vectors are $\mathbf{b}_1 = (-0.8894247, -0.4570817), \mathbf{b}_2 = (0.8916017, -0.4528206)$. Such a dependence is difficult to interpret (almost impossible) once the number of term functions is more than two.

The same dependence for single predictor variables (the first two plots in the Figure 11.3) can be extract also from the regression surface fitted by MARS but MARS algorithm implicitly includes all order interactions (in this case MARS includes the second order interactions) therefore the regression surface fitted by MARS is more precise once the interactions are present.

Of course, one can include the interaction terms in the additive model too (up to the same order as in MARS) - then the regression surfaces fitted by MARS algorithm and the additive regression method are very similar.

To compare four regression methods used to estimate the dependence of the average Life Expectancy on both predictor variables (people per TV, people per physicians) the residual sum of squares are displayed below for each method used to estimate the regression function:

```
Additive Regression:                   MARS Algorithm:
   [1]  11.89261                         [2] 11.32777

Polynomial Regression:                 PPR Regression:
   [3] 10.56021                          [4] 6.129001
```

Regarding to the fact that Projection Pursuit Regression usually gives the best fit to the data the residual sum of squares is nearly always the lowest for PPR. However, it is rather due the overfitting the data, therefore, if one considers the ability to interpret the researched dependence, residual sum of squares is not an optimal decision criterion anyway.

Conclusion

In this work I tried to give an insight on using of nonparametric regression estimation techniques with a special emphasis on spline approaches used in multivariate regression. I have mentioned problems which arise in multivariate regression (curse of dimensionality, low interpretation ability) and I proposed some recommended methods how to avoid most of problems. The multivariate additive regression splines approach was presented as a successful tool for multivariate regression which offers a great interpretability, simpleness of fitting a regression model and a convenient limitation of many multivariate problems. Function estimates computed by additive splines are smooth functions of predictor variables even with continuous derivatives up to the specific order. A proof that spline estimates are consistent (with the same rate of convergence for any dimension) was given for additive regression splines. Moreover, additive splines are great at simple interpretability of single dependencies between the response variable and predictor variables. Another advantage is a straightforward identification of interactions as a simple function of included variables.

I have discussed also some sophisticated methods for multivariate regression analysis as Projection Pursuit Regression which can be thought as spline regression approach if one defines the estimates of the term functions to be splines. However, using of Projection Pursuit Regression algorithm has an unseemly property of very low interpretation ability once that more than just one term function is used to estimate the primary dependence. Another discussed algorithm used for multivariate regression is MARS algorithm which can be also thought as splines regression strategy moreover, the estimate given by MARS can be thought as an additive spline approximation.

In the end some simple examples have been shown to present the using of multivariate regression techniques in real situations. The examples are not supposed to be a manual for solving regression problems it just should serve as a brief overview on different procedures and their properties.

Software

Software **R 2.2.0** and the following packages:

- STATS - the standard package includes splines and PPR algorithm
- SPLINES, PSPLINE - smoothing splines, interpolation splines
- LMESPLINES, XGOBI - some additional spline functions
- MGCV - additive regression methods
- RPART - recursive partitioning regression and regression trees
- MDA - Multivariate Adaptive (Additive) Regression Splines

All another functions and procedures which have been used for this diploma thesis and are not listed in packages mentioned above were programmed by myself. The source codes are attached on CD on the last cover (together with PDF version, figures used in diploma thesis and all required packages for R 2.2.0).

Data which has been used to present some multivariate regression techniques in Chapter 11 are also uploaded on the attached CD together with the original description from the Database Archive of the University of Massachutsetts, USA.

Bibliography

[1] R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[3] C. De Boor. *A Practical Guide to Splines.* Springer, Berlin, 1978.

[4] L. Breiman, J. Friedman, R. Olshen, and Ch. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, California, 1984.

[5] L. Brieman and W. Meisel. General estimates of the intrinsic variability of data in nonlinear regression models. *Journal of the American Statistical Association*, (No. 71):301–307, 1976.

[6] P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Journal of Numerical Mathematics*, (No. 31):377–403, 1979.

[7] P. Diaconis and M. Shahshahani. On nonlinear functions of linear combinations. *SIAM J. Sci. Statist. Comput.*, (No. 5):175–191, 1984.

[8] P. Dierckx. *Curve and Surface Fitting with Splines.* Clarendon, Oxford, 1993.

[9] B. Efron and C. Stein. The jackknife estimate of the variance. *The Annals of Statistics*, Vol. 9:586–596, 1981.

[10] P. H. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, Vol. 11(No. 2):89–102, 1996.

[11] R. L. Eubank. *Spline Smoothing and Nonparametric Regression.* Dekker, New York, 1988.

[12] J. H. Friedman. A tree-structured approach to nonparametric multiple regression: In smoothing techniques fot curve estimation. 1979.

[13] J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, Vol. 19(No. 1):1–141, 1991.

[14] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, Vol. 76(No. 376):817–823, Dec. 1981.

[15] L. Gyorfi, A. Krzyzak, M. Kohler, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, 2002.

[16] M. Hegland and V. Pestov. Additive models in high dimensions. *ANZIAM Journal*, (No. 46):1205–1221, 2005.

[17] J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of The American Statistical Association*, Vol. 58:415–434, 1963.

[18] F. O'Sullivan. A statistical perspective on ill-posed inverse problem. *Statictics Sci.*, (No. 1):505–527, 1986.

[19] B. W. Silverman. Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, Vol. 12(No. 3):898–916, 1984.

[20] P. L. Smith. Splines as a useful and convenient statistical tool. *The American Statistician*, Vol. 33(No. 2):57–62, 1979.

[21] Ch. J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, (No. 5):549–645, 1977.

[22] Ch. J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, (No. 8):1348–1360, 1980.

[23] Ch. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, Vol. 10(No. 4):1040–1053, 1982.

[24] Ch.s J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, Vol. 13(No. 2):689–705, 1985.

[25] E. J. Wegman and I. W. Wright. Splines in statistics. *Journal of the American Statistical Association*, Vol. 78, Jun 1983.

[26] S. Wold. Spline function in data analysis. *Technometrics*, Vol. 16(No. 1), February 1974.

[27] S. Wold and G. Wahba. Completely automatic french curve: Fitting spline functions by cross-validation. *Communications in Statistics*, Vol. 4:1–17, 1975.

Index