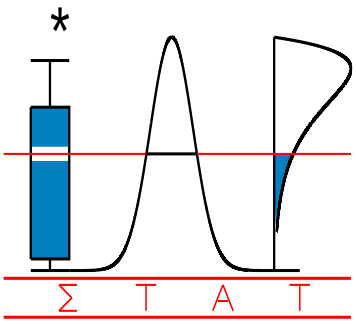
T E C H N I C A L
R E P O R T

0513

A SURVEY ABOUT SINGLE-INDEX MODELS THEORY

GEENENS, G. and M. DELECROIX



I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A survey about Single-Index Models theory

Gery Geenens*
Institut de Statistique, Université Catholique de Louvain, Belgium.
tel. : +3210473053, fax : +3210473032, mail : geenens@stat.ucl.ac.be

Michel Delecroix
ENSAI - CREST, Rennes
tel. : +33299053242, fax : +33299053205, mail : delecroi@ensai.fr

February 25, 2005

Abstract

One of the most refered semiparametric regression models in literature is certainly the Single-Index Model. It can be seen as a generalization of the Generalized Linear Model, where the link function is kept unknown and has to be estimated via nonparametric techniques. In this paper we propose a complete summary of the theory of the SIM : identification conditions, estimation of the link, estimation of the index and goodness of fit tests, as well as a simulation study permitting to compare the practical performances of different estimators of the index.

1 Introduction

Consider the model

$$Y = m(X) + \varepsilon, \tag{1}$$

with Y a scalar outcome, X a p-dimensional vector of regressors, m a function from a subset of $\mathbf{R}^p$ to the real line and ε an univariate random disturbance. Suppose we are interested in estimating the conditional mean function $E(Y|X = x)$ from a sample $\{(X_i, Y_i), i = 1, ..., n\}$. It is clear that this function, also known as the regression function, is equal to $m(x)$ provided that $E(\varepsilon|X = x)$ is zero, which is in general assumed.

It is well known that several main approaches can be considered to tackle the problem. First, one can adopt the parametric point-of-view, assuming that $m(.)$ and the distribution of ε are known up to a finite set of parameters. These parameters can easily be estimated, e.g. solving a least squares problem or by a maximum likelihood method. In spite of the easiness that this model presents in terms of computation and interpretation of the results, it admits an important drawback : the lack of flexibility. Indeed, in some situations, to force $m(x)$ to belong to a parametric family of functions can be too restrictive, even totally inappropriate, and this can lead to an important modelling bias and wrong conclusions about the link between X and Y (unconsistent estimation of m). On the other hand, the nonparametric approach releases such restrictive functional hypotheses about m. Only mild smoothness conditions are required, e.g. one typically assumes that $m(x)$

*Financial support from the IAP research network nr. P5/24 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

is twice differentiable, what allows a great flexibility for data analysis. Nevertheless, this greater flexibility has a high cost : when increasing the number of regressors, these methods get very demanding with respect to the number of observations. Specifically, the fastest achievable rate of convergence in probability of nonparametric regression function estimators towards the true regression function decreases as the number of components of X increases (Stone (1980)). Hence, the larger the number of 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 regressors is greater than 2 or 3. This phenomenon is known as the "curse of dimensionality" and motivates the need of dimension reduction methods.

Semiparametric models propose a mix of both approaches, which permits to compensate for their respective drawbacks. They are characterized by a twofold parametrization, say θ and γ, where θ lies in a finite-dimensional space Θ and γ lies in infinite-dimensional space Γ. For example, it should be the case if m belonged to a parametric family and the distribution of ε was totally unknown in model (1). One of the most popular semiparametric model is precisely the single-index model. It will be shown that it relaxes some of the restrictive assumptions of parametric models, thus ensuring some flexibility, while avoiding the curse of dimensionality.

In section 2 a formal definition of single-index models and identification conditions are given. Sections 3 and 4 propose a survey of the main estimation procedures in the SIM context while section 5 is concerned with testing the single-index hypothesis. Finally, section 6 presents a small simulation study that permits to compare the practical performances of these procedures.

2 Definition of a Single-Index Model and identification conditions

2.1 Definition

Ichimura (1993) gives the following definition of a single-index model :

Definition 1 *Let p and M be positive integers. The model*

$$Y = g[h(X, \theta_0)] + \varepsilon$$

where

- *the random vector (X', Y) is such that $Y \in \mathbf{R}$ and $X \in \mathbf{R}^p$;*
- *$\varepsilon \in \mathbf{R}$ is an unobserved random disturbance, with $E(\varepsilon|X) = 0$;*
- *$\theta_0 \in \mathbf{R}^M$ is an unknown parameter vector to be estimated ;*
- *the function $h : \mathcal{S} \times \Theta \to R$, for some $\mathcal{S} \times \Theta \subset \mathbf{R}^p \times \mathbf{R}^M$, is known up to a parameter θ ;*
- *the function $g : \mathbf{R} \to \mathbf{R}$ is not known ;*

is a single index model.

Note that the model defined like above is indeed semiparametric : θ_0 lies in a finite-dimensional space and the function g belongs to a functional space so that it can be seen as an infinite-dimensional set of parameters. Moreover, the conditional probability of ε is not specified, except for $E(\varepsilon|X) = 0$.

Great simplifications in most of the results can be obtained by fixing

$$h(X,\theta) = \theta' X = \sum_{k=1}^{p} \theta^{(k)} X^{(k)},$$

where $\theta^{(k)}$ and $X^{(k)}$ represent the kth components of vectors θ and X. Ichimura calls such a model a "linear single-index model". By sake of simplicity, and as this form for the function h is almost ever supposed, "single-index model" will always refer to linear single-index model in this paper. Hence, all the methods examined hereafter rest on the main following hypothesis :

$$E(Y|X=x) = g(\theta_0' x). \tag{2}$$

An equivalent formulation of hypothesis (2) is given by the following condition :

$$\exists \theta_0 \in \mathbf{R}^p : E(Y|X=x) = E(Y|\theta_0' X = \theta_0' x),$$

so that we can note that the single-index modelling is strongly related to the first step of the projection pursuit approach defined by Hall (1989).

2.2 Identification conditions

Restrictions must be imposed in order to make θ_0 and g uniquely determined by the population distribution of (X,Y). It is quite clear that such conditions are needed. Suppose for example that g is a constant function on $\mathbf{R}$: any vector of $\mathbf{R}^p$ should be acceptable as estimator of θ_0. It is also clear that, as in a linear model, no identification is possible if there is an exact linear relation among the components of X.

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

$$g^*(\alpha + \beta u) = g(u)$$

for all u in the support of $\theta_0' X$. We have

$$\begin{aligned} E(Y|X=x) &= g(\theta_0' x) && (3) \\ &= g^*(\alpha + \beta \theta_0' x). && (4) \end{aligned}$$

Models (3) and (4) are equivalent : they could not be distinguished, even if the whole population (X,Y) was known. Therefore, restrictions on α (location) and β (scale) have to be imposed in order to make θ_0 and g uniquely defined. In the remainder, these will be : X contains no intercept (location restriction) and the first component of θ_0 is equal to one (scale restriction)[1].

Besides, g must be differentiable. Indeed, note that the single-index hypothesis imposes that $E(Y|X=x)$ remains constant if x changes in such a way that $\theta_0' x$ stays constant. However, if $\theta_0' X$ is continuously distributed, the set of X values on which $\theta_0' X = c$ has probability 0 for any c, so that no identification is possible. But if g is differentiable, then $g(\theta_0' X)$ is close to $g(c)$ provided that $\theta_0' X$ is close to c. Therefore, the set of X values on which $\theta_0' X$ is within any specified non-zero distance of c has non-zero probability and identification of θ_0 gets possible through "approximate" constancy of $\theta_0' X$.

Based on Ichimura's observations, it can be stated :

[1] another common scale restriction is to fix the euclidean norm of θ_0 equal to one.

Theorem 2 *θ_0 and g are identified if :*

- *g is differentiable and not constant on the support of $\theta_0' X$;*
- *X admits at least one continuously distributed component ;*
- *the support of X is not contained in any proper linear subspace of R^p ;*
- *$\theta_0 \in \Theta$, with $\Theta = \{\theta \in \mathbf{R}^p : \theta^{(1)} = 1\}$.*

The second condition ensures that for any p-vector θ, the linear combination $\theta' X$ is continuously distributed. Note that if some covariates are actually discrete, two extra conditions are needed : (1) varying the values of the discrete components must not divide the support of $\theta_0' X$ into disjoints subsets ; (2) g must not be a periodic function.

3 Estimating g

Suppose at first that θ_0 is known. Then g can be estimated by classical means of univariate nonparametric regression of Y on $U = \theta_0' X$. Many various methods are proposed in Härdle (1990). Although it is well known that it is not the more efficient one, the Nadaraya-Watson kernel estimator is used in many situations because of its easiness of implementation and interpretation and its mathematical tractability. It is defined the following way.

Let $\{(X_i, Y_i), i = 1, \ldots, n\}$ be the sample and define $U_i = \theta_0' X_i$. Let K be the kernel function, usually taken to be a bounded symmetric probability density, and h a bandwidth, i.e. a smoothing parameter. Then the Nadaraya-Watson estimator of the regression function is

$$\widehat{g}^{\theta_0,h}(u) = \frac{1}{nh\widehat{p}^{\theta_0,h}(u)} \sum_{i=1}^{n} K(\frac{u - U_i}{h}) Y_i \tag{5}$$

where $\widehat{p}^{\theta_0,h}$ is the usual kernel estimator of the density p of U :

$$\widehat{p}^{\theta_0,h}(u) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{u - U_i}{h}). \tag{6}$$

Of course, these estimators cannot be implemented since θ_0 is not known. If an estimator $\widehat{\theta}$ of θ_0 is known, the estimator of g becomes :

$$\widehat{g}^{\widehat{\theta},h}(u) = \frac{1}{nh\widehat{p}^{\widehat{\theta},h}(u)} \sum_{i=1}^{n} K(\frac{u - \widehat{U}_i}{h}) Y_i \tag{7}$$

with $\widehat{U}_i = \widehat{\theta}' X_i$ and

$$\widehat{p}^{\widehat{\theta},h}(u) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{u - \widehat{U}_i}{h}). \tag{8}$$

Methods of estimating θ_0 will be described in the next section. It will be shown that θ_0 can be estimated with a $n^{-1/2}$ rate of convergence in probability, i.e. there exist estimators $\widehat{\theta}$ such that

$$(\widehat{\theta} - \theta_0) = O_P(n^{-1/2}),$$

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 $\widehat{\theta}$ is thus faster than the fastest possible rate of any nonparametric estimator of g. Therefore, it is intuitively clear that the difference between the estimators $\widehat{g}^{\theta_0,h}(u)$ and $\widehat{g}^{\widehat{\theta},h}(u)$ is asymptotically negligible. Specifically, we have

$$(nh)^{1/2}[\widehat{g}^{\widehat{\theta},h}(u) - g(u)] = (nh)^{1/2}[\widehat{g}^{\theta_0,h}(u) - g(u)] + o_P(1)$$

for any u in the support of $\theta_0'X$, which shows that root-n estimation of θ_0 has no effect on the asymptotic distribution of the Nadaraya-Watson estimator. See Horowitz (1998) for a complete argument of this result. Hence, the estimation of g is direct via standard methods once an estimator of θ_0 is known, so that this point will no more be explicitly developed in the next sections. Note however an important attraction of the SIM : because the nonparametric estimation of g, the model remains flexible enough, but as this nonparametric estimation is done on an univariate index U, the curse of dimensionality is avoided.

4 Estimating θ_0

Many methods of estimating θ_0 have been proposed in the literature. We resume most of them in this section. First, notice that estimators of θ can be classified in two main groups, according to whether they require solving nonlinear optimization problem (M-estimators) or not (direct estimators).

4.1 M-estimators

If g was known, a M-estimator of θ should typically have the form

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, g(\theta' X_i)), \tag{9}$$

where $\Psi : \mathbf{R}^2 \to \mathbf{R}$ is a function verifying some mild regularity conditions. In the SIM context, we substitute the unknown g for its leave-one-out Nadaraya-Watson estimator, so that the criterion to maximize becomes

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \widehat{g}^{\theta,h}_{(-i)}(\theta' X_i)). \tag{10}$$

The leave-one-out estimator of g at point $\theta' X_i$, denoted by $\widehat{g}^{\theta,h}_{(-i)}(\theta' X_i)$, is equal to the Nadaraya-Watson estimator (7) based on all the observations but the ith, and is used for evident bias reasons. Note that a trimming function $\tau(X_i)$ is often added to (10) for technical reasons. It is essentially useful in the proofs in order to guard against too small values for the denominator appearing in the expression of $\widehat{g}^{\theta,h}_{(-i)}(\theta' X_i)$. Nevertheless, in practice, one may often take $\tau \equiv 1$, so that we will not refer again to this problem.

Delecroix and Hristache (1999) give sufficient conditions on Ψ in order to make the estimator $\widehat{\theta}$ a.s. consistent and asymptotically normal, for any joint law of (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, i.e. if there exist differentiable functions A, B and $C : \mathbf{R} \to \mathbf{R}$ satisfying

$$A'(x) + C'(x)x \equiv 0$$

and

$$C'(x) \geq 0 \quad \forall x \in \mathbf{R},$$

such that

$$\Psi(y,x) = A(x) + C(x)y + B(y). \tag{11}$$

Theorem 3 *$\widehat{\theta}$ converges a.s. towards θ_0, and*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Sigma_M)$$

if

- *$\{(X_i, Y_i), i = 1, \ldots, n\}$ is an i.i.d. sample ;*
- *the support $\mathcal{S}$ of X is compact ;*
- *$\inf_{x \in \mathcal{S}} p(\theta' x) > 0$;*
- *p and g are three times continuously differentiable functions, and their third derivatives satisfy suitable Lipschitz conditions ;*
- *$G(x, \theta) = g(\theta' x)$ is twice continuously differentiable function with respect to θ, on $\mathcal{S} \times \Theta$;*
- *$P(x, \theta) = p(\theta' x)$ is continuous on $\mathcal{S} \times \Theta$;*
- *$E(|Y|^m) < \infty$, $m \geq 4$;*
- *$\sigma^2(x) = var(Y|X = x)$ is bounded and positive on $\mathcal{S}$;*
- *the kernel K is a twice continuously differentiable function with support $[-1, 1]$ and the second derivative satisfies a suitable Lipschitz condition. Moreover,*

$$\int v^j K(v) dv = \begin{cases} 1 \text{ if } j = 0 \\ 0 \text{ if } j = 1 \end{cases} ;$$

- *the bandwidth sequence is such that $h \sim n^{-\gamma}$, with $\gamma \in (\frac{1}{8}, \frac{1}{7})$;*
- *$\Psi(y, x)$ is as (11) with the functions A and C twice continuously differentiable and their second derivatives satisfy a suitable Lipschitz condition ;*
- *the matrix*

$$M = E\{C'[g(\theta_0' X)] \frac{\partial g(\theta' X)}{\partial \theta}|_{\theta=\theta_0} \frac{\partial g(\theta' X)}{\partial \theta'}|_{\theta=\theta_0}\}$$

is positive definite.

Note that $\Sigma_M = M^{-1} V M^{-1}$, with

$$V = E\{[C'(g(\theta_0' X))]^2 \sigma^2(X) \frac{\partial g(\theta' X)}{\partial \theta}|_{\theta=\theta_0} \frac{\partial g(\theta' X)}{\partial \theta'}|_{\theta=\theta_0}\}$$

and that this matrix can be consistently estimated. An important observation is the following : since $\theta^{(1)}$ is known to be equal to 1 for any θ belonging to Θ, the matrix M and V are degenerate, so that their first row and their first column are equal to zero. This also holds for Σ_M. Another remark is that the estimator $\widehat{\theta}$ can be made efficient via a slight modification of the criterion, for any function Ψ satisfying (11).

We examine hereafter two important particular M-estimators : generalizations of the parametric least squares and maximum likelihood estimators.

4.1.1 Semiparametric Least Squares (SLS)

Description 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 $g(\theta' X_i)$. If g was known, we should have the classical least squares estimator given by

$$\theta^* = \arg\min \frac{1}{n}\sum_{i=1}^{n} w(X_i)[Y_i - g(\theta' X_i)]^2, \tag{12}$$

where w is a positive bounded weight function. Under mild regularity conditions, least squares theory shows that this estimator θ^* is root-n consistent (see o.a. Amemiya (1985)). In the single-index context, the least squares criterion to minimize becomes

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)[Y_i - \widehat{g}^{\theta,h}_{(-i)}(\theta' X_i)]^2, \tag{13}$$

where the leave-one-out Nadaraya-Watson estimator is used as a trite function of θ.

Ichimura (1993) studied this method in details. He pointed out a important modification : in a general framework, it is needed to weight the terms in the calculation of $\widehat{g}^{\theta,h}$ the same way it was done in the calculation of (12). The estimator $\widehat{g}^{\theta,h}_{(-i)}(\theta' X_i)$ appearing in criterion (13) now becomes

$$\widehat{g}^{\theta,h}_{(-i)}(\theta' X_i) = \frac{\sum_{j\neq i} w(X_j) K(\frac{\theta' X_i - \theta' X_j}{h}) Y_j}{\sum_{j\neq i} w(X_j) K(\frac{\theta' X_i - \theta' X_j}{h})}. \tag{14}$$

However, if the variance function $\sigma^2(x)$ depends on x only through the index $\theta' x$, then this correction is not necessary.

Weighting scheme 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

$$w(x) = 1/\sigma^2(x). \tag{15}$$

If $\sigma^2(x)$ is unknown, a consistent estimator $\widehat{s}^2(x)$ has to be used in (15). Such an estimator can be obtained by using the following two-steps procedure : first, estimate θ_0 by $\widehat{\theta}_1$, the minimizer of the unweighted version of (13), which is a root-n but inefficient estimator. Then let e_i be the ith residual from the estimated model, i.e.

$$e_i = Y_i - \widehat{g}^{\widehat{\theta}_1,h}_{(-i)}(\widehat{\theta}_1' X_i),$$

and set $\widehat{s}^2(x)$ equal to a nonparametric estimator (e.g. the Nadaraya-Watson estimator) of the mean regression of e_i^2 on x. Note that if we know a function V such that

$$\sigma^2(x) = V[g(\theta_0' x)],$$

we can also act as follows : first compute $\widehat{\theta}_1$ taking the weight function w to be identically equal to 1, as above. Then in expression (13) replace $w(X_i)$ by $\{V[\widehat{g}^{\widehat{\theta}_1,h}(\widehat{\theta}_1' X_i)]\}^{-1}$ in order to compute $\widehat{\theta}$.

4.1.2 Semiparametric Maximum Likelihood (SML)

Description Another optimization based method is inspired by the parametric maximum likelihood methods. In our single-index context, the joint distribution of X and Y and the conditional density of Y given X clearly depend on θ and on g. Suppose this conditional density depends upon X only through $\theta_0'X$ and denote these two $l_{g,\theta}(.,.)$ and $l_{g,\theta}(.|.)$, respectively. The likelihood is

$$\begin{aligned} L_g(\theta) &= \prod_{i=1}^{n} l_{g,\theta}(X_i, Y_i) \\ &= \prod_{i=1}^{n} l_{g,\theta}(Y_i|\theta'X = \theta'X_i)\ f(X_i) \end{aligned}$$

where f is the marginal density of X. Hence the log-likelihood is

$$LL_g(\theta) = \sum_{i=1}^{n} \log l_{g,\theta}(Y_i|\theta'X = \theta'X_i) + \sum_{i=1}^{n} \log f(X_i).$$

Of course the term $\sum_{i=1}^{n} \log f(X_i)$ does not depend on g and θ, so that maximizing $LL_g(\theta)$ amounts to maximizing $\sum_{i=1}^{n} \log l_{g,\theta}(Y_i|X = X_i)$. If g was known, the maximum likelihood estimator of θ should be given by the following maximization problem

$$\theta^* = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log l_{g,\theta}(Y_i|\theta'X = \theta'X_i). \tag{16}$$

Standard theory of maximum likelihood estimation implies that θ^* is root-n consistent, efficient and asymptotically normal under regularity conditions.

Nevertheless, since g is unknown, θ^* is not feasible. If the conditional distribution of Y given X is known up to g and θ (what we call the "nonignorant" semiparametric maximum likelihood), we simply overcome the problem by replacing g in problem (16) with its Nadaraya-Watson leave-one-out estimator, thus forming a pseudo-likelihood. The estimator is finally given by

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log l_{\widehat{g}^{\theta,h}_{(-i)},\theta}(Y_i|\theta'X = \theta'X_i). \tag{17}$$

If the conditional distribution of Y given X is not known (the "ignorant" SML), we estimate it in a fully nonparametric way :

$$\widehat{l}_{g,\theta}(Y_i|\theta'X = \theta'X_i) = \frac{\sum_{j\neq i} K(\frac{Y_i - Y_j}{h^*}) K(\frac{\theta'X_i - \theta'X_j}{h})}{\sum_{j\neq i} K(\frac{\theta'X_i - \theta'X_j}{h})},$$

i.e. nothing else but a kernel estimator of the joint distribution of $(\theta'X, Y)$ divided by the classical kernel estimator of the marginal density of $\theta'X$. Delecroix et al. (2003) show that the resulting estimator is asymptotically efficient : it keeps the most important property of the parametric maximum likelihood estimators.

Illustration To illustrate this method, consider the case where the only possible values of Y are 0 and 1 (binary-response models). See Klein and Spady (1993) for a detailed study of the problem. With the constraint of Y being binary, we have directly that

$$g(\theta_0' x) = E(Y|X = x) = P(Y = 1|X = x)$$

which leads to the following conditional distribution of Y given X :

$$l_{g,\theta}(Y|X) = g(\theta' X)^Y (1 - g(\theta' X))^{1-Y}.$$

The estimator of θ_0 is thus given by[2]

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \left\{ Y_i \log[\widehat{g}_{(-i)}^{\theta,h}(\theta' X_i)] + (1 - Y_i) \log[1 - \widehat{g}_{(-i)}^{\theta,h}(\theta' X_i)] \right\}.$$

4.1.3 Bandwidth selection

In order to construct the estimator (14), 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. Define

$$\widehat{S}(\theta, h) = \frac{1}{n} \sum_{i=1}^{n} \Psi(Y_i, \widehat{g}_{(-i)}^{\theta,h}(\theta' X_i)) \tag{18}$$

the criterion to be maximized. One way to select the bandwidth is to consider it as an extra parameter of the model, and to maximize the aim function with respect to it as well. That is :

$$(\widehat{\theta}, \widehat{h}) = \arg\max_{\theta \in \Theta, h \in \mathbf{R}^+} \widehat{S}(\theta, h). \tag{19}$$

An important feature is that the semiparametric criterion $\widehat{S}(\theta, h)$ can be split into a purely parametric part $\widetilde{S}(\theta)$, a purely nonparametric part $T(h)$ and some negligible reminders terms, where

$$\widetilde{S}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\Psi(Y_i, g(\theta' X_i)) - \Psi(Y_i, g(\theta_0' X_i)) \right]$$

and

$$T(h) = \frac{1}{n} \sum_{i=1}^{n} \Psi(Y_i, \widehat{g}_{(-i)}^{\theta_0,h}(\theta_0' X_i)).$$

$\widetilde{S}(\theta)$ is an approximation of $S(\theta) = E(\Psi(Y, g(\theta' X))$ and $T(h)$ is the usual cross-validation criterion for choosing h when θ_0 is known. This result leads to a simple way of simultaneously maximizing with respect to both θ and h as it is very much like separately maximizing $\widetilde{S}(\theta)$ with respect to θ and $T(h)$ with respect to h. It is proved that this method produces a root-n consistent estimator of θ and an asymptotically optimal estimator of h, in the sense that

$$\widehat{h}/h_{opt} \to 1,$$

where h_{opt} is the theoretical optimal value of the bandwidth.

[2]here the above mentioned eventual trimming term should be useful in order to keep $\widehat{g}_{(-i)}^{\theta,h}$ away from 0 or 1 as well.

4.2 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 (see e.g. (19)). In spite of slightly worst theoretical properties, direct estimators are highly attractive, as they provide the estimator on an analytic form.

4.2.1 Average Derivatives Estimator (ADE)

Recall we set $u = \theta_0' x$ and $m(x) = g(\theta_0' x)$. Average derivatives method rests on the fact that

$$\nabla m(x) = \frac{\partial g}{\partial u}(\theta_0' x)\ \theta_0,$$

which induces that

$$\delta_w \doteq E[w(X)\nabla m(X)] = E[w(X)\frac{\partial g}{\partial u}(\theta_0' X)]\ \theta_0 \tag{20}$$

for any bounded continuous weight function w. The quantity δ_w is called a weighted average derivative of g with weight function w. It appears from (20) that any weighted average derivative is proportional to θ_0, provided $E[w(X)\frac{\partial g}{\partial u}(\theta_0' X)]$ is not zero. Note that this condition is in particular violated when $w \equiv 1$, g is an even function and X is symmetrically distributed. Remark also that considering the gradient of m implies that X is a continuously distributed random vector. However, an extension of the method to the case where some components of X are discrete is possible. See Horowitz and Härdle (1996).

Unweighted Average Derivatives (UADE) Härdle and Stoker (1989) take $w \equiv 1$ and use nonparametric estimation of the marginal density of X. Let $f(x)$ be this marginal density, $\nabla f = \partial f / \partial x$ its gradient vector and let $l = -\nabla f / f$ the negative log-density derivative. By definition, we have

$$\delta = \int \nabla g(\theta_0' x) f(x) dx.$$

Assuming that $f(x) = 0$ on the boundary of x values, integration by parts gives

$$\begin{aligned} \delta &= -\int g(\theta_0' x) \nabla f(x) dx \\ &= \int g(\theta_0' x) l(x) f(x) dx \\ &= E[Y\ l(X)]. \end{aligned}$$

The proposed estimator is a sample analog of this last expression, using a nonparametric estimator of $l(x)$, that is

$$\widehat{\delta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{l}^h_{(-i)}(X_i)\ Y_i \tag{21}$$

where

$$\widehat{l}^h_{(-i)}(x) = -\frac{1}{\widehat{f}^h_{(-i)}(x)} \nabla \widehat{f}^h_{(-i)}(x).$$

$\widehat{f}^h$ can be the classical leave-one-out multivariate kernel density estimator

$$\widehat{f}^h_{(-i)}(x) = \frac{1}{nh^p}\sum_{j\neq i} K(\frac{x - X_j}{h}) \tag{22}$$

with K a multivariate kernel function. Remark that dividing by $\widehat{f}^h_{(-i)}$ can lead to erratic behavior when its value becomes too small, what can motivate once again the introduction of a trimming term in (21). By dividing this vector $\widehat{\delta}$ by its first component, one gets an estimate of θ_0. Härdle and Stoker (1989) show :

Theorem 4 *$\widehat{\delta}$ is a consistent estimator of δ and*

$$\sqrt{n}(\widehat{\delta} - \delta) \xrightarrow{d} N(0, \Sigma_u)$$

if

- *X is a continuously distributed vector of size p with density f being smooth and having partial derivatives of order $q \geq p + 2$;*
- *the support Ω of f is a convex subset of $\mathbf{R}^p$ and $f(x) = 0$ for all $x \in \partial\Omega$, where $\partial\Omega$ is the boundary of Ω ;*
- *the function g is twice differentiable ;*
- *the second moments of $\frac{\partial g}{\partial u}$ and gl exist ;*
- *the function $m_2(x) = E(Y^2|X = x)$ is continuous in x ;*
- *$f(x)$ and $g(x)$ obey suitable local Lipschitz conditions and any partial derivative of f is locally Hölder continuous ;*
- *the bandwidth sequence is such that $h \to 0$, $nh^{2p-2} \to 0$;*
- *the multivariate kernel K has support $\{u \mid ||u|| \leq 1\}$ and is such that $K(u) = 0$ if $||u|| = 1$ and $\int K(u)du = 1$. K is of order q, i.e.*

$$\int u_1^{l_1} u_2^{l_2} ... u_p^{l_p} K(u)du \begin{cases} = 0 \text{ if } 0 < l_1 + ... + l_p < q \\ \neq 0 \text{ if } l_1 + ... + l_p = q \end{cases} .$$

This latter condition implies that K must be a higher-order kernel, meaning that it must take on positive and negative values. Such kernel is usually used to reduce bias. In other words, it is needed here to insure that the asymptotic distribution of $\sqrt{n}(\widehat{\delta} - \delta)$ is centered at 0. Remark also that $\widehat{\delta}$ achieves $O_p(n^{-1/2})$ rate of convergence although nonparametric kernel estimators of $l(x)$ cannot achieve it. This is due to the sum over i in (21), that makes $\widehat{\delta}$ an average of kernel estimators. It is well known that averages of kernel estimators can achieve faster rates of convergence than kernel estimators that are not averaged. The variance-covariance matrix Σ_u is the covariance matrix of $r(Y, X)$, where

$$r(Y, X) = \nabla m(X) + [Y - m(X)]l(X),$$

and can be consistently estimated.

Several remarks can be expressed. First, the method is based on a fully nonparametric estimation of the multivariate density $f(x)$, which is severely subject to the curse of dimensionality. Hence, we loose the main advantage of the single-index modelling.

Second, experiments show that the method is very sensitive to the choice of h. It turns out to be very important to have automatic methods for setting the bandwidth that assures good small sample behavior of the estimator. Härdle et al. (1992) state that the best choice of h is substantially smaller than the typical bandwidth for density estimation. In fact, it appears that the behavior of $\widehat{l}^h$ is essentially dictated by the estimation of $\nabla \widehat{f}^h$, what makes the optimal bandwidth close to the typical bandwidth for density derivatives estimation. They show that the best bandwidth is of order $n^{-2/(2q+p)}$, and propose a plug-in estimator based on the minimization of the MSE of $\widehat{\delta}$. The idea is very similar to what will be done for density–weighted ADE, see below.

Third, remark that since $\sqrt{n}(\widehat{\delta} - \delta) \to N(0, \Sigma_u)$, the delta method provides

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to N(0, \Omega),$$

with

$$\Omega = \phi \Sigma_u \phi',$$

where

$$\phi = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ -\frac{\delta^{(2)}}{\delta^{(1)}} & \frac{1}{\delta^{(1)}} & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ -\frac{\delta^{p)}}{\delta^{(1)}} & 0 & \cdots & 0 & \frac{1}{\delta^{(1)}} \end{pmatrix}. \tag{23}$$

Note that the first row and the first column of the obtained matrix Ω are equal to zero, which obviously marks the fact that the first component of $\widehat{\theta}$ is equal to one, without uncertainty.

Density-weighted Average Derivatives (DWADE) 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 $w(x) = f(x)$. With this weight function, we have, from (20),

$$\delta_f = \int \nabla g(\theta_0' x) f^2(x) dx.$$

Assuming again that $f(x) = 0$ on the boundary of the support of X, we find

$$\begin{aligned} \delta_f &= -2 \int g(\theta_0' x) \nabla f(x) \; f(x) dx \\ &= -2 \; E[Y \; \nabla f(X)] \end{aligned} \tag{24}$$

so that we can estimate δ_f with

$$\widehat{\delta}_f = -\frac{2}{n} \sum_{i=1}^{n} Y_i \nabla \widehat{f}^h_{(-i)}(X_i).$$

As announced, only the gradient of f has to be estimated. This can be done using the formula

$$\nabla \widehat{f}^h_{(-i)}(x) = \frac{1}{(n-1)h^{p+1}} \sum_{j \neq i} \nabla K(\frac{x - X_j}{h})$$

which is nothing else but the gradient of a leave-one-out version of the kernel density estimator (22). The estimator of δ_f is thus directly given by

$$\widehat{\delta}_f = -\frac{2}{n(n-1)h^{p+1}} \sum_{i=1}^{n} \sum_{j \neq i} \nabla K(\frac{X_i - X_j}{h}) Y_i. \tag{25}$$

Remark that no denominator appears in this latter expression, what is another great advantage of this weighting scheme, compared with the unweighted ADE. Nevertheless, the estimator is again based on a nonparametric estimation in a multidimensional space. Powell et al. (1989) prove the following theorem. Let $P = (p+4)/2$ if p is even and $P = (p+3)/2$ if p is odd.

Theorem 5 *$\widehat{\delta}_f$ is a consistent estimator of δ_f, and*

$$\sqrt{n}(\widehat{\delta}_f - \delta_f) \xrightarrow{d} N(0, \Sigma_f).$$

if

- *X is a continuously distributed vector of size p with density f being smooth and having partial derivatives up to order $P+1$;*
- *the support Ω of f is a convex subset of $\mathbf{R}^p$ and $f(x) = 0$ for all $x \in \partial\Omega$, where $\partial\Omega$ is the boundary of Ω ;*
- *g is continuously differentiable in the components of x ;*
- *the components of the random vector ∇g and random matrix $[\nabla f][Y, X']$ have finite second moments ;*
- *∇f and $\nabla(gf)$ satisfy suitable Lipschitz conditions ;*
- *$E(Y^2|X = x)$ is continuous in x ;*
- *h obeys $nh^{p+2} \to \infty$ and $nh^{2P} \to 0$ as $n \to \infty$;*
- *the multivariate kernel K has support $\{u \in \mathbf{R}^p \mid ||u|| \leq 1\}$ and is such that $K(u) = 0$ if $||u|| = 1$ and $\int K(u)du = 1$. All moments of order P of K exist. Besides, we have*

$$\int u_1^{l_1} u_2^{l_2} ... u_p^{l_p} K(u)du \begin{cases} = 0 \text{ if } 0 < l_1 + ... + l_p < P \\ \neq 0 \text{ if } l_1 + ... + l_p = P \end{cases}.$$

Note that this latter condition requires K to be a higher-order kernel, as $P > 2$ once $p > 1$. The matrix Σ_f is 2 times the variance-covariance matrix of $r(Y, X)$, where

$$r(X, Y) = f(X)\nabla m(X) - [Y - m(X)]\nabla f(X),$$

and can be consistently estimated. As previously, this result implies that $\widehat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ_0, with asymptotic normal distribution.

Based on observations of Härdle and Tsybakov (1993), Powell and Stoker (1996) propose a plug-in bandwidth selection rule for this estimator. They develop the mean squared error of $\widehat{\delta}_f$ as

$$E[||\widehat{\delta}_f - \delta||^2] = Q_1 n^{-1} + Q_2 n^{-2} h^{-p-2} + Q_3 h^{2P} + \text{lower order terms}, \tag{26}$$

where Q_1, Q_2 and Q_3 are constants depending on the unknown parameters of the model. It is easy to see that the value which minimizes the leading terms of (26) is

$$h^* = h_0 \, n^{-2/(2P+p+2)}$$

where

$$h_0 = \left(\frac{Q_2(p+2)}{PQ_3} \right)^{1/(2P+p+2)}.$$

The constants Q_2 and Q_3 can be consistently estimated provided there is an initial bandwidth estimate h_1 available. This pilot bandwidth has less influence on the final result, so that it can be determined via less elaborate methods. Consistent estimators are

$$\begin{aligned} \widehat{Q}_2 &= \frac{1}{n(n-1)h_1^p} \sum_{i,j} (Y_j - Y_i)^2 \, ||\nabla K(\frac{X_j - X_i}{h_1})||^2, \qquad (27) \\ \widehat{Q}_3 &= ||\frac{\widehat{\delta}_f^{\tau h_1} - \widehat{\delta}_f^{h_1}}{[(\tau h_1)^P - h_1^P]}||^2, \\ \widehat{h}_0 &= \left(\frac{\widehat{Q}_2(p+2)}{P\widehat{Q}_3} \right)^{1/(2P+p+2)}, \end{aligned}$$

where τ is any positive number $\neq 1$. It is shown that

$$\widehat{h}^* = \widehat{h}_0 \, n^{-2/(2P+p+2)} \qquad (28)$$

is such that

$$\widehat{h}^* - h^* = o_p(n^{-2/(2P+p+2)}).$$

Iterative average derivative estimator (IADE) The major drawback of the previous two procedures is the need to estimate the density of X and/or its gradient in a fully nonparametric way, what can lead to very poor performance due to the curse of dimensionality. Hristache et al. (2001) propose another type of direct estimate of θ_0, which can be regarded as an iterative improvement of the average derivative estimator. The idea is the following. Suppose for the moment that $p = 2$ and that the observations X_i are scattered uniformly over the square $[0,1]^2$. The expected gradient of m, appearing in (20) with $w \equiv 1$, will be estimated by a sample average of estimates of this ∇m at each point X_i. At X_i, a kind of local least squares problem is used :

$$\begin{pmatrix} \widehat{m}(X_i) \\ \widehat{\nabla m}(X_i) \end{pmatrix} = \underset{c \in \mathbf{R}, \beta \in \mathbf{R}^p}{\arg\min} \sum_{j=1}^{n} [(Y_j - c) - \beta'(X_j - X_i)]^2 K(\frac{X_j - X_i}{h}). \qquad (29)$$

As kernel, it is here recommended to choose a function depending only on the squared euclidean norm of its argument, that is

$$K\left(\frac{X_j - X_i}{h} \right) = K_0 \left(\frac{||X_j - X_i||^2}{h^2} \right),$$

so that the weights of all points X_j outside a spherical neighborhood $V_h(X_i)$ of diameter h around X_i vanish. Hence, the expected gradient

$$\beta^* = E[\nabla m(X)]$$

can be estimated by

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\nabla m}(X_i)$$

leading to an estimate of θ_0 being

$$\check{\theta} = \frac{\widehat{\beta}}{\widehat{\beta}^{(1)}}. \tag{30}$$

It is shown that the following upper bound for the error of the estimate $\widehat{\beta}$ holds :

$$||\widehat{\beta} - \beta^*|| \leq C_1 h + C_2 \frac{||\xi||}{\sqrt{n}h}, \tag{31}$$

where C_1 and C_2 are constants and ξ is a gaussian random vector with zero mean. The first term is the deterministic error, due to the local approximation of m by a linear function in (29). It is thus a bias term and therefore proportional to h. The second term is the stochastic error, independent of m and of order $(\sqrt{n}h)^{-1}$. The balance between these two terms gives[3] $h \sim n^{-1/4}$ and the error is then

$$||\widehat{\beta} - \beta^*|| = O(n^{-1/4}), \tag{32}$$

just as well for $\widehat{\beta}$ as for $\check{\theta}$. Of course, we are far from the achievable $n^{-1/2}$ rate of convergence, so that an improvement is needed.

Recall that we are working for the moment in a two-dimensional space for X. As the bias term of (31) is in fact proportional to the width of the projection of the spheric neighborhood $V_h(X_i)$ on the direction θ_0, we can stretch $V_h(X_i)$ along the direction orthogonal to θ_0 without increasing this bias term. This reflection is based on a well known property of the gradient : it points towards the direction in which the function increases most, and this function is locally constant in the orthogonal direction. Although θ_0 is not known, we can use the first estimate (30) : at any X_i define an elliptic window $V_{h,\rho}(X_i)$, centered at X_i, with small axis of size $O(\rho h)$ (with $\rho < 1$) oriented along $\check{\theta}$, and large axis of size $O(h)$ orthogonal to $\check{\theta}$. If ρ is small and $\check{\theta}$ a good approximation of θ_0, we can expect that the approximation error of m by a linear function in the neighborhood $V_{h,\rho}$ would be small. We can deal with such an elliptic window by replacing the weights $K_0(h^{-2}||X_j - X_i||^2)$ in (29) with $K_0(h^{-2}||\Lambda_{\rho,\check{\theta}}(X_j - X_i)||^2)$, where the positive definite symmetric matrix

$$\Lambda_{\rho,\check{\theta}} = I + \rho^{-1}\check{\theta}\check{\theta}' \tag{33}$$

defines the elliptic geometry of the window. The estimate of the gradient at X_i is now given by

$$\begin{pmatrix} \widehat{m}(X_i) \\ \widehat{\nabla m}(X_i) \end{pmatrix} = \underset{c\in\mathbf{R},\beta\in\mathbf{R}^p}{\arg\min} \sum_{j=1}^{n}[(Y_j - c) - \beta'(X_j - X_i)]^2 K_0\left(\frac{||\Lambda_{\rho,\check{\theta}}(X_j - X_i)||^2}{h^2}\right), \tag{34}$$

that is, from classical least-squares theory,

$$\widehat{\beta}_i = W^{-1}_{h,\Lambda_{\rho,\check{\theta}}}(X_i)\sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_j - X_i \end{pmatrix} K_0\left(\frac{||\Lambda_{\rho,\check{\theta}}(X_j - X_i)||^2}{h^2}\right)$$

[3] when $p > 4$, one cannot take the optimal initial window $h = O(n^{-1/4})$ in(29) because there will not be enough observations points in the neighborhood $V_h(X_i)$ to compute the p-dimensional vector $\widehat{\nabla m}(X_i)$ (problem related to the curse of dimensionality). One has to take $h = O(n^{-1/p})$.

with

$$W_{h,\Lambda_{\rho,\widehat{\theta}}}(x) = \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_j - x \end{pmatrix} \begin{pmatrix} 1 \\ X_j - x \end{pmatrix}' K_0 \left(\frac{||\Lambda_{\rho,\check{\theta}}(X_j - X_i)||^2}{h^2} \right).$$

After averaging, we can compute the estimator $\widehat{\theta}$ of θ_0 the same way as in (30). It can be shown that this estimate satisfies

$$||\widehat{\theta} - \theta_0|| = C_3 h \rho^2 + C_4 \frac{||\xi||}{\sqrt{n}h}$$

if there exists some $\gamma > 0$ such that $||\widehat{\theta} - \theta_0|| \leq \gamma$ and $\rho \geq \gamma$. From (32), we have $\gamma = O(n^{-1/4})$, so that we can take $h = O(1)$ and $\rho = O(n^{-1/4})$ leading to

$$||\widehat{\theta} - \theta_0|| = O(n^{-1/2}). \tag{35}$$

The estimator $\widehat{\theta}$ is thus root-n consistent. The procedure can be iterated, each time flattening the elliptic window in the direction of the current estimate and stretching it in the orthogonal direction : the bias term then rapidly becomes negligible with respect to the stochastic term. Note that thanks to the matrix notations (33), the argument easily extends to $p > 2$.

The assumptions needed to attain (35) are the following :

- *the kernel function K is nonnegative and bounded by 1, is positive on $[0, 1[$ and vanishes outside, and is continuously differentiable;*
- *g is twice differentiable with bounded second derivative;*
- *the value*

  $$E[\nabla m(X)]$$

  is separated away from 0;
- *the design points are "well diffused" throughout the support of X (see the reference for more details).*

This latter condition ensures that the matrix $W_{h,\Lambda_{\rho,\widehat{\theta}}}(.)$ is well-conditioned at any point X_i. Note that the asymptotic distribution of this estimator is not mentioned as such. It is simply stated that $\sqrt{n}(\widehat{\theta} - \theta_0)$ is close in distribution to a gaussian vector.

4.2.2 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 θ_0 and the inverse regression $E(X|Y = y)$. Unfortunately, their results require an important design condition :

Condition 6 *For any $\theta \in \mathbf{R}^p$, the conditional expectation $E(\theta' X|\theta_0' X = u)$ is linear in u.*

If it is not the case, a bias has to be taken into account. It can be shown that this condition is fulfilled if X is sampled randomly from any nondegenerate elliptically symmetric distribution, e.g. the normal distribution.

Description The enormous advantage of considering the inverse regression function $\xi(y) = E(X|Y = y)$ is to avoid the curse of dimensionality. Indeed, $\xi(y)$ can be nonparametrically estimated in a reliable way since Y is a scalar. It is noteworthy that we have

$$\xi(y) = \mu + \Sigma\theta_0\kappa(y), \tag{36}$$

where $\mu = E(X)$, $\Sigma = cov(X)$ and

$$\kappa(y) = \frac{E[\theta_0'(X-\mu)|Y=y]}{\theta_0'\Sigma\theta_0}.$$

This follows from the fact that

$$E(X|\theta_0'X = \theta_0'x) = \mu + \frac{\theta_0'(x-\mu)\Sigma\theta_0}{\theta_0'\Sigma\theta_0} \tag{37}$$

due to the design condition on X, and $\xi(y) = E[E(X|\theta_0'X)|Y=y]$.

Hence, from (36), it appears that θ_0 is proportional to $\Sigma^{-1}(\xi(y)-\mu)$, with the proportionality constant being $1/\kappa(y)$. For any y such that $\kappa(y) \neq 0$, we can thus estimate θ_0 by suitably scaling an estimate of $\Sigma^{-1}(\xi(y)-\mu)$. In order to combine the information from all y's, consider $\Gamma = cov(\xi(Y))$. We have, according to (36), that

$$\Gamma = var(\kappa(Y))\Sigma\theta_0\theta_0'\Sigma.$$

This matrix has clearly rank one : the only degree of freedom of ξ is y. From Cauchy's inequality, it is found that θ_0 solves the maximization problem

$$\theta_0 = \arg\max_{\theta\in\Theta} \frac{\theta'\Gamma\theta}{\theta'\Sigma\theta}. \tag{38}$$

θ_0 is thus the suitably scaled principal eigenvector for Γ, with respect to the inner product

$$< a, b >= a'\Sigma b. \tag{39}$$

The maximum value of the quotient is the principal eigenvalue. Remark that the spectral decomposition for Γ is trivial : all eigenvalues except the first are zero, since the rank of Γ is one.

Sampling case The estimation of θ_0 is of course based on relation (38), where the matrix Γ and Σ are estimated from the data. First of all we need to estimate the inverse regression function $\xi(y)$. For simplicity, a step function is used for this estimation. It might be not very efficient if the interest is the function $\xi(y)$ itself, but here it is just used at the first step of the procedure and it can be shown that the whole method is almost insensitive[4] to the smoothing degree of $\widehat{\xi}$. The estimation is done as follows : first, the range of Y is partitioned into, say, Q slices $\{s_1, ..., s_Q\}$. For each slice, $\xi(y)$ is estimated by the sample average of the corresponding X_i's, that is

$$\widehat{\xi}(y) = \widehat{\xi}_q = \frac{\sum_{i=1}^{n} X_i I(Y_i \in s_q)}{\sum_{i=1}^{n} I(Y_i \in s_q)} \quad \text{if } y \in s_q.$$

[4]this is emphasized in Zhu and Fang (1996), where a kernel estimate of the inverse regression function is used.

Similarly to (36), we have

$$\xi_q = E(\widehat{\xi}_q) = \mu + \Sigma\theta_0 k_q$$

where

$$k_q = E[\kappa(Y)|Y \in s_q].$$

Thus, if the scalar k_q for the qth slice is nonzero, we can estimate the direction of θ_0 using the direction of $\widehat{\Sigma}^{-1}(\widehat{\xi}_q - \overline{X})$, where $\overline{X}$ and $\widehat{\Sigma}$ are the sample average and variance-covariance matrix of vector X.

Since there is usually more than one slice for which k_q is nonzero, we can combine the information from all these slices to estimate the direction of θ_0. Let

$$\begin{aligned} \xi &= (\xi_1, ..., \xi_Q)',\ \widehat{\xi} = (\widehat{\xi}_1, ..., \widehat{\xi}_Q)',\ k = (k_1, ..., k_Q)', \\ p_q &= P(Y \in s_q),\ p = (p_1, ..., p_Q)' \text{ and } P = D(p) - pp' \end{aligned}$$

where $D(p)$ denotes the diagonal matrix with elements being the elements of vector p. p_q can obviously be estimated by $\widehat{p}_q$ the sample proportion of Y_i's in the qth slice, which leads to an estimate $\widehat{P}$ of P. Γ is then estimated by

$$\widehat{\Gamma} = \widehat{\xi}'\widehat{P}\widehat{\xi},$$

which is nothing else but a weighted sample variance-covariance matrix for the vector $\widehat{\xi}$. Remark that, by using $\widehat{P}$ as weight matrix, each slice is weighted by the sample proportion of observations falling inside the slice. By the strong law of large numbers, $\widehat{\Gamma}$ converges almost surely to

$$\xi' P \xi = k' P k \Sigma\theta_0\theta_0'\Sigma,$$

which is proportional to Γ. The estimator of θ_0 is then the answer of a maximization problem similar to (38) :

$$\widehat{\theta}_{sir} = \arg\max_{\theta\in\Theta} \frac{\theta'\widehat{\Gamma}\theta}{\theta'\widehat{\Sigma}\theta}. \tag{40}$$

This sliced inverse regression estimator is thus the principal eigenvector for $\widehat{\Gamma}$ with respect to the inner product $(a, b) = a'\widehat{\Sigma}b$. If the matrix $\widehat{\Sigma}$ is well-conditioned, it is well known that it amounts to the (suitably scaled) principal eigenvector of the matrix $\widehat{\Sigma}^{-1}\widehat{\Gamma}$.

If $\widehat{\delta}_{sir}$ is the unit principal eigenvector of $\widehat{\Sigma}^{-1}\widehat{\Gamma}$ and δ_0 the unit principal eigenvector of $\Sigma^{-1}\Gamma$, Duan and Li (1991) showed that

Theorem 7 *$\widehat{\delta}_{sir}$ is a consistent estimator of δ_0 and*

$$\sqrt{n}(\widehat{\delta}_{sir} - \delta_0) \xrightarrow{d} N(0, \Sigma_{sir})$$

if

- *the regressor variable X is sampled randomly from a nondegenerate elliptically symmetric distribution;*
- *$k'Pk > 0$.*

This latter condition can be difficult to check in practice, because very few is usually known about k. A sufficient condition is that $k \neq 0$, which is much easier to verify. For example, this condition holds if $\kappa(y)$ is monotonic. On the other hand, this condition prevents the inverse regression curve from being degenerate. This could be the case if g is a symmetric function about the mean of X. The matrix Σ_{sir} can be found in the reference. Of course, this result implies that $\widehat{\theta}_{sir}$ is a root-n consistent estimator of θ_0 and is asymptotically normal, with variance-covariance matrix $\phi\Sigma_{sir}\phi'$. ϕ is as in (23).

Violation of the design condition A natural question is to ask whether the slicing regression still provides a good estimate of θ_0 when the hypothesis about the design of X is not fulfilled anymore. Let θ^* be the solution to the maximization problem of (38) and λ be the value of the quotient for this θ^*. θ^* is the population version of the sliced inverse regression estimate. However, if the condition is not met, θ^* might not be collinear with θ_0. Let $\Lambda = cov[E(X|\theta_0'X = \theta_0'x)]$. With respect to the inner product (39), the principal eigenvector of Λ is θ_0 and its principal eigenvalue is one. Let τ be the second eigenvalue. Since, under the condition, $E(X|\theta_0'X = \theta_0'x)$ falls along the straight line (37), $\tau = 0$. If the hypothesis is violated, $E(X|\theta_0'X = \theta_0'x)$ is a curve which meanders around (37) and τ measures the largest mean squared deviation from this line. Duan and Li (1991) show that a bound for the noncollinearity between θ^* and θ_0 is given by

$$\sin^2(\theta_0, \theta^*) \leq \frac{\tau(1-\tau)}{\lambda(1-\lambda)}.$$

We see that θ_0 and θ^* are collinear either if $\tau = 0$ (the condition is not violated) or if $\lambda = 1$ ($Y = g(\theta_0'X)$, where g is invertible). As τ depends on Λ which itself depends on θ_0, this bound can be estimated if we have an initial estimate of this parameter. If it is not the case, we can replace τ by $\sup_{\theta\in\Theta}\tau(\theta)$, obviously leading to a more conservative bound.

4.3 Other estimators

The four previous ideas (SLS, SML, ADE, SIR) are historically the most popular ones in order to estimate θ_0. Nevertheless, this list is far to be exhaustive. There are much more estimators which have been proposed in the literature. 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). Unfortunately, this method requires the link function g to be strictly monotonic. In a dimension reduction purpose, Li (1992) suggests a method called Principal Hessian Directions, which can be adapted to the Single-Index context. Cook (1998) revisits it. However, the main results are based on Stein's lemma[5], which assumes that X has a normal distribution. Naik and Tsai (2000) extend the method of Partial Least Squares, well suited in parametric regression, to the case of Single-Index models. Xia et al. (2002) propose an adaptive approach for dimension reduction, called the Minimum Average Variance Estimation (MAVE). This is a kind of M-estimation method, inspired by the SIR method, the IADE 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. Finally, Huh and Park (2002) derive an extension of ADE, where the gradient of the regression function is evaluated in any X_i via local polynomial fits based on kernel weighted conditional likelihoods. A problem is that the method requires the maximization of locally weighted log-likelihood, that is it looses the main advantage of direct estimators. Besides, the conditional distribution of Y given X is assumed to belong to the exponential family.

[5] lemma 4, in Stein (1981).

5 Testing for the single-index hypothesis

Although the unknown nature of g provides a great flexibility for the model, hypothesis (2) remains somewhat restrictive : geometrically, it amounts to say that the regression function is constant when the regressors are varying along any direction orthogonal to θ_0. Therefore, it seems crucial to check the validity of this assumption. Specifically, we have to test

$$H_0 : \exists g : \mathbf{R} \to \mathbf{R}, \exists \theta_0 \in \mathbf{R}^p \text{ such that } E(Y|X = x) = g(\theta_0' x) \text{ a.e.}$$

versus

$$H_1 : \forall g : \mathbf{R} \to \mathbf{R}, \forall \theta_0 \in \mathbf{R}^p \; E(Y|X = x) \neq g(\theta_0' x).$$

5.1 Fan and Li's test

The first consistent test was proposed by Fan and Li (1996). Their procedure was based on the analysis of the residuals. With

$$\varepsilon_i = Y_i - g(\theta_0' X_i),$$

we see that

$$\begin{aligned} E(\varepsilon_i|X = X_i) &= E(Y_i|X = X_i) - E(g(\theta_0' X_i)|X = X_i) \\ &= m(X_i) - g(\theta_0' X_i), \end{aligned}$$

which equals zero if and only if H_0 is true. Hence, $E\{[E(\varepsilon_i|X = X_i)]^2\} \geq 0$ and the equality holds only under H_0. This property will be the basis of the test. Based on $\widehat{\theta}$ a root-n consistent estimator of θ_0 and $\widehat{g}^{\widehat{\theta},h}$ the associated kernel estimator of g with bandwidth h (see expression (7)), an empirical version I_n of $E\{[E(\varepsilon_i|X = X_i)]^2\}$ can be computed and compared to 0. Of course, the larger I_n, the more evidence there is to reject H_0. Asymptotic properties of this test statistic are given and it is shown that the test can detect sequences of local alternatives that differ from the null by $O((nh^{p/2})^{-1/2})$.

5.2 Delecroix, Hall and Vial's test

This latter remark constitutes the main drawback of the procedure : it fails to detect alternatives distant from $n^{-1/2}$ from H_0. Delecroix et al. (2004) develop such a test, based on geometrical arguments. As explained in introduction of this section, under H_0, the regression function m is a p-variate function of x which varies only with the value of the trace of x on θ_0.The proposed test statistic can be built on a measure of the variability of integral averages of m over hypercylinders whose axes are orthogonal to the potential index vector of the model. In this case, if the hypercylinders have congruent bases and represent axial slices of a space orthogonal to θ_0, then the integral averages will not vary with either the thickness or the location of the hypercylinders. Formally, let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two such hypercylinders (see the figures 1 and 2 for a representation in the case $p = 2$) and let $I_{\mathcal{S}}(m)$ be the integral average of m over an hypercylinder $\mathcal{S}$:

$$I_{\mathcal{S}}(m) = \frac{\int_{\mathcal{S}} m(x)dx}{\int_{\mathcal{S}} dx}. \tag{41}$$

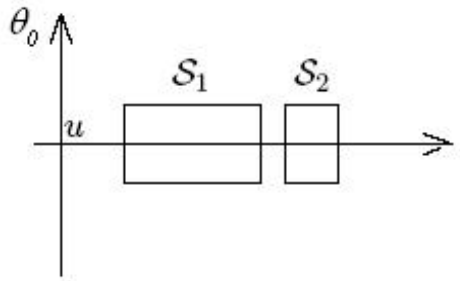


Fig.1 : hypercylinders whose axes are orthogonal to the index vector.

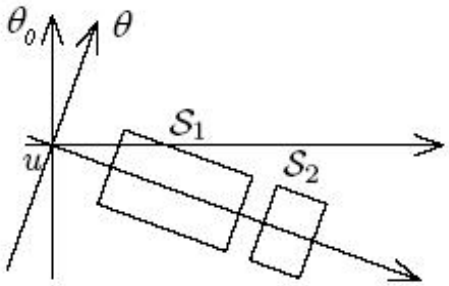


Fig.2 : hypercylinders whose axes are not orthogonal to the index vector.

Assuming the single-index model hypothesis, we have[6]

$$I_{\mathcal{S}_1}(m) = I_{\mathcal{S}_2}(m).$$

Building an empirical version of such $I_{\mathcal{S}}(m)$, depending on the orientation of the cylinders, the variability of integral averages can be estimated by[7]

$$\widehat{W}(\theta) = \int (\widehat{I}_{\mathcal{S}_1}(m) - \widehat{I}_{\mathcal{S}_2}(m))^2 d\mathcal{S},$$

where the integral is taken over the "set" of all pairs of hypercylinders defined as above. Naturally, if $\widehat{W}(\widehat{\theta})$ exceeds a certain critical value, H_0 is rejected. The proposed calibration is based on a bootstrap algorithm.

5.3 Other tests

It is probably worth mentioning Li (1992)'s approach. He tackles the problem in a totally different way, considering the multiple-index model, that is

$$Y = g(\theta_1'X, \theta_2'X, ..., \theta_k'X) + \varepsilon,$$

and proposes a significance test for determining the number k of components needed in the model. Of course, finding that one component is enough amounts to accept the single-index hypothesis and vice-versa. The clue is to check how many eigenvalues of an estimate of a matrix that we know being of rank k are significantly different of 0. Based on Stein's lemma, the method unhappily requires the normality of the regressors X. However, the idea is interesting.

[6] a formal mathematical statement of the equivalence of single index-ness and this property can be found in Delecroix et al. (2004).

[7] Note that $\widehat{\theta} = \arg\min \widehat{W}(\theta)$ can be shown to be a root-n consistent estimator of θ_0.

Finally, one can also mentioned Stute and Zhu (2002)'s test. As for the test of Fan and Li, this one is based on the residuals. The difference is that they transform, via an estimated quantile function, the estimated index $\widehat{\theta}' X_i$ to a variable which is approximately uniform on the unit interval $[0,1]$, which makes everything distribution-free, no matter the distributions of $\widehat{\theta}$ and X. The obtained test can detect local alternatives when they approach the null at the rate $O(n^{-1/2})$.

6 A simulation study

In this section, we investigate the practical performances of the different methods proposed in the previous sections from the perspective of estimating the unknown index vector θ_0. The data $\{(X_i, Y_i), i = 1, ..., n\}$ are independent and drawn from the following model :

$$Y_i = g(\theta_0' X_i) + \varepsilon_i,$$

with

$$g(x) = \sin(x),$$

$$p = 2,$$

$$\theta_0 = (1, 2)',$$

$$\varepsilon_i \sim N(0, 0.1)$$

and

$$X_i \sim N(0, I).$$

First of all remark that the normal distribution of the error term implies that the nonignorant SML exactly amounts to the SLS. Therefore, the results given in the table below for the SML are the results for an ignorant SML. Second, note that as the first component of θ_0 is known to be 1, the only parameter to be estimated is $\theta_0^{(2)}$, equal to 2.

The sample size is set to $n = 50$, $n = 100$ or $n = 200$ and 500 Monte-Carlo replications are drawn in each case. Each time a kernel estimate was needed, we used the Nadaraya-Watson estimator with Epanechnikov kernel. The needed bandwidths have been directly determined from (19) for SLS and SML, and has been fixed after prior experiments for AD estimators. So was set the number of slices for SIR. The expectation and the standard deviation of $\widehat{\theta}$, for each method, are given in the table below.

	$n = 50$		$n = 100$		$n = 200$	
	mean$(\widehat{\theta})$	st.dev.$(\widehat{\theta})$	mean$(\widehat{\theta})$	st.dev.$(\widehat{\theta})$	mean$(\widehat{\theta})$	st.dev.$(\widehat{\theta})$
SLS	1.6923	2.6458	2.2453	1.1564	2.1021	0.4738
SML	1.6447	2.4647	2.1965	1.1842	2.1076	0.5792
UADE	0.8398	2.8443	1.4219	2.5131	1.5704	3.1159
DWADE	1.7231	2.3047	2.0650	2.0697	2.2435	1.0359
IADE	1.0358	2.7653	0.9291	2.6166	0.8698	2.6036
SIR	0.0421	2.6013	0.5222	2.7310	0.7634	2.6060

First of all, it can be seen that the standard deviation of the estimates remains very important for $n = 50$ and $n = 100$, so that no significant effect of $X^{(2)}$ on Y could be detected, in an inference purpose. We also note that once the sample size grows, the M-estimators becomes clearly better than the direct ones, what was expected as it is stated that M-estimators are asymptotically efficient. On the other hand, for small sample size ($n = 50$), the DWAD Estimator seems to be the best choice : smallest bias and smallest variance. Nevertheless, as this estimator is subject to the curse of dimensionality, one could expect that its behavior should be worse with p larger than 2. A surprising feature is that the variance of UADE, IADE and SIR does not decrease with n. Actually, these methods are numerically very unstable, so that no reliable estimation seems to be possible from them, even when n becomes large. Finally, recalling that the SLS method is here totally equivalent to the nonignorant SML method, it is seen that the first one is slightly better than the latter, for n sufficiently large, what could also be expected.

7 Concluding remarks

The aim of this paper was not to make out an exhaustive survey about the whole existing theory about Single-Index Models. We simply wanted to provide the main basic ideas of this theory to the interested reader. This latter can found detailed and rigorous mathematical developments in the mentioned references, if needed. The principal part of the work is devoted to the estimation of the index coefficients : six methods among the most popular ones have been set out. All proposed estimators are root-n consistent, and the most have been proven to be asymptotically normal. Theoretically, M-estimators get some very nice properties. In particular, they are efficient and they do not require strong assumptions on the distribution of X, contrary to the direct methods. Besides, they supply the bandwidth needed for the nonparametric estimation of the link function. On the other hand, direct estimators are much more easy and fast to compute, as they do not require solving an nonlinear optimization problem. Practically, the simulation study shows that the M-estimators again outperform the direct ones. Among the direct estimators, the only one which gives satisfactory results seems to be the density-weighted average derivative estimator. One can take advantage of this fact by using it as an initial estimate in the optimization problems arising in M-estimation. However, the DWAD estimation lies on a bandwidth, and the choice of this bandwidth remains problematical. Once the index coefficients have been estimated, the estimation of the link function is done via standard nonparametric regression methods, with any danger of curse of dimensionality, as the index is an univariate random variable. The rate of convergence of this estimate is the usual rate of univariate nonparametric regression : $O((nh)^{-1/2})$.

References

[1] Amemiya, T. (1985), Advanced econometrics. Cambridge, MA, Harvard University Press.

[2] Cavanagh, C., Sherman, R.P. (1998), Rank estimators for monotonic index models, J. Econometrics 84, no. 2, 351–381.

[3] Cook, R.D. (1998), Principal Hessian directions revisited, J. Amer. Statist. Assoc. 93, no. 441, 84–100.

[4] Delecroix, M., Hall, P., Vial, C. (2004), Test for single-index models that are powerful against local alternatives, submitted to Econometrica.

[5] Delecroix, M., Härdle, W., Hristache, M. (2003), Efficient estimation in conditional single-index regression, J. Multivariate Anal. 86, no. 2, 213–226.

[6] Delecroix, M., Hristache, M. (1999), M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique, Bull. Belg. Math. Soc. 6, no. 2, 161–185.

[7] Delecroix, M., Hristache, M., Patilea, V. (2003), On semiparametric M-estimation in single-index regression, to appear in JSPI.

[8] Duan, N., Li, K.C. (1991), Slicing regression: a link-free regression method, Ann. Statist. 19, no. 2, 505–530.

[9] Fan, Y., Li, Q. (1996), Consistent model specification tests: omitted variables and semiparametric functional forms, Econometrica 64, no. 4, 865–890.

[10] Han, A.K. (1987), Nonparametric analysis of a generalized regression model : the maximum rank correlation estimator, J. Econometrics 35, no. 2-3, 303–316.

[11] Hall, P. (1989), On projection pursuit regression, Ann. Statist. 17, no. 2, 573–588.

[12] Härdle, W. (1990), Applied nonparametric regression, Econometric Society Monographs, 19. Cambridge University Press, Cambridge.

[13] Härdle, W., Hall, P., Ichimura, H. (1993), Optimal smoothing in single-index models, Ann. Statist. 21, no. 1, 157–178.

[14] Härdle, W., Hart, J.; Marron, J.S.; Tsybakov, A.B. (1992), Bandwidth choice for average derivative estimation, J. Amer. Statist. Assoc. 87, no. 417, 218–226.

[15] Härdle, W., Stoker, T.M. (1989), Investigating smooth multiple regression by the method of average derivatives, J. Amer. Statist. Assoc. 84, no. 408, 986–995.

[16] Härdle, W., Tsybakov, A.B. (1993), How sensitive are average derivatives?, J. Econometrics 58, no. 1-2, 31–48.

[17] Horowitz, J.L. (1998), Semiparametric methods in econometrics, Lecture Notes in Statistics, 131. Springer-Verlag, New York.

[18] Horowitz, J.L., Härdle, W. (1996), Direct semiparametric estimation of single-index models with discrete covariates, J. Amer. Statist. Assoc. 91, no. 436, 1632–1640.

[19] Hristache, M., Juditsky, A., Spokoiny, V. (2001), Direct estimation of the index coefficient in a single-index model, Ann. Statist. 29, no. 3, 595–623.

[20] Huh, J., Park, B. U. (2002), Likelihood-based local polynomial fitting for single-index models, J. Multivariate Anal. 80, no. 2, 302–321.

[21] Ichimura, H. (1993), Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, J. Econometrics 58, no. 1-2, 71–120.

[22] Klein, R.W., Spady, R.H. (1993), An efficient semiparametric estimator for binary response models, Econometrica 61, no. 2, 387–421.

[23] Li, K.C. (1991), Sliced inverse regression for dimension reduction, J. Amer. Statist. Assoc. 86, no. 414, 316–342.

[24] Li, K.C. (1992), On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma, J. Amer. Statist. Assoc. 87, no. 420, 1025–1039.

[25] Naik, P., Tsai, C.L. (2000), Partial least squares estimator for single-index models, J. R. Stat. Soc. Ser. B Stat. Methodol. 62, no. 4, 763–771.

[26] Newey, W.K., Stoker, T.M. (1993), Efficiency of weighted average derivative estimators and index models, Econometrica 61, no. 5, 1199–1223.

[27] Powell, J.L., Stock, J.H., Stoker, T.M. (1989), Semiparametric estimation of index coefficients, Econometrica 57, no. 6, 1403–1430.

[28] Powell, J.L., Stoker, T.M. (1996), Optimal bandwidth choice for density-weighted averages. J. Econometrics 75, no. 2, 291–316.

[29] Sherman, R.P. (1993), The limiting distribution of the maximum rank correlation estimator, Econometrica 61, no. 1, 123–137.

[30] Stein, C.M. (1981), Estimation of the mean of a multivariate normal distribution, Ann. Statist. 9, no. 6, 1135–1151.

[31] Stone, C.J. (1980), Optimal rates of convergence for nonparametric estimators, Ann. Statist. 8, no. 6, 1348–1360.

[32] Stute, W., Zhu, L.X. (2002), Model checks for generalized linear models, Scand. J. Statist. 29, no. 3, 535–545.

[33] Xia, Y., Tong, H., Li, W. K., Zhu, L.X. (2002), An adaptive estimation of dimension reduction space, J. R. Stat. Soc. Ser. B Stat. Methodol. 64, no. 3, 363–410.

[34] Zhu, L.X., Fang, K.T. (1996), Asymptotics for kernel estimate of sliced inverse regression, Ann. Statist. 24, no. 3, 1053–1068.