UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI" – MILANO

Facoltà di Economia

Dottorato di Ricerca in Statistica

XIX Ciclo

Penalized Estimation in Single Index Models under Monotonicity Constraint

Coordinatore:

Ch.mo Prof. Pietro Muliere

Tesi di:

Gabriella Mostallino

N.935809

UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"
ISTITUTO DI METODI QUANTITATIVI

The thesis "**Penalized Estimation in Single Index Models under Monotonicity Constraint**" by **Gabriella Mostallino N.935809** is recommended for acceptance by the members of the delegated committee, as stated by the enclosed reports, in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Dated: January 2007

Research Supervisor: **Jon A. Wellner**

External Examiners: **Walter Racugno**
Piero Veronese

UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"

Date: **January 2007**

Author: **Gabriella Mostallino N.935809**

Title: **Penalized Estimation in Single Index Models under Monotonicity Constraint**

Department: **Istituto di Metodi Quantitativi**

Signature of Author

Alla mia famiglia

Table of Contents

Preface

Semiparametric regression models are widely theoretically studied, but are much underused in the applications. Indeed, in comparison with the linear regression model $Y = \underline{\theta}'\underline{X}+\epsilon$, semiparametric techniques are theoretically sophisticated and often require substantial programming experience. These models are of interest for statistically oriented scientist as biostatisticians, econometricians or epidemiologist who usually have a good working knowledge of linear models but desire to use more flexible semiparametric models.

In this thesis we discuss estimation in special models of multivariate semiparametric regression: the single-index models $Y = r(\underline{\theta}'\underline{X}) + \epsilon$, which applies a nonparametric function to the linear single-index $\underline{\theta}'\underline{X}$. This regression model is a natural extension to the linear model $Y = \underline{\theta}'\underline{X} + \epsilon$. It allows greater flexibility and overcomes the problem of the "curse of dimensionality" that we have in nonparametric regression with the inclusion of multiple explanatory variables. We will study the special single-index model in which the regression function is such that its second derivative is square integrable and is subject to the monotonicity constraint.

The theory of the procedures in semiparametric estimation is necessarily based on asymptotic approximations, while actual performance for finite sample sizes is often gauged best by simulations. Therefore our focus is, first of all, on asymptotic theory. We limit ourselves to models for independent, identically distributed observations, the basic building blocks of most models for data.

We begin by explaining the reasons that lead to a choice of semiparametric regression, such as non flexibility of parametric regression and the "curse of dimensionality" in nonparametric regression. Continuing in the description of the single-index model, we illustrate that, as a compromise between parametric and nonparametric models, it overcomes the problems that the other approaches present. Moreover, we present a review of the main estimation procedures known in the literature.

Chapters 2 and 3 are devoted to give a background about some notions of semiparametric and nonparametric theory, needed in the study of the proposed estimator, in the single-index model.

For the model, object of study in this thesis, chapter 4 introduces the penalized maximum likelihood estimator $(\hat{\underline{\theta}}, \hat{r})$. Theoretical results are illustrated about the existence of this estimator $(\hat{\underline{\theta}}, \hat{r})$, when the regression function r to be estimated, is restricted to be a monotone function. Moreover some properties of consistency for $(\hat{\underline{\theta}}, \hat{r})$ and ($\hat{\underline{\theta}}$ and $\hat{r}$) separately investigated and the asymptotic efficiency of the Euclidean parameter $\hat{\underline{\theta}}$ are shown.

Chapter 5 explains the role of the smoothing spline estimators in nonparametric regression and how it is possible to use them in the case of penalized single index models when the regression function is supposed to be monotone. We will see that, surprisingly, these estimators minimizing the penalized maximum likelihood criterion over infinite-dimensional classes of smooth functions, can be obtained by solving finite-dimensional optimization problems.

In the chapter 6, the attention is focused in computational studies and numerical results. Indeed an algorithm is described, that gives an approximated value for the estimator $(\hat{\underline{\theta}}, \hat{r})$ and some simulation studies illustrate how this algorithm works. Finally, it is exposed an example with real data for an investigation in an environmental study on how the concentration of the air pollutant ozone depends on three other meteorological variables.

Acknowledgements

I would like to thank and acknowledge my gratitude to Professor Jon Wellner for his helpful suggestions, stimulating advice in making research for my thesis. Not only he orientated me toward the topic of the thesis, but he patiently drove me in the development of the work, during my visiting period at the University of Washington. Particular thanks to the Department of Statistics of Washington University for its kind hospitality and the Istituto di Metodi Quantitativi of Università Commerciale “Luigi Bocconi" for the opportunity to follow my studies of PhD. I am also grateful to Professor Walter Racugno for his support and encouragements during my Ph.D. studies and to Professor Maria Luisa Targhetta, that provided me thought-provoking interest for the Statistics. Finally I am indebted with all special people that I met and knew, making these years a glad period of my life. Among the others, thanks to Chiara, Davide, Lisa, Massimiliano, Petra and Raffaele.

Chapter 1

Single-Index Models

1.1 Introduction

Regression analysis concerns the relationship between two random vectors $\underline{Y} \in \mathbb{R}^k$ and $\underline{X} \in \mathbb{R}^m$ where $\underline{Y}$ is a vector of response variables and $\underline{X}$ is a vector of explanatory variables. Statistical inference in parametric, nonparametric and finally semiparametric regression are particular aspects of regression analysis.

Considering one-dimensional response variable $Y \in \mathbb{R}$, we can explicitly summarize the estimation in parametric, nonparametric and semiparametric regression, in the estimation of the *regression function* $r : \mathbb{R}^m \to \mathbb{R}$ when the relationship between the outcome Y and the vector of regressors $\underline{X}$ is stated in the form

$$Y = r(\underline{X}) + \epsilon, \tag{1.1.1}$$

where ϵ is, for example, a random variable with zero mean and finite variance. It is clear that the regression function is equal to $\mathrm{E}(Y|\underline{X} = \underline{x})$ provided that $\mathrm{E}(\epsilon|\underline{X} = \underline{x})$ is zero, which is in general assumed.

It is well known that several approaches can be considered to tackle the problem, in fact the determination of a suitable inferential methodology for model (1.1.1) will hinge on the assumptions it is possible to make about r.

When the form of r and the distribution of ϵ are known except for finitely many unknown parameters, the model (1.1.1) is a *parametric regression model*. In this case , it is considered a vector of parameters $\underline{\theta} \in \underline{\Theta} \subseteq \mathbb{R}^M$, $M > 0$ and a parametric regression model inference about r is therefore tantamount to inference about $\underline{\theta}$.

Regression analysis techniques for parametric models represent one approach to conducting inference about r. Using an appropriate estimation methodology, such as least-squares or maximum likelihood methods, it is possible to utilize the data to estimate the parameters $\underline{\theta}$ and thereby estimate the regression function r. The result is a fitted curve that has been selected from the family of curves allowed under the model and conforms to the data in some fashion.

Parametric regression model is undoubtedly the most relevant approach to regression analysis because of the easiness that this model presents in terms of computation and interpretation of the results, but it admits an important drawback in the lack of flexibility. This is one of the reasons for which we use other methods of fitting curve of data. One collection of procedures that can be used for this purpose are nonparametric techniques. These methods give estimates of r that allow great flexibility in the possible form of the regression curve and, in particular, make no assumptions about a parametric form. In fact, in some situations, to force the regression function to belong to a parametric family of functions can be too restrictive and this can lead to an important modeling bias and wrong conclusions about the link function between Y and $\underline{X}$. On the other hand, the nonparametric approach releases such restrictive functional hypothesis about r.

A *nonparametric regression model* generally only assumes that the regression curve belongs to some infinite-dimensional collection of functions. For example, r may be assumed to be differentiable or differentiable with a square integrable second derivative or with constraints of monotonicity. Assumptions of this type are concerned with

qualitative properties of r and are in contrast to the assumptions used in parametric modeling that entails a much greater level of specificity about the regression function. Note that the concepts of smoothing and monotonicity are central ideas in statistics. Theirs role is to extract structural elements of variable complexity from patterns of random variation. The nonparametric smoothing concept and the concept of monotonicity are designed to simultaneously estimate and model the underlying structure. This involves high-dimensional objects, like regression surfaces. Such objects are difficult to estimate for data sets with mixed, high-dimensional and partially unobservable variables. Moreover in nonparametric regression, smoothness conditions (in particular, the existence of bounded derivatives) play a central role in ensuring consistency of the estimator. They are also critical in determining the rate of convergence as well as certain distributional results. Additionally, with sufficient smoothness, derivatives of the regression function can be estimated consistently, sometimes by differentiating the estimator of the function itself. About constraints such as monotonicity or concavity, we have that they do not improve the (large sample) rate of convergence if enough smoothness is imposed in the model. They can improve performance of the estimator chosen, if strong smoothness assumptions are not made or if the data set is of moderate size.

The greater flexibility in nonparametric estimation has nevertheless, a high cost that is when the number of the regressors is increasing, these methods get very demanding with respect to the number of the observations. Specifically, from Stone (1980) we have that the larger the number of the regressors, the larger the dimension of data samples needed in order to achieve reasonable estimates. For sample sizes that we have to face in practice, this is translated into critically bad estimates once the number of the regressors is greater than two or three. This phenomenon that is known as the *curse of dimensionality* and that motivates the need of dimension reduction methods,

is explained in more detail in the following section.

Consequently, researchers have tried to develop models and estimators which offer more flexibility than standard parametric regression but overcome the curse of dimensionality by employing some form of dimension reduction. Examples of dimension reduction methods are the semiparametrics methods.

An important difference between parametric and nonparametric regression methodologies is then their respective degree of reliance on the information about r obtained from the experimenter and from the data. Indeed with the purpose to specify a nonparametric regression model, an appropriate function space that is believed to contain the unknown regression curve, will need to be chosen. Is the experimenter that usually motivates this choice only by smoothness or of monotonicity properties, the regression function can be assumed to possesses. The data is then utilized to determine an element of this function space that is representative of the unknown regression curve. In contrast, under a parametric model, the experimenter chooses one possible family of curves, from the collection of all curves, and inputs this choice into the inferential process. The information the data can supply concerning model development is then restricted to what can be extracted from the data under this assumed parametric form. Then, nonparametric regression techniques rely more heavily in the data for information about r than their parametric counterparts, where the choice of the experimenter predominates.

If it is selected in an appropriate way, parametric models have some definite advantages. The corresponding inferential methods usually have nice efficiency properties. Also, the parameters may have physical meaning which makes them interpretable and of interest in their own right.

Unfortunately parametric models are often used when there is a little available information concerning the functional form of r. In these cases, the function r is assumed

to have some exact parametric representations with little or no knowledge concerning the accuracy of the assumptions. Therefore, if the assumed parametric model is in error, we don't have the mentioned advantages of the parametric approach. The use of an inappropriate parametric model can be quite dangerous in the sense that it can produce misleading or incorrect inference about the regression curve. This happens in particular, when there is little that is known about regression function. In such cases, the information about r lies in the data rather than in the subjective assumptions. Accordingly, it seems more reasonable to use inferential techniques which rely heavily on the data. For this reason, the nonparametric techniques are ideally suited to problems of inference when the available knowledge about r is limited.

Nonparametric regression overcomes the difficulty with parametric techniques that require that the functional form of r must be known. But, generally, nonparametric estimators are less efficient than the parametric variety when the parametric model is valid.

For most parametric estimators the risk, or expected squared error of estimation, will decay to zero at a rate of n^{-1}. On the other hand, the corresponding rate for nonparametric estimators is usually $n^{-\delta}$, for $\delta \in (0, 1)$ that depends on the smoothness of r. For example, if r is twice differentiable $n^{-4/5}$ is an often quoted rate. Thus, nonparametric regression techniques suffer a loss of efficiency when compared to parametric methods. Then, nonparametric estimators become candidates for estimation of r only when there is some question about an appropriate parametric form for r.

The result of a nonparametric regression analysis is a curve fitted to a set of data. Since this curve is produced without assuming a parametric form for r there will be some loss in the interpretability of estimators obtained. Indeed, in this case there will no longer be quantities such as estimated regression coefficients to be interpreted. However, the fitted curve itself is an estimate of the infinite-dimensional parameter r

and any functional of r is also a parameter which can be estimated using an estimate of the regression curve. This type of parameters can be estimated and interpreted from either a nonparametric or parametric viewpoint. Thus, there are regression curve based parameters, such as functionals of r, about which inference can be made and interpretations can be drawn using either parametric or nonparametric techniques.

To summarize, parametric methods require very specific, quantitative information about the form of r and place restrictions on what the data can tell us about the regression function. In contrast, nonparametric regression techniques rely only on qualitative information about r and let the data speak for itself concerning the actual form of the regression curve. These methods are best suited for inference in situations where there is little or no prior information available about regression curve.

Semiparametric estimation is, first of all, about estimation in situations when we believe we have enough knowledge to model some features of the data parametrically, but are unwilling to assume anything for other features, then the semiparametric modeling technique compromises the two aims, flexibility of nonparametric theory and simplicity of statistical procedures, by introducing partial parametric components. Further advantages of semiparametric methods are the possible inclusion of categorical variables (which can often only be included in a parametric way) and an easy interpretation of the results, besides the possibility of a part specification of a model.

Then semiparametric models propose a mix of parametric and nonparametric approaches, which permits to compensate for their respective drawbacks. They are characterized by a twofold parametrization, say $\underline{\delta}$ and s, where $\underline{\delta}$ lies in a finite-dimensional space $\underline{\Delta}$ and s lies in an infinite-dimensional space. For example, it should be the case if r belong to a parametric family and the distribution of ϵ was totally unknown in model (1.1.1) or in the case of the single-index model, object of interest in this thesis.

1.2 The curse of dimensionality

A problem that occurs with nonparametric regression and smoothing methods is the *curse of dimensionality*, a term usually attributed to Bellman (1961). For an extensive discussion on the curse of dimensionality, see Hastie *et al.* (2001).

Roughly speaking, this means that estimation gets harder very quickly as the dimension of the observations increases.

There are at least two versions of this curse. The first is the computational curse of dimensionality. This refers to the fact that the computational burden of some methods can increase exponentially with dimension. Our focus here is however, with the second version, which we call the *statistical curse of dimensionality*: if the data have dimension m, then we need a sample size n that grows exponentially with m.

To gain an appreciation of the problem, we begin with a deterministic framework. The objective is obviously approximate the regression function r. If it is known to be linear in one variable, two observations are sufficient to determine the entire function precisely. Three are sufficient if r is linear in two variables. If r is of the form $r(\underline{X}, \underline{\theta})$, where r is known and $\underline{\theta}$ is an unknown m-dimensional vector, then m judiciously selected points are usually sufficient to solve for $\underline{\theta}$. No further observations on the function are necessary.

Let us turn to the pure nonparametric case and suppose r, defined on the unit interval, is known only to have a first derivative bounded by a constant K. If we sample r at n equidistant points and approximate r at any point by the closest point at which we have an evaluation, then the approximation error cannot exceed $K/2n$. Increasing the number of the points reduces the approximation error at a rate $O(n^{-1})$.

Now suppose r is a function on the unit square and that it has derivatives bounded in all directions by K. To approximate the function, we need to sample throughout its

domain. If we distribute n points uniformly on the unit square, each will occupy an area $1/n$, and the typical distance between points will be $n^{-1/2}$ so that the approximation error is now $O(n^{-1/2})$. If we repeat this argument for function in m variables, the typical distance between points becomes $n^{-1/m}$ and the approximation error is $O(n^{-1/m})$. In general , this method of approximation yields errors proportional to the distance to the nearest observation.

Indeed the mean squared error of any nonparametric estimator of a smooth (for example, twice differentiable) curve has typically mean squared error of the form

$$MSE \approx \frac{c}{n^{4/(4+m)}}$$

for some $c > 0$. If we want the MSE to be equal to some small number δ, we can set $MSE = \delta$ and solve for n. We find that

$$n \propto \left(\frac{c}{\delta}\right)^{m/4}$$

which grows exponentially with the dimension m.

The reason for this phenomenon is that smoothing involves estimating a function $r(x)$ using data points in a local neighborhood of x. But in high-dimensional problem, the data are very sparse, so local neighborhoods contain very few points. However, even if we can overcome the computational problems, we are still left with the statistical curse of dimensionality. You may be able to compute a smooth nonparametric estimator but it will not be accurate.

It is possible consider different type of restrictions that substantially reduce approximation error, such as partial linear structure, additive separability or smoothness assumptions and the index model specification.

1.3 The Single-Index Model

In this thesis we are therefore interested in a special topic of regression analysis that is in the estimation in one of the most important regression model and one of the most referred semiparametric regression models in the literature: the *single-index model* (Stoker (1986), Härdle and Stoker (1989), Li (1991), Ichimura (1993)).
Ichimura (1993) gives the following definition of a single-index model:

Definition 1.3.1. *Let m and M be positive integers. The model*

$$Y = r[h(\underline{X}, \underline{\theta})] + \epsilon$$

where

- the random vector $(\underline{X}', Y)$ is such that $Y \in \mathbb{R}$ and $\underline{X} \in \mathbb{R}^m$;

- $\epsilon \in \mathbb{R}$ is an unobserved random disturbance, with $E(\epsilon|\underline{X}) = 0$;

- $\underline{\theta} \in \mathbb{R}^M$ is an unknown parameter vector to be estimated;

- the function $h : \mathcal{S} \times \underline{\Theta} \to \mathbb{R}$, for some subset $\mathcal{S} \times \underline{\Theta} \subset \mathbb{R}^m \times \mathbb{R}^M$, is known up to the parameter $\underline{\theta}$;

- the function $r : \mathbb{R} \to \mathbb{R}$ is not known

is a single-index model.

Note that the model defined like above is indeed semiparametric: $\underline{\theta}$ lies in a finite-dimensional space and the link function r belongs to a functional space so that it can be seen as an infinite-dimensional parameter and has to be estimated via nonparametric techniques. Moreover, the conditional probability of ϵ is not specified, except for $\mathrm{E}(\epsilon|\underline{X}) = 0$.
Great simplifications in most of the results can be obtained by fixing

$$h(\underline{X}, \underline{\theta}) = \underline{\theta}' \underline{X} = \sum_{k=1}^{m} \theta_k X_k$$

where θ_k and X_k represent the kth components of vectors $\underline{\theta}$ and $\underline{X}$.
Ichimura calls such a model a "linear single-index model", but by sake of simplicity and as it is frequently referred in the literature, "single-index model" will always refer to linear single-index model in this thesis.

By reducing the dimensionality from multivariate predictors to a univariate index $(\underline{\theta}'\underline{X})$, single-index models avoid the curse of dimensionality while still capturing important features in high-dimensional data. Because a nonlinear link function r is applied to the index $(\underline{\theta}'\underline{X})$, interactions between the covariates can be modeled. Thus single-index models are a useful alternative to additive models, which also reduce dimensionality but do not incorporate interactions.
Single-index model therefore relaxes some of the restrictive assumptions of parametric models, thus ensuring some flexibility and mitigating the risk of misspecifying the link function, while avoiding the curse of dimensionality. Horowitz and Härdle (1996) have shown that misleading results are obtained if a binary probit model is estimated by specifying the cumulative normal distribution function as the link function rather than estimating r by nonparametric methods.

The motivation, importance, and a broad potential applications of single-index model are widely discussed in the literature. Härdle and Stoker (1989) and Ichimura (1993) have given examples of classical regression, discrete regression, and censored regression that can be classified as single-index models. Single-index model can be seen as a generalization of linear regression models by replacing the linear combination $(\underline{\theta}'\underline{X})$ with a nonparametric component, $r(\underline{\theta}'\underline{X})$. Also, the single-index model generalizes both the generalized linear model (GLIM) (McCullagh and Nelder (1983)) and the missing-link problem in GLIM (Weisberg and Welsh (1994)).

We can summarize the remarks of Li (1991) in this three points:

1. In practice, lowering dimensionality before fitting data is important and in many cases crucial for further analysis.

2. If r is monotone, as in the case of this thesis, then $\underline{\theta}$ has the same general interpretation as effect parameters as in ordinary linear models.

3. Given an estimated value of $\underline{\theta}$, the multivariate model fitting is reduced to a more manageable low-dimensional modeling problem.

Applications of single-index models lie in a variety of fields, such as discrete choice analysis in econometrics and dose-response models in biometrics studies (Härdle *et al.* (1993)) as a reasonable compromise between fully parametric and fully nonparametric modeling. They are also extensively used in projection pursuit regression (Friedman and Stuetzle (1981) and Hall (1989)).

1.4 Identification conditions

Restrictions must be imposed in order to make the finite-dimensional vector of parameters $\underline{\theta}$ and the regression function r uniquely determined by the population distribution of $(Y, \underline{X})$. It is quite clear that such conditions are needed. Suppose for example that r is a constant function on $\mathbb{R}$; in this case, any vector of $\mathbb{R}^m$ should be acceptable as estimator of $\underline{\theta}$. Moreover, as in linear model, no identification is possible if there is an exact linear relation among the components of $\underline{X}$.

More formally, let α be any constant and β be any non-zero constant. Define the function r^* by

$$r^*(\alpha + \beta t) = r(t)$$

for all t in the support of $\underline{\theta}'\underline{X}$. We have

$$\begin{aligned} \mathrm{E}(Y|\underline{X}=\underline{x}) &= r(\underline{\theta}'\underline{x}) && (1.4.1)\\ &= r^*(\alpha+\beta\underline{\theta}'\underline{x}). && (1.4.2)\end{aligned}$$

Models (1.4.1) and (1.4.2) are equivalent: they could not be distinguished, even if the whole population $(Y,\underline{X})$ was known. Indeed, the use of the vector $\beta\underline{\theta}'$ and of the rescaled link function $r_\beta(t)=r(t/\beta)$ leads to the same regression function r. Therefore, restrictions on α (location) and β (scale) have to be imposed in order to make $\underline{\theta}$ and r uniquely defined. In the remainder, $\underline{X}$ will contain no intercept (location restriction) and $\underline{\theta}$ needs to be with the first component equal to one or with Euclidean norm equal to one (scale restriction).

Besides, r must be differentiable. Indeed, note that the single-index hypothesis imposes that $\mathrm{E}(Y|\underline{X}=\underline{x})$ remains constants if $\underline{x}$ changes in such a way that $\underline{\theta}'\underline{x}$ stays constant. However, if $\underline{\theta}'\underline{X}$ is continuously distributed, the set of $\underline{X}$ values on which $\underline{\theta}'\underline{X}=c$ has probability zero for any c, so that no identification is possible. But, as in the case of this thesis, if r is differentiable then $r(\underline{\theta}'\underline{X})$ is close to $r(c)$ provided that $\underline{\theta}'\underline{X}$ is close to c. Therefore, the set of $\underline{X}$ values on which $\underline{\theta}'\underline{X}$ is within any specified non-zero distance of c has non-zero probability and identification of $\underline{\theta}$ gets possible through approximate constancy of $\underline{\theta}'\underline{X}$.

Based on the observation in Ichimura (1993), it can be stated:

Theorem 1.4.1. *$\underline{\theta}$ and r are identified if:*

- r is differentiable and not constant in the support of $\underline{\theta}'\underline{X}$;

- $\underline{X}$ admits at least one continuously distributed component;

- The support of $\underline{X}$ is not contained in any proper linear subspace of $\mathbb{R}^m$;

- $\underline{\theta}\in\underline{\Theta}$, with $\underline{\Theta}=\{\underline{\theta}\in\mathbb{R}^m : \|\underline{\theta}\|_m=1\}$.

1.5 Survey of the main known estimation procedures

1.5.1 Estimating the link function r

Suppose at first that $\underline{\theta}$ is known. Then r can be estimated by classical means of univariate nonparametric regression of Y on $\underline{\theta}'\underline{X}$. Many various methods are proposed in Härdle (1990). Although it is well known that it is not the more efficient one, the Naradaya-Watson kernel estimator is used in many situations because of its easiness of implementation and interpretation and its mathematical tractability.
Of course, these estimators $\hat{r}(\underline{\theta})$ cannot be implemented since $\underline{\theta}$ is not known. Then, if an estimator $\hat{\underline{\theta}}$ of $\underline{\theta}$ is known, r is estimated by $\hat{r}(\hat{\underline{\theta}})$ obtained using $\hat{\underline{\theta}}$ in place of $\underline{\theta}$.
Methods of estimating $\underline{\theta}$ will be described in the next section. It will be shown that $\underline{\theta}$ can be estimated with $n^{-1/2}$ rate of convergence in probability, i.e. there exist estimators $\hat{\underline{\theta}}$ such that

$$(\hat{\underline{\theta}} - \underline{\theta}) = O_P\left(n^{-1/2}\right),$$

which is the typical rate of convergence for parametric estimators. Besides, it is well known that no nonparametric estimator of regression functions can achieve this rate. The convergence of $\hat{\underline{\theta}}$ is thus faster than the fastest possible rate of convergence of any nonparametric estimator of r. Therefore, it is intuitively clear that the difference between the estimators $\hat{r}(\underline{\theta})$ and $\hat{r}(\hat{\underline{\theta}})$ is asymptotically negligible namely root-n estimation of $\underline{\theta}$ has non effect on the asymptotic distribution of the Naradaya-Watson estimator. See Horowitz (1998) for a complete argument of this result. Hence, the estimation of r is direct via standard methods once an estimator of $\underline{\theta}$ is known, so no more will be explicitly developed in the next sections.

1.5.2 Estimating the Euclidean parameter

Many methods of estimating $\underline{\theta}$ have been proposed in the literature. We resume most of them in this section. First, notice that estimators of $\underline{\theta}$ can be classified in two main groups, according to whether they require solving nonlinear optimization problem (M-estimators) or not (direct estimators).

M-estimators

If r was known, a M-estimator of $\underline{\theta}$ should typically have the form

$$\hat{\underline{\theta}} = \arg\max_{\underline{\theta}\in\underline{\Theta}} \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, r(\underline{\theta}'\underline{X}_i)),$$

where $\Psi : \mathbb{R}^2 \to \mathbb{R}$ is a function verifying some mild regularity conditions. In the single index model context, we substitute the unknown r for its leave-one-out estimator. Delecroix and Hristache (1999) give sufficient conditions on Ψ in order to make the estimator $\hat{\underline{\theta}}$ a.s. consistent and asymptotically normal, for any joint law of $(\underline{X}, Y)$. They show that it is the case if Ψ is equal to the log-likelihood of a density belonging to the exponential family.

• Semiparametric Least Squares (SLS)

As in a parametric least squares problem, the idea is to minimize the mean square distance between the observed values Y_i and the values given by the model $r(\underline{\theta}'\underline{X}_i)$. If r was known, we should have the classical least squares estimator given by

$$\underline{\theta}^* = \arg\min_{\underline{\theta}\in\underline{\Theta}} \frac{1}{n}\sum_{i=1}^{n} w(\underline{X}_i)[Y_i - r(\underline{\theta}'\underline{X}_i)]^2,$$

where w is a positive bounded weight function. Under some regularity conditions, least squares theory shows that this estimator $\underline{\theta}^*$ is root-n consistent (see Amemiya (1985)). In the single-index context, the least squares criterion to minimize uses the leave-one-out Naradaya-Watson estimator instead of r, as a trite function of $\underline{\theta}$. Ichimura (1993)

studied this method in details.

The choice of the weight function w affects the efficiency of the estimator. Newey and Stoker (1993) found the efficiency bound for semiparametric models. In a single-index context, the SLS estimator achieves this bound if

$$w(\underline{x}) = 1/\sigma^2(\underline{x}),$$

where $\sigma^2(\underline{x}) = var(Y|\underline{X} = \underline{x})$. If $\sigma^2(\underline{x})$ is unknown, a consistent estimator of $\sigma^2(\underline{x})$ has to be used.

• Semiparametric Maximum Likelihood (SML)

Another optimization based method is inspired by the parametric maximum likelihood methods. If r was known, the maximum likelihood estimator of $\underline{\theta}$ should be given by the following maximization problem

$$\underline{\theta}^* = \arg\max_{\underline{\theta}\in\underline{\Theta}} \frac{1}{n}\sum_{i=1}^{n} \log l_{r,\underline{\theta}}(Y_i|\underline{\theta}'\underline{X} = \underline{\theta}'\underline{X}_i)), \tag{1.5.1}$$

where $l_{r,\underline{\theta}}(\cdot|\cdot)$ is the conditional density of Y given $\underline{X}$.

Standard theory of maximum likelihood estimation implies that $\underline{\theta}^*$ is root-n consistent, efficient and asymptotically normal, under regularity conditions.

Nevertheless, since r is unknown, $\underline{\theta}^*$ is not feasible. If the conditional distribution of Y given $\underline{X}$ is known up to r and $\underline{\theta}$, we overcome the problem by replacing r in problem (1.5.1) with its Naradaya-Watson leave-one-out estimator, thus forming a pseudo-likelihood.

If the conditional distribution of Y given $\underline{X}$ is not known, we estimate it in a fully nonparametric way. Delecroix *et al.* (2003) show that the estimator obtained with the kernel estimator of the joint distribution of $(\underline{\theta}'\underline{X}, Y)$ divided by the classical kernel estimator of the marginal density of $\underline{\theta}'\underline{X}$, is asymptotically efficient.

• Bandwidth selection

In order to construct these estimators via kernel estimators, a bandwidth h is needed. The choice of that bandwidth is crucial, because practical performance of the method can depend significantly on it. Delecroix *et al.* (2003) propose an empirical rule for selecting it. Actually, they extend the methodology first introduced by Härdle *et al.* (1993) for the SLS estimator. This rule can be brought back to the usual cross-validation criterion.

Direct estimators

Although their many advantages (efficiency, asymptotic normality, automatic selection of the bandwidth), M-estimators admit an important drawback: they require solving an intricate optimization problem in a high dimensional space. In spite of slightly worst theoretical properties, direct estimators are highly attractive, as they provide the estimator on an analytic form.

• Average Derivative Estimator (ADE)

If we set $u = \underline{\theta}'\underline{x}$ and $m(\underline{x}) = r(\underline{\theta}'\underline{x})$, average derivatives method rests on the fact that

$$\nabla m(\underline{x}) = \frac{\partial r}{\partial u}(\underline{\theta}'\underline{x})\underline{\theta},$$

which induces that

$$\delta_w \doteq \mathrm{E}[w(\underline{X})\nabla m(\underline{X})] = \mathrm{E}[w(\underline{X})\frac{\partial r}{\partial u}(\underline{\theta}'\underline{X})]\underline{\theta}$$

for any bounded continuous weight function w. The quantity δ_w is called a weight average derivative of r with weight function w. From the last expression, we can see that δ_w is proportional to $\underline{\theta}$, provided $\mathrm{E}[w(\underline{X})\frac{\partial r}{\partial u}(\underline{\theta}'\underline{X})]$ is not zero. This condition is in particular violated if $w = 1$, r is an even function and $\underline{X}$ is symmetrically distributed. Moreover, considering the gradient of m, this implies that $\underline{X}$ is a continuously

distributed random vector. However, Horowitz and Härdle (1996) show an extension of the method to the case where some components of $\underline{X}$ are discrete is possible.

Stoker (1986) and Härdle and Stoker (1989) take $w = 1$ and use a nonparametric estimation of the marginal density of $\underline{X}$. They show that $\hat{\delta}$ is a consistent estimator of δ and $\sqrt{n}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \Sigma_u)$, under some regularity conditions.
Note that the method is based on a fully nonparametric estimation of the multivariate density of $\underline{X}$, which is severely subject to the curse of dimensionality. Hence, we loose the main advantage of the single-index modeling.

The previous method requires the estimation of both the density f and its gradient. To avoid this twofold estimation, Powell *et al.* (1989) have proposed to set $w(\underline{x}) = f(\underline{x})$. In this way, only the gradient of f has to be estimated. Also in this case, $\hat{\delta}_f$ is a consistent estimator of δ_f and $\sqrt{n}(\hat{\delta}_f - \delta_f) \xrightarrow{d} N(0, \Sigma_f)$, under some regularity conditions.
As previously, this result implies that $\hat{\underline{\theta}}$ is a $\sqrt{n}$-consistent estimator of θ, with asymptotic normal distribution.

The major drawback of the previous two procedures is the need to estimate the density of $\underline{X}$ and its gradient in a fully nonparametric way, what can lead to very poor performance due to the curse of dimensionality. Hristache *et al.* (2001) introduced a new type of direct estimator of the index coefficient that can be viewed as an iterative improvement of the average derivative estimator. They showed that $||\hat{\underline{\theta}} - \underline{\theta}|| = O(n^{-1/2})$, where $\|\cdot\|$ is the squared Euclidean norm. Thus $\hat{\underline{\theta}}$ is a $\sqrt{n}$-consistent. The asymptotic distribution of the estimator is not mentioned, but it is stated that $\sqrt{n}(\hat{\underline{\theta}} - \underline{\theta})$ is close in distribution to a gaussian vector.

• Sliced Inverse Regression (SIR)

In a dimension reduction purpose, Li (1991) proposed a simple and easy to implement algorithm. Duan and Li (1991) adapted this method in the single-index context. It is based on the relationship between $\underline{\theta}$ and the inverse regression $\mathrm{E}(\underline{X}|Y = y)$. Unfortunately, their results require an important design condition, not always satisfied.

Other estimators

The four previous ideas (SLS, SML, ADE, SIR) are historically the most popular ones in order to estimate $\underline{\theta}$. Nevertheless, this list is far to be exhaustive. There are much more estimators which have been proposed in the literature. Naik and Tsai (2000) extend the method of Partial Least Squares to the case of single-index models. Xia *et al.* (2002) propose an adaptive approach for dimension reduction. This is a kind of M-estimation method, inspired by the SIR method and the idea of local linear smoothers, but with fewer restrictions on the distribution of the covariates. A drawback is that no asymptotic distribution for the estimator is provided. Huh and Park (2002) derive an extension of ADE. A problem is that the method requires the maximization of locally weighted log-likelihood, that is it looses the main advantage of direct estimators. Finally, Han (1987) proposes an estimator based on the rank correlation between the observed values and the values fitted by the model. Asymptotic theory for this estimator is completed in Sherman (1993) and a generalization is given in Cavanagh and Sherman (1998). In particular, this last method cited, requires the link function to be strictly monotonic.

1.6 Single Index Model studied and estimation procedure proposed

The single-index model studied in this thesis is of the form:

$$Y = r(\underline{\theta}' \underline{X}) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2), \tag{1.6.1}$$

where Y is a scalar response variable, $\underline{X}$ is a m-dimensional explanatory variable and ϵ is the error variable assumed to be standard normal distributed, with finite variance. We are interested in the estimation of the finite-dimensional vector of parameters $\underline{\theta} \in \underline{\Theta} \subset \mathbb{R}^m$ and the univariate regression function $r : \mathbb{R} \to \mathbb{R}$, that is completely unknown except for monotonicity and smoothness assumptions. Single-index models will be then studied in semiparametric context.

In the previous section, we showed a large number of methods to estimate the link function r and the parameter $\underline{\theta}$. Because of its easiness of implementation and interpretation and its mathematical tractability, the link function is usually estimated by Naradaya-Watson kernel estimator, although it is well known that it is not the more efficient one. In the majority of these methods, $\underline{\theta}$ is proved to be root-n consistent, efficient and asymptotically normal, under regularity conditions, that is under the assumption that the link function is twice o three times differentiable with bounded second derivative.

In this thesis, we study the couple $(\underline{\theta}, r)$ that will be estimated by penalized maximum likelihood estimator in order to take into account and exploit the conditions of the existence and boundness of the second derivative of r. Note that for the model described in (1.6.1), this estimator coincides with the penalized version of the nonlinear least squares estimator.

Moreover, we decided to make this study imposing the constraint of monotonicity for

the regression function r because of the importance of monotonicity in a variety of statistical models. In the particular case of single-index model, we have also that when the regression function r is assumed to be monotone, the role of the parameters $\underline{\theta}$ is unchanged with respect to the role in the linear regression model. In this way, an easy interpretation of the meaning of the parameters is preserved. On the other hand, with the introduction of the function r in the linear regression model, we have a more flexible model. Withal, because this nonparametric estimation is done on a univariate index, the curse of dimensionality is avoided. Besides, we would liked to impose a restriction in the class of functions to be estimated, with the intent to obtain a faster rate of convergence for the regression function estimator, which usually is stated in $n^{-4/5}$, when the link function is supposed to be twice differentiable.

For a computational study of the performance for finite sample size, smoothing spline estimation is proposed, for linear single-index models. This approach may be classified as an M-estimation problem. Indeed, we will model the regression function r by smoothing splines and the parameter $\underline{\theta}$ minimizing the penalized maximum likelihood criterion, as we will see later.

Chapter 2

Theory for Euclidean Parameters in Infinite-Dimensional Models

2.1 Introduction and overview

When we have observations that are random sample from a common distribution P, a *model* is the set $\mathcal{P}$ of all possible values of P, that is a collection of probability measures in the sample space.

The *nonparametric model* is that model in which we observe a random sample from a completely unknown distribution. Are interesting, intermediate models that are not parametrized by a Euclidean parameter as in parametric models, but do restrict the distribution in important ways. Such models are often parametrized by infinite-dimensional parameters, such as distribution functions or densities or regression functions, that express the structure under study. Many aspects of these parameters are estimable by the same order of accuracy as classical parameters, and efficient estimators are asymptotically normal.

These models may have a natural parametrization $(\underline{\theta}, r) \to P_{\underline{\theta},r}$ where $\underline{\theta}$ is the Euclidean parameter and r runs through a nonparametric class of functions. This give a *semiparametric model* in which we aim at estimating $\underline{\theta}$ and consider r as a *nuisance parameter*. More generally, we focus on estimating the value $\psi(P)$ of some function

$\psi : \mathcal{P} \to \mathbb{R}^m$ on the model.

To set up the semiparametric regression model, the model of interest in this thesis, suppose $(Y, \underline{X})$ is distributed as $P_{\underline{\theta},r}$, indexed by a finite dimensional parameter $\underline{\theta} \in \underline{\Theta} \subset \mathbb{R}^m$ and an infinite dimensional parameter r belonging to some class of functions $\mathcal{R}$. The semiparametric model is then

$$\mathcal{P} = \{P_{\underline{\theta},r} : \underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}\}.$$

The finite dimensional parameter $\underline{\theta}$ is usually called the *regression parameter*, which measures the influence of the explanatory variable $\underline{X}$ on the response variable Y.

2.2 Tangent spaces and information bounds

Harking to the general theory of semiparametric models, suppose that we observe a random sample $\underline{X}_1, \ldots, \underline{X}_n$ from a distribution P that is known to belong to a set $\mathcal{P}$ of probability measures on the sample space $(\underline{\mathcal{X}}, \mathcal{A})$. In this section it is given a notion of information for estimating $\psi(P)$ given the model $\mathcal{P}$, which extends the notion of Fisher information for parametric models.

To estimate the parameter $\psi(P)$ given the model $\mathcal{P}$ is harder than to estimate this parameter given that P belongs to a submodel $\mathcal{P}_0 \subset \mathcal{P}$. For every smooth parametric submodel $\mathcal{P}_0 = \{P_{\underline{\theta}} : \underline{\theta} \in \underline{\Theta}\} \subset \mathcal{P}$, we can calculate the Fisher information for estimating $\psi(P_{\underline{\theta}})$. Then the information for estimating $\psi(P)$ in the whole model is not bigger than the infimum of the informations over all submodels. Then we shall define the *information bound* for the whole model as this infimum. A submodel for which the infimum is taken, is called *least favorable submodel*.

In most situations, it suffices to consider one-dimensional submodels $\mathcal{P}_0$. These should pass through the true distribution P of the observations and be differentiable in quadratic mean at P. Thus, we consider maps $t \to P_t$ from a neighborhood of $t \in [0, \infty)$ to $\mathcal{P}$

such that, for some measurable function $g : \underline{\mathcal{X}} \to \mathbb{R}$,

$$\int \left(\frac{dP_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2} g dP^{1/2} \right)^2 \longrightarrow 0.$$

Then, the parametric submodel $\{P_t : 0 < t < \epsilon\}$ is differentiable in quadratic mean at $t = 0$ with *score function* g. Letting $t \to P_t$ range over a collection of submodels, we obtain a collection of score functions, called *tangent set of the model* $\mathcal{P}$ *at* P and denoted by $\dot{\mathcal{P}}_P$. The tangent set is often a linear space, in which case is called *tangent space*.

Usually, we construct the submodels $t \to P_t$ such that, for every x,

$$g(x) = \frac{\partial}{\partial t}_{|t=0} \log dP_t(x).$$

For defining the information for estimating $\psi(P)$, only those submodels $t \to P_t$ along which the parameter $t \to \psi(P_t)$ is differentiable are of interest. Thus, we consider only submodels $t \to P_t$ such that $t \to \psi(P_t)$ is differentiable at $t = 0$. Then we define $\psi : \mathcal{P} \to \mathbb{R}^m$ to be *differentiable at* P *relative to a given tangent set* $\dot{\mathcal{P}}_P$ if there exist a continuous linear map $\dot{\psi}_P : L_2(P) \to \mathbb{R}^m$ such that for every $g \in \dot{\mathcal{P}}_P$ and a submodel $t \to P_t$ with score function g,

$$\frac{\psi(P_t) - \psi(P)}{t} \to \dot{\psi}_P g.$$

This requires that the derivative of the map $t \to \psi(P_t)$ exists in the ordinary sense, and also that it has a special representation.

By the Riesz representation theorem for Hilbert space, the map $\dot{\psi}_P$ can always be written in the form of an inner product with a fixed vector-valued, measurable function $\tilde{\psi}_P : \underline{\mathcal{X}} \to \mathbb{R}^m$,

$$\dot{\psi}_P g = \int \tilde{\psi}_P g dP.$$

Here the function $\tilde{\psi}_P$ is not uniquely defined by the functional ψ and the model $\mathcal{P}$. However, it is always possible to find a candidate $\tilde{\psi}_P$ whose coordinate functions are

contained in $\overline{\text{lin}}\dot{\mathcal{P}}_P$, the closure of the linear span of the tangent set. This function is unique and is called *efficient influence function*. It can be found as the projection of any other influence function onto the closed linear span of the tangent set.

As usual, an estimator T_n is a measurable function $T_n(\underline{X}_1, \ldots, \underline{X}_n)$ of the observations. An estimator T_n is called *regular at P for estimating $\psi(P)$* (relative to $\dot{\mathcal{P}}_P$) if there exists a probability measure L such that

$$\sqrt{n}(T_n - \psi(P_{1/\sqrt{n},g})) \overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} L, \qquad \text{for every } g \in \dot{\mathcal{P}}_P.$$

We shall say that an estimator sequence is *asymptotically efficient at P*, if it is regular at P with limit distribution $L = N(0, P\tilde{\psi}_P\tilde{\psi}_P^T)$.

The efficient influence function $\tilde{\psi}_P$ plays the same role as the normalized score function in parametric models. In particular, a sequence of estimators T_n is asymptotically efficient at P if

$$\sqrt{n}(T_n - \psi(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_P(\underline{X}_i) + o_P(1).$$

This justifies the name "efficient influence function".

2.3 Efficient score functions

A function $\psi(P)$ of particular interest is the parameter $\underline{\theta}$ in a semiparametric model $\{P_{\underline{\theta},r} : \underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}\}$, where $\underline{\Theta}$ is a subset of $\mathbb{R}^m$ and $\mathcal{R}$ is an arbitrary set, typically of infinite dimension. The information bound for the functional of interest $\psi(P_{\underline{\theta},r}) = \underline{\theta}$ can be expressed in an efficient score function.

As submodels, we use paths of the form $t \to P_{\underline{\theta}+t\underline{a},r_t}$, for given paths $t \to r_t$ in the parameter set $\mathcal{R}$. The score functions for such submodels typically have the form of a sum of partial derivative with respect to $\underline{\theta}$ and r. If $\dot{\ell}_{\underline{\theta},r}$ is the ordinary score function

for $\underline{\theta}$ in the model in which r is fixed, then we expect

$$\frac{\partial}{\partial t}_{|t=0} \log dP_{\underline{\theta}+t\underline{a},r_t} = \underline{a}^T \dot{\ell}_{\underline{\theta},r} + g.$$

The function g has the interpretation of a score function for r if $\underline{\theta}$ is fixed and runs through an infinite-dimensional set if we are concerned with a "true" semiparametric model. We refer to this set as the *tangent set for* r and denote it by ${}_r\dot{\mathcal{P}}_{P_{\underline{\theta},r}}$.

The parameter $\psi(P_{\underline{\theta}+t\underline{a},r_t}) = \underline{\theta} + t\underline{a}$ is differentiable with respect to t in the ordinary sense but is, by definition, differentiable as a parameter on the model if and only if there exists a function $\tilde{\psi}_{\underline{\theta},r}$ such that

$$\underline{a} = \frac{\partial}{\partial t}_{|t=0} \psi(P_{\underline{\theta}+t\underline{a},r_t}) = \left\langle \tilde{\psi}_{\underline{\theta},r}, \underline{a}^T \dot{\ell}_{\underline{\theta},r} + g \right\rangle_{P_{\underline{\theta},r}}, \qquad \underline{a} \in \mathbb{R}^m,\ g \in_r \dot{\mathcal{P}}_{P_{\underline{\theta},r}}.$$

Setting $\underline{a} = 0$, we see that $\tilde{\psi}_{\underline{\theta},r}$ must be orthogonal to the tangent set ${}_r\dot{\mathcal{P}}_{P_{\underline{\theta},r}}$ for the nuisance parameter. Define $\Pi_{\underline{\theta},r}$ as the orthogonal projection onto the closure of the linear span of ${}_r\dot{\mathcal{P}}_{P_{\underline{\theta},r}}$ in $L_2(P_{\underline{\theta},r})$.

The function defined by

$$\tilde{\ell}_{\underline{\theta},r} = \dot{\ell}_{\underline{\theta},r} - \Pi_{\underline{\theta},r}\dot{\ell}_{\underline{\theta},r}$$

is called the *efficient score function for* $\underline{\theta}$, and its covariance matrix $\tilde{I}_{\underline{\theta},r} = P_{\underline{\theta},r}\tilde{\ell}_{\underline{\theta},r}\tilde{\ell}^T_{\underline{\theta},r}$ is the *efficient information matrix*. The efficient score function for $\underline{\theta}$ is then the score function for $\underline{\theta}$ minus its projection on the set of nuisance score functions. Therefore, as a consequence, we refer to

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{\underline{\theta},r}(\underline{X}_i) = 0$$

as the *efficient score equation*.

Lemma 2.3.1. *Suppose that for every* $\underline{a} \in \mathbb{R}^m$ *and every* $g \in_r \dot{\mathcal{P}}_{P_{\underline{\theta},r}}$ *there exists a path* $t \to r_t$ *in* $\mathcal{R}$ *such that*

$$\int \left(\frac{dP^{1/2}_{\underline{\theta}+t\underline{a},r_t} - dP^{1/2}_{\underline{\theta},r}}{t} - \frac{1}{2}\left(\underline{a}^T \dot{\ell}_{\underline{\theta},r} + g\right) dP^{1/2}_{\underline{\theta},r} \right)^2 \longrightarrow 0.$$

If $\tilde{I}_{\underline{\theta},r}$ is nonsingular, then the functional $\psi(P_{\underline{\theta},r}) = \underline{\theta}$ is differentiable at $P_{\underline{\theta},r}$ relative to the tangent set $\dot{\mathcal{P}}_{P_{\underline{\theta},r}} = lin\dot{\ell}_{\underline{\theta},r} +_r \dot{\mathcal{P}}_{P_{\underline{\theta},r}}$ with efficient influence function $\tilde{\psi}_{\underline{\theta},r} = \tilde{I}^{-1}_{\underline{\theta},r}\tilde{\ell}_{\underline{\theta},r}$.

Consequently, an estimator sequence is asymptotically efficient for estimating $\underline{\theta}$ if

$$\sqrt{n}\,(T_n - \underline{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}^{-1}_{\underline{\theta},r}\tilde{\ell}_{\underline{\theta},r}(\underline{X}_i) + o_{P_{\underline{\theta},r}}(1).$$

2.4 Score and information operators

The method to find the efficient influence function of a parameter given in the preceding section is the most convenient method if the model can be naturally partitioned in the parameter of interest and the nuisance parameter. For many parameters such a partition is not possible or not natural. We can then give another useful description of the tangent set for the nuisance parameter, in term of a *score operator*.

Consider first the situation that the model $\mathcal{P} = \{P_r : r \in \mathcal{R}\}$ is indexed by a parameter r that is itself a probability measure on some measurable space. We are interested in estimating a parameter of the type $\psi(P_r) = \chi(r)$ for a given function $\chi : \mathcal{R} \to \mathbb{R}^m$ on the model $\mathcal{R}$.

The model $\mathcal{R}$ gives rise to a tangent set $\dot{\mathcal{R}}_r$ at r. If the map $r \to P_r$ is differentiable in an appropriate sense, then its derivative maps every score $\underline{b} \in \dot{\mathcal{R}}_r$ into a score g for the model $\mathcal{P}$. To be precise, we assume that a smooth parametric submodel $t \to r_t$ induces a smooth parametric submodel $t \to r_t$ are related by

$$g = A_r\underline{b}.$$

Then $A_r\dot{\mathcal{R}}_r$ is a tangent set for the model $\mathcal{P}$ at P_r. Because A_r turns scores for the model $\mathcal{R}$ into scores for the model $\mathcal{P}$ it is called *score operator*.

So far, we have assumed that the parameter r is a probability distribution, but we

can consider the more general case of a model $\mathcal{P} = \{P_r : r \in \mathcal{R}\}$ indexed by a parameter r running through an arbitrary set $\mathcal{R}$. In this general setting, let $\mathbb{R}_r$ be a subset of a Hilbert space that indexes direction $\underline{b}$ in which r can be approximated within $\mathcal{R}$.
We can now illustrate the particular case of a semiparametric model $\{P_{\underline{\theta},r} : \underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}\}$, where the pair $(\underline{\theta}, r)$ plays the role of the single r and $\mathbb{R}^k \times \mathbb{R}_r$ plays the role of the earlier $\mathbb{R}_r$. The two parameters can be perturbed independently, and the score operator can be expected to take the form

$$A_{\underline{\theta},r}(\underline{a}, \underline{b}) = \underline{a}^T \dot{\ell}_{\underline{\theta},r} + B_{\underline{\theta},r}\underline{b}.$$

Here $B_{\underline{\theta},r} : \mathbb{R}_r \to L_2(P_{\underline{\theta},r})$ is the score operator for the nuisance parameter. The domain of the operator $A_{\underline{\theta},r} : \mathbb{R}^k \times \text{lin}\mathbb{R}_r \to L_2(P_{\underline{\theta},r})$ is a Hilbert space relative to the inner product $\big\langle (\underline{a}, \underline{b}), (\underline{\alpha}, \underline{\beta}) \big\rangle_\eta = \underline{a}^T \underline{a} + \big\langle \underline{b}, \underline{\beta} \big\rangle_{\mathbb{R}_r}$.
The efficient influence function for estimating $\underline{\theta}$ is expressed in the *efficient score function for* $\underline{\theta}$ in Lemma 2.3.1, which is defined as the ordinary score function minus its projection onto the score-space for r. Presently, the latter space is the range of the operator $B_{\underline{\theta},r}$. Denoted by $B^*_{\underline{\theta},r}$ the adjoint score operator of $B_{\underline{\theta},r}$, if the operator $B^*_{\underline{\theta},r} B_{\underline{\theta},r}$ is continuously invertible, then the operator $B_{\underline{\theta},r}(B^*_{\underline{\theta},r} B_{\underline{\theta},r})^{-1} B^*_{\underline{\theta},r}$ is the orthogonal projection onto the nuisance score space, and

$$\tilde{\ell}_{\underline{\theta},r} = \big(I - B_{\underline{\theta},r}(B^*_{\underline{\theta},r} B_{\underline{\theta},r})^{-1} B^*_{\underline{\theta},r}\big)\, \dot{\ell}_{\underline{\theta},r}.$$

This means that $\underline{b} = -B_{\underline{\theta},r}(B^*_{\underline{\theta},r} B_{\underline{\theta},r})^{-1} B^*_{\underline{\theta},r} \dot{\ell}_{\underline{\theta},r}$ is a least favorable direction in $\mathcal{R}$, for estimating $\underline{\theta}$.

2.5 Efficient score equations

The most important method of estimating the parameter in a parametric model is the method of maximum likelihood, and it can be reduced to solving the score equations

$\sum_{i=1}^n \dot{\ell}_{\underline{\theta}}(\underline{X}_i) = 0$. A natural generalization to estimating the parameter $\underline{\theta}$ in a semi-parametric model $\{P_{\underline{\theta},r} : \underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}\}$ is to solve $\underline{\theta}$ from the *efficient score equation*

$$\sum_{i=1}^n \tilde{\ell}_{\underline{\theta},\hat{r}}(\underline{X}_i) = 0.$$

Here we use the efficient score function instead of the ordinary score function, and we substitute an estimator $\hat{r}$ for the unknown nuisance parameter.

A disadvantage of this method is that it requires an explicit form of the efficient score function. In general the efficient score function is defined only implicitly as an orthogonal projection.

A variation of this approach is then to obtain an estimator $\hat{r}(\underline{\theta})$ of r for each given value of $\underline{\theta}$, and next to solve $\underline{\theta}$ from the equation

$$\sum_{i=1}^n \tilde{\ell}_{\underline{\theta},\hat{r}(\underline{\theta})}(\underline{X}_i) = 0.$$

If $\hat{\underline{\theta}}$ is a solution, then it is also a solution of the preceding display, for $\hat{r} = \hat{r}(\hat{\underline{\theta}})$. The asymptotic normality of $\hat{\underline{\theta}}$ can therefore be proved by the same methods as applying to this estimating equation. Due to our special choice of estimating function, the nature of the dependence of $\hat{r}(\underline{\theta})$ on $\underline{\theta}$ should be irrelevant for the limiting distribution of $\sqrt{n}(\hat{\underline{\theta}} - \underline{\theta})$.

For simplicity, we shall adopt the notation as in the first estimating equation, even though for the construction of $\hat{\underline{\theta}}$ the two-step procedure, which profile out the nuisance parameter, may be necessary.

Often the nuisance parameter r, which is infinite-dimensional, cannot be estimated within the usual order $O(n^{-1/2})$ for parametric models. Then the classical approach to derive the asymptotic behavior of Z-estimators is impossible. Instead, we utilize the notion of a Donsker class.

The auxiliary estimator for the nuisance parameter should satisfy

$$P_{\hat{\underline{\theta}},r}\tilde{\ell}_{\hat{\underline{\theta}},\hat{r}} = o_P\left(n^{-1/2} + \left\|\hat{\underline{\theta}} - \underline{\theta}\right\|_m\right) \tag{2.5.1}$$

and

$$P_{\underline{\theta},r}\left\|\tilde{\ell}_{\hat{\underline{\theta}},\hat{r}} - \tilde{\ell}_{\underline{\theta}_n,r}\right\|^2 \to_P 0, \qquad P_{\hat{\underline{\theta}},r}\left\|\tilde{\ell}_{\hat{\underline{\theta}},\hat{r}}\right\|^2 = O_P(1). \tag{2.5.2}$$

This last condition requires that the plug-in estimator $\tilde{\ell}_{\underline{\theta}_n,\hat{r}}$ is a consistent estimator for the true efficient influence function. Because $P_{\underline{\theta},r}\tilde{\ell}_{\hat{\underline{\theta}},r} = 0$, the first condition (2.5.1) requires that the bias of the plug-in estimator, due to estimating the nuisance parameter, converge to zero faster than $n^{-1/2}$.

A first theorem that we can enunciate about the asymptotic efficiency of the Euclidean parameter $\hat{\underline{\theta}}$ is as follows

Theorem 2.5.1. *Suppose that the model $\{P_{\underline{\theta},r} : \underline{\theta} \in \underline{\theta}\}$ is differentiable in quadratic mean with respect to $\underline{\theta}$ at $(\underline{\theta}, r)$ and let the efficient information matrix $\tilde{I}_{\underline{\theta},r}$ be non-singular. Assume that (2.5.1) an (2.5.2) hold. Let $\hat{\underline{\theta}}$ satisfy $\sqrt{n}\mathbb{P}_n\tilde{\ell}_{\hat{\underline{\theta}},\hat{r}} = o_{\hat{P}}(1)$ and be consistent for $\underline{\theta}$. Furthermore, suppose that there exists a Donsker class with square-integrable envelope function that contains every function $\tilde{\ell}_{\hat{\underline{\theta}},\hat{r}}$ with probability tending to one. Then the sequence $\hat{\underline{\theta}}$ is asymptotically efficient at $(\underline{\theta}, r)$.*

An improvement of this theorem will be stated and applied in chapter 4 to prove the asymptotic efficiency of the parameter $\hat{\underline{\theta}}$.

Chapter 3

Some Concepts of Nonparametric Theory

The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible or keeping the number of underlying assumptions as weak as possible. Usually, this means using statistical models that are infinite-dimensional.

The estimation of the distribution function, density estimation, nonparametric regression or curve estimation are examples of problem that frequently occur in nonparametric inference. Typically, we will assume that the distribution function, the density or the regression function r, lies in some large set $\mathcal{R}$ called *statistical model*. For example, we might assume that

$$r \in \mathcal{R} = \left\{ f : \int \left(f''(x) \right)^2 dx \leq \rho \right\}, \qquad \rho \geq 0,$$

which is the set of functions that are not too wiggly, or we might assume that

$$r \in \mathcal{R} = \left\{ f : f \text{ is monotone and } \int \left(f''(x) \right)^2 dx \leq \rho \right\}, \qquad \rho \geq 0,$$

which is the set of functions that are not too wiggly and monotone. These are, actually, the restrictions in which we are interested in this thesis. In the next section we review some basic concepts on empirical processes that are used repeatedly in the later chapter.

3.1 Overview of the basic theory

Empirical processes, main tool of nonparametric theory, are applied to the study of nonparametric aspect of the single index model.

Let $\mathcal{F}$ a class of functions $f : \underline{\mathcal{X}} \to \mathbb{R}$ defined on the probability space $(\underline{\mathcal{X}}, \mathcal{A}, P)$.

Suppose that $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. P on $\underline{\mathcal{X}}$. Then the *empirical measure*

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\underline{X}_i}$$

is the discrete random measure that puts mass $1/n$ at every observation. Thus for any Borel set $A \subset \mathbb{R}$

$$\mathbb{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(\underline{X}_i) = \frac{\sharp \{i \leq n : \underline{X}_i \in A\}}{n}.$$

Moreover, for $f : \underline{\mathcal{X}} \to \mathbb{R}$, we write the empirical measure evaluated at f

$$\mathbb{P}_n(f) = \int f d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(\underline{X}_i).$$

The *empirical process* $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ evaluated at the function f is

$$\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f(\underline{X}_i) - \int f dP \right).$$

If $\mathcal{F}$ is a class of functions for which

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n(f) - P(f)| \longrightarrow_{a.s.} 0$$

then we say that $\mathcal{F}$ is a *P-Glivenko-Cantelli class* of functions.

The class $\mathcal{F}$ is called *Donsker class* if the empirical process $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ converges in distribution in the metric space $l^\infty(\mathcal{F})$ of all bounded function $z : \mathcal{F} \to \mathbb{R}$, which is equipped with the supremum norm, that is

$$l^\infty(\mathcal{F}) = \left\{ z : \mathcal{F} \to \mathbb{R} : \|z\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |z(f)| < \infty \right\}.$$

When a class of functions $\mathcal{F}$ is a P-Glivenko-Cantelli class and when it is a P-Donsker class? Answers to these questions began during the 1970's especially with Vapnik and Červonenkis (1971) and Dudley (1978) and continued with contributions from Pollard, Giné and Zinn, and Gaenssler.

Statements about Glivenko-Cantelli and Donsker classes are frequently phrased in terms of bracketing numbers and covering numbers for a class $\mathcal{F}$ of functions $f : \underline{\mathcal{X}} \to \mathbb{R}$. The *covering number* $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \epsilon\}$ of radius ϵ needed to cover $\mathcal{F}$. The centers of the balls need not belong to $\mathcal{F}$, but they should have finite norms.

Given a pair of functions $l \leq u$ the *bracket* $[l, u]$ consist of all functions f with $l \leq f \leq u$. An ϵ-*bracket* is a bracket $[l, u]$ with $\|u - l\| < \epsilon$. The *bracket number* $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. The upper and lower bounds u and l of the brackets need not belong to $\mathcal{F}$ themselves but are assumed to have finite norms.

The *entropy with bracketing* is the logarithm of the bracketing number.

A sufficient condition for a class $\mathcal{F}$ to be a P-Glivenko-Cantelli class is that the bracketing numbers $N_{[]}(\epsilon, \mathcal{F}, L_1(P))$ are finite for every $\epsilon > 0$.

According to a theorem of Ossiander (1987) a sufficient condition for $\mathcal{F}$ to be a P-Donsker class is that

$$\int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty.$$

Important examples of classes such that this last inequality is satisfied are classes of smooth functions on Euclidean spaces.

Chapter 4

Theoretical Properties of Index and Regression Function Estimators

4.1 Introduction

Two estimation problems in single-index models are intensively discussed in the literature. As is explained in section 1.5, the first consists of estimation of the unknown regression function $r : \mathbb{R} \to \mathbb{R}$ and the objective of the second, more intensively problem studied, is to recover the index vector $\underline{\theta} \in \underline{\Theta} \subset \mathbb{R}^m$. In this thesis instead, we will study the estimation of the couple of parameters $(\underline{\theta}, r)$ in the single-index model

$$Y = r(\underline{\theta}' \underline{X}) + \epsilon, \tag{4.1.1}$$

where Y is the one-dimensional response variable, $\underline{X}$ is an observed m-dimensional vector of independent variables.
We suppose that $\epsilon \sim N(0, \sigma^2)$ is an unobserved error with finite variance and that the regression function r is completely unknown, except for monotonicity and qualitative smoothness assumptions.

Several methods to estimate $(\underline{\theta}, r)$ have been developed in the theory of semiparametric estimation.

In the M-estimation approach the estimator $(\hat{\underline{\theta}}, \hat{r})$ of $(\underline{\theta}, r)$, constructed by minimization of an M-functional with respect to $(\underline{\theta}, r)$, could be expressed in the form

$$(\hat{\underline{\theta}}, \hat{r}) = \arg\min_{\underline{\theta}, r_{\underline{\theta},\lambda}} \sum_{i=1}^{n} \psi\left(y_i, r_{\underline{\theta},\lambda}(\underline{\theta}'\underline{x}_i)\right)$$

where $r_{\underline{\theta},\lambda}(t) = \mathrm{E}[Y|\underline{\theta}'\underline{X} = t]$, λ is a smoothing parameter for the regression function and ψ is a so called contrast function. Typical examples are the semiparametric maximum likelihood estimator, with $-\psi$ being the log-likelihood of the errors ϵ, and the semiparametric least squares estimators, with $\psi(y, r) = |y - r|^2$. With the hypothesis of standard normality for the error variable, these two methods are coincidental.
In this thesis we study a particular case of these two methods, where the smoothing parameter is introduced by addition of a penalization term with which it is possible to take into account the assumptions of smoothness for the regression function r. Moreover, if the regression function r is supposed to be monotone, then the estimator $\hat{r}$ is constrained to belong to a class of monotone functions.

4.2 Discussion on the assumptions of smoothing and monotonicity

Data smoothing, consisting in fitting a smooth function to filter out noise in data, is one of the basic tools in statistical applications. This has been reflected by the large amount of recent literature on nonparametric regression estimation. A number of smoothing techniques have been proposed, including kernel smoothing, nearest neighbors, smoothing splines, local polynomials and B-splines approximations. The statistical theory, often asymptotic, and computational issues have been rather extensively studied by supporters of each method.

In this thesis the attention is focused on the less-discussed problem of estimating regression curves that are known or required to be monotone. In many applications, monotonicity is an integrated part of the function being fitted and it is natural to expect a monotonic relationship between a response variable and an associated index. For a simple example, growth curves describing weight or height of growing objects over time, are known to be increasing. We expect wages to be increasing in an index of human capital. In the item response theory, the item characteristic curve, which measures probability of getting a correct answer for an examine with given latent ability parameter, is generally believed to be monotone. Such example are abound in economics, medical sciences and psychometrics. Considerations of both efficiency and interpretability would lead us to constrained smoothing. The articles by Friedman and Tibshirani (1984), Hawkins (1994), and Ramsay (1988) include several other examples where monotone smoothing is useful.

While it may be reasonable to assume a monotonic relationship between a response and a linear index, it is usually difficult to specify the exact nature of the monotonicity. It is therefore desirable to develop estimators of $(\underline{\theta}, r)$, for semiparametric monotonic linear index models.

Arguably, in nonparametric context, the best known method for preserving monotonicity is isotonic regression, which provides the fitted values at the observed predictor with monotonicity. However, this method undersmoothes the data and is very sensitive to outlying observations at the endpoints of the design space. One natural idea is to combine smoothing with isotonic regression. In theory it is possible to incorporate the monotonicity in every smoothing method, but a satisfactory solution is not always easy to come by. Friedman and Tibshirani (1984) used this approach with local averaging. Mammen (1991), Mukerjee (1988) and Wright (1982) investigated the asymptotic rates

of convergence for kernel estimators in conjunction with isotonic regression. Smoothing can be done before or after isotonization. However, if we isotonize before, it is not guaranteed that the estimate is monotone and if we isotonize after smoothing, it is not guaranteed that for the estimated curve is preserved the initial degree of smoothness. Monotonicity can also be imposed on smoothing splines (Villalobos and Wahba (1987)). An important article by Ramsay (1988) proposed using I-splines defined on a suitably chosen set of knots. I-splines are obtained by integrating B-splines with positive coefficients to ensure monotonicity. B-splines have long been known to have computational efficiency and great approximation power. But it is not difficult to see that the class of I-splines is relatively small compared to the class of all monotone splines, and that there is always a possibility that the fit to the data could be improved by allowing more general monotone splines. A different method based on a characterization of monotone functions through differentiation operators was recently studied by Ramsay (1998).

In this thesis we propose a penalized maximum likelihood estimator $(\hat{\underline{\theta}}, \hat{r})$, searched in the class of the monotone functions. We will prove that $(\hat{\underline{\theta}}, \hat{r})$ exists, that it is consistent, that the rate of convergence for $\hat{\underline{\theta}}$ is $n^{-1/2}$ as in parametric context and for $\hat{r}$ is $n^{-4/5}$ as in nonparametric context when the assumption of bounded second derivative is made, but without the constraint of monotonicity of the regression function r. Finally we will prove the efficiency and asymptotic normality for the first estimator of the couple $(\hat{\underline{\theta}}, \hat{r})$.

About the computational implementation of this monotone and penalized maximum likelihood estimator, we propose a simple but effective monotone smoothing method based on B-splines with which construct the searched estimator that is the maximizer, in the class of monotone functions, of a penalized likelihood, introduced in the next section. Indeed, a monotone B-spline is obtained by forcing the spline coefficients to be monotone. In this way, the estimator is monotone and maintains the required

smoothness properties.

4.3 Description of the smoothing monotone estimator

In this section we introduce the penalized maximum likelihood estimator, with the constraint of monotonicy, that we propose in this thesis in order to estimate $(\underline{\theta}, r)$ in single index models.

Suppose that n vectors of observations $(y_i, \underline{x}_i), i = 1, \ldots, n$ are available to estimate the monotone and smoothing regression function and in its argument, the finite-dimensional vector of parameter in the single-index model (4.1.1), in order to summarize how the response variable depends on $\underline{X}$.

Because it will be assumed that $\underline{X}$ is independent of r, $\underline{\theta}$ and ϵ, the response variable is then distributed as $Y \sim N(r(\underline{\theta}' \underline{X}), \sigma^2)$ and the density of the random vector $(Y, \underline{X})$ is

$$\begin{aligned} f_{Y\underline{X}}(y, \underline{x}; \underline{\theta}, r, \sigma^2) &= f_{Y|\underline{X}}(y|\underline{x}; \underline{\theta}, r, \sigma^2) \cdot f_{\underline{X}}(\underline{x}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y - r(\underline{\theta}' \underline{x})}{\sigma}\right)^2} \cdot f_{\underline{X}}(\underline{x}), \end{aligned} \tag{4.3.1}$$

and its log-likelihood function is

$$\begin{aligned} l_n(\underline{\theta}, r, \sigma^2) &= \frac{1}{n} \log \prod_{i=1}^{n} f_{Y\underline{X}}(y_i, \underline{x_i}; \underline{\theta}, r, \sigma^2) \\ &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2n} \sum_{i=1}^{n} \frac{\left(y_i - r(\underline{\theta}' \underline{x_i})\right)^2}{\sigma^2}. \end{aligned}$$

Here the distribution of $\underline{X}$ does not appear in the likelihood, because is assumed independent on r, $\underline{\theta}$ and ϵ. It is considered fixed throughout the dissertation, but need not be known.

Note that maximizing $l_n(\underline{\theta}, r, \sigma^2)$ w.r.t. $(\underline{\theta}, r)$ is the same as consider the following

maximization criterion from which obtain the maximum likelihood estimator

$$(\tilde{\underline{\theta}}, \tilde{r}) = \arg\max_{\underline{\theta}, r} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \left(y_i - r(\underline{\theta}' \underline{x_i}) \right)^2 \right\} \equiv \arg\max_{\underline{\theta}, r} \; h_n(\underline{\theta}, r).$$

It is possible that the $(\tilde{\underline{\theta}}, \tilde{r})$ estimator is suboptimal, perhaps even asymptotically, particularly if r is a priori thought to be smooth. The roughness of the "profile" function

$$\underline{\theta} \to \sup_{r} \; h_n(\underline{\theta}, r),$$

is caused by the fact that the estimator $\tilde{r}$ for r is "not smooth" and is the cause too of the possible suboptimality of the estimator $(\tilde{\underline{\theta}}, \tilde{r})$. Then we must smooth the data in some way and, for example, we can control for the roughness of $\tilde{r}$ by adding a *penalty term* defined as

$$J^2(r) = \int_D r''(u)^2 du,$$

where the domain D of the integral is taken to be a finite interval in $\mathbb{R}$, that contains the support of $(\underline{\theta}' \, \underline{X})$ for every $\underline{\theta} \in \underline{\Theta}$.

The addition of a penalty term to the criterion we are optimizing is sometimes called *regularization.*

In this thesis, we investigate the *penalized maximum likelihood estimator*

$$(\hat{\underline{\theta}}, \hat{r}) = \arg \max_{\underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}} \left[-\frac{1}{n} \sum_{i=1}^{n} \left(y_i - r(\underline{\theta}' \underline{x_i}) \right)^2 - \hat{\lambda}_n^2 J^2(r) \right]. \tag{4.3.2}$$

For a future study in single index models where the error variable ϵ is suppose to have mean zero e finite variance, but its distribution is unknown, it could be useful to note that, in this case, because of the hypothesis of standard normality for the error ϵ, $(\hat{\underline{\theta}}, \hat{r})$ is obviously equivalent to the *penalized least squares estimator*

$$(\hat{\underline{\theta}}, \hat{r}) = \arg \min_{\underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(y_i - r(\underline{\theta}' \underline{x_i}) \right)^2 + \hat{\lambda}_n^2 J^2(r) \right]. \tag{4.3.3}$$

In the expressions (4.3.2) and (4.3.3), we have that the smoothing parameter $\hat{\lambda}_n \in \mathbb{R}$ and the set $\mathcal{R} = \{r : D \to \mathbb{R} : r \text{ is monotone and } J^2(r) < \infty\}$.

Of course, other penalty terms could be used as well, for instance the L_2-norm of a higher derivative, that is

$$J^2(r) = \int_D r^{(p)}(u)^2 du, \qquad p \in \mathbb{Z}.$$

If certain smoothness properties are known or believed to hold for r, this can suggest a value for p. Using the second derivative appears to yield the minimal smoothness of the estimator $\hat{r}$ needed to make our arguments go through. The standard choice of $p = 2$ corresponds to the assumption that r is continuously differentiable with a square integrable second derivative.

The parameter $\hat{\lambda}_n$ in (4.3.2) or (4.3.3), governs the trade-off between smoothness and goodness of fit and is referred to $\hat{\lambda}_n$ as the *smoothing parameter*.
What is the meaning of the smoothing parameter? Note that when $\hat{\lambda}_n$ is large a premium is being placed on smoothness and potential estimators with large, in the general case, pth derivatives are penalized. In the limiting case with $\hat{\lambda}_n^2 = \infty$, the estimator $\hat{r}$ is an pth order polynomial regression fit to the data. In particular, when $p = 2$, $\hat{r}$ converges to the least squares line. Conversely, a small value of $\hat{\lambda}_n$ corresponds to more emphasis on goodness-of-fit and $\hat{\lambda}_n = 0$ produces an estimator $\hat{r}$ that interpolates the data in that $r(x_i) = y_i$, $i = 1, \ldots, n$. This corresponds to no smoothing the data at all. Then, the size of the smoothing parameter $\hat{\lambda}_n$ determines the importance of the penalty.
The main challenge in smoothing is to determine how much smoothing to do, in order moreover to find a solution for the called *bias-variance trade-off* problem. We have

indeed that large values of $\hat{\lambda}_n$ leads to an estimator with large bias and small variance, called *oversmoothing*. On the other hand, small values of the smoothing parameter leads to an estimator with small bias but large variance, called *undersmoothing*.
The smoothing parameter may be data-dependent, but as we will see later, it should satisfy

$$\hat{\lambda}_n^2 = o_P\left(\frac{1}{n^{1/2}}\right), \qquad \hat{\lambda}_n^{-1} = O_P\left(n^{2/5}\right). \tag{4.3.4}$$

Conditions (4.3.4) are needed to be valid the results of consistency and asymptotic efficiency of the Euclidean parameter. As a matter of fact, Theorem 4.4.5 is demonstrated provided that $\hat{\lambda}_n^{-1} = O_P\left(n^{2/5}\right)$. Moreover, in order to verify the equation (4.4.5) and so have that Theorem 4.4.14 holds, we need that $\hat{\lambda}_n^2 = o_P\left(n^{-1/2}\right)$.
Note that condition (4.3.4) leaves some freedom in choosing $\hat{\lambda}_n$, that is $\hat{\lambda}_n$ can be conveniently chosen such that $C_1 n^{-2/5} \leq \hat{\lambda}_n \leq C_2 n^{-1/4}$, for positive constants C_1, C_2. Any choice satisfying (4.3.4) will result in an asymptotically efficient estimator $\hat{\underline{\theta}}$. The best convergence rate for the estimator $\hat{r}$ is obtained by choosing $\hat{\lambda}_n$ exactly of the order $n^{-2/5}$, but this may not be optimal (in terms of higher order properties) for estimating $\underline{\theta}$.

4.4 Theoretical results

In this section some properties of the penalized maximum likelihood estimator (4.3.2) or penalized least squares estimator (4.3.3) are investigated.
We begin recalling the single-index model in study, that is:

$$Y = r(\underline{\theta}' \underline{X}) + \epsilon, \tag{4.4.1}$$

where $\epsilon \sim N(0, \sigma^2)$ is an unobserved error with finite variance and the regression function r is completely unknown, except for the assumptions of monotonicity and of boundness of its second derivative.

The following assumptions are used:

Assumption A1. Let $\underline{X}$ be independent of r, $\underline{\theta}$ and ϵ;

Assumption A2. Let the support of $(\underline{\theta}'\,\underline{X})$ be strictly contained in the finite interval D for every $\underline{\theta} \in \underline{\Theta}$.

Assumption A3. Let $\underline{\Theta}$ be compact and $\underline{\theta}_0$ be an interior point of $\underline{\Theta}$.

Assumption A4. Let the support of $(\underline{\theta}_0'\,\underline{X})$ be the closure of its interior.

Assumption A5. Let $(\underline{\theta}'\,\underline{X})$ and $\underline{X}$ have densities uniformly bounded (also in $\underline{\theta}$).

Assumption A6. Let $\underline{\Theta} = \{\underline{\theta} \in \mathbb{R}^m : \|\underline{\theta}\|_m = 1\}$.

Assumption A7. Let r be not constant in the support of $(\underline{\theta}'\underline{X})$;

Assumption A8. Let the distribution of $\underline{X}$ be continuous and have compact support, not contained in any proper linear subspace of $\mathbb{R}^m$ and that contains an interior point.

Assumptions A1. and A2. are made in order to individualize the single-index model and the penalized criterion defining the estimator.
Assumptions A3.-A5. will be useful in the prof of the following theorems.
Finally, assumptions A6.-A8. are essentially the usual identification conditions for $\underline{\theta}$ and r, in single-index models.

The main results in this chapter are about existence and consistency of the estimator $(\hat{\underline{\theta}}, \hat{r})$ and mainly about the asymptotic efficiency of the Euclidean estimator $\hat{\underline{\theta}}$.

To do that, in the following, we shall use the particular notations:
- We shall use $P_{\underline{\theta},r}$ for the distribution of $(Y, \underline{X})$ under $(\underline{\theta}, r)$.
- Let abbreviate $P_{\underline{\theta}_0,r_0}$ to P_0 and in particular, let $P_0 f$ denote the mean of $f = f(Y, \underline{X})$ under the law of $(Y, \underline{X})$ under $(\underline{\theta}_0, r_0)$.
- Let $\|\cdot\|_2$ denote the L_2-norm under $P^{Y,\underline{X}}$, given by the product of the dominating

measure for $\underline{X}$ and the conditional distribution of Y given X.

- Let $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\underline{X}_i}$ denote the empirical distribution and with $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$, the empirical process and finally,

- $\lesssim$, $\gtrsim$ mean smaller than, bigger than, up to a constant. This constant may depend on the true parameter of the model, but not on any other parameter values.

Later on, we shall prove the results only for monotone increasing function r. The same results hold and can be proved in the same way for monotone decreasing function.

4.4.1 Existence of the estimator $(\hat{\theta}, \hat{r})$

In this section we shall attend to prove the existence of the estimator $(\hat{\underline{\theta}}, \hat{r})$ of $(\underline{\theta}, r)$. We begin with the statement and the proof of the following useful lemma.

Lemma 4.4.1. *Let $r : \mathbb{R} \to \mathbb{R}$ be a monotone function with $J(r) < \infty$. Then*
(i) $|r'(s) - r'(s_0)| \leq J(r)\,|s - s_0|^{1/2}$ for every $s, s_0 \in D$;
(ii) $\sup_{s\in D} |r'(s)| \lesssim 1 + J(r)$.

Proof. By the Cauchy-Schwarz inequality $|r'(s) - r'(s_0)| = \left|\int_{s_0}^{s} r''(t)dt\right| \leq J(r)\,|s - s_0|^{1/2}$ for every $s, s_0 \in D$. Integrating this w.r.t. s we see that $|r(s) - r(s_0) - r'(s_0)(s - s_0)| \leq J(r)\,|D|^{3/2}$. Since r is bounded, we conclude that $|r'(s_0)|$ and hence $\|r'\|_\infty$ is bounded by a multiple of $1 + J(r)$. □

Recalling that we defined $\mathcal{R} = \{r : D \to \mathbb{R} : r \text{ is monotone and } J^2(r) < \infty\}$, we can enunciate the following

Theorem 4.4.2. *If the assumptions A1. and A2. hold and if $r \in \mathcal{R}$, then $(\hat{\underline{\theta}}, \hat{r})$ exists (but it is not unique).*

Proof. For a given $\underline{\theta}$ and a given vector $\underline{p} \in \mathbb{R}^n$ such that $-\infty < p_1 \leq p_2 \leq \ldots \leq p_n < \infty$, let $\mathcal{R}_{\underline{\theta},\underline{p}}$ be the set of all functions r obtained by first ordering the points

$\underline{\theta}' \underline{x}_i$, yielding the points $t_1 \leq t_2 \leq \ldots \leq t_n$, and next requiring that $J(r) < \infty$, that $r(t_i) = p_i$, for every i, and that r is monotone on each of the intervals $[t_i, t_{i+1}]$.

Then, the searched $\arg\max_{\underline{\theta},r} \left[h_n(\underline{\theta}, r) - \hat{\lambda}_n^2 J^2(r)\right]$ is equal to

$$\sup_{\underline{\theta}} \sup_{\underline{p}} \sup_{r \in \mathcal{R}_{\underline{\theta},\underline{p}}} \left[h_n(\underline{\theta}, r) - \hat{\lambda}_n^2 J^2(r)\right].$$

To see this, about the inner supremum, note that $h_n(\underline{\theta}, r)$ is constant on $\mathcal{R}_{\underline{\theta},\underline{p}}$, and suppose that $r_m \in \mathcal{R}_{\underline{\theta},\underline{p}}$ is a sequence with

$$J^2(r_m) \to G(\underline{\theta}, \underline{p}) := \inf_{r \in \mathcal{R}_{\underline{\theta},\underline{p}}} J^2(r).$$

By the parallelogram law applied to the Hilbert space norm $J(r)$, we have

$$J^2(r_m - r_n) + J^2(r_m + r_n) = 2J^2(r_m) + 2J^2(r_n).$$

Since $\frac{1}{2}(r_m + r_n) \in \mathcal{R}_{\underline{\theta},\underline{p}}$, we have $J^2\left(\frac{1}{2}(r_m + r_n)\right) \geq G(\underline{\theta}, \underline{p})$. Combined with the preceding display, this shows that $J^2(r_m - r_n) \to 0$. Thus r_m'' is a Cauchy sequence and hence has a converging subsequence.

By Lemma 4.4.1(ii) and the Ascoli-Arzela theorem, the sequence r_m also has a subsequence that converges uniformly to a function $r_{\underline{\theta},\underline{p}}$. Conclude that $r_{\underline{\theta},\underline{p}} \in \mathcal{R}_{\underline{\theta},\underline{p}}$ and $J^2(r_{\underline{\theta},\underline{p}}) = G(\underline{\theta}, \underline{p})$.

The function $\underline{p} \mapsto J(r_{\underline{\theta},\underline{p}})$ is convex. Indeed, since $\frac{1}{2}\left(r_{\underline{\theta},\underline{p}_1} + r_{\underline{\theta},\underline{p}_2}\right) \in \mathcal{R}_{\underline{\theta},(\underline{p}_1+\underline{p}_2)/2}$ and the semi-norm $r \mapsto J(r)$ is convex, we have

$$J\left(r_{\underline{\theta},(\underline{p}_1+\underline{p}_2)/2}\right) \leq J\left(\frac{1}{2}\left(r_{\underline{\theta},\underline{p}_1} + r_{\underline{\theta},\underline{p}_2}\right)\right) \leq \frac{1}{2}J\left(r_{\underline{\theta},\underline{p}_1}\right) + \frac{1}{2}J\left(r_{\underline{\theta},\underline{p}_2}\right).$$

We conclude that $\underline{p} \mapsto J(r_{\underline{\theta},\underline{p}})$ is continuous in $\mathbb{R}^n$ and in particular on the compact set $-\infty < p_1 \leq p_2 \leq \ldots \leq p_n < \infty$. Therefore, the function $\underline{p} \mapsto h_n(\underline{\theta}, r_{\underline{\theta},\underline{p}}) - \hat{\lambda}_n^2 J^2(r_{\underline{\theta},\underline{p}})$ is continuous as well and hence attains its maximum. It follows that the supremum on the right side of

$$g(\underline{\theta}) := \sup_{r} \left[h_n(\underline{\theta}, r) - \hat{\lambda}_n^2 J^2(r)\right]$$

is taken for some $r_{\underline{\theta}}$.

Since $h_n(\underline{\theta}, r) \leq 0$ and $\inf_{\underline{\theta}} g(\underline{\theta}) > -\infty$ there must exist a finite constant M such that

$$g(\underline{\theta}) := \sup_{r:J(r)\leq M} \left[h_n(\underline{\theta}, r) - \hat{\lambda}_n^2 J^2(r) \right].$$

The function $\underline{\theta} \mapsto h_n(\underline{\theta}, r)$ are equicontinuous when r satisfies $J(r) \leq M$, since

$$\begin{aligned} |r(\underline{\theta}_1' \underline{x}) - r(\underline{\theta}_2' \underline{x})| &\leq \|r'\|_\infty |\underline{\theta}_1' \underline{x} - \underline{\theta}_2' \underline{x}| \lesssim (1 + J(r)) \left| (\underline{\theta}_1 - \underline{\theta}_2)' \underline{x} \right| \\ &\lesssim K \|\underline{\theta}_1 - \underline{\theta}_2\|_m , \end{aligned}$$

in view of Lemma 4.4.1(ii). It follows that the functions $\underline{\theta} \mapsto g(\underline{\theta})$ are continuous and hence attains their maximum at some point $\hat{\underline{\theta}}$. One can now conclude that $\left(\hat{\underline{\theta}}, r_{\hat{\underline{\theta}}}\right)$ maximizes the penalized likelihood. □

4.4.2 Results of consistency

In this section it is shown that, under some regularity conditions, the penalized maximum likelihood estimator $(\hat{\underline{\theta}}, \hat{r})$, with, as known, $\hat{r}$ restricted to be monotone and with bounded second derivative, is consistent. It is also shown that these estimators are consistent separately.

We begin quoting the useful Lemma 3.5 in Murphy *et al.* (1999) that is in this thesis renumerated in the following way:

Lemma 4.4.3. *Let $\mathcal{F}$ be a class of functions $f : D \to \mathbb{R}$ on a interval $D \subset \mathbb{R}$ such that $\|f\|_\infty \leq M$ and such that the $(k-1)$th derivative is absolutely continuous with $\int f^{(k)}(x)^2 dx \leq M$, for some constant M. Then there exists a constant C such that*

$$\log N_{[]} (\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq C \left(\frac{M}{\epsilon} \right)^{1/k}, \qquad 0 < \epsilon \leq M.$$

Proof. See Birman and Solomjak (1967). □

Let denote here by $p_{\underline{\theta},r}$ the density of $(Y, \underline{X})$ under $(\underline{\theta}, r)$, with respect to $P^{Y,\underline{X}}$ and $\mathcal{P} = \{p_{\underline{\theta},r} : \underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}\}$. With this notation, an expression for $(\underline{\theta}, r)$, equivalent to (4.3.2) is

$$(\hat{\underline{\theta}}, \hat{r}) = \arg \max_{\underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}} \left[\int \log p_{\underline{\theta},r} d\mathbb{P}_n - \hat{\lambda}_n^2 J^2(r) \right]. \tag{4.4.2}$$

Moreover, recalling that the Hellinger distance between two densities, p_1 and p_2, with respect to a σ-finite measure μ, is defined as

$$h(p_1, p_2) = \left(\frac{1}{2} \int \left(p_1^{1/2} - p_2^{1/2} \right)^2 d\mu \right)^{1/2},$$

we can now enunciate the

Lemma 4.4.4.

$$h^2(p_{\hat{\underline{\theta}},\hat{r}}, p_{\underline{\theta}_0,r_0}) + 4\hat{\lambda}_n^2 J^2(\hat{r}) \leq 16 \int g_{p_{\hat{\underline{\theta}},\hat{r}}} d(\mathbb{P}_n - P) + 4\hat{\lambda}_n^2 J^2(r_0),$$

where $g_p = \frac{1}{2} \log \frac{p + p_{\underline{\theta}_0,r_0}}{2p_{\underline{\theta}_0,r_0}}$.

Proof.

$$4 \int g_{p_{\hat{\underline{\theta}},\hat{r}}} d\mathbb{P}_n - \hat{\lambda}_n^2 J^2(\hat{r}) \geq \int \log \left(\frac{p_{\hat{\underline{\theta}},\hat{r}}}{p_{\underline{\theta}_0,r_0}} \right) d\mathbb{P}_n - \hat{\lambda}_n^2 J^2(\hat{r}) \geq -\hat{\lambda}_n^2 J^2(r_0),$$

so

$$\begin{aligned} \int g_{p_{\hat{\underline{\theta}},\hat{r}}} d(\mathbb{P}_n - P) - 4\hat{\lambda}_n^2 J^2(\hat{r}) &\geq -16 \int g_{p_{\hat{\underline{\theta}},\hat{r}}} dP - 4\hat{\lambda}_n^2 J^2(r_0) \\ &\geq 16h^2 \left(\frac{p_{\hat{\underline{\theta}},\hat{r}} + p_{\underline{\theta}_0,r_0}}{p_{\underline{\theta}_0,r_0}}, p_{\underline{\theta}_0,r_0} \right) - 4\hat{\lambda}_n^2 J^2(r_0) \\ &\geq h^2 \left(p_{\hat{\underline{\theta}},\hat{r}}, p_{\underline{\theta}_0,r_0} \right) - 4\hat{\lambda}_n^2 J^2(r_0). \end{aligned}$$

□

The next theorem shows that, under some regularity conditions, the estimator $(\hat{\underline{\theta}}, \hat{r})$ is consistent and it is given a rate of convergence.

Theorem 4.4.5. *Under the assumptions A1.-A8. listed previously, if* $r_0 \in \mathcal{R}$ *and* $p_{\underline{\theta}_0,r_0} \geq \eta_0^2$ *for a constant* η_0*, then*

$$h(p_{\underline{\hat{\theta}},\hat{r}}, p_{\underline{\theta}_0,r_0}) = O_P(\hat{\lambda}_n) \tag{4.4.3}$$

and

$$J(\hat{r}) = O_P(1). \tag{4.4.4}$$

Proof. Let $\mathcal{P}_M = \{p_{\underline{\theta},r} \in \mathcal{P} : J(r) \leq M\}$ and $\overline{\mathcal{P}}_M^{\frac{1}{2}} = \left\{\overline{p}_{\underline{\theta},r}^{\frac{1}{2}} = \sqrt{\frac{p_{\underline{\theta},r}+p_{\underline{\theta}_0,r_0}}{2}} : p_{\underline{\theta},r} \in \mathcal{P}_M\right\}$, with $M \geq 1$. Since $p_{\underline{\theta}_0,r_0} \geq \eta_0^2$, Theorem 2.4 of van de Geer (2000) implies for the bracketing entropy in L_2-norm

$$H_{[]}\left(\epsilon, \overline{\mathcal{P}}_M^{\frac{1}{2}}, L_2\left(P^{Y,\underline{X}}\right)\right) \leq A\left(\frac{M}{\epsilon}\right), \quad M \geq 1, \epsilon > 0.$$

Then, if $\mathcal{P}'_M = \{p_{\underline{\theta},r} \in \mathcal{P} : 1 + J(r) + J(r_0) \leq 2M\}$, we have that $\mathcal{P}'_M \subseteq \mathcal{P}_{2M}$ and if $\overline{\mathcal{P}'}_M^{\frac{1}{2}} = \left\{\overline{p}_{\underline{\theta},r}^{\frac{1}{2}} = \sqrt{\frac{p_{\underline{\theta},r}+p_{\underline{\theta}_0,r_0}}{2}} : p_{\underline{\theta},r} \in \mathcal{P}'_M\right\}$, we have $\overline{\mathcal{P}'}_M^{\frac{1}{2}} \subseteq \overline{\mathcal{P}}_{2M}^{\frac{1}{2}}$, then

$$\begin{aligned} H_{[]}\left(\epsilon, \overline{\mathcal{P}'}_M^{\frac{1}{2}}, L_2\left(P^{Y,\underline{X}}\right)\right) &\leq H_{[]}\left(\epsilon, \overline{\mathcal{P}}_{2M}^{\frac{1}{2}}, L_2\left(P^{Y,\underline{X}}\right)\right) \\ &\leq A'\left(\frac{M}{\epsilon}\right), \quad M \geq 1, \forall \epsilon > 0. \end{aligned}$$

Moreover,

$$\left\|\overline{p}_{\underline{\theta},r}^{1/2} - \overline{p}_{\underline{\theta}_0,r_0}^{1/2}\right\|_2 \leq h(p_{\underline{\theta},r}, p_{\underline{\theta}_0,r_0}), \quad \forall \overline{p}_{\underline{\theta},r}^{1/2} \in \overline{\mathcal{P}'}_M^{1/2},$$

in fact this inequality is satisfied iff

$$\int \left(\overline{p}_{\underline{\theta},r}^{1/2} - \overline{p}_{\underline{\theta}_0,r_0}^{1/2}\right)^2 dP^{Y,\underline{X}} \leq \frac{1}{2}\int \left(p_{\underline{\theta},r}^{1/2} - p_{\underline{\theta}_0,r_0}^{1/2}\right)^2 dP^{Y,\underline{X}}$$

that is verified iff

$$\int \left[\frac{1}{2}\left(p_{\underline{\theta},r}^{1/2} - p_{\underline{\theta}_0,r_0}^{1/2}\right)^2 - \frac{1}{2}\left((p_{\underline{\theta},r} + p_{\underline{\theta}_0,r_0})^{1/2} - (2p_{\underline{\theta}_0,r_0})^{1/2}\right)^2\right] dP^{Y,\underline{X}} \geq 0.$$

The last inequality holds because of the first inequality in exercise 4, page 337 of van der Vaart and Wellner (1996).

In order to apply Lemma 5.14 of van de Geer (2000) with $\mathcal{G}_M = \overline{\mathcal{P}'}_M^{\frac{1}{2}}$, $I(g) = 1 + J(r) + J(r_0)$, $d(\overline{p}_{\underline{\theta},r}^{\frac{1}{2}}, \overline{p}_{\underline{\theta}_0,r_0}^{\frac{1}{2}}) = h(p_{\underline{\theta},r}, p_{\underline{\theta}_0,r_0})$ and $\alpha = \frac{1}{2}$, $\beta = 0$, we have still to verify that there exists a constant $c_0 > 0$ such that for all $M \geq 1$ we have

$$\sup_{\overline{p}_{\underline{\theta},r}^{1/2} \in \overline{\mathcal{P}'}_M^{1/2}} d(\overline{p}_{\underline{\theta},r}^{1/2}, \overline{p}_{\underline{\theta}_0,r_0}^{1/2}) \leq c_0 M.$$

This inequality holds seeing that $h^2(p,q) = 1 - \int \sqrt{pq} d\mu = 1 - \rho(p,q) \leq 1$ since $0 \leq \rho(p,q) \leq 1$.

Now, from Lemma 5.14 of van de Geer (2000) we have

$$\sup_{h(p_{\underline{\theta},r}, p_{\underline{\theta}_0,r_0}) > n^{-\frac{2}{5}}[1+J(r)+J(r_0)]} \frac{\int g_{p_{\underline{\theta},r}} d(\mathbb{P}_n - P)}{h^{3/4}(p_{\underline{\theta},r}, p_{\underline{\theta}_0,r_0})[1 + J(r) + J(r_0)]} = O_P(n^{-\frac{1}{2}}),$$

$$\sup_{h(p_{\underline{\theta},r}, p_{\underline{\theta}_0,r_0}) \leq n^{-\frac{2}{5}}[1+J(r)+J(r_0)]} \frac{\int g_{p_{\underline{\theta},r}} d(\mathbb{P}_n - P)}{1 + J(r) + J(r_0)} = O_P(n^{-\frac{2}{5}}).$$

Thus, proceeding as in Theorem 10.6 of van de Geer (2000), for m=2, we find (4.4.3) and (4.4.4), in view of (4.3.4). □

By theorem 4.4.5, the density $p_{\hat{\underline{\theta}},\hat{r}}$ is consistent for $p_{\underline{\theta}_0,r_0}$ also for the $\|\cdot\|_2$-norm and $J(\hat{r}) = O_P(1)$. Here we prove that these statements carry over into the consistency of $\hat{\underline{\theta}}$ and $\hat{r}$ separately, using for $\hat{\underline{\theta}}$, the Euclidean norm in $\mathbb{R}^m$.

Since the true $h_n(\underline{\theta}, r)$ evaluates the functions r only at the points $\underline{\theta}_0' \underline{x}$, we do not control over r off the support of the variable $\underline{\theta}_0' \underline{X}$. To assert that r is consistent, we may use the norm $\|\cdot\|_D$ for D the support of $\underline{\theta}_0' \underline{X}$, and, for a given set D,

$$\|r\|_D = \sup_{z \in D} |r(z)| + \sup_{z \in D} |r'(z)|.$$

For the consistency of the derivative $\hat{r}'$, we avail ourself of the assumption A4., letting D be the closure of its interior, which is true, for instance, if D is an interval.

Let, first of all, enunciate some lemmas, remarkable for their own importance and, above all, for their usefulness in the proof of the main following theorem in this section, about the separate consistency of the estimators $\hat{\underline{\theta}}$ and $\hat{r}$.

Lemma 4.4.6. *For every fixed M, the set of the restrictions $r_{|D}$ of the monotone functions r with $J(r) \leq M$ is precompact relatively to $\|\cdot\|_D$.*

Proof. By Lemma 4.4.1 (i) the class of functions $r'_{|D}$ is uniformly Lipschitz of order 1/2, hence equicontinuous, and $\left\|r'_{|D}\right\|_\infty$ is uniformly bounded, as soon as $J(r)$ is uniformly bounded. Applying the Ascoli-Arzela theorem, we see that every sequence of function r_n with $J(r_n) = O(1)$ has a subsequence such that both r_n and r'_n converge uniformly on D to limits. The limit of r'_n must necessarily be the derivative of the limit of r_n. □

Lemma 4.4.7. *If $\left\|p_{\underline{\theta},r} - p_{\underline{\theta}_0,r_0}\right\|_2 = 0$ for r such that $J(r) < \infty$, then $\underline{\theta} = \underline{\theta}_0$ and $r = r_0$ on the support of $\underline{\theta}'_0 \underline{X}$.*

Proof. By hypothesis and by equality (4.3.1) we have that $r(\underline{\theta}' \underline{x}) = r_0(\underline{\theta}'_0 \underline{x})$ almost surely under the distribution of $(Y, \underline{X})$. By continuity and by the assumptions for the identifiability of the estimators, it is possible conclude that the functions must be equal on the support of $\underline{X}$. Differentiating partially the identity with respect to $\underline{x}$, we find

$$\theta_i r'_i(\theta_i \underline{x}_i) = \theta_{0i} r'_{0i}(\theta_{0i} \underline{x}_i), \qquad \text{for } i = 1, \ldots, n.$$

These identities are valid on the interior of the support of $\underline{X}$. Once more, by the assumptions for identifiability of the parameters, since r'_0 is nonzero, one can conclude that $\underline{\theta} = \underline{\theta}_0$. Next conclude that $r = r_0$ almost surely under the distribution of $\underline{\theta}'_0 \underline{X}$ and hence $r = r_0$ on the support of $\underline{\theta}'_0 \underline{X}$. □

Lemma 4.4.8. $\hat{\underline{\theta}} \xrightarrow{P} \underline{\theta}_0$ *and* $\|\hat{r} - r_0\|_D \xrightarrow{P} 0$.

Proof. Suppose that $p_{\underline{\theta}_m, r_m} \to p_{\underline{\theta}_0, r_0}$ in $\|\cdot\|_2$ and $J(r_m) = O(1)$. By Lemma 4.4.6, every subsequence of $(\underline{\theta}_m, r_m)$ has a further subsequence such that $\underline{\theta}_m \to \underline{\theta}$ and $\|r_m \to r\|_D \to 0$ for some $\underline{\theta}$ and r. Then $\left\|p_{\underline{\theta}_m, r_m} - p_{\underline{\theta}, r}\right\|_2 \to 0$ by the continuity of the map $(\underline{\theta}, r) \longmapsto p_{\underline{\theta}, r}$. Thus, $\left\|p_{\underline{\theta}, r} - p_{\underline{\theta}_0, r_0}\right\|_2 = 0$ and hence $\underline{\theta} = \underline{\theta}_0$ and $r = r_0$ on the support of $\underline{\theta}_0 X$ by the Lemma 4.4.7. Under the assumption that D is the closure of its interior, this implies that r' and r_0' agree on D as well. It follows that $r_m \to r_0$ and $r_m' \to r_0'$ uniformly on D. Combined with the preceding lemmas and Theorem 4.4.5, this yields the lemma. □

It is now formulate the main theorem of consistency, in this section.

Theorem 4.4.9. *Under the assumptions A1.-A8. listed previously, if the conditional distribution of* $\underline{X}$ *given* $\underline{\theta}_0' \underline{X}$ *is nondegenerate, then this implies that both* $\left\|\hat{\underline{\theta}} - \underline{\theta}_0\right\|_m$ *and* $\|\hat{r}(\underline{\theta}_0' \underline{x}) - r_0(\underline{\theta}_0' \underline{x})\|_2$ *are* $O_P(\hat{\lambda}_n)$.

Proof. To see that the rate of convergence of $\hat{r}(\hat{\underline{\theta}}' \underline{x})$ in the $\|\cdot\|_2$-norm carries over into a rate for $\hat{r}$ in the $L_2(P^{\underline{\theta}_0' \underline{X}})$-distance, we prove, first of all , the differentiability of $r(\underline{\theta}' \underline{x})$ in $(\underline{\theta}, r)$, as it is shown in the following Lemma 4.4.10.
Moreover, since $\left\|\hat{\underline{\theta}} - \underline{\theta}_0\right\|_m \xrightarrow{P} 0$, that $P_0(\hat{r}' - r_0')^2 (\underline{\theta}_0 \underline{x}) \xrightarrow{P} 0$ and that $J(\hat{r})$ is bounded, we see that

$$
\begin{aligned}
& P_0 \left[\hat{r}\left(\hat{\underline{\theta}}' \underline{x}\right) - r_0\left(\underline{\theta}_0' \underline{x}\right)\right]^2 \\
& \gtrsim P_0 \left[-\left(\hat{\underline{\theta}} - \underline{\theta}_0\right)' \underline{x}\, r_0'\left(\underline{\theta}_0' \underline{x}\right) + \left(\hat{r} - r_0\right)\left(\underline{\theta}_0' \underline{x}\right)\right]^2 - o_P(1) \left\|\hat{\underline{\theta}} - \underline{\theta}_0\right\|_m^2 .
\end{aligned}
$$

By the assumptions that the conditional distribution of $\underline{X}$ given $\underline{\theta}_0 \underline{X}$ is nondegenerate and r_0' is nonzero, the expectation on the right is bounded (below) by a constant times

$\left\|\underline{\hat{\theta}}-\underline{\theta}_0\right\|_m^2 + P_0\left(\hat{r}-r_0\right)^2\left(\underline{\theta}_0\underline{x}\right)$ by Lemma 4.4.11 applied with $g_1 = \left(\underline{\hat{\theta}}-\underline{\theta}_0\right)'\underline{x}\, r_0'\left(\underline{\theta}_0'\underline{x}\right)$ and $g_2 = \left(\hat{r}-r_0\right)\left(\underline{\theta}_0'\underline{x}\right)$. Indeed, by the Cauchy-Schwartz inequality, for any function g, we have

$$\begin{aligned}\left(P_0\underline{x}\, r_0'\left(\underline{\theta}_0'\underline{x}\right) g(\underline{\theta}_0'\underline{x})\right)^2 &= \left(E_0E_0(\underline{X}|\underline{\theta}_0'\underline{X})r_0'(\underline{\theta}_0'\underline{X})g(\underline{\theta}_0'\underline{X})\right)^2 \\ &\leq E_0\left[E_0(\underline{X}|\underline{\theta}_0'\underline{X})^2(r_0')^2(\underline{\theta}_0'\underline{X})\right]E_0g^2(\underline{\theta}_0'\underline{X}))^2.\end{aligned}$$

The first term on the right is strictly smaller than $E_0\underline{X}^2(r_0')^2(\underline{\theta}_0'\underline{X})$ unless $\underline{X}r_0'(\underline{\theta}_0'\underline{X})$ is a function of $\underline{\theta}_0'\underline{X}$, which is excluded by our assumptions. This concludes the proof of the theorem. □

Lemma 4.4.10.

$$\begin{aligned}&P_0\left[r\left(\underline{\theta}'\underline{x}\right)-r_0\left(\underline{\theta}_0'\underline{x}\right)-\left(-\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}\, r_0'\left(\underline{\theta}_0'\underline{x}\right)+\left(r-r_0\right)\left(\underline{\theta}_0'\underline{x}\right)\right)\right]^2 \\ &\lesssim \left\|\underline{\theta}-\underline{\theta}_0\right\|_m^{5/2} J(r) + \left\|\underline{\theta}-\underline{\theta}_0\right\|_m^2 P_0(r'-r_0')^2\left(\underline{\theta}_0\underline{x}\right).\end{aligned}$$

Proof. The left side is equal to

$$\begin{aligned}&P_0\left[r\left(\underline{\theta}'\underline{x}\right)-r\left(\underline{\theta}_0'\underline{x}\right)+\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}\, r_0'\left(\underline{\theta}_0'\underline{x}\right)\right]^2 \\ &= P_0\left[r\left(\underline{\theta}'\underline{x}\right)-r\left(\underline{\theta}_0'\underline{x}\right)-\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}\, r'\left(\underline{\theta}_0'\underline{x}\right)+\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}\left(r'+r_0'\right)\left(\underline{\theta}_0'\underline{x}\right)\right]^2 \\ &= P_0\left[\left(\underline{\theta}-\underline{\theta}_0\right)'\nabla_{\underline{\theta}}\, r(\tilde{\underline{\theta}}'\underline{x})-\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}\, r'\left(\underline{\theta}_0'\underline{x}\right)+\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}\left(r'+r_0'\right)\left(\underline{\theta}_0'\underline{x}\right)\right]^2 \\ &= P_0\left[\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}r'(\tilde{\underline{\theta}}'\underline{x})-\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}\, r'\left(\underline{\theta}_0'\underline{x}\right)+\left(\underline{\theta}-\underline{\theta}_0\right)'\underline{x}\left(r'+r_0'\right)\left(\underline{\theta}_0'\underline{x}\right)\right]^2 \\ &\lesssim \left\|\underline{\theta}-\underline{\theta}_0\right\|_m^2\left\|\underline{x}\right\|_\infty^2 P_0\left[r'(\tilde{\underline{\theta}}'\underline{x})-r'\left(\underline{\theta}_0'\underline{x}\right)+\left(r'+r_0'\right)\left(\underline{\theta}_0'\underline{x}\right)\right]^2 \\ &\lesssim \left\|\underline{\theta}-\underline{\theta}_0\right\|_m^2\left\|\underline{x}\right\|_\infty^2\left\{P_0[r'(\tilde{\underline{\theta}}'\underline{x})-r'\left(\underline{\theta}_0'\underline{x}\right)]^2+P_0(r'-r_0')^2\left(\underline{\theta}_0'\underline{x}\right)\right\}\end{aligned}$$

for $\tilde{\underline{\theta}}$ in the line segment $L(\underline{\theta},\underline{\theta}_0)$.

Now $\left|r'(\tilde{\underline{\theta}}\,\underline{x})-r'\left(\underline{\theta}_0\underline{x}\right)\right| \lesssim J(r)\left\|\tilde{\underline{\theta}}-\underline{\theta}_0\right\|_m^{1/2}\left\|\underline{x}\right\|_\infty$ and $\left\|\underline{x}\right\|_\infty$ is bounded. The result follows. □

We write up here Lemma 5.7 of Murphy *et al.* (1999) that is now renumerated in the following way:

Lemma 4.4.11. *Let g_1 and g_2 be measurable functions such that $(Pg_1g_2)^2 \leq cPg_1^2g_2^2$ for a constant $c < 1$. Then*

$$P(g_1 + g_2)^2 \geq (1 - \sqrt{c})(Pg_1^2 + Pg_2^2).$$

4.4.3 Asymptotic efficiency for the Euclidean parameter

In this section we show that the first component of the estimator $(\hat{\underline{\theta}}, \hat{r})$ is asymptotically normal with covariance matrix equal to the inverse of the efficient information matrix, that is it is asymptotically efficient in the semiparametric sense. It is obviously assumed that the efficient information matrix is nonsingular.

Before to enunciate the main theorem of this section, it is illustrated the general setting in which the following Proposition 4.4.12 (van der Vaart (1996)) works.
Suppose that the observations are i.i.d. sample from a density $p_{\theta,\eta}$ indexed by a Euclidean parameter θ and an arbitrary parameter η. For every parameter (θ, η) let $\tilde{l}_{\theta,\eta}$ be an arbitrary measurable vector-valued function such that $\tilde{l}_{\theta_0,\eta_0}$ is the efficient score function for the parameter θ at (θ_0, η_0). We consider estimators $(\hat{\theta}_n, \hat{\eta}_n)$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{\hat{\theta}_n,\hat{\eta}_n}(X_i) = o_{P}\left(\frac{1}{n^{1/2}}\right). \tag{4.4.5}$$

Therefore, such estimators need not necessarily satisfy the efficient score equation in its full strength. Really, note that the equation (4.4.5) may be satisfied even if the efficient score function is not an actual score function, in which case the approach still holds. A further note is that the estimators $(\hat{\theta}_n, \hat{\eta}_n)$ need not be the maximum

likelihood estimators. As this in our case, any consistent estimators for which (4.4.5) is valid could be used.

The following proposition yields the asymptotic normality of the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ under regularity conditions and structural "no bias"-condition

$$P_{\hat{\theta}_n,\eta_0}\tilde{l}_{\hat{\theta}_n,\hat{\eta}_n} = o_P\left(\frac{1}{n^{1/2}}\right). \tag{4.4.6}$$

It is formulated in terms of empirical process theory reviewed in the preceding chapter.

Proposition 4.4.12. *Suppose that the model $\theta \mapsto p_{\theta,\eta_0}$ is differentiable in quadratic mean at θ_0, that $(P_{\theta_0,\eta_0} + P_{\theta,\eta_0})\left\|\tilde{l}_{\theta,\eta}\right\|^2 = O(1)$, and that $(\theta,\eta) \mapsto \tilde{l}_{\theta,\eta}$ is continuous in P_{θ_0,η_0}-probability at (θ,η_0). Furthermore, suppose that the class of functions $\tilde{l}_{\theta,\eta}$ is P_{θ_0,η_0}-Donsker for (θ,η) ranging over a neighborhood of (θ_0,η_0). If $(\hat{\theta}_n,\hat{\eta}_n)$ is consistent for (θ_0,η_0) and (4.4.5) and (4.4.6) are satisfied, then the sequence $\sqrt{n}(\hat{\theta}-\theta_0)$ is asymptotically m-variate Gaussian with mean zero and covariance matrix the inverse of the efficient information matrix $\tilde{I}_{\underline{\theta}_0,\eta_0} = E_{\underline{\theta}_0,\eta_0}\left[\tilde{l}_{\underline{\theta}_0,\eta_0}\tilde{l}'_{\underline{\theta}_0,\eta_0}\right]$, where $\tilde{l}_{\underline{\theta}_0,\eta_0}$ is the efficient score function at (θ_0,η_0).*

Proof. See van der Vaart (1996). □

Lemma 4.4.13. *For every $M \geq 1$,*

$$\log N_{[]}\left(\epsilon, \left\{r\left(\underline{\theta}'\,\underline{x}\right) : \underline{\theta} \in \underline{\Theta}, J(r) \leq M\right\}, L_2(P_0)\right) \lesssim \left(\frac{M}{\epsilon}\right)^{1/2}.$$

Proof. Let $r_{|D}$ be the restriction of r to D and using Lemmas 4.4.1 and 4.4.3, we have

$$\log N_{[]}\left(\epsilon, \left\{r_{|D} : J(r) \leq M\right\}, \|\cdot\|_\infty\right) \lesssim \left(\frac{M}{\epsilon}\right)^{1/2}.$$

Now we may construct a net over the class of functions of interest, by first choosing an ϵ/M-net $\underline{\theta}_1,\ldots,\underline{\theta}_p$ over $\underline{\Theta}$ (for the Euclidean distance in $\mathbb{R}^m$), next choosing an

ϵ-net $r_1, \ldots, r_q$ over the functions $r_{|D}$ (for the supremum metric), and finally forming all functions $r_i(\underline{\theta}_j\, \underline{x})$. Then for every $(\underline{\theta}, r)$ there exists $(\underline{\theta}_j, r_i)$ such that

$$\left|r(\underline{\theta}\, \underline{x}) - r_i(\underline{\theta}_j\, \underline{x})\right| \leq \|r'\|_\infty \left\|\underline{\theta} - \underline{\theta}_j\right\|_m \|\underline{x}\|_\infty + \left\|r_{|D} - r_i\right\|_\infty \lesssim M\frac{\epsilon}{M} + \epsilon \lesssim \epsilon.$$

If $B > 0$ is such that $\underline{\Theta} \subseteq [-B, B]^m$, then we need at most $\left(\frac{2BM}{\epsilon}\right)^m$ points $\underline{\theta}_j$ to get $\left\|\underline{\theta} - \underline{\theta}_j\right\|_m < \epsilon/M$, for some $\underline{\theta}_j$ and we need at most $\exp\left(C\,(M/\epsilon)^{1/2}\right)$ points r_i for some C. Thus, the entropy for the uniform norm of the class of function in the lemma is bounded by a multiple of $(M/\epsilon)^{1/2} + m\log(M/\epsilon)$. Consequently, the bracketing entropy for the L_2-norm is bounded similarly. □

Theorem 4.4.14. *Suppose that the assumptions A1.-A8. listed previously hold, that r_0 is a monotone function (e.g. increasing), that the functions $u \mapsto E(\underline{X}|\underline{\theta}'\underline{X} = u)$ can be chosen three times continuously differentiable and form a P_0-Donsker class when $\underline{\theta}$ range over $\underline{\Theta}$, and that r_0 is three times continuously differentiable on the interval D with nonzero derivative. Then $\sqrt{n}(\hat{\underline{\theta}} - \underline{\theta}_0)$ is asymptotically m-variate Gaussian with mean zero and covariance matrix the inverse of the efficient information matrix $\tilde{I}_{\underline{\theta}_0, r_0} = E_{\underline{\theta}_0, r_0}\left[\tilde{l}_{\underline{\theta}_0, r_0}\tilde{l}'_{\underline{\theta}_0, r_0}\right]$, for $\tilde{l}_{\underline{\theta}, r}$ given by (4.4.7).*

Proof. We shall apply the proposition 4.4.12 with $(\theta, \eta) = (\underline{\theta}, r)$ and $\tilde{l}_{\theta,\eta}$ the efficient score function for the model, for every $(\underline{\theta}, r)$, that we denote with $\tilde{l}_{\underline{\theta}, r}$. We construct suitable m-dimensional submodels in order to show that the efficient score equation (4.4.5) is satisfied. For this choice of functions $\tilde{l}_{\underline{\theta}, r}$, the bias condition (4.4.6) is satisfied trivially, with the left side vanishing, as will follow from the direct calculation later in this section.

We begin with the construction of the efficient score function for $\underline{\theta}$, $\tilde{l}_{\underline{\theta}, r}$. The ordinary

score function for $\underline{\theta}$ of the model is the function

$$\dot{l}_{\underline{\theta},r}(y,\underline{x}) = \frac{1}{\sigma^2}\left(y - r\left(\underline{\theta}'\underline{x}\right)\right) r'\left(\underline{\theta}'\underline{x}\right)\underline{x}.$$

The score function for the submodel given by $r_t = r + \underline{t}'\underline{B}$ is equal to

$$A_{\underline{\theta},r}\underline{B}(y,\underline{x}) = \frac{1}{\sigma^2}\left(y - r\left(\underline{\theta}'\underline{x}\right)\right)\underline{B}\left(\underline{\theta}'\underline{x}\right)$$

Of course, the surface $r_t = r + \underline{t}'\underline{B}$ defines a true submodel only for perturbations $\underline{B}$ such that r_t is nondecreasing and $B_i(-\infty) = B_i(\infty) = 0$ for $i = 1,\ldots,m$. If we restrict the model by requiring that $J(r) < \infty$, then B_i should also have $J(B_i) < \infty$ for $i = 1,\ldots,m$. Comparing the formulas $\dot{l}_{\underline{\theta},r}$ and $A_{\underline{\theta},r}\underline{B}$, we see that minimizing $P_{\underline{\theta},r}\left(\dot{l}_{\underline{\theta},r} - A_{\underline{\theta},r}\underline{B}\right)^2$ over $\underline{B}$ is a weighted least squares problem that is solved by

$$\underline{B}_{\underline{\theta},r}(u) = r'(u)\underline{h}_{\underline{\theta}}(u),$$

for

$$\underline{h}_{\underline{\theta}}(u) = E\left[\underline{X}|\underline{\theta}'\underline{X} = u\right].$$

Then the efficient score function for $\underline{\theta}$ is given by

$$\tilde{l}_{\underline{\theta},r}(y,\underline{x}) = \frac{1}{\sigma^2}\left(y - r\left(\underline{\theta}'\underline{x}\right)\right) r'\left(\underline{\theta}'\underline{x}\right)\left(\underline{x} - E\left[\underline{X}|\underline{\theta}'\underline{X} = \underline{\theta}'\underline{x}\right]\right). \tag{4.4.7}$$

Note, however, that the present functions $\underline{B}_{\underline{\theta},r}$ satisfy $J(\underline{B}_{\underline{\theta},r}) < \infty$ only if r is three times differentiable, which is more than we initially assume for every r in the model. Therefore, we will construct and use a more complicated type of surface r_t, which is well defined as soon as $J(r) < \infty$. From this it is clear that $\tilde{l}_{\underline{\theta},r}$ is the efficient score function already under the condition that $J(r) < \infty$. The construction of this surface is necessary because our proof of asymptotic normality of $\hat{\underline{\theta}}$ uses a perturbation of $\hat{r}$, for which the finiteness of $J(\hat{r})$ is guaranteed by definition, but possibly not a smooth third derivative.

By assumption the support $D_{\underline{\theta}}$ of the variables $\underline{\theta}'\underline{X}$ (under P_0) is contained strictly within the interval D, for every $\underline{\theta}$. Therefore, for every $(\underline{\theta}, \underline{t})$ such that $\|\underline{\theta} - \underline{t}\|_m$ is sufficiently close to zero, there exists a strictly increasing , infinitely often differentiable function $u \mapsto \psi_{\underline{\theta},\underline{t}}(u)$ with

$$\begin{aligned} \psi_{\underline{\theta},\underline{t}}(u) &= u, \qquad u \in D_{\underline{\theta}}, \\ \psi_{\underline{\theta},\underline{t}}(u + (\underline{\theta} - \underline{t})'\underline{h}_{\underline{\theta}}(u)) &= u, \qquad u \in \delta D. \end{aligned}$$

Moreover, we can ensure that $(u, \underline{t}) \mapsto \psi_{\underline{\theta},\underline{t}}(u)$ is infinitely often differentiable at $u \in D, \underline{t} = \underline{\theta}$ as well. The second identity in the preceding display ensures that $\psi_{\underline{\theta},\underline{t}}(D) = D$ and will be used to control the partial derivative of $J(r_{\underline{t}}(\underline{\theta}, r))$ with respect to $\underline{t}$ in the argument below.

For a given pair $(\underline{\theta}, r)$, we now define a least favorable submodel as

$$r_{\underline{t}}(\underline{\theta}, r)(u) = r \circ \psi_{\underline{\theta},\underline{t}}(u + (\underline{\theta} - \underline{t})'\underline{h}_{\underline{\theta}}(u)).$$

Then $r_{\underline{\theta}}(\underline{\theta}, r)(\underline{\theta}'\underline{x}) = r(\underline{\theta}'\underline{x})$ for every $\underline{x}$ in the support of $\underline{X}$, and, with a dot denoting differentiation with respect to $\underline{t}$,

$$\begin{aligned} r_{\underline{t}}(\underline{\theta}, r)'(u) = r' &\circ \psi_{\underline{\theta},\underline{t}}(u + (\underline{\theta} - \underline{t})'\underline{h}_{\underline{\theta}}(u)) \\ &\times \psi'_{\underline{\theta},\underline{t}}(u + (\underline{\theta} - \underline{t})'\underline{h}_{\underline{\theta}}(u))(1 + (\underline{\theta} - \underline{t})'\underline{h}'_{\underline{\theta}}(u)), \\ \dot{r}_{\underline{t}}(\underline{\theta}, r)(u) = r' &\circ \psi_{\underline{\theta},\underline{t}}(u + (\underline{\theta} - \underline{t})'\underline{h}_{\underline{\theta}}(u)) \\ &\times \left[\dot{\psi}_{\underline{\theta},\underline{t}}(u + (\underline{\theta} - \underline{t})'\underline{h}_{\underline{\theta}}(u)) - \psi'_{\underline{\theta},\underline{t}}(u + (\underline{\theta} - \underline{t})'\underline{h}_{\underline{\theta}}(u))\underline{h}_{\underline{\theta}}(u)\right], \end{aligned}$$

$$\frac{\partial}{\partial \underline{t}} \log p_{\underline{t}, r_{\underline{t}}(\underline{\theta}, r)}(y, \underline{x}) = \frac{1}{\sigma^2} \left(y - r_{\underline{t}}(\underline{\theta}, r)\left(\underline{t}'\underline{x}\right)\right) \left(\underline{x} r_{\underline{t}}(\underline{\theta}, r)'(\underline{t}'\underline{x}) + \dot{r}_{\underline{t}}(\underline{\theta}, r)(\underline{t}'\underline{x})\right).$$

Evaluated at $\underline{t} = \underline{\theta}$ this yields to the efficient score function $\tilde{l}_{\underline{\theta},r}$. Next, with $\phi_{\underline{\theta},\underline{t}}(u) = \psi_{\underline{\theta},\underline{t}}(u + (\underline{\theta} - \underline{t})'\underline{h}_{\underline{\theta}}(u))$,

$$r_{\underline{t}}(\underline{\theta}, r)''(u) = r'' \circ \phi_{\underline{\theta},\underline{t}}(u)\phi'_{\underline{\theta},\underline{t}}(u)^2 + r' \circ \phi_{\underline{\theta},\underline{t}}(u)\phi''_{\underline{\theta},\underline{t}}(u).$$

For sufficiently small $\|\underline{\theta} - \underline{t}\|_m$ the map $\phi_{\underline{\theta},\underline{t}}$ is a strictly increasing, three times differentiable bijection on D. Therefore,

$$J^2(r_{\underline{t}}(\underline{\theta}, r)) = \int_D \left[r''(v)(\phi'_{\underline{\theta},\underline{t}} \circ \phi^{-1}_{\underline{\theta},\underline{t}}(v))^2 + r'(v)(\phi''_{\underline{\theta},\underline{t}} \circ \phi^{-1}_{\underline{\theta},\underline{t}}(v)) \right]^2 \frac{dv}{\phi'_{\underline{\theta},\underline{t}} \circ \phi^{-1}_{\underline{\theta},\underline{t}}(v)}.$$

It follows that $J(r_{\underline{t}}(\underline{\theta}, r)) < \infty$ whenever $J(r) < \infty$ and $\|\underline{\theta} - \underline{t}\|_m$ is sufficiently close to zero. Furthermore, this quantity is partially differentiable with respect to $\underline{t}$ in a neighborhood of $\underline{\theta}$, with derivative at $\underline{t} = \underline{\theta}$ bounded in absolute value by a multiple of $\int_D (r''(v)^2 + r'(v)^2)dv \lesssim J^2(r)$. Since $\hat{\underline{\theta}}, \hat{r}$ maximizes the likelihood and $r_{\hat{\underline{\theta}}}(\hat{\underline{\theta}}, \hat{r}) = \hat{r}$, the value $\hat{\underline{\theta}}$ maximizes the function $\underline{t} \mapsto \log \prod p_{\underline{t}, r_{\underline{t}}(\hat{\underline{\theta}}, \hat{r})}(y_i, \underline{x}_i) - \hat{\lambda}^2 J^2(r_{\underline{t}}(\hat{\underline{\theta}}, \hat{r}))$. It follows that

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{l}_{\hat{\underline{\theta}}, \hat{r}}(Y_i, \underline{X}_i) - \widehat{\lambda}^2 \frac{\partial}{\partial \underline{t}_{|\underline{t}=\hat{\underline{\theta}}}} J^2 \left(r_{\underline{t}}(\hat{\underline{\theta}}, \hat{r}) \right) = \underline{0}.$$

In view of (4.3.4) and the fact that $J(\hat{r}) = O_P(1)$, the condition (4.4.5) is verified, that is

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{l}_{\hat{\underline{\theta}}, \hat{r}}(Y_i, \underline{X}_i) = o_{P}(n^{-1/2}).$$

Moreover, the condition (4.4.6) is also verified and we have

$$P_{\hat{\underline{\theta}}, r_0} \tilde{l}_{\hat{\underline{\theta}}, \hat{r}} = o_{P} \left(\frac{1}{n^{1/2}} \right).$$

In order to verify the regularity conditions of Proposition 4.4.12, note first, by Lemma 4.4.1, that

$$\begin{aligned} |\hat{r}'(\hat{\underline{\theta}}\underline{x}) - \hat{r}'(\hat{\underline{\theta}}_0\underline{x})| &\lesssim J(\hat{r}) \left\| \hat{\underline{\theta}} - \underline{\theta}_0 \right\|_m^{1/2}, \\ |\hat{r}(\hat{\underline{\theta}}\underline{x}) - \hat{r}(\hat{\underline{\theta}}_0\underline{x})| &\lesssim (1 + J(\hat{r})) \left\| \hat{\underline{\theta}} - \underline{\theta}_0 \right\|_m. \end{aligned}$$

By Lemma 4.4.8, since $\hat{\underline{\theta}} \xrightarrow{P} \underline{\theta}_0$, the right sides converge to zero in probability. Combined with the convergence $\hat{r} \xrightarrow{P} r_0$ with respect to the uniform norm on the closure of the support of $\underline{\theta}_0\underline{X}$, and the assumption that r_0 is bounded, the functions $(1/\sigma^2)(y - \hat{r}(\hat{\underline{\theta}}'\underline{x}))$

are seen to be bounded with probability tending to one. Furthermore, the functions $\hat{r}'(\underline{\hat{\theta}}'\underline{x})$ are uniformly bounded. Since also the functions $\underline{h}_{\underline{\hat{\theta}}}$ are uniformly bounded, it follows that the functions $\tilde{l}_{\underline{\hat{\theta}},\hat{r}}(y,\underline{x})$ are uniformly bounded with probability tending to one. Thus $(P_{\underline{\theta}_0,r_0} + P_{\underline{\theta},r_0})||\tilde{l}_{\underline{\theta},r}||^2 = O(1)$ is bounded trivially. Furthermore, $\tilde{l}_{\underline{\hat{\theta}},\hat{r}}(y,\underline{x}) \to \tilde{l}_{\underline{\hat{\theta}}_0,\hat{r}_0}(y,\underline{x})$ for $P_{\underline{\theta}_0,r_0}$-almost every $(y,\underline{x})$.

It is straightforward to check that the model $\underline{\theta} \mapsto p_{\underline{\theta},r_0}$ is differentiable in quadratic mean at $\underline{\theta}_0$ with score function $\dot{l}_{\underline{\theta}_0,r_0}$ as given previously.

By Lemma 4.4.13 and the bracketing-central-limit-theorem of Ossiander (Cf. Theorem 2.5.6 of van der Vaart and Wellner (1996)), the class of functions $r(\underline{\theta}'\underline{x})$, with r ranging over the monotone (e.g. increasing) function with $J(r) \leq M$ and $\underline{\theta} \in \underline{\Theta}$, is P_0-Donsker. For the functions $r(\underline{\theta}'\underline{x})$ restricted to be bounded, the functions $(1/\sigma^2)(y - r(\underline{\theta}'\underline{x}))$ are Lipschitz transformations of the functions $(r(\underline{\theta}'\underline{x}), y)$. Thus, under this restriction, this class is P_0-Donsker by Theorem 2.10.6 of van der Vaart and Wellner (1996). It is also uniformly bounded.

By Lemma 4.4.15 (below) and the bracketing-central-limit-theorem, the class of functions $r'(\underline{\theta}'\underline{x})$ with $J(r) \leq M$ is P_0-Donsker. It is also uniformly bounded.

The class of function $\underline{x} - \underline{h}_{\underline{\theta}}(\underline{\theta}'\underline{x})$ is P_0-Donsker by assumption.

Combining these results, we conclude by Theorem 2.10.6 of van der Vaart and Wellner (1996) that the class of functions $\tilde{l}_{\underline{\theta},r}$ with $\underline{\theta}$ ranging over $\underline{\Theta}$ and r over the monotone functions such that $J(r) \leq M$ and such that $\|r - r_0\|_D$ is sufficiently small, is P_0-Donsker. □

Lemma 4.4.15. *For every $M \geq 1$,*

$$\log N_{[]}\left(\epsilon, \left\{r'(\underline{\theta}'\underline{x}) : \underline{\theta} \in \underline{\Theta}, J(r) \leq M\right\}, L_2(P_0)\right) \lesssim \left(\frac{M}{\epsilon}\right).$$

Proof. By Lemma 4.4.1 the class of derivative r' of functions r with $J(r) \leq M$ is

uniformly bounded by a multiple of $1 + J(r) \lesssim M$ on D. Clearly, $\int_D (r')'^2(u)du = J^2(r) \leq M^2$. Therefore, by Lemma 4.4.3

$$\log N\left(\epsilon, \left\{r'_{|D} : J(r) \leq M\right\}, \|\cdot\|_\infty\right) \lesssim \left(\frac{M}{\epsilon}\right).$$

Furthermore, by Lemma 4.4.1 $|r'(s) - r'(s_0)| \leq |s - s_0|^{1/2} J(r)$ for every $s, s_0 \in D$.
Now we can construct a net over the class of functions of interest, by first choosing an $(\epsilon/M)^2$-net $\underline{\theta}_1, \ldots, \underline{\theta}_p$ over $\underline{\Theta}$ (for the Euclidean distance in $(R)^m$) next choosing an ϵ-net $s_1, \ldots, s_q$ over the functions $r'_{|D}$ (for the supremum metric), and finally forming all functions $s_i(\underline{\theta}_j \underline{x})$. Then for every $(\underline{\theta}, r)$ there exists $(\underline{\theta}_j, s_i)$ such that

$$\left|r'(\underline{\theta}\,\underline{x}) - s_i(\underline{\theta}_j\,\underline{x})\right| \leq M \left\|\underline{\theta} - \underline{\theta}_j\right\|_m^{1/2} \|\underline{x}\|_\infty + ||r'_{|D} - s_i||_\infty \lesssim M\frac{\epsilon}{M} + \epsilon \lesssim \epsilon.$$

If $B > 0$ is such that $\underline{\Theta} \subseteq [-B, B]^m$, then we need at most $\left(\frac{2BM^2}{\epsilon^2}\right)^m$ points $\underline{\theta}_j$ to get $\left\|\underline{\theta} - \underline{\theta}_j\right\|_m < (\epsilon/M)^2$, for some $\underline{\theta}_j$ and we need at most a power of (M/ϵ) points s_i. Thus, the entropy for the uniform norm of the class of function in the lemma is bounded by a multiple of $(M/\epsilon) + m\log(M/\epsilon)$. Consequently, the bracketing entropy for the L_2-norm is bounded similarly. □

From these theorems, we argue that, under the regularity conditions listed previously, $\underline{\theta}_0$ can be estimated with a $n^{-1/2}$ rate of convergence in probability, which is the typical rate of convergence achieved by parametric estimators under i.i.d. sampling. The $n^{-1/2}$ convergence rate of the estimator $\hat{\underline{\theta}}$ implies that the estimator is not infinitely inefficient compared with conventional parametric approaches even though the model is not restricted within a finite-dimensional space.
This rate of convergence for $\hat{\underline{\theta}}$ and its efficiency, is not a surprise in fact, as we saw in section 1.5, a lot of the exposed estimators of $\underline{\theta}$ achieve the same rate of convergence, under, more o less, similar assumptions as these made in the study of the penalized

maximum likelihood estimator.

Particular attention and a special comment we owe to the paper of Yu and Ruppert (2002). They studied partially linear single-index models, a generalization of single-index models examinated in this thesis, even without assumptions of monotonicity for the link function and of normality for the error variable ϵ. In their work, they focused the attention in the estimation of $\underline{\theta}$ by means of penalized least squares and by means a P-splines estimation which is a generalization of smoothing splines, allowing a more flexible choice of knots and penalty. As a direct least squares fitting method this approach is computationally stable and, unlike our estimation in the following chapters, by natural cubic splines, their approach has the benefit to be rapid. About the asymptotic properties, they proved consistency and asymptotic normality for their penalized least squares estimator of $\underline{\theta}$. The only drawback in that paper is in the strong assumption that the regression function r to be estimated, is supposed to be a spline function. The contribution of this thesis with respect to the article of Yu and Ruppert (2002) is in a similar study, but about any regression function provided that it is monotone, as a sentence written in this work explain: "*Moreover asymptotic results with an increasing number of knots* (i.e. when the link function is not supposed to be a spline function) *are limited to rates of convergence; at least this is true of all results of which we are aware.*"

An interesting generalization of the study in this thesis, could be then removing the assumptions of normality for the error variable and of monotonicity for the regression function.

About the rate of convergence of the estimator $\hat{r}$, this is also not completely a surprise, even if usually in literature, we don't care about consistency and rate of convergence for the estimator of the regression function in single-index models. Anyway, our study in this thesis, is about the couple $(\underline{\theta}, r)$ and we investigated also in the properties of $\hat{r}$.

Our aim and our hope were that, imposing the constraint of monotonicity, it could have been possible to obtain a faster rate of convergence than $n^{-4/5}$, which is the rate of nonparametric estimators of r, when the regression function r is only supposed to be twice differentiable. Indeed, in solely nonparametric regression context, it happens that if the true regression function is strictly monotone (e.g. if the first derivative is bounded away from zero), then with sufficient smoothness assumptions, the monotonicity restrictions become nonbinding as the sample size increases. The constrained estimator then has the same convergence rate as the unconstrained estimator (see Utreras (1985)). This negative finding, however, does not imply that monotonicity will be uninformative in small samples. Indeed, one could argue that, given the paucity of a priori information present in nonparametric estimation, any additional constraints should be exploited as far as possible particularly in moderately sized samples.

Chapter 5

Nonparametric Regression with Smoothing Splines

5.1 Introduction

There are a lot of methods to study nonparametric regression functions such as local regression methods and penalization methods or series-based smoothers, including wavelets (Tarter and Lock (1993), Ogden (1996)). The former includes kernel regression(Wand and Jones (1995)) and local polynomial regression(Fan and Gijbels (1996)). Penalization methods lead to methods based on splines (Eubank (1988, 1999a), Wahba (1990), Friedman (1991), Green and Silverman (1994), Stone *et al.* (1997) and Hansen and Kooperberg (2002)). All these estimators are linear smoothers, in the sense as it is explained later.

In this chapter we will present some of the known theory on smoothing spline estimators for the nonparametric regression curve, useful for our purposes. Smoothing splines are indeed used when we wish to fit a data set using a function that reflects the key features of the data but retains some degree of smoothness.

Let us suppose that observations are taken on the random variable Y at n predetermined values of the independent variable X. Let (y_i, x_i), $i = 1, \ldots, n$, be the values of

Y and X and assume that y_i and x_i are related by the regression model

$$y_i = r(x_i) + \epsilon_i \qquad i = 1, \ldots, n \tag{5.1.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$ and $r(x_i)$ are values of some unknown function $r \in W_2^p[a, b]$, with $a, b \in \mathbb{R}$, at the *design points* $x_1, \ldots, x_n$. Without loss of generality, we will assume here, that $a \leq x_1 \leq \cdots \leq x_n \leq b$.

A natural measure of smoothness associated with a function $r \in W_2^p[a, b]$ is $\int_a^b (r^{(p)}(x))^2 dx$ while a standard measure of goodness-of-fit to the data is the average residual sum-of-squares $ARSS_n(r) = n^{-1} \sum_{i=1}^n [y_i - r(x_i)]^2$. Thus, a good estimator of r could then be obtained by the function $\hat{r}$ that minimizes

$$\frac{1}{n} \sum_{i=1}^{n} [y_i - r(x_i)]^2 + \hat{\lambda}_n \int_a^b (r^{(p)}(x))^2 dx, \quad \hat{\lambda}_n > 0, \tag{5.1.2}$$

over $r \in W_2^p[a, b]$. Note that this minimization problem can be led back to that in the preceding chapter when we consider the one-dimensional case with $m = 1$.

The result of the minimization of (5.1.2), is the smoothing spline estimator of the nonparametric regression function to be studied.

5.2 Meaning of the penalization term

In chapter 4 a first explanation of the use of the penalized criterion (5.1.2) is given. To have a better insight into the meaning of the penalization, we provide now another motivation that can be obtained from polynomial regression. Using Taylor's theorem, we can write the model (5.1.1) as

$$y_i = \sum_{j=1}^{p} \alpha_j x_i^{j-1} + Rem(x_i) + \epsilon_i \qquad i = 1, \ldots, n$$

for constants $\alpha_1, \ldots, \alpha_p$ and where

$$Rem(x_i) = \frac{1}{(p-1)!} \int_a^b r^{(p)}(x)(x_i - x)_+^{p-1} dx.$$

From the Cauchy-Schwarz inequality we have

$$\max_{1\leq i\leq n} Rem(x_i)^2 \leq \frac{J_p(r)}{(2p-1)[(p-1)!]^2},$$

where

$$J_p(r) = \int_a^b r^{(p)}(x)^2 dx$$

is the smoothness measure from the criterion (5.1.2). The value of $J_p(r)$ can be viewed as providing a bound on how far (5.1.1) departs from a polynomial model.
If we knew, for example, that

$$J_p(r) \leq \rho\,, \qquad \rho \geq 0, \tag{5.2.1}$$

then this would provide us with information on how far the regression curve could depart from a polynomial form and we could input this information into the estimation process. One way of accomplishing this would be to estimate r by the minimizer of

$$RSS_n(r) = \sum_{i=1}^{n}(y_i - r(x_i))^2 \tag{5.2.2}$$

over all functions $r \in W_2^p[a,b]$ which satisfy (5.2.1). This is equivalent to minimizing $(1/n)RSS_n(r) + \hat{\lambda}_n(J_p(r) - \rho)$, where $\hat{\lambda}_n$ is now the Lagrange multiplier for the constraint. But this is essentially the criterion (5.1.2) and produces the same estimator of r as (5.1.2). The relationship between the estimators obtained from (5.2.1)-(5.2.2) and (5.1.2) is made precise in the following theorem of Schoenberg (1964).

Theorem 5.2.1. *Assume that $n \geq p$ and let $r(\cdot,\rho)$ the minimizer of (5.2.2) in $W_2^p[a,b]$ subject to the constraint (5.2.1). Let $r_{\hat{\lambda}_n}$ denote the minimizer of (5.1.2) in $W_2^p[a,b]$. Then, there is a computable constant ρ_0 such that the sets $\{r(\cdot,\rho) : 0 \leq \rho \leq \rho_0\}$ and $\left\{r_{\hat{\lambda}_n}(\cdot) : 0 \leq \hat{\lambda}_n \leq \infty\right\}$ are identical in that for any value of $\hat{\lambda}_n$ there is a unique ρ such that $r_{\hat{\lambda}_n}(\cdot) = r(\cdot,\rho)$ and conversely. If $\rho \leq \rho_0$, then $J(r(\cdot,\rho)) = \rho$.*

Theorem 5.2.1 has the consequence that, as we will see later, the solution to the constrained minimization problem posed by (5.2.1)-(5.2.2) is a smoothing spline estimator of r corresponding to some value of $\hat{\lambda}_n$. On the other hand, the choice of a particular value for $\hat{\lambda}_n$ corresponds to the assumption that $J_p(r) \leq \rho_{\hat{\lambda}_n}$ with $\rho_{\hat{\lambda}_n} = J_p(r)$ reflecting the beliefs about the magnitude of the remainder terms $Rem(x_i)$, $i = 1, \ldots, n$. Since the choice $\rho = 0$ produces a polynomial regression estimator, it follows that smoothing splines give an extension of polynomial regression that attempts to guard against departures from an idealized polynomial regression model.

5.3 Smoothing splines

Smoothing splines are a popular and effective technique for data smoothing. The origin of smoothing splines appears to lie in work on graduating data by Whittaker (1923). However, spline smoothing techniques were generally regarded as numerical analysis methods until extensive research by Grace Wahba demonstrated their utility for solving a host of statistical estimation problems. It has now become clear that smoothing splines provide extremely flexible data analysis tools. Indeed piecewise polynomials or smoothing splines extend the advantages of polynomials to include greater flexibility, local effects of parameter changes and the possibility of imposing useful constraints on estimated functions. Among these constraints is monotonicity, which can be an important property in many curve estimation problems. As a result, smoothing splines have become quite popular and have found applications in such diverse areas as the analysis of growth data, medicine, remote sensing experiments and economics. Extensive developments of spline smoothing methods and related techniques can now be found in several books including Schumaker (1981), de Boor (2001), Wahba (1990), Green and Silverman (1994) and Ruppert *et al.* (2003). The review of splines in statistics by

Wegman and Wright (1983) is particularly recommended. Silverman (1985) provides a very readable review of the application of spline smoothing in nonparametric regression estimation.

Polynomial, $r(x) = \sum_{i=1}^{k} \alpha_i x^{i-1}$, owe their central role in practical mathematics to two features: they are linear in the parameters α_i to be estimated, and the functions x^{i-1} that are linearly combined are easy to manipulate algebraically and numerically, especially with respect to differentiation and integration. However, polynomials do have one serious limitation: a lack of flexibility in the sense that changing the behavior of r near to one value x_1 has radical implications for its behavior for any other value x_2. This poses the problem of how to retain flexibility where it is needed, while leaving the function elsewhere either relatively unaffected or constrained as desired.

We want now give a definition and some properties of smoothing splines and in particular of natural splines. First of all, let us define a *spline of order k with knots at* $x_1, \ldots x_n$ to be any function s of the form

$$s(x) = \sum_{j=0}^{k-1} \alpha_j x^j + \sum_{j=1}^{n} \beta_j (x - x_j)_+^{k-1} \tag{5.3.1}$$

for some set of coefficients $\alpha_0, \ldots, \alpha_{k-1}, \beta_1, \ldots, \beta_n$, called *spline coefficient vector*. Here $(x - x_j)_+^{k-1}$ means $\{\max(x - x_j), 0\}^{k-1}$. Let $S^k(x_1, \ldots x_n)$ denote the vector space of all functions of the form (5.3.1).

The functions $1, x, x^2, \ldots, x^{k-1}, (x - x_1)_+^{k-1}, \ldots, (x - x_n)_+^{k-1}$ form a basis for the set of splines of order k at the knots $x_1, \ldots x_n$, called *truncated power basis*. Although the simplicity of this basis makes it attractive for statistical work, and Smith (1979) and others have used it effectively in applications, it has the rather serious disadvantage of generating considerable rounding error except for very low values of k.

The expression in (5.3.1) is then a particular form for splines expressed in term of truncated power basis. We will see that it is possible rewrite that same spline in relation to other basis as, for example, B-splines basis.

This definition in (5.3.1), of spline of order k with knots at $x_1, \dots x_n$, is equivalent to saying that

1) s is a piecewise polynomial of order k on any subinterval $[x_i, x_{i+1})$;

2) s has $k-2$ continuous derivatives and;

3) s has a discontinuous $(k-1)$st derivative with jumps at $x_1, \dots x_n$.

According to this definition, a spline is a piecewise polynomial whose different polynomial segments have been joined at the knots in such a way that are ensured certain continuity properties. Notice, in particular, that a spline is the smoothest possible piecewise polynomial which still retains a segmented nature.

In this thesis, it is of interest the set of spline functions that are *natural splines of order* $2p$ *with knots at* $x_1, \dots x_n$. These are splines of order $2p$ with knots at $x_1, \dots x_n$ that, in addition to the properties 1)-3), satisfies the further property

4) s is a polynomial of order p outside of $[x_1, x_n]$.

The name "natural spline" stems from the fact that, as a result of the property 4), s satisfies the natural boundary condition

$$r^{(p+j)}(a) = r^{(p+j)}(b) = 0, \quad j = 1, \dots, p-1.$$

Denoting by $NS^{2p}(x_1, \dots x_n)$ the set of all natural splines of order $2p$ with knots at $x_1, \dots x_n$, we have that $NS^{2p}(x_1, \dots x_n) \subset S^{2p}(x_1, \dots x_n)$, in fact $NS^{2p}(x_1, \dots x_n)$ is obtained from $S^{2p}(x_1, \dots x_n)$ taking only the splines in $S^{2p}(x_1, \dots x_n)$ satisfying the property 4). About the dimension of these spaces, we have that the dimension of $S^{2p}(x_1, \dots x_n)$ is $n + 2p$ and the dimension of the vector space $NS^{2p}(x_1, \dots x_n)$ is n. In

particular, we see that in order to be s a natural spline, we must have in (5.3.1)

$$\alpha_0 = \cdots = \alpha_{2p-1} = 0$$

since s must be a polynomial of order p for $x < x_1$. The estimation of the regression function r is then led back from an infinite dimensional estimation problem, to a finite dimensional estimation problem: the number of parameter to estimate is $n + 2p$ for splines of order $2p$ and n for natural splines with knots at $x_1, \ldots x_n$.

We can now enunciate a known theorem (see, for example, Wahba (1990)) in spline regression theory, that relates natural splines and the estimator obtained from the minimization of the the penalized maximum likelihood function or average residual sum-of-squares in (5.1.2).

Theorem 5.3.1. *If $n \geq p$, the function $\hat{r}$ that minimizes*

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - r(x_i)]^2 + \hat{\lambda}_n \int_a^b (r^{(p)}(x))^2 dx, \quad \hat{\lambda}_n > 0,$$

over $r \in W_2^p[a, b]$, is a natural spline of order $2p$ with knots at the data points $x_1, \ldots x_n$. The estimator $\hat{r}$ is called smoothing spline.

We have that $\hat{r}$ is then a $2p$th order piecewise polynomial with $2p - 2$ continuous derivatives that consists of different polynomial segments over each of the intervals $[x_i, x_{i+1}]$, $i = 1, \ldots, n-1$, and is a polynomial of order p outside of $[x_1, x_n]$. As a matter of fact, for any function $r \in W_2^p[a, b]$ criterion (5.1.2) can only become smaller if we replace r by the natural spline which agrees with r at the design points.
Cubic splines are the most common splines used in practice. They arise naturally in the penalized regression framework when in the penalization term we consider $p = 2$. The theorem above does not give an explicit form for $\hat{r}$. We will study this aspect of the smoothing splines in the next section.

5.4 Form of the estimator

In this section we will recall how to obtain an explicit expression for the smoothing spline estimator.

As we remarked at the start of this chapter, the smoothing spline estimators are linear smoothers. So, to begin, we provide some definitions.

Definition 5.4.1. *An estimator $\hat{r}$ of r is a linear smoother if, for each x, there exists a vector $\underline{\ell}(x) = (\ell_1(x), \ldots, \ell_n(x))^T$ such that*

$$\hat{r}(x) = \sum_{i=1}^{n} \ell_i(x) Y_i.$$

Define the vector of fitted values

$$\underline{\hat{r}} = (\hat{r}(x_1), \ldots, \hat{r}(x_n))^T$$

where $\underline{Y} = (Y_1, \ldots, Y_n)^T$. It then follows that

$$\underline{\hat{r}} = \mathbf{L}\underline{Y}$$

where $\mathbf{L}$ is an $n \times n$ matrix whose i^{th} row is $\ell(x_i)^T$; thus, $L_{ij} = \ell_j(x_i)$. The entries of the i^{th} row show the weights given to each Y_i in forming the estimate $\hat{r}(x_i)$.

Definition 5.4.2. *The matrix $\boldsymbol{L}$ is called the smoothing matrix or the hat matrix and the effective degrees of freedom is usually defined by*

$$\nu = tr(\boldsymbol{L}).$$

Now, we can start with the search of the form of the smoothing spline estimator.

We will consider the case where the x_i are distinct. Also we will assume that a value for $\hat{\lambda}_n$ has been selected so that the smoothing parameter value is fixed. In Section 5.6, we will show how to obtain a right value for $\hat{\lambda}_n$, completely automatically from the

data.

Since the solution of the minimization of (5.1.2) over all functions in $W_2^p[a,b]$ is a natural spline $\hat{r}$, we have then a reduction of complexity of the problem, from a infinite-dimensional to the finite-dimensional problem of minimization over the n dimensional set of natural splines. A closed form for the estimator can be derived from the following theorem (see, for example, Eubank (1988)).

Theorem 5.4.3. *Let $b_1, \dots b_n$ be a basis for the set of natural splines of order $2p$ with knots at $x_1, \dots x_n$ and define $\boldsymbol{B} = b_j(x_i)_{i,j=1,n}$. Then, if $n \geq p$ the unique minimizer of (5.1.2) is*

$$\hat{r} = \sum_{j=1}^{n} \alpha_j b_j, \tag{5.4.1}$$

where $\underline{\alpha}$ is the unique solution with respect to $\underline{\gamma}$ of the equation system

$$\left(\boldsymbol{B}^T\boldsymbol{B} + \hat{\lambda}_n\Omega\right)\underline{\gamma} = \boldsymbol{B}^T\underline{y} \tag{5.4.2}$$

where

$$\Omega = \left\{\int_a^b b_i^{(p)}(x) b_j^{(p)}(x)dx\right\}_{i,j=1,n}. \tag{5.4.3}$$

Then the smoothing spline estimators are actually linear estimators and the vector of fitted values corresponding to the smoothing spline estimator is

$$\hat{r} = \mathbf{S}_{\hat{\lambda}_n}\underline{y} \tag{5.4.4}$$

where, in analogy with linear regression, the hat matrix is

$$\mathbf{S}_\lambda = \mathbf{B}\left(\mathbf{B}^T\mathbf{B} + \hat{\lambda}_n\Omega\right)^{-1}\mathbf{B}^T. \tag{5.4.5}$$

The problems are now in the choice of a suitable basis for the set of natural splines of order $2p$ with knots at $x_1, \dots x_n$ and, in particular, of a suitable basis with which one can construct natural cubic splines (that is of fourth order, for $p = 2$) at which

it could be possible to impose the monotonicity constraint of the regression function. Moreover, it remains the problem of the selection of a right value for the smoothing parameter $\hat{\lambda}_n$.

5.5 Natural cubic splines with B-spline basis

In this section we will see how to construct natural cubic spline estimator. For the moment, let's assume that the value of $\hat{\lambda}_n$ has been specified and that the interest is now in the evaluation of the corresponding vector of fitted values $\underline{\hat{r}} = (\hat{r}(x_1), \ldots, \hat{r}(x_n))^T$. From the preceding theorem 5.4.3 and equations (5.4.1) and (5.4.2) we have that $\hat{r}$ is given by the expression

$$\hat{r} = \mathbf{B}\left(\mathbf{B}^T\mathbf{B} + \hat{\lambda}_n\Omega\right)^{-1}\mathbf{B}^T\underline{y}.$$

In order to accomplish these calculations efficiently, in the literature is suggested to use the natural spline basis functions such that $\mathbf{B}$ and the system (5.4.2) are band limited and thereby allow the fitted values to be computed in $O(n)$ calculations.

In his book, de Boor (2001), using a piecewise polynomial representation of the estimator, provides a code which implements an efficient approach that gives a $O(n)$ algorithm for computing $\hat{r}$.

We now introduce a different basis for the set of splines called the *B-spline basis* that is particularly well suited for computation. These are defined as follows.

Let $x_0 = a$ and $x_{n+1} = b$. Define new knots $\tau_1, \ldots, \tau_p$ such that

$$\tau_1 \leq \tau_2 \leq \cdots \leq \tau_p \leq x_0,$$

$\tau_{j+p} = x_j$ for $j = 1, \ldots, n$, and

$$x_{n+1} \leq \tau_{n+p+1} \leq \cdots \tau_{n+2p}.$$

The choice of extra knots is arbitrary; usually one takes $\tau_1 = \cdots = \tau_p = x_0$ and $x_{n+1} = \tau_{n+p+1} = \cdots = \tau_{n+2p}$. We define the basis functions recursively as follows. First we define

$$B_{i,1} = \begin{cases} 1 & \text{if } \tau_i \leq x \leq \tau_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n + 2p - 1$. Next, for $q \leq p$ we define

$$B_{i,q} = \frac{x - \tau_i}{\tau_{i+q-1} - \tau_i} B_{i,q-1} + \frac{\tau_{i+q} - x}{t_{i+q} - \tau_{i+1}} B_{i+1,q-1}$$

for $i = 1, \ldots, n+2p-q$. It is understood that if the denominator is 0, then the function is defined to be 0.

Theorem 5.5.1. *The functions $\{B_{i,4}, i = 1, \ldots, n\}$ are a basis for the set of cubic splines. They are called the B-spline basis functions.*

The advantage of the B-spline basis functions is that they have compact support which makes it possible to speed up calculations. See Hastie *et al.* (2001) for details.
Figure 5.1 shows the cubic B-splines basis using nine equally spaced knots on $(0, 1)$.

According to Theorem 5.3.1, $\hat{r}$ is a natural cubic spline. Hence we can write

$$\hat{r}(x) = \sum_{j=1}^{n} \hat{\alpha}_j B_j(x)$$

where $B_1, \ldots, B_n$ are basis for the natural splines (such as the B-splines). Thus, we only need to find the coefficients $(\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^T$. By expanding r in the basis we can rewrite the problem in the minimization of:

$$(Y - \mathbf{B}\underline{\alpha})^T (Y - \mathbf{B}\underline{\alpha}) + \hat{\lambda}_n \underline{\alpha}^T \Omega \underline{\alpha} \tag{5.5.1}$$

where $B_{ij} = B_j(X_i)$ and $\Omega_{jk} = \int B_j''(x) B_k''(x) dx$.
The value of $\underline{\alpha}$ that minimizes (5.5.1) is then

$$\hat{\underline{\alpha}} = (\mathbf{B}^T \mathbf{B} + \hat{\lambda}_n \Omega)^{-1} \mathbf{B}^T \underline{Y}.$$

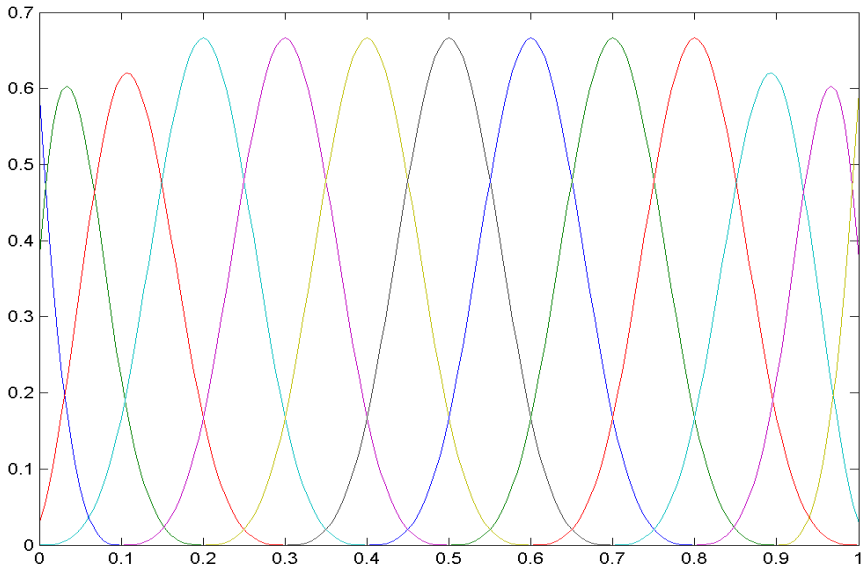

Figure 5.1: Cubic B-spline basis using nine equally spaced knots on $(0, 1)$.

Cubic B-splines are then an example of linear smoothers, indeed we can rewrite that for the cubic B-spline spline $\hat{r}(x)$ there exist weights $\underline{\ell}(x)$ such that $\hat{r}(x) = \sum_{i=1}^{n} Y_i \ell_i(x)$. In particular, the smoothing matrix $\mathbf{L}$ is

$$\mathbf{L} = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \hat{\lambda}_n\Omega)^{-1}\mathbf{B}^T \tag{5.5.2}$$

and the vector $\underline{\hat{r}}$ of fitted values is given by

$$\underline{\hat{r}} = \mathbf{L}\underline{Y}.$$

If we had done ordinary linear regression of $\underline{Y}$ on $\mathbf{B}$, the hat matrix would be $\mathbf{L} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ and the fitted values would interpolate the observed data. The effect of the term $\hat{\lambda}_n\Omega$ in (5.5.2) is to shrink the regression coefficients toward a subspace, which results in a smoother fit. As before, we define the effective degrees of freedom by $\nu = tr(\mathbf{L})$ and we choose the smoothing parameter $\hat{\lambda}_n$ by minimizing either the cross-validation score CV or the generalized cross-validation score GCV.

5.6 Selection of the smoothing parameter λ

The spline smoothers depend on some smoothing parameter λ and we will need some way of choosing λ.

In this section we will show how in literature is suggested to select a value for λ in (5.1.2) and solve the problem to select a suitable level of smoothing for a set of data. It is possible to find formulas, definition and theorems written in this section, in Wasserman (2006).

There are a lot of ways to choose a value for the smoothing parameter, even to try the estimation with arbitrary values of λ until one is found which gives a visually satisfactory fit. This can be time consuming and it may be preferable to use some data driven value that can be used subsequently for the estimation by the suitable chosen splines basis. We would not expect the same value of λ to work for every value data set.

We begin thus defining the risk

$$R(\lambda) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}(x_i) - r(x_i)\right)^2\right).$$

As one can understand, we would like to select that value of λ which minimizes $R(\lambda)$. From this, it is clear that a good choice of the smoothing parameter depends both on the unknown true regression curve as well as the inherent variability of the estimator. The minimization of $R(\lambda)$ is however not possible because it depends on the unknown function $r(x)$. Because of this inconvenient, we choose to minimize an estimate $\hat{R}(\lambda)$ of $R(\lambda)$. The risk $R(\lambda)$ is then estimated using the leave-one-out cross-validation score which is defined as follows.

Definition 5.6.1. *The leave-one-out cross-validation score is defined by*

$$CV(\lambda) = \hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{r}_{(-i)}(x_i)\right)^2 \tag{5.6.1}$$

where $\hat{r}_{(-i)}$ *is the estimator obtained by omitting the* i^{th} *pair* (x_i, Y_i).

In this definition, we need to say that

$$\hat{r}_{(-i)}(x) = \sum_{j=1}^{n} Y_j \ell_{j,(-i)}(x)$$

where

$$\ell_{j,(-i)}(x) = \begin{cases} 0 & \text{if } j = i \\ \frac{\ell_j(x)}{\sum_{k \neq i} \ell_k(x)} & \text{if } j \neq i. \end{cases}$$

In other words we set the weight on x_i to 0 and renormalize the other weights to sum to one.

Note that, since $\mathrm{E}(\hat{R}) \approx R + \sigma^2$, the cross-validation score is nearly an unbiased estimate of the risk.

The estimation of $\hat{R}(\lambda)$ by means of the expression (5.6.1), is not convenient since we need to recompute the estimator every time that each observation is dropped out. We will use, in the implementation of the algorithm in the next chapter, a modification of the following, more practical formula for computing $\hat{R}$ for linear smoothers.

Theorem 5.6.2. *Let* $\hat{r}_n$ *be a linear smoother. Then the leave-one-out cross-validation score* $\hat{R}(\lambda)$ *can be written as*

$$\hat{R}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{r}(x_i)}{1 - L_{ii}} \right)^2 \tag{5.6.2}$$

where $L_{ii} = \ell_i(x_i)$ *is the* i^{th} *diagonal element of the smoothing matrix* $\boldsymbol{L}$.

The smoothing parameter λ can be now chosen by minimizing $\hat{R}(\lambda)$.

The method that, actually, has been chosen in order to be applied in this thesis is the *generalized cross-validation criterion*. Indeed rather than minimize the cross-validation score, we will minimize an approximation, that is the generalized cross-validation in which each L_{ii} is replaced with its average $n^{-1} \sum_{i=1}^{n} L_{ii} = \nu/n$ where $\nu = tr(\mathbf{L})$ is the

effective degree of freedom. Thus, we would minimize

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{r}(x_i)}{1-\nu/n}\right)^2. \tag{5.6.3}$$

As one can expect, the bandwidth that minimize the generalized cross-validation score is close to the bandwidth that minimizes the cross-validation score.

The evaluation of the GCV(λ) function, implies then the computation of the residual sum of squares and the trace of the smoothing spline hat matrix, corresponding to any particular value of λ. The residual sum of squares can be evaluated on $O(n)$ operations once r_λ is available. About the trace of the hat matrix, it is possible to find several algorithms for the calculation of the trace, in an exact or approximated way. Suggestions for these calculations are given, for example, in Utreras (1981) or in Silverman (1984a).

It is sometimes feasible to conduct the minimization of the quantities on (5.6.2) or in (5.6.3) through a global search by evaluating the criterion of interest, over a grid of λ. We will choose that value of λ in the grid, that minimize the CV or GCV criterion. With the use of the grid, it is easy to produce plots of the criterion function and then find its local minima.

The values of λ selected from (5.6.2) or (5.6.3) provide estimators of the smoothing level that minimizes the loss $L(\lambda) = (1/n)\sum_{i=1}^{n}(r(x_i) - r_\lambda(x_i))^2$ or the risk E$[L(\lambda)]$ corresponding to r_λ. Results about consistency properties of $\hat{\lambda}_n$, the data driven value of λ estimated in these ways, in relation to the loss and the risk functions can be found in works as Craven and Wahba (1979), Cox (1983), Nychka (1991) or in Li (1986) where is proved, for example, that

$$\frac{L(\hat{\lambda}_n)}{\inf_{\lambda>0} L(\lambda)} \longrightarrow_P 1 \qquad \text{as } n \to \infty,$$

where $\hat{\lambda}_n$ could be the GCV estimator of the smoothing parameter.

5.7 The constraint of monotonicity

A typical example of shape restriction in nonparametric estimation of a regression function r is the monotonicity constraint. Standard reference in monotonicity constraints are Barlow *et al.* (1972) and Robertson *et al.* (1988).

Our study is about monotone smoothing based on spline functions. We now confine the discussion to cubic splines, solution of the main minimization problem in this thesis.

The B-splines are a convenient basis for the space of splines of interest, indeed, as it is known, such spline function $s(x)$ can be uniquely expressed as a liner combination of the B-splines $B_j(x)$ in the form

$$s(x) = \sum_{j=1}^{n} \alpha_j B_j(x).$$

The B-splines coefficients α_j have the interesting properties that there are no more sign changes in $s(x)$ than there are in the sequence α_j. Then, if the α_j are nonnegative, so is $s(x)$; if α_j is a nondecreasing sequence, $s(x)$ is not decreasing. This is exactly that kind of restriction in study. The basic idea is to use a spline function for smoothing and to enforce the monotonicity constraints by placing constraints in the B-splines coefficients. For smoothing a set of data subject to the constraint that the curve be nondecreasing, for example, the smoothing function that we use is then, the natural cubic spline function with knots at the design points and with coefficient $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n$.

In order to have an increasing sequence of the coefficient and then an estimated monotone regression function, we use the *pooled-adjacent-violator* (PAV) *algorithm* in order to obtain the *greatest convex minorant* of the coefficients of the spline. Let $P_0 = (0, 0)$ and $P_j = (j, \sum_{i=1}^{j} \alpha_i)$, for j=1,...,n. The greatest convex minorant $G(x)$ is the supremum of all convex functions that lie below the points $P_0, \ldots, P_n$. The left derivative of G gives the searched monotone sequence of coefficient.

These function $G(x)$ can be found quickly using the PAV algorithm. This algorithm starts by joining all the points $P_0, P_1, \ldots$ with line segments. If the slope between P_0 and P_1 is greater than the slope between P_1 and P_2, replaces these two segments with one line segment joining P_0 and P_2. If the slope between P_0 and P_2 is greater than the slope between P_2 and P_3, the algorithm replaces these two other segments with one line segment joining P_0 and P_3. The process is continued in this way and the result is the function $G(x)$.

Chapter 6

Computational Studies and Numerical Results

6.1 Introduction

In this chapter we will investigate the computational aspect of the semiparametric estimation of the single-index model. We are then interested in the implementation of an algorithm with which to find a good approximation of the estimate of the monotone regression function r with finite second derivative and the Euclidean parameter $\underline{\theta}$ of the single-index model

$$Y = r(\underline{\theta}' \underline{X}) + \epsilon, \tag{6.1.1}$$

where the unobserved error is assumed to be $\epsilon \sim N(0, \sigma^2)$, where thevariance σ^2 is finite.

In order to find these quantities, we search that values of $(\hat{\underline{\theta}}, \hat{r})$ such that

$$(\hat{\underline{\theta}}, \hat{r}) = \arg \min_{\underline{\theta} \in \underline{\Theta}, r \in \mathcal{R}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(y_i - r(\underline{\theta}' \underline{x_i}) \right)^2 + \hat{\lambda}_n^2 J^2(r) \right]. \tag{6.1.2}$$

We suppose at first that $\underline{\theta}_0$ is known. Then r can be estimated by classical means of univariate nonparametric regression of Y on $T = \underline{\theta}_0' \underline{X}$. The method that we will use to estimate r, is by cubic splines estimator with knots at $\underline{\theta}_0' \underline{x}_1, \ldots, \underline{\theta}_0' \underline{x}_n$, explained in

the preceding chapter. We will use this estimator because the easiness of implementation and interpretation, because with these spline estimators is possible to construct a monotone regression estimator that preserve the required hypothesis of smoothness and because essentially, (6.1.2) is the definition of the cubic spline with knots at $\underline{\theta}_0'\underline{x}_1, \ldots, \underline{\theta}_0'\underline{x}_n$. It is possible to obtain the searched estimator, choosing B-spline basis as basis for the spline estimator and then monotonizing the spline coefficients.
Of course, these estimators cannot be implemented since $\underline{\theta}_0$ is not known. If an estimator $\hat{\underline{\theta}}$ of $\underline{\theta}$ is known, we will find $\hat{r}$ by cubic splines estimator with knots at $\hat{\underline{\theta}}'\underline{x}_1, \ldots, \hat{\underline{\theta}}'\underline{x}_n$. Many methods of estimating $\underline{\theta}$ have been proposed in the literature. The resulting estimators of $\underline{\theta}$ can be classified in two main groups, according to whether they require solving nonlinear optimization problem, such as M-estimator or not, with direct estimators. Although their many advantages such as efficiency and asymptotic normality, M-estimators require solving an intricate optimization problem in a high dimensional space. On the other hand, in spite of slightly worst theoretical properties, direct estimator are highly attractive, as they provide the estimator on an analytic form.
In this thesis, denoting by $\hat{r}_\theta$ the estimate of the regression function r relative to the fixed value of $\underline{\theta}$, we will find $\hat{\underline{\theta}}$ as solution of the minimization problem:

$$\hat{\underline{\theta}} = \arg\min_{\underline{\theta}\in\underline{\Theta}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{r}_\theta(\underline{\theta}'\underline{x}_i) \right)^2 + \hat{\lambda}_n^2 J^2(\hat{r}_\theta) \right], \tag{6.1.3}$$

where the smoothing parameter $\hat{\lambda}_n^2$ is chosen by generalized cross-validation, as explained in the next section.

6.2 The estimation algorithm

In this section we explain the steps of the algorithm used for calculating, by means of cubic splines, the penalized least-squares estimator $(\hat{\underline{\theta}}, \hat{r})$ in (6.1.2) with the constraints that r is monotone and $\left\| \hat{\underline{\theta}} \right\|_m = 1$.

Step 1. We start choosing randomly k points $\hat{\underline{\theta}}_i$ for $i = 1, \ldots, k$, in the unit m-dimensional sphere by generating independent standard normal components and then normalizing in order to obtain vectors with norm equal to one. These points are the possible values that the vector of the parameters $\underline{\theta}$ can take.

Step 2. For each $\hat{\underline{\theta}}_i$, we construct the B-spline basis functions $B_1, \ldots, B_n$ with knots at the design points $\hat{\underline{\theta}}_i'\underline{x}_1, \ldots, \hat{\underline{\theta}}_i'\underline{x}_n$. To do that, I use the construction of natural cubic splines, exposed in section 5.5.
With the obtained B-spline basis, we can find now the natural cubic spline by means of

$$\hat{r}(x) = \sum_{j=1}^{n} \hat{\beta}_j B_j(x).$$

$\hat{\beta}$ is estimated by

$$\hat{\beta} = (\mathbf{B}'\mathbf{B} + \hat{\lambda}_n^2 \Omega)^{-1} \mathbf{B}' Y \tag{6.2.1}$$

where the matrices are $B_{ij} = B_j(x_i)$ and $\Omega_{ij} = \int B_j''(x) B_i''(x) dx$.

Step 3. Because we are interested in the estimation of monotone regression function, the estimated monotone regression function $\hat{r}(x)$ is obtained monotonizing the sequence of spline coefficients $\hat{\beta}$ by an implementation of the pool adjacent violator (PAV) algorithm, in a Matlab program available on the Web.

Step 4. In the expression (6.2.1) the only quantity unknown is the smoothing parameter $\hat{\lambda}_n^2$. This is selected minimizing the generalized cross-validation score over a grid of value of $\hat{\lambda}_n^2$. In the formula of the GCV score is then used the monotone function $\hat{r}(x)$. The necessity to search a value for the smoothing parameter for each value of the

single index, derives from the fact that the parameter $\hat{\lambda}_n^2$ depends on the index values estimated.

Step 5. Now we have an estimated regression function for each $\hat{\underline{\theta}}_i$. The best seven values among the $\hat{\underline{\theta}}_i$ are chosen. The criterion used to do that is selecting the seven first values minimizing the penalized sum of squares (6.1.3). A first approximated values of the final estimate of $\hat{\underline{\theta}}$ is obtained.

Step 6. The procedure in the points 1.-5. is repeated taking as initial points in the unit sphere, 200 points chosen randomly from a m-dimensional normal distribution with mean in the seven selected points and variance 0.1. Repeating the step 5., five points are now selected and subsequently three points and finally one point. With that we have then the estimate $(\hat{\underline{\theta}}, \hat{r})$.

6.3 Simulation study

In this section, we investigate with simulated data, the practical performance of the methods analyzed so far in order to minimize (6.1.2).

We show at first, how the approximation by natural cubic splines works, in nonparametric regression. Then the following figures give plots of natural cubic splines that fit sets of data generated from the nonparametric regression model (5.1.1). The error is $\epsilon \sim N(0, 0.1^2)$ and the regression function is the logistic function $r(x) = [1/(1 + 6e^{-5x})] \cdot I_{-1 \leq x \leq 1}(x)$ for $I_A(x)$ the indicator function for x falling in the set A. These functions find applications in a range of fields, from biology to economics.
We present results for the cases where the sample size is $n = 20$ and $n = 111$.
Figure 6.1, shows the curve fits to a random sample, of size $n = 20$, of the simulated

data and each display corresponds to the increasing values $\hat{\lambda}_n^2 = 10^{-7}$, $\hat{\lambda}_n^2 = 10^{-6}$, $\hat{\lambda}_n^2 = 10^{-5}$ for the smoothing parameter. In every graph it is possible to find drawed, the simulated data, the real logistic function from which the data are extracted which is plotted with a blue line, a red line that is the natural cubic spline, estimate of the logistic function and the green line which is the cubic spline estimator with the restriction of monotonicity. As we can see, this monotone function preserve the property of smoothness imposed in the definition of the estimator by penalized sum of squares. Moreover one can see that the algorithm described in the section 6.2 works effectively in fitting the data, because the cubic splines fit is very close to the true mean function even with a little sample of data.
As expected from our discussion above, the smaller value $\hat{\lambda}_n^2 = 10^{-7}$ (figure up) produces an estimator that is more subject to the data while the larger value $\hat{\lambda}_n^2 = 10^{-5}$ (figure down) gives a more smooth fit ignoring many of the features coming from the data. With the first value of $\hat{\lambda}_n^2$, the line sketched overfits and with the latter underfits the sample data. The value of $\hat{\lambda}_n^2 = 10^{-6}$, chosen by generalized cross-validation (GCV) criterion, is a good compromise between the other two cases.

The figure 6.2 is analogous to the preceding figure, but in this case the sample size is set to $n = 111$. Because of the bigger sample size, the value of the smoothing parameter suggested by generalized cross validation score is $\hat{\lambda}_n^2 = 5 \cdot 10^{-7}$. As expected, this value is smaller than that with smaller sample size.
As one can see, the curves are now more detailed giving a better approximation of the true logistic function. The three graphs are respectively for $\hat{\lambda}_n^2 = 5 \cdot 10^{-9}$ that overfits the sample, for $\hat{\lambda}_n^2 = 5 \cdot 10^{-7}$ and for $\hat{\lambda}_n^2 = 5 \cdot 10^{-6}$ that underfits the simulated data from the logistic regression function.

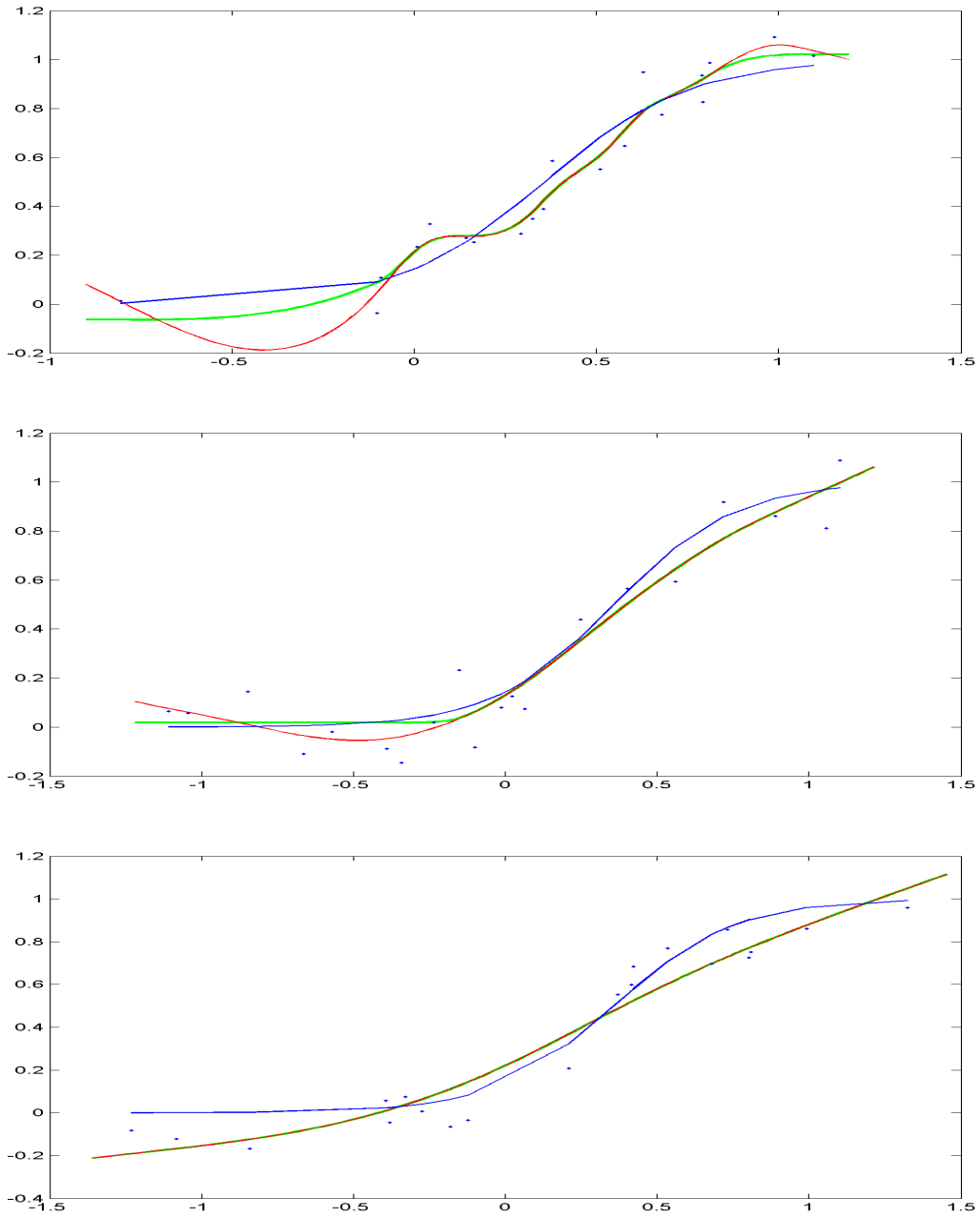

Figure 6.1: Nonparametric natural cubic spline estimators for a logistic regression function, for different values of the smoothing parameter. The $n = 20$ sample data are represented by blue points and the blue curve is the true mean function. The red curve is the spline estimate of the regression function and the green curve is the spline estimate subject to the monotonicity constraints.

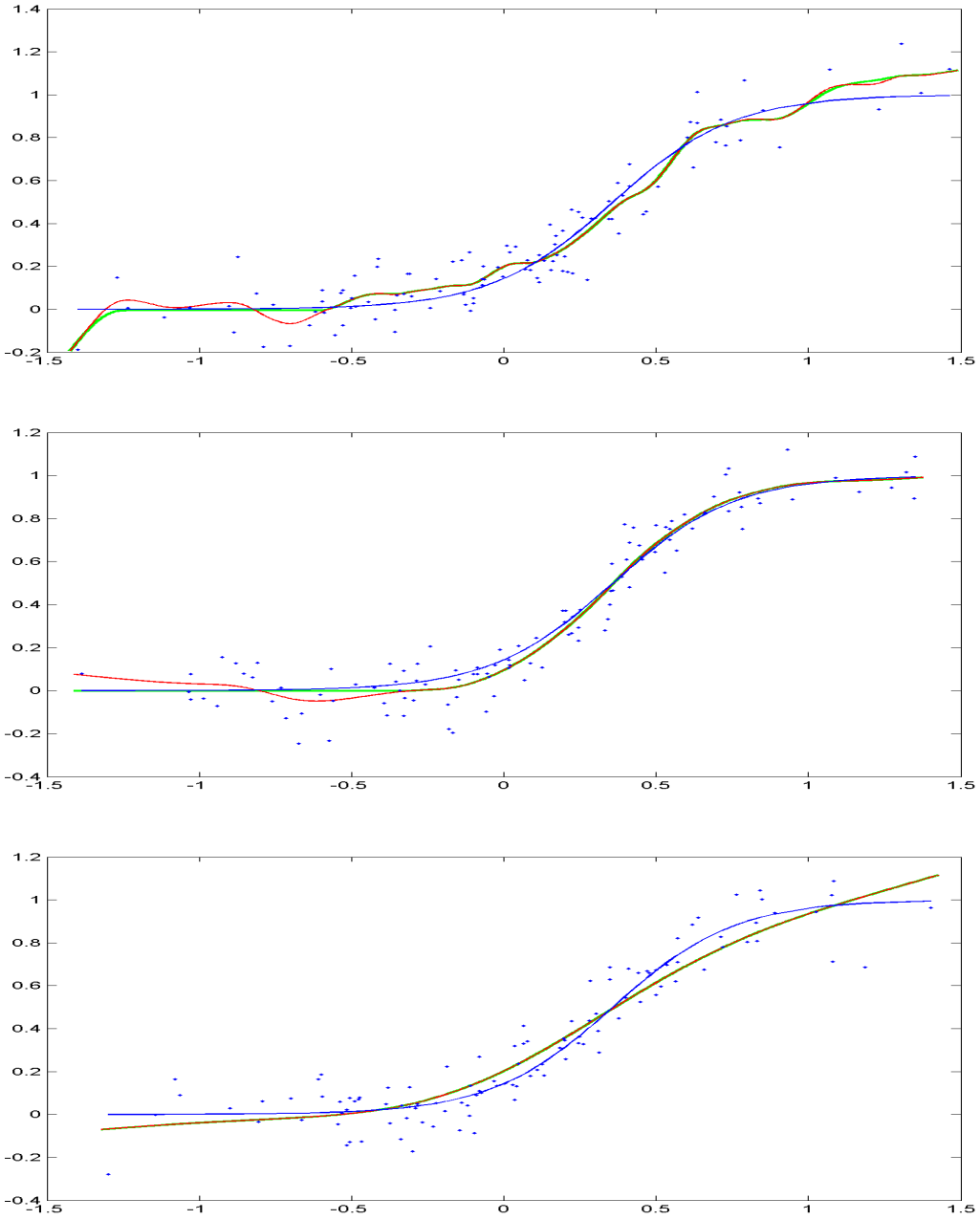

Figure 6.2: Nonparametric natural cubic spline estimators for a logistic regression function, for different values of the smoothing parameter. The $n = 111$ sample data are represented by blue points and the blue curve is the true mean function. The red curve is the spline estimate of the regression function and the green curve is the spline estimate subject to the monotonicity constraints.

Figure 6.3 shows the semiparametric estimation of the logistic function when the argument of the function is the single-index $\underline{\theta}'\underline{x}$, where $\underline{\theta}$ and $\underline{x}$ are 4-dimensional vectors. In particular, for the simulation study, $\underline{\theta}_0 = (\,1\;1\;1\;3\,)'/\sqrt{7}$ is fixed as the true value to estimate, that is

$$\underline{\theta}_0 = \begin{pmatrix} 0.3780 \\ 0.3780 \\ 0.3780 \\ 0.7559 \end{pmatrix},$$

Note that $\underline{\theta}_0$ is chosen such that $\|\underline{\theta}_0\|_4 = 1$. The graph at the top of the page is relative to the sample size $n = 20$ and the other is given from the sample size $n = 111$.

The estimated value of $\hat{\underline{\theta}}$, for $n = 20$ after 2000 replications in the first step of the algorithm, is

$$\hat{\underline{\theta}} = \begin{pmatrix} 0.3529 \\ 0.3894 \\ 0.4393 \\ 0.7286 \end{pmatrix}$$

Generalized cross-validation criterion suggests a value for the smoothing parameter $\hat{\lambda}_n^2 = 10^{-6}$.

The estimated value of $\hat{\underline{\theta}}$, for $n = 111$ after 2000 replications in the first step of the algorithm, is

$$\hat{\underline{\theta}} = \begin{pmatrix} 0.3554 \\ 0.3863 \\ 0.3364 \\ 0.7819 \end{pmatrix}$$

Generalized cross-validation criterion suggests a value for the smoothing parameter $\hat{\lambda}_n^2 = 10^{-7}$.

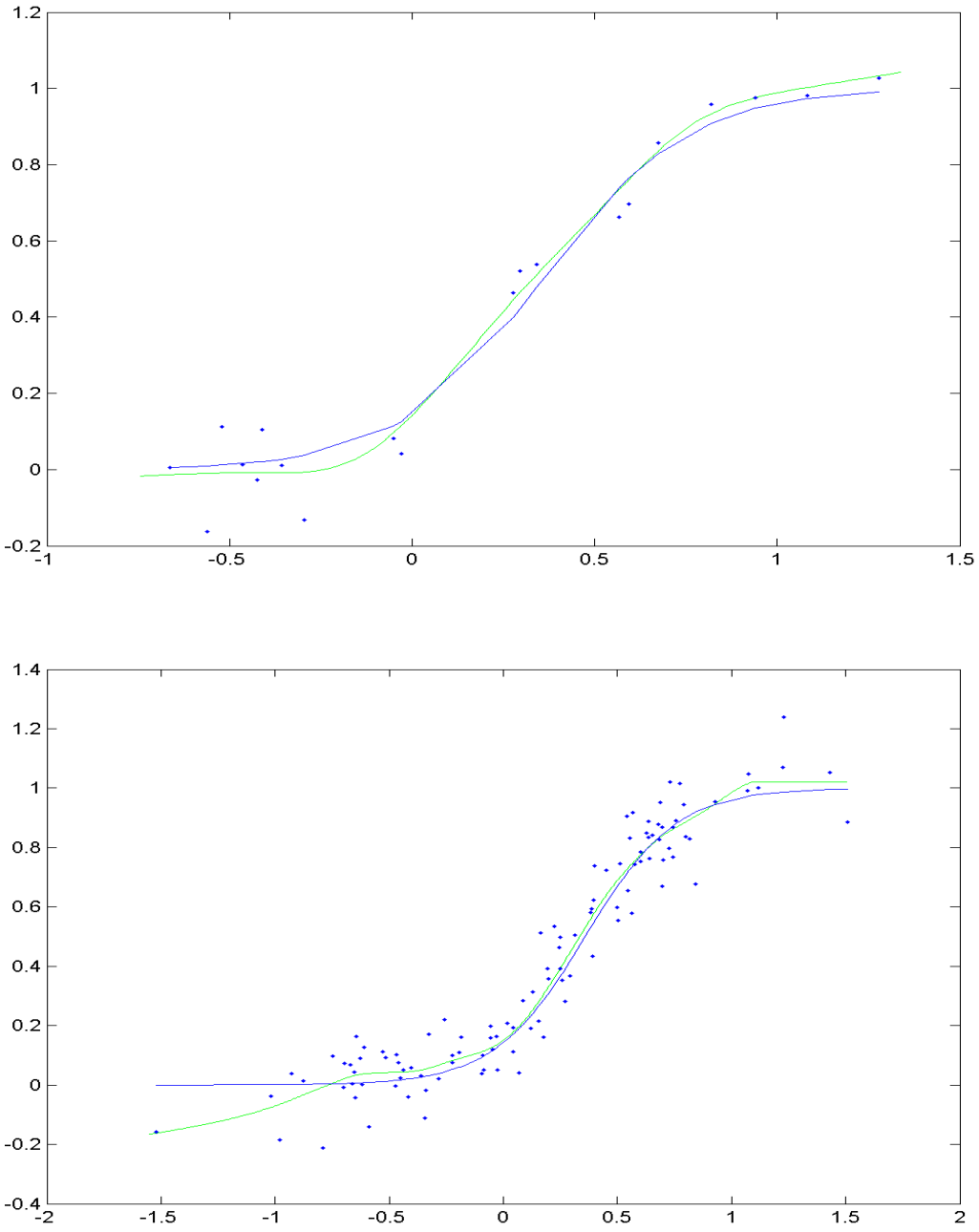

Figure 6.3: Semiparametric natural cubic spline estimators for a logistic regression function. The sample data (n=20 above and N=111 below) are represented by blue points and the blue curve is the true mean function. The green curve is the spline estimate subject to the monotonicity constraints.

6.4 An example from real data

In this section an application of monotone smoothing splines to real data analytic problem is presented. Here we consider an environmental study of how the concentration Y of the air pollutant ozone depends on three meteorological variables: the wind speed x_1, the temperature x_2 and the solar radiation x_3 (Chambers and Hastie (1992) and Yu and Ruppert (2004)). The data are daily measurements of the four variables for $n = 111$ days. Because there are three predictor variables, a full nonparametric fit with only 111 data points might not be desirable.

In order to understand the interaction of each singular predictor with the variable response Y = ozone, in figure 6.4 there are three panels, one for each of the three atmospheric variables. In each panel, the regressive lines are fitted

$$\begin{aligned} ozone &= 0.0041 \cdot radiation + 2.4860 \\ ozone &= 0.0704 \cdot temperature - 2.2260 \\ ozone &= -0.1498 \cdot wind + 4.7369. \end{aligned}$$

From these, one can understand that the air pollutant ozone increases as the solar radiation increases and more as the temperature increases and more strongly as the wind speed decreases. The other curve in each panel is the curve estimates obtained with natural cubic spline, by smoothing the variables separately against the variable ozone.

In the figure 6.5 we find plotted the data relative to the estimated value of the single-index parameter

$$\underline{\hat{\theta}} = \begin{pmatrix} 0.0252 \\ 0.5382 \\ -0.8424 \end{pmatrix}$$

for $\hat{\lambda}_n^2 = 10^{-2}$ and it is shown also the natural cubic spline curve estimate of the data.

Note that the normalized value of $\underline{\hat{\theta}}_{lin} = (\,0.0041\ 0.0704\ -0.1498\,)$, the vector of the slopes in the line regressions for each of the three variables, is

$$\underline{\hat{\theta}}'_{lin} = \begin{pmatrix} 0.0248 \\ 0.4252 \\ -0.9048 \end{pmatrix}.$$

This value is modified with respect that in our estimated value of $\underline{\hat{\theta}}$ but reflects a similar role of the three covariates in the interaction with the air pollutant ozone. In the figure 6.5 the presence of the curvature in $\hat{r}$ is observed. Of course, this curvature can not be caught by a linear model.

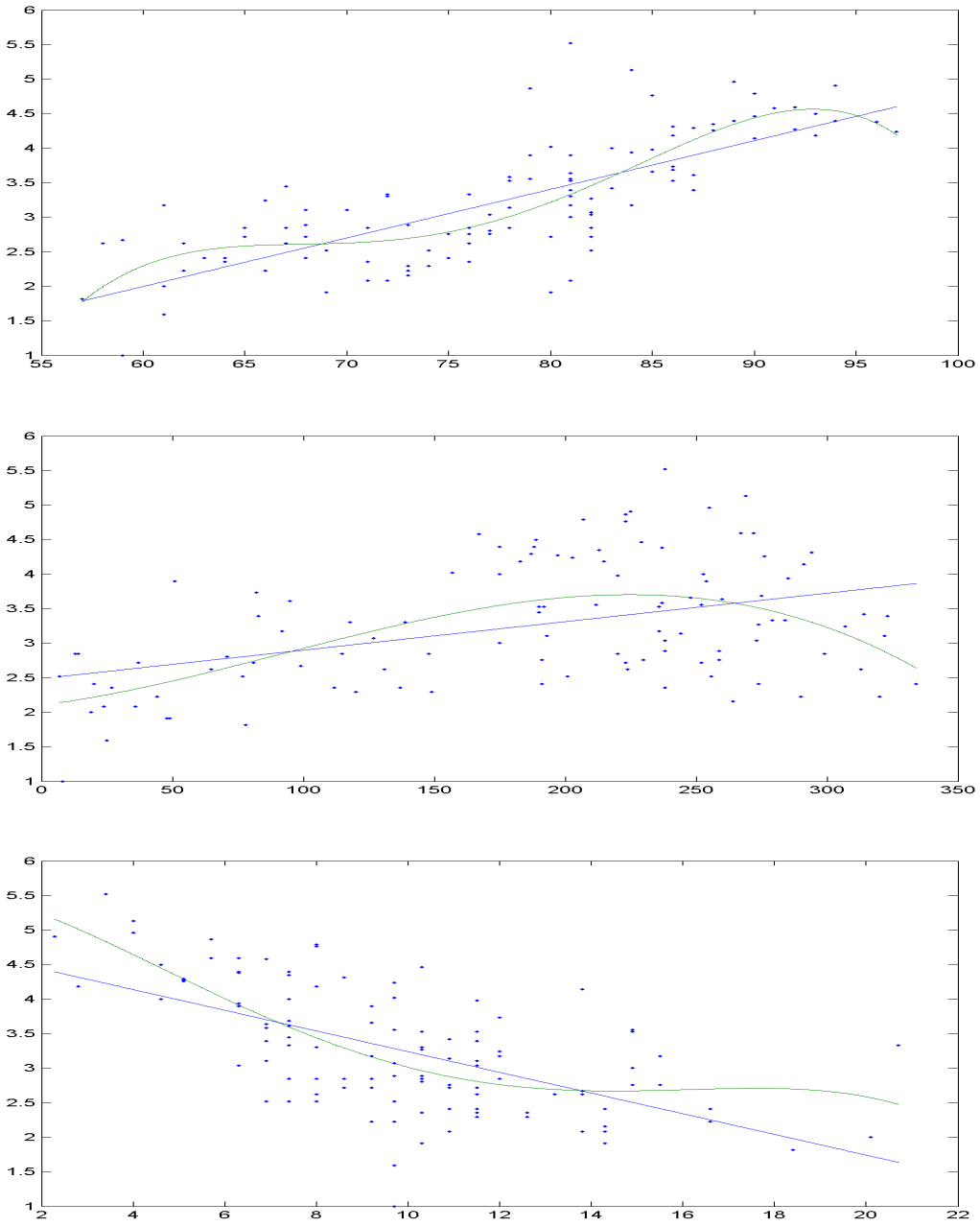

Figure 6.4: In each panel it is plotted the scatterplot, the regression line and the curve estimates obtained with natural cubic spline, of the variables separately (temperature, solar radiation and wind speed, respectively) against the variable ozone

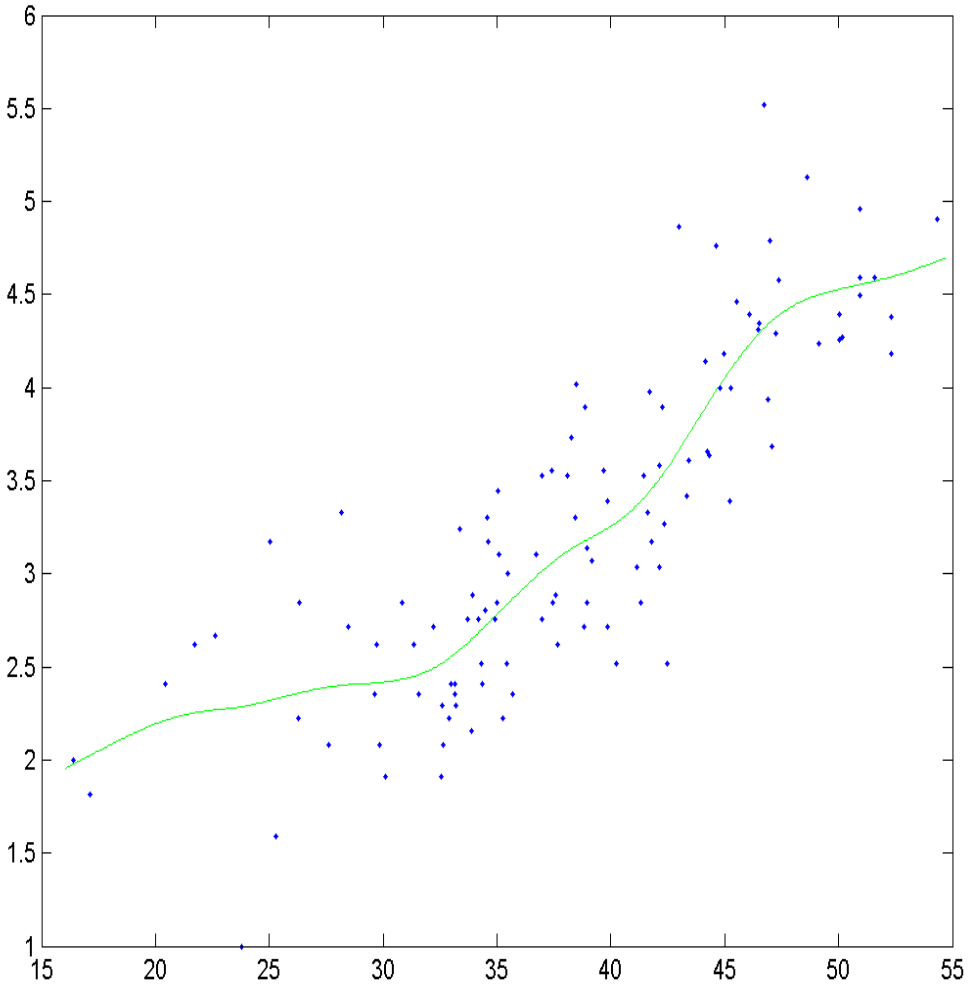

Figure 6.5: Curve estimates for the air pollution data. The data are represented by points and the solid curve corresponds to the semiparametric natural cubic spline estimator, with the constraint of monotonicity.

Bibliography

Abramovich, F. and Grinshtein, V. (1999). Derivation of equivalent kernel for general spline smoothing: a systematic approach. *Bernoulli*, **5**(2), 359–379.

Akhiezer, N. I. and Glazman, I. M. (1963). *Theory of linear operators in Hilbert space. Vol. II.* Translated from the Russian by Merlynd Nestell. Frederick Ungar Publishing Co., New York.

Amemiya, T. (1985). *Advanced econometrics.* Cambridge, MA, Harvard University Press.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, **26**, 641–647.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Bellman, R. (1961). *Adaptive control processes: A guided tour.* Princeton University Press, Princeton, N.J.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models.* Springer-Verlag, New York. Reprint of the 1993 original.

Billingsley, P. (1986). *Probability and measure*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition. A Wiley-Interscience Publication.

Birman, M. Š. and Solomjak, M. Z. (1967). Piecewise polynomial approximations of functions of classes $W_p{}^{\alpha}$. *Mat. Sb. (N.S.)*, **73 (115)**, 331–355.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.*, **92**(438), 477–489.

Cavanagh, C. and Sherman, R. P. (1998). Rank estimators for monotonic index models. *J. Econometrics*, **84**(2), 351–381.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*. Pacific Grove, Calif. : Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica*, **51**(3), 765–782.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. *Ann. Statist.*, **11**(2), 530–551.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**(4), 377–403.

de Boor, C. (2001). *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition.

Delecroix, M. and Hristache, M. (1999). M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *Bull. Belg. Math. Soc. Simon Stevin*, **6**(2), 161–185.

Delecroix, M., Härdle, W., and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.*, **86**(2), 213–226.

Delecroix, M., Hristache, M., and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference*, **136**(3), 730–769.

Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *Ann. Statist.*, **19**(2), 505–530.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.*, **6**(6), 899–929 (1979).

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*, volume 90 of *Statistics: Textbooks and Monographs*. Marcel Dekker Inc., New York.

Eubank, R. L. (1999a). *Nonparametric regression and spline smoothing*, volume 157 of *Statistics: Textbooks and Monographs*. Marcel Dekker Inc., New York, second edition.

Eubank, R. L. (1999b). A simple smoothing spline. II. *J. Statist. Plann. Inference*, **81**(2), 229–235.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.*, **19**(1), 1–141. With discussion and a rejoinder by the author.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, **76**(376), 817–823.

Friedman, J. H. and Tibshirani, R. (1984). The monotone smoothing of scatterplots. *Technometrics*, **26**, 243–250.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Internat. Statist. Rev.*, **55**(3), 245–259.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*, volume 58 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London. A roughness penalty approach.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19 of *DMV Seminar*. Birkhäuser Verlag, Basel.

Hall, P. (1989). On projection pursuit regression. *Ann. Statist.*, **17**(2), 573–588.

Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, **92**(1), 105–118.

Han, A. K. (1987). Nonparametric analysis of a generalized regression model. The maximum rank correlation estimator. *J. Econometrics*, **35**(2-3), 303–316.

Hansen, M. H. and Kooperberg, C. (2002). Spline adaptation in extended linear models. *Statist. Sci.*, **17**(1), 2–51. With comments and a rejoinder by the authors.

Härdle, W. (1990). *Applied nonparametric regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, **84**(408), 986–995.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.*, **21**(1), 157–178.

Härdle, W., Spokoiny, V., and Sperlich, S. (1997). Semiparametric single index versus fixed link function modelling. *Ann. Statist.*, **25**(1), 212–243.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York. Data mining, inference, and prediction.

Hawkins, D. (1994). Fitting monotonic polynomials to data. *Computational Statistics Quarterly*, **9**, 233–247.

He, X. and Shi, P. (1998). Monotone B-spline smoothing. *J. Amer. Statist. Assoc.*, **93**(442), 643–650.

Heckman, N. E. and Ramsay, J. O. (2000). Penalized regression with model-based penalties. *Canad. J. Statist.*, **28**(2), 241–258.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, **64**(1), 103–137.

Horowitz, J. L. (1998). *Semiparametric methods in econometrics*, volume 131 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.*, **91**(436), 1632–1640.

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, **29**(3), 595–623.

Huh, J. and Park, B. U. (2002). Likelihood-based local polynomial fitting for single-index models. *J. Multivariate Anal.*, **80**(2), 302–321.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*, **58**(1-2), 71–120.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, **46**(4), 1071–1085.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, **27**, 887–906.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, **61**(2), 387–421.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.*, **14**(3), 1101–1112.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86**(414), 316–342. With discussion and a rejoinder by the author.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.*, **19**(2), 724–740.

Mammen, E. and Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions. *Scand. J. Statist.*, **26**(2), 239–252.

Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.*, **25**(3), 1014–1035.

Mammen, E., Marron, J. S., Turlach, B. A., and Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statist. Sci.*, **16**(3), 232–248.

McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Mukerjee, H. (1988). Monotone nonparametric regression. *Ann. Statist.*, **16**(2), 741–750.

Müller, H.-G. (1988). *Nonparametric regression analysis of longitudinal data*, volume 46 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin.

Murphy, S. A., van der Vaart, A. W., and Wellner, J. A. (1999). Current status regression. *Math. Methods Statist.*, **8**(3), 407–425.

Naik, P. and Tsai, C.-L. (2000). Partial least squares estimator for single-index models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **62**(4), 763–771.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, **61**(5), 1199–1223.

Nychka, D. (1991). Choosing a range for the amount of smoothing in nonparametric regression. *J. Amer. Statist. Assoc.*, **86**(415), 653–664.

Nychka, D. (1995). Splines as local smoothers. *Ann. Statist.*, **23**(4), 1175–1197.

Ogden, R. (1996). *Essential Wavelets for Statistical Applications and Data Analysis*. Bikhäuser, Boston.

Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.*, **15**(3), 897–919.

Pfanzagl, J. (1990). *Estimation in semiparametric models*, volume 63 of *Lecture Notes in Statistics*. Springer-Verlag, New York. Some recent developments.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**(6), 1403–1430.

Prakasa Rao, B. L. S. (1983). *Nonparametric functional estimation*. Probability and Mathematical Statistics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statist. Sci.*, **3**(4), 425–461.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **60**(2), 365–375.

Rice, J. and Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Statist.*, **11**(1), 141–156.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.*, **11**(4), 735–757.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Schimek, M. G., editor (2000). *Smoothing and regression*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. Approaches, computation, and application, A Wiley-Interscience Publication.

Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. U.S.A.*, **52**, 947–950.

Schumaker, L. L. (1981). *Spline functions: basic theory*. John Wiley & Sons Inc., New York. Pure and Applied Mathematics, A Wiley-Interscience Publication.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, **61**(1), 123–137.

Silverman, B. W. (1984a). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Amer. Statist. Assoc.*, **79**(387), 584–589.

Silverman, B. W. (1984b). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.*, **12**(3), 898–916.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B*, **47**(1), 1–52. With discussion.

Smith, P. (1979). Splines as a useful and convenient statistical tool. *Amer. Statist.*, **33**, 57–62.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, **13**(3), 970–983.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, **54**(6), 1461–1481.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, **8**(6), 1348–1360.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.*, **25**(4), 1371–1470. With discussion and a rejoinder by the authors and Jianhua Z. Huang.

Tarter, M. E. and Lock, M. D. (1993). *Model-free curve estimation*, volume 56 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York.

Utreras, F. I. (1981). Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Stat. Comput.*, **2**, 349–362.

Utreras, F. I. (1985). Smoothing noisy data under monotonicity constraint: existence, characterization and convergence rates. *Numer. Math.*, **47**(4), 611–625.

van de Geer, S. A. (2000). *Empirical process theory in M-estimation*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.*, **24**(2), 862–878.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

Vapnik, V. N. and Červonenkis, A. J. (1971). The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, **16**, 264–279.

Villalobos, M. and Wahba, G. (1987). Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Amer. Statist. Assoc.*, **82**(397), 239–248.

Wahba, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.*, **24**(5), 383–393.

Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics*, **20**, 595–601.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York.

Wegman, E. J. and Wright, I. W. (1983). Splines in statistics. *J. Amer. Statist. Assoc.*, **78**(382), 351–365.

Weisberg, S. and Welsh, A. H. (1994). Adapting for the missing link. *Ann. Statist.*, **22**(4), 1674–1700.

Whittaker, E. (1923). On a new method of graduation. *Proc. Edinburgh Math. Soc.*, **41**, 63–75.

Wright, F. T. (1982). Monotone regression estimates for grouped observations. *Ann. Statist.*, **10**(1), 278–286.

Wright, I. W. and Wegman, E. J. (1980). Isotonic, convex and related splines. *Ann. Statist.*, **8**(5), 1023–1035.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**(3), 363–410.

Yatchew, A. (2003). *Semiparametric regression for the applied econometrician*. Themes in Modern Econometrics. Cambridge University Press, Cambridge.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.*, **97**(460), 1042–1054.

Yu, Y. and Ruppert, D. (2004). Root-n consistency of penalized spline estimator for partially linear single-index models under general Euclidean space. *Statist. Sinica*, **14**(2), 449–455.

Zhang, J.-T. (2004). A simple and efficient monotone smoother using smoothing splines. *J. Nonparametr. Stat.*, **16**(5), 779–796.