Dimension-reduction in reference curves estimation

Ali Gannoun[1], Stéphane Girard[1],
Christiane Guinot[2] and Jérôme Saracco[1]

[1] Laboratoire de Probabilités et Statistique,
Université Montpellier II,
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
Phone: +33-4-67-14-3505; Fax: +33-4-67-14-4974
e-mail: {gannoun, girard, saracco}@stat.math.univ-montp2.fr

[2] CE.R.I.E.S,
20, Rue Victor Noir, 92 521 Neuilly sur Seine Cedex, France
Phone: +33-1-46-43-43-59; Fax: +33-1-46-43-46-00
e-mail: christiane.guinot@ceries-lab.com

In order to get reference curves for data sets when the covariate is multidimensional, we propose a new methodology based on dimension-reduction and nonparametric estimation of conditional quantiles. This semiparametric approach combines Sliced Inverse Regression (SIR) and a kernel estimation of conditional quantiles. The asymptotic convergence of the derived estimator is shown. By a simulation study, we compare this procedure to the classical kernel nonparametric one for different dimensions of the covariate. The semiparametric estimator shows the best performance. The usefulness of this estimation procedure is illustrated on a real data set collected in order to establish reference curves for biophysical properties of the skin of healthy French women.

Key Words: Conditional quantiles, dimension-reduction subspace, kernel estimation, semiparametric method, reference curves, Sliced Inverse Regression (SIR).

1 Introduction

The reference intervals are a tool of some importance in clinical medecine. They provide a guideline to clinicians or clinical chemists seeking to interpret a measurement obtained from a new patient. Many experiments, in particular biomedical studies, are conducted to establish the range of values that a variable of interest, say Y whose values are in $\mathbb{R}$, may normally take in a target population. Here "normally" refers to values that one can expect to see with a given probability under normal conditions or for typical individuals, and the corresponding ranges are often referred to as norms or reference values. The conventional definition of a reference range is a pair of numbers (the reference limits) that bind, for example, the central 90% of a set of values (called the reference values) obtained from a specified group of subjects (the reference subjects).

The need for reference curves, rather than a simple reference range, arises when a covariate, say X whose values are in $\mathbb{R}$, is simultaneously recorded with Y. Norms are then constructed by estimating a set of conditional quantile (also called regression quantile) curves. Conditional quantiles are widely used for screening biometrical measurement (height, weight, circonferences and skinfold) against an appropriate covariate (age, time). For details, the readers may refer, for example, to the work of Healy, Rasbash and Yang (1988), Cole (1988), Goldstein and Pan (1992) or Royston and Altman (1992). Some extreme (high or low) quantiles of underlying distributions of measurement are particularly useful for industrial applications (see for example Magee, Burbidge and Robb (1991), Hendricks and Koenker (1992)).

Let $\alpha \in (0, 1)$, the conditional quantile of Y given $X = x$, denoted by $q_\alpha(x)$, is naturally defined as the the root of the equation

$$F(y|x) = \alpha, \tag{1}$$

where $F(y|x) = P(Y \leq y \mid X = x)$ denotes the conditional distribution function of Y given $X = x$. For $\alpha > 0.5$, reference curves are defined, when x varies, by

$$I_\alpha(x) = [q_{1-\alpha}(x), q_\alpha(x)]. \tag{2}$$

Thus, estimating reference curves is reduced to estimating conditional quantiles.

When the distribution of (X, Y) is unknown and n observations $\{(X_i, Y_i), i = 1, \ldots, n\}$ of (X, Y) are available, two basic estimation methods can be considered. An usual parametric method goes as follows: *i)* Assume that the error term in the regression model is homogeneous and normal, that is $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$;

ii) Estimate σ^2 by $\hat{\sigma}^2$ in the usual way; *iii)* Estimate the $\alpha-$quantile by $\hat{\mu}(x)+\hat{\sigma}N_\alpha$, where $\hat{\mu}(x)$ is the ordinary least square estimated regression (of the conditional expectation $\mu(x) = E(Y \mid X = x)$) and N_α is the $\alpha-$quantile of the standard normal distribution $\mathcal{N}(0,1)$. This method concords reasonably well with the conditional quantile when Y given X is normally distributed.

One nonparametric approach is to group data along the X-axis, estimate the α-quantile value for each group and connect the values between groups by the use of smoothing device. This method will usually be practical only for simple regression where the data can be displayed as a two-dimensional scattergram.

More generally, in the last decade a nonparametric theory has been developped in order to estimate the conditional quantiles. From (1), an estimator of the conditional distribution induces an estimator of corresponding quantiles. For instance, a *Nadaraya-Watson* estimator, $\hat{F}_{NW}(y|x)$, can be affected to $F(y|x)$. If we write $Y^\star = \mathbb{I}_{\{Y\leq y\}}$, then $F(y|x) = E(Y^\star|X = x)$. So, the estimation problem may be viewed as a regression of $Y^\star$ given X. This estimator is given by

$$\hat{F}_{NW}(y|x) = \sum_{i=1}^{n} K\{(x - X_i)/h_n\}\mathbb{I}_{\{Y_i\leq y\}} \Big/ \sum_{i=1}^{n} K\{(x - X_i)/h_n\} \ , \tag{3}$$

where h_n and K are respectively a bandwidth and a bounded (kernel) function. The estimator of $q_\alpha(x)$ is then deduced from $\hat{F}_{NW}(y|x)$ as the root of the equation

$$\hat{F}_{NW}(y|x) = \alpha. \tag{4}$$

Many authors are interested in this estimator, see for example Stute (1986), Samanta (1989), Gannoun (1990) and Berlinet et al. (2001).

Various other nonparametric methods are explored in order to estimate $q_\alpha(x)$. Among them we can cite the *local polynomial*, the *double kernel*, the *weighted Nadaraya-Watson* methods. For motivation, discussion and theoretical results on these estimation methods, the reader can see for example Stone (1977), Tsybakov (1986), Fan et al. (1994), Jones and Hall (1990), Chaudhuri (1991), Yu and Jones (1998), Poirot-Casanova (2000), Mint el Mouvid (2000) and Cai (2001).

Although theoretically the extension of conditional quantiles to higher dimension p of X is obvious, its practical success depends on the number of observations and suffers from the so-called *curse of dimensionality*. Further, because reference curves are, in this case, a pair of p-dimensional hyper-surfaces, their visual display is difficult making it less useful directly for exploratory purposes unlike the one-dimensional case. When $p \leq 2$, two- and three-dimensional plots can provide useful information on such changes, as discussed by Cook and Weisberg (1994). It is now easy to view three-dimensional plots with widely available software. Nevertheless it is very complicated to detect *graphically* if an individual is normal or not, even if we rotate the axis in the right direction. When $p > 2$, graphical methods are more difficult, because viewing all the data in single $(p+1)$-dimensional plot may no longer be possible.

Motivated by this, the key is then to reduce the dimension of the predictor vector X without loss of information on the conditional distribution of Y given X and without requiring a prespecified parametric model. Sufficient dimension-reduction leads naturally to the idea of sufficient summary plot that contains the whole information on the regression available from the sample. Moreover, it is a very helpful step in nonparametric estimation for circumvening the curse of dimensionality.

In this paper, we investigate the idea of the dimension-reduction in order to get efficient estimator of conditional quantiles from which we deduce reference curves. In Section 2, we present the dimension-reduction context. We derive in Section 3 a semiparametric estimator of conditional quantiles based on a dimension-reduction method. Section 4 is devoted to some asymptotic results. Simulations are conducted in Section 5 to assess the performance of this estimator in finite-sample situation. Numerical examples involving real-data application are reported in Section 6. All proofs and technical arguments are given in the Appendix.

2 Theoretical dimension-reduction context

This section is devoted to the presentation of the dimension-reduction subspaces as well as the construction of associated base by the SIR method. We emphasize the consequence of the dimension-reduction on the definition of the conditional quantile.

2.1 Dimension reduction subspaces

Definitions. Herein, the notation $U \perp V \mid Z$ means that the random vectors U and V are independent given any value for the random vector Z. Similarly, $U \perp V$ means that U and V are independent. Subspaces are denoted by the letter S. For any $p \times q$ matrix W, let $S(W)$ the subspace of $\mathbb{R}^p$ spanned by the columns of the matrix W. We suppose that there exists a $p \times q$ matrix $(q \leq p)$ B such that

$$Y \perp X \mid B^T X. \tag{5}$$

Such a B always exists, because (5) is trivially true when $B = I_p$, the $p \times p$ identity matrix. This implies that the $p \times 1$ predictor vector X can be replaced by $q \times 1$ predictor vector $B^T X$ without loss of regression information. If $q < p$ the reduction in the dimension of the predictor vector is effective. If (5) holds, then it also holds when B is replaced with any matrix whose columns form a basis for $S(B)$. Thus (5) is appropriately viewed as a statement about $S(B)$, which is called a *dimension-reduction subspace for the regression of Y on X*. Clearly knowledge of the smallest dimension-reduction subspace would be useful for parcimoniously characterizing how the distribution of Y given X changes with values of X. There are several ways to define such a subspace. The focus is on the *central dimension-reduction subspace*, denoted by $S_{Y|X}$ (see Cook (1994, 1996)) which is the intersection of all dimension-reduction subspaces. Under various reasonable conditions, this subspace exists and is unique. This central subspace is then the smallest dimension-reduction subspace such that

$$Y \perp X \mid \beta^T X, \tag{6}$$

where the columns of the $p \times d$ matrix β ($d \leq q$) form a basis of this subspace.

Consequence. In the current study, the statement (6) is equivalent to

$$F(y|x) = F(y|\beta^T x), \tag{7}$$

for all values of x in the sample space. Straightforwardly, it follows that

$$q_\alpha(x) = q_\alpha(\beta^T x).$$

Caracterization. Let $S_{E(X|Y)}$ denote the subspace spanned by $\{E(X|Y) - E(X) : Y \in \Omega_Y\}$ where $\Omega_Y \in \mathbb{R}$ is the sample space of Y. Given (6), assume that the marginal distribution of the predictors X satisfies the following *linearity condition*:

(LC) For all $b \in \mathbb{R}^p$, $E(b^T X|\beta^T X)$ is linear in $\beta^T X$.

Let Σ be the variance matrix of X, supposed positive-definite. Under **(LC)**, Li (1991) shows that the centered inverse regression curve $E(X|Y) - E(X)$ satisfies

$$E(X|Y) - E(X) \in S(\Sigma\beta).$$

Thus,

$$S_{E(X|Y)} \subseteq S(\Sigma\beta) = \Sigma S_{Y|X}. \tag{8}$$

Let Z denote the standardized version of the predictor X defined by $Z = \Sigma^{-1/2}(X - E(X))$, where $\Sigma^{-1/2}$ is the unique symmetric positive-definite square root of Σ^{-1}. Let $\eta = \Sigma^{1/2}\beta$. There is no loss of generality working on the Z scale, because any basis for $S_{Y|Z}$ can be back-transformed to a basis for $S_{Y|X}$:

$$S_{Y|X} = \Sigma^{-1/2} S_{Y|Z}.$$

Therefore, in view of (8), we have

$$S_{E(Z|Y)} \subseteq S(\eta) = S_{Y|Z}. \tag{9}$$

This does not guarantee equality between $S_{E(Z|Y)}$ and $S_{Y|Z}$ and, thus, inference about $S_{E(Y|Y)}$ possibly covers only part of $S_{Y|Z}$. Moreover, it is clear that

$$S\{Var(E(Z|Y))\} = S_{E(Z|Y)}, \tag{10}$$

exept on a set of measure zero.

Using results (9) and (10), the estimation of the inverse regression curve $E(Z|Y)$ serves to estimate the central dimension-reduction subspace via estimating the matrix $Var(E(Z|Y))$. Methods are available for estimating portions of the central subspace. In the next subsection, we mainly focus on the classical SIR method introduced by Li (1991) which is a simple non-smooth nonparametric estimation method for $S_{Y|Z}$.

Remark. The linearity condition is required to hold only for the basis β of the central subspace. Since β is unknown, in practice we may require that it holds for all possible β, which is equivalent to elliptical symmetry of the distribution of X (see Eaton (1986)). This condition holds when X is normally distributed. Li (1991) mentioned that the linearity condition is not a severe restriction, since most low-dimensional projections of high-dimensional data clouds are close to being normal (see Diaconis and Freedman (1984), Hall and Li(1993)). Experience indicates that linearizing predictor transformations often result in relatively simple models. In addition, it is possible to use the reweighting procedure proposed by Cook and Nachtsheim (1994) after predictor transformations to remove gross nonlinearities.

2.2 Slice Inverse Regression

The idea is based on partitionning the range of the one-dimensional response variable Y into a fixed number H of slices denoted $\mathcal{S}_1, \dots, \mathcal{S}_H$. Then, the p components of Z are regressed on $\widetilde{Y}$, the discrete version of Y resulting from slicing its range, giving p one-dimensional regression problems, instead of the possibly high-dimensional forward regression of Y on Z. Let M denote the matrix $Var(E(Z|\widetilde{Y}))$. Using (9) and (10), it is clear that

$$S(M) = S_{E(Z|\tilde{Y})} \subseteq S_{\tilde{Y}|Z} \subseteq S_{Y|Z}. \qquad (11)$$

The last inclusion in (11) holds because $\widetilde{Y}$ is a function of Y, which implies that $S_{Y|Z}$ is a dimension-reduction subspace for the regression of $\widetilde{Y}$ on Z. Note that, using the slicing $\mathcal{S}_1, \ldots, \mathcal{S}_H$, M is written

$$M = \sum_{h=1}^{H} p_h m_h m_h^T,$$

where $p_h = \mathrm{P}(Y \in \mathcal{S}_h)$ and $m_h = E[Z|Y \in \mathcal{S}_h]$.
Let $s_1 \geq s_2 \geq \ldots \geq s_p$ denote the singular values of M, and $u_1, u_2, \ldots, u_p$ denote the corresponding left singular vectors. Assuming that $d = \dim(S(M))$, $S\,(M) = S(u_1, u_2, \ldots, u_d)$. Transforming back to the X scale, $b_k = \Sigma^{-1/2} u_k$, $k = 1, \ldots, d$ form an basis of $S(\beta)$. Following SIR vocabulary, the reduction subspace $S(\beta)$ is called the *effective dimension-reduction* (EDR) space. The vectors b_k are named EDR directions.

Remark. More details and comments on the SIR estimation procedure can be found in Li (1991) and Chen and Li (1998). SIR has been discussed in several articles with emphasis on its asymptotic properties (Hsing and Carroll (1992), Kötter (1996), Zhu and Fang (1996), Saracco (1997, 1999) among others) and on the estimation of the dimension of the EDR space (Schott (1994), Ferré (1995, 1998)). Carroll and Li (1992) use SIR in a nonlinear regression model with measurement error in the covariates. The situation of small sample sizes has been studied by Aragon and Saracco (1997). Bura (1997) uses a multivariate linear model for the inverse regression curve. Li *et al* (1999) consider the case of censored regression data.

Note that various methods for estimating the EDR space have been developped. They are based on second moments (Li (1991) for SIR-II and SIR$_\alpha$ approaches, Cook and Weisberg (1991) for SAVE method, Saracco (2001) and Gannoun and Saracco (2000, 2001) for the pooled slicing version of these methods and the choice of α) and on the properties of so called principal Hessian directions (Li (1992)).

3 Estimation procedure

In this section, we describe the practical implementation of the dimension-reduction step. Then, we present the kernel method for estimating the conditional quantiles.

To this end, let Y_i denote the ith observation on the univariate response and let X_i denote the corresponding $p \times 1$ vector of observed covariate values, $i = 1, \ldots, n$. The $(p+1) \times 1$ data (X_i, Y_i) are assumed to be independent and identically distributed observations from the multivariate random vector (X, Y) with finite moment.

3.1 SIR estimation step

Let $\overline{X}$ and $\widehat{\Sigma}$ be the sample mean and the sample variance matrix of the X_i's. Let $\widehat{Z}_i$ be the estimated standardized predictor defined by $\widehat{Z}_i = \widehat{\Sigma}^{-1/2}(X_i - \overline{X})$, $i = 1, \ldots, n$. Then the SIR estimate of M is given by

$$\widehat{M} = \sum_{h=1}^{H} \widehat{p}_h \widehat{m}_h \widehat{m}_h^T$$

where H is the fixed number of slices, $\widehat{p}_h = n_h/n$ with n_h being the number of observations in the hth slice, and $\widehat{m}_h$ is the p-vector of the average of $\widehat{Z}$ within slice h. Let $\hat{s}_1 \geq \hat{s}_2 \geq \ldots \geq \hat{s}_p$ denote the singular values of $\widehat{M}$, and $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_p$ denote the corresponding left singular vectors.
Assuming that the dimension d of $S(M)$ is known,

$$S\left(\widehat{M}\right) = S(\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_d)$$

is a consistent estimate of $S(M)$.
In practice, the dimension d is replaced with an estimate $\hat{d}$ equal to the number of singular values that are inferred to be nonzero in the population (see Li (1991,1992), Schott (1994) and Ferré (1998)).

If $\dim(S(\eta)) = d$, $\widehat{M}$ clearly provides an estimated basis of $S(\eta)$. Transforming back to the original scale, $\hat{b}_k = \widehat{\Sigma}^{-1/2}\hat{u}_k$, $k = 1, \ldots, d$ form an estimated basis of $S(\beta)$. Similarly to the population version, the vectors $\hat{b}_k$ are the estimated EDR directions and they span the estimated EDR space.

We are now able to construct a conditional quantile estimator based on these SIR estimates.

3.2 Conditional quantile estimation step

Using the SIR estimates obtained in the previous subsection, we now give an estimator of the conditional distribution function from which we derive an estimator of the conditional quantile.

For the sake of convenience of statement, we assume that $d = 1$. Let us recall that, in the present dimension-reduction context, we have

$$F(y|x) = F(y|\beta^T x) = F(y|b^T x). \tag{12}$$

Using the notation $\hat{b} = \hat{b}_1$, $\hat{b}$ is an estimated basis of $S_{Y|X} = S(\beta)$. Then, the corresponding estimated index values of X_i and x are as follows:

$$\widehat{V}_i = \hat{b}^T X_i, i = 1, \ldots, n \quad \text{and} \quad \hat{v} = \hat{b}^T x.$$

Following (3), from the data $\{(Y_i, \hat{V}_i), i = 1, \ldots, n\}$, we define a kernel estimator of $F(y|x)$ by

$$F_n\left(y \,\middle|\, \hat{b}^T x\right) = F_n(y|\hat{v}) = \sum_{i=1}^{n} K\{(\hat{v} - \widehat{V}_i)/h_n\} \mathbb{I}_{\{Y_i \leq y\}} \bigg/ \sum_{i=1}^{n} K\{(\hat{v} - \widehat{V}_i)/h_n\}. \tag{13}$$

Then, as in (4), we derive from (13) an estimator of $q_\alpha(x)$ by

$$q_{n,\alpha}\left(\hat{b}^T x\right) = q_{n,\alpha}(\hat{v}) = F_n^{-1}(\alpha \mid \hat{v}). \tag{14}$$

As a consequence of the above result, for $\alpha > 0.5$, the corresponding estimated reference curves are given by the following

$$I_{n,\alpha}(x) = [q_{n,1-\alpha}(\hat{b}^T x), q_{n,\alpha}(\hat{b}^T x)], \tag{15}$$

as x varies.

Remark. The above definitions have been presented in the context of single index ($d = 1$). A natural extension is to consider the general multiple indices ($d > 1$) and to work with $\hat{b}_j = \widehat{\Sigma}^{-1/2}\hat{u}_j, j = 1, \ldots, d$ and $\hat{V}_i = (\hat{b}_1^T X_i, \hat{b}_2^T X_i, \ldots, \hat{b}_d^T X_i), i = 1, \ldots, n$. Then we follow the classical multi-kernel estimation to get $q_{n,\alpha}(\hat{b}_1^T x, \ldots, \hat{b}_d^T x)$ as in (14).

In the next section, we study the consistency of $q_{n,\alpha}(\hat{b}^T x)$.

4 Asymptotic properties

The additional assumptions under which the results in this paper are derived are gathered together below for easy reference. It will be usually indicated which of these assumptions are used in the various steps of the derivations.

Assumptions

(A1) The random vectors $(X_i, Y_i), i \geq 1$, are defined on probability space $(\Omega, \mathcal{A}, \mathcal{P})$ and constitute a strictly stationary process.

(A2) The kernel $K : \mathbb{R} \longrightarrow \mathbb{R}$ is a probability density function such that:
(i) K is bounded
(ii) $\mid v \mid K(v) \longrightarrow 0$ as $\mid v \mid \longrightarrow \infty$.
(iii) $\int vK(v)dv = 0$ and $\int v^2 K(v)dv < \infty$.

(A3) The sequence of bandwidth $h = h_n$ tends to zero such that

$$nh/\ln n \longrightarrow \infty.$$

(A4) The variable X admits a continuous marginal density.

(A5) $F(. \mid b^T x)$ and $F(y \mid .)$ are both continuous.

(A6) For $\alpha \in (0, 1)$ and $x \in \mathbb{R}^p$, $F(. \mid b^T x)$ has a unique α-quantile.

Comments on the Assumptions. Asumptions **(A2)** and **(A3)** are quite usual in kernel estimation. As a direct consequence of Assumption **(A4)**, the variable $b^T X$ admits a continuous marginal density. Assumption **(A5)** is used to prove the uniform convergence (in probability) of $F_n(. \mid \hat{b}^T x)$ to $F(. \mid b^T x)$. Assumption **(A6)** is used in the proof of the convergence of $q_{n,\alpha}(\hat{b}^T x)$ to $q(x)$. Note that if there is no unicity, we can define $q_\alpha(x) = \inf\{y : F(y \mid b^T x) \geq \alpha\}$.

With this in mind, we have then:

Theorem 1 *Under Assumptions **(LC)**, **(A1)**-**(A5)**, for a fixed x in $\mathbb{R}^p$, we have*

$$\sup_{y \in \mathbb{R}} \mid F_n(y \mid \hat{b}^T x) - F(y \mid x) \mid \longrightarrow 0 \quad \textit{in probability, as } n \to +\infty.$$

Theorem 2 *Under Assumptions **(LC)**, **(A1)**-**(A6)**, for a fixed x in $\mathbb{R}^p$, we have*

$$q_{n,\alpha}(\hat{b}^T x) \longrightarrow q_\alpha(x) \quad \textit{in probability, as } n \to +\infty.$$

5 Simulation study

In this section, we study the numerical performances of the proposed method on simulated data. In particular, we compare our method with the classical nonparametric estimation method which does not include a dimension reduction step.

5.1 Estimation methods

Let us introduce the following estimators of the α-conditional quantile at point x :

(a) $q_{n,\alpha}^{(a)}(x) \stackrel{\text{def}}{=} q_{n,\alpha}(\widehat{b}^T x)$, the estimator defined in (14). The direction $\widehat{b}$ is estimated with the SIR procedure, and the corresponding conditional quantile by numerically inversing the estimated conditional c.d.f. (3). The one-dimensional kernel is the density of the standard normal distribution $\mathcal{N}(0,1)$, and the bandwidth is chosen by a cross-validation technique (see for instance Härdle (1990) or Schimek (2000)).

(b) $q_{n,\alpha}^{(b)}(x) \stackrel{\text{def}}{=} q_{n,\alpha}(\beta^T x)$ has no practical interest, it is only introduced for the sake of comparison. It is similar to **(a)** except the dimension-reduction direction is not estimated but fixed to the theoretical one.

(c) $q_{n,\alpha}^{(c)}(x) \stackrel{\text{def}}{=} q_{n,\alpha}(x)$ is the classical conditional nonparametric quantile estimator. It is computed by numerically inversing the conditional c.d.f. (1) estimated with a multidimensional kernel. This kernel is the density of the standard multinormal distribution $\mathcal{N}_p(0, I_p)$ and the bandwidth is chosen by the same cross-validation technique as in **(a)**.

5.2 Simulated models

We consider the following regression model

$$Y = f(\beta^T X) + \varepsilon, \tag{16}$$

where X follows the standard multinormal distribution $\mathcal{N}_p(0, I_p)$ and where ε is normally distributed $\varepsilon \sim \mathcal{N}(0,1)$ and is independent from X. f is a $\mathbb{R}^p \to \mathbb{R}$ function and $\beta \in \mathbb{R}^p$. We examine three situations:

(M1) $p = 3$, $f(t) = 1 + 2t/3$ and $\beta^T = 2^{-1/2}[1, -1, 0]$.

(M2) $p = 10$, $f(t) = 1 + 2t/3$ and $\beta^T = 3^{-1}[1, 1, 1, 1, 1, -1, -1, -1, -1, 0]$.

(M3) $p = 3$, $f(t) = 1 + \exp(2t/3)$ and $\beta^T = 2^{-1/2}[1, -1, 0]$.

Our motivation for considering the pair of models **(M1,M2)** is to investigate the behavior of the estimation methods when the dimension increases. The pair of models **(M1,M3)** is introduced to evaluate the influence of the link function f on the accuracy of the estimation methods. Let us note that, in all cases, the true conditional α-quantile associated to the model (16) can be written:

$$q_\alpha(x) = f(\beta^T x) + N_\alpha, \tag{17}$$

where N_α is the α-quantile of the standard normal distribution.

5.3 Evaluation of the results

Our goal is to compare successively the three estimators **(a)**, **(b)** and **(c)** to the true quantile (17) in the situations **(M1)**, **(M2)** and **(M3)**. To this end, $N = 100$ data sets with size $n = 200$ are simulated in each of the above situations. The conditional quantiles are estimated for $\alpha = 5\%$ and $\alpha = 95\%$ on a p dimensional grid. This grid is composed of 125 points $\{z_\ell,\ \ell = 1, \ldots, 125\}$ randomly generated with a uniform distribution on $[-3/2, 3/2]^p$. Then, the performance of the estimators can be assessed on each of the N simulated data sets by a mean square error criterion

$$E_{n,\alpha}^{(\Theta)} = \frac{1}{125} \sum_{\ell=1}^{125} \left(q_{n,\alpha}^{(\Theta)}(z_\ell) - q_\alpha(z_\ell) \right)^2,$$

where $\Theta \in \{a, b, c\}$.

5.4 Results

An example of graphical representation of the estimated quantiles for the **(M3)** model is presented Figure 1. The superimposition in Figure 1.1 of the theoretical quantile (17) with the estimated one with the theoretical index **(b)** shows the accuracy of the estimation. The estimation **(a)** is presented on a separate sheet (see Figure 1.2) since it is parametrized with a different axis (directed by the estimated direction $\widehat{b}$) from the previous ones.

A more quantitative comparison is possible by studying the empirical distribution of the mean square error $E_{n,\alpha}^{(\Theta)}$ for $\Theta \in \{a, b, c\}$ and $\alpha \in \{0.05, 0.95\}$ on each model **(M1)**, **(M2)** and **(M3)**. The corresponding boxplots are represented on Figure 2.

Let us consider first the results of Figure 2.1. It appears that there is no difference between the distribution of $E^{(a)}_{n,\alpha}$ and $E^{(b)}_{n,\alpha}$. The estimation of the direction β by $\widehat{b}$ has no significative consequence on the accuracy of the estimation of the reference curves. This result is confirmed by Figure 3.1. Directions β and $\widehat{b}$ remain very close to each other on the whole N simulations. On the contrary, results obtained by the estimators **(a)** and **(c)** are very different. The proposed estimator **(a)** gives better results than the estimator without dimension-reduction **(c)**. Besides, this difference of quality increases with the number p of covariates (see Figure 2.3). In this case, the curse of dimensionality becomes an essential limitation to the use of estimator **(c)**, and thus estimator **(a)** is particularly interesting in such situations. Moreover, we can see on Figure 3.2 that the quality of the estimation of β is not severely affected by the covariates number. Finally, in view of Figure 2.2, the nature of the link function f does not seem to have any influence on the relative behaviors of the three estimators.

6 Application to real data

When studying biophysical skin properties of healthy women, knowledge about reference "curves" of certain parameters is lacking. Information concerning these parameters refers to few studies. The aim is to establish 90% reference "curves" for some biophysical properties of the skin of healthy Caucasian women on two facial areas and one forearm area, using the age and a set of covariates.

We organize this section in the following manner. The first subsection gives an overview of the considered data set. Next, the estimation procedure is detailed. Finally, we describe the corresponding results.

6.1 Data

This study was conducted from November 1998 to March 1999 on $n = 322$ Caucasian women between 20 and 80 years old with healthy apparent skin (i.e. without any sign of ongoing skin disease or general disease with proven cutaneous manifestations), and living in the Ile de France area. The volunteers were preselected by a subcontractor company. Each healthy volunteer was examined at CE.R.I.E.S ("CEntre de Recherches et d'Investigations Epidermiques et Sensorielles" or Epidermal and Sensory Research and Investigation Centre) in a controlled environment (temperature $23 \pm 1°c$ and a relative humidity of $50 \pm 5\%$). This evaluation included self-administered questionnaires on skin-related habits, a medical examination and a biophysical evaluation. Some biophysical properties of the skin were performed on two areas of the face (forehead and cheeks) and on the left volar forearm.

Three independent studies are made on each area. The variable of interest is the conductance, denoted by KBRAS (resp. KFRONT, KJOUE) for the forearm (resp. forehead, cheek) area. The variables AGE (age of the volunteer), TEMP (temperature of the controlled environment) and HYGRO (relative humidity of the controlled environment) occur in each study as covariates. For the forearm (resp. forehead, cheek) area, the other available covariates included are:

- the skin temperature denoted by TBRAS (resp. TFRONT, TJOUE),
- the transepidermal water loss denoted by BRAS1 (resp. FRONT1, JOUE1),
- the skin pH denoted by PBRAS (resp. PFRONT, PJOUE),
- the skin hydration by capacitance denoted by C2BRAS (resp. C2FRONT, C2JOUE),
- the skin color expressed using L*, a*, b* system, where L* expresses brightness, a* the red/green chromacity coordinate and b* the yellow/blue chromacity coordinate. Three variables are then measured: ABRAS, BBRAS and LBRAS (resp. AFRONT, BFRONT, LFRONT, AJOUE, BJOUE and LJOUE).
- the sebum casual level which is only measured on the forehead and cheek areas, the corresponding variables are SFRONT and SJOUE.

The goal is to estimate the 90% reference "curves" for the variables of interest KBRAS, KFRONT and KJOUE, using their corresponding set of covariates.

6.2 Methodology

Three steps are necessary to describe the methodology. For convenience, let us denote by $\mathcal{X}$ the set of the p covariates measured on the considered area ($p = 10$ or 11 in the current study).

- **Step 1.** We apply SIR method using the variable of interest Y and the covariates of $\mathcal{X}$. From the eigenvalues scree plot, we determine the number d of EDR directions to keep, that is the number of eigenvalues significantly different from zero in theory. From a practical point of view, we look for a visible jump in the scree plot and K is then the number of the eigenvalues located before this jump. Note that if no jump is detected, no reduction of dimension is possible.

The corresponding estimated EDR directions are therefore $\hat{b}_1, \ldots, \hat{b}_d$.

We visualize the structure of the "reduced" data $\left\{(Y_i, \hat{b}_1^T X_i, \ldots, \hat{b}_d^T X_i)\right\}$ and we check visually if there is no structure on $\left\{(Y_i, \hat{b}_{d+1}^T X_i, \ldots, \hat{b}_p^T X_i)\right\}$.

- **Step 2.** The aim here is to "simplify" the indices $\hat{b}_k^T X$ in order to get an easier interpretation. To this end, for each index, we make a forward-selected linear regression model of $\hat{b}_k^T X$ on the covariates of $\mathcal{X}$ (based on the AIC criterion). Note that we impose the presence of the covariate AGE in the model. We then obtain $\mathcal{X}_1, \ldots, \mathcal{X}_d$, the corresponding subsets of selected covariates. The final subset is then $\widetilde{\mathcal{X}} = \cup_{k=1}^d \mathcal{X}_k$. Let us remark that the selection of covariates is effective if $\widetilde{\mathcal{X}}$ is strictly included in $\mathcal{X}$.

  We apply SIR again with the covariates of $\widetilde{\mathcal{X}}$ and we obtain the corresponding d estimated EDR directions $\tilde{b}_1, \ldots, \tilde{b}_d$. Finally, we graphically check that each plot $\left\{(\hat{b}_k^T X_i, \tilde{b}_k^T X_i)\right\}$ has a linear structure.

- **Step 3.** We are now able to estimate the reference "curves" (which are hypersurfaces when $d > 1$) on the sample $\left\{(Y_i, \tilde{b}_1^T X_i, \ldots, \tilde{b}_K^T X_i)\right\}$, by the kernel method described in subsection 3.2.

Note on the computational implementation. We implement the SIR estimation procedure in Splus. All the graphics are made with Splus too. As in the simulation study, the kernel used is the multivariate normal density $\mathcal{N}(0, I_K)$ and the bandwidth is selected by cross validation. The estimation of the reference curves are obtained using C codes. If $K = 1$, each reference curve is evaluated on 50 points equidistributed on the range of the estimated index, otherwise the reference hyperfursaces are evaluated on an appropriate grid depending on the dimension K.

6.3 Results

We are now going to apply the methodology developped in the previous subsection in order to get the 90% reference "curves" for the variables KBRAS, KFRONT and KJOUE.

- **Study of KBRAS response variable.** Following the step 1 of the methodology, from the eigenvalues scree plot (see Figure 4.1), we select the dimension $d = 1$. With graphic tools, we clearly observe on Figure 5.1.1 a structure between the first estimated index and KBRAS. No structure is detected with the second estimated index (see Figure 5.1.2). These observations confirm the choice $d = 1$.

  Step 2 is summarized in Table 1 which gives the selected covariates (first column), the value of the AIC criterion in the corresponding single term addition step (second column). The estimated EDR direction is provided in the last column.

  Figure 6.1 confirms that the index based on the remaining selected covariates is very similar to the one based on the whole covariates.

  In the third step, we construct the 90% reference curves for KBRAS using this estimated index, see Figure 7.1.

  The results of the analysis on the forearm show that apart from age five covariates enter in the model. Two of these represent the environmental conditions of the measurements, which is to be expected, the three other covariates are directly clinically related with skin hydration: skin pH, capacitance and transepidermal water loss.

- **Study of KJOUE response variable.**

  As for the previous variable, Figure 4.2, Figure 5.2.1 and Figure 5.2.2 allow us to select $d = 1$.

  Table 2 is devoted to the selection of the covariates for the index of step 2.

  The estimated 90% reference curves for KJOUE using this estimated index are plotted in Figure 7.2.

  Regarding the cheek, the results of the analysis show that the environmental conditions are represented only by room temperature in the model, and that sebum casual level plays a major role on skin hydration. These results lead to a slightly different model than those obtain for the forearm. It is, however, perfectly consistent with the physiological topographic specificity of the skin area studied.

- **Study of KFRONT response variable.**

  In a similar way, we select $d = 1$ direction, see the eigenvalues scree plot of Figure 4.3 and the plots of Figure 5.3.1 and Figure 5.3.2.

  The selection of the covariates for the index of step 2 is exhibited in Table 3.

  Surprisingly, the variable AGE is not significant in the final model selected by the forward selection procedure, but let us recall that we impose it in the index.

  Figure 7.3 shows the estimated 90% reference curves for KFRONT using this estimated index.

  Interestingly the results of the forehead lead to a different conclusion as only one covariate enters in the model and even the effect of the age is not significant. Although these results might appear inconsistent with those of the two previous analyses, they reflect in fact once again the anatomo- physiological specificity of the skin area studied. It is indeed well known that the bone structure underlying the forehead plays a major role in masking the effect of ageing in this specific area.

Global comment. Altogether we consider that the results provided by our approach enable the construction of references curves indicating skin hydration assessed by conductance. The results are consistent with the physiological specificity of the various skin areas studied.

Acknowledgement

We are grateful to Dr Mohamed Mellouk and Dr Ludovic Menneteau, University Montpellier II, for useful comments and helpful suggestions during the preparation of the manuscript, to Pr. Denis Malvy, University Bordeaux 2, for his advice and clinical expertise, to Pr. Erwin Tschachler for his encouragement, to Isabelle Le Fur and Frédérique Morizot and all the CE.R.I.E.S.' team for their important contribution to the data.

APPENDIX

Proof of Theorem 1

From Li (1991), under **(LC)**, we have that $\hat{b}$ converges to b with the rate $\sqrt{n}$. Then for fixed $x \in \mathbb{R}^p$ we have $\hat{b}^T x$ converges to $b^T x$. Let us also recall that $F(y \mid x) = F(y \mid b^T x)$.

Now $\forall y \in \mathbb{R}$,

$$\mid F_n(y \mid \hat{b}^T x) - F(y \mid x) \mid \leq \mid F_n(y \mid \hat{b}^T x) - F(y \mid \hat{b}^T x) \mid + \mid F(y \mid \hat{b}^T x) - F(y \mid x) \mid .$$

Consequently,

$$\begin{aligned} \sup_{y \in \mathbb{R}} \mid F_n(y \mid \hat{b}^T x) - F(y \mid x) \mid \quad &\leq \quad \underbrace{\sup_{y \in \mathbb{R}} \mid F_n(y \mid \hat{b}^T x) - F(y \mid \hat{b}^T x) \mid}_{\textbf{(P1)}} \\ &+ \quad \underbrace{\sup_{y \in \mathbb{R}} \mid F(y \mid \hat{b}^T x) - F(y \mid x) \mid}_{\textbf{(P2)}} . \end{aligned}$$

Let us first show that **(P1)** converges to 0.

Let $\gamma > 0$ and $J(\gamma) := [b^T x - \gamma, b^T x + \gamma]$, then

$$\begin{aligned} \sup_{y \in \mathbb{R}} \mid F_n(y \mid \hat{b}^T x) - F(y \mid \hat{b}^T x) \mid \quad &\leq \quad \sup_{y \in \mathbb{R}} \mid F_n(y \mid \hat{b}^T x) - F(y \mid \hat{b}^T x) \mid \mathbb{I}_{\{\hat{b}^T x \in J(\gamma)\}} \\ &+ \quad \sup_{y \in \mathbb{R}} \mid F_n(y \mid \hat{b}^T x) - F(y \mid \hat{b}^T x) \mid \mathbb{I}_{\{\hat{b}^T x \in \bar{J}(\gamma)\}}, \end{aligned}$$

and, it is clear that

$$\sup_{y \in \mathbb{R}} \mid F_n(y \mid \hat{b}^T x) - F(y \mid \hat{b}^T x) \mid \mathbb{I}_{\hat{b}^T x \in J(\gamma)} \leq \sup_{y \in \mathbb{R}} \quad \sup_{z \in J(\gamma)} \mid F_n(y \mid z) - F(y \mid z) \mid$$

which converges to 0, under Assumptions **(A1)**-**(A4)**, by the use of Samanta (1989), Härdle (1990) or Gannoun (1990).

To prove that the other term converges to 0, we note that

$$\sup_{y \in \mathbb{R}} \mid F_n(y \mid \hat{b}^T x) - F(y \mid \hat{b}^T x) \mid \mathbb{I}_{\hat{b}^T x \in \bar{J}(\gamma)} \leq \mathbb{I}_{\{\hat{b}^T x \in \bar{J}(\gamma)\}}$$

which converges to 0 under **(LC)** and because $\mathbb{I}_{\{\hat{b}^T x \in \bar{J}(\gamma)\}} = 0$ for n sufficently large.

We may now tend γ to 0 to conclude that **(P1)** converges to 0.

Now let us prove that **(P2)** converges to 0.

By continuity of $F(y \mid .)$ (Assumption **(A5)**) and because $\hat{b}^T x$ converges to $b^T x$, we have that

$$F(y \mid \hat{b}^T x) \longrightarrow F(y \mid b^T x).$$

Now by continuity of $F(. \mid b^T x)$ (Assumption **(A5)**) and using Prakasa Rao (1983) (Theorem 9.2.1), we have

$$\sup_{y \in \mathbb{R}} \mid F(y \mid \hat{b}^T x) - F(y \mid b^T x) \mid \longrightarrow 0,$$

which proves that **(P2)** converges to 0. This finishes the proof of the Theorem 1.

Proof of Theorem 2

It is similar to that of Theorem 2.2 in Berlinet et al (2001).

Since $F(. \mid b^T x)$ is a distribution function with a unique quantile of order α (Assumption **(A6)**), for any $\varepsilon > 0$, there exists $\xi(\varepsilon) > 0$ defined by

$$\begin{aligned}\xi(\varepsilon) = \min \{ \; & F(q_\alpha(b^T x) + \varepsilon | b^T x) - F(q_\alpha(b^T x) \mid b^T x), \\ & F(q_\alpha(b^T x) \mid b^T x) - F(q_\alpha(b^T x) - \varepsilon \mid b^T x) \}\end{aligned}$$

such that

$$\forall \varepsilon > 0, \forall y \in \mathbb{R}, \mid q_\alpha(b^T x) - y \mid > \varepsilon \Longrightarrow \mid F(q_\alpha(b^T x) \mid b^T x) - F(y \mid b^T x) \mid > \xi(\varepsilon).$$

Now, the expansion

$$\begin{aligned} & F(q_{n,\alpha}(\hat{b}^T x) \mid b^T x) - F(q_\alpha(b^T x) \mid b^T x) \\ = \; & F(q_{n,\alpha}(\hat{b}^T x) \mid b^T x) - F_n(q_{n,\alpha}(\hat{b}^T x) \mid \hat{b}^T x) \\ + \; & \underbrace{F_n(q_{n,\alpha}(\hat{b}^T x) \mid \hat{b}^T x)}_{\alpha} - \underbrace{F(q_\alpha(b^T x) \mid b^T x)}_{\alpha} \end{aligned}$$

yields

$$\mid F(q_{n,\alpha}(\hat{b}^T x) \mid b^T x) - F(q_\alpha(b^T x) \mid b^T x) \mid \leq \sup_{y \in \mathbb{R}} \mid F(y \mid b^T x) - F_n(y \mid \hat{b}^T x) \mid .$$

We thus get

$$P(\mid q_{n,\alpha}(\hat{b}^T x) - q_\alpha(b^T x) \mid > \varepsilon) \leq P(\sup_{y \in \mathbb{R}} \mid F(y \mid b^T x) - F_n(y \mid \hat{b}^T x) \mid > \xi(\varepsilon)),$$

and this bound allows us to apply Theorem 1 to conclude that Theorem 2 holds.

REFERENCES

Aragon, Y., and Saracco, J. (1997), "Sliced Inverse Regression (SIR): An Appraisal of Small Sample Alternatives to Slicing", *Computational Statistics*, 12, 109-130.

Berlinet, A., Gannoun, A., and Matzner-Løber, E. (2001), "Asymptotic Normality of Convergent Estimates of Conditional Quantiles", *Statistics*, 35, 139-169.

Bura, E. (1987), "Dimension Reduction via Parametric Inverse Regression", *L_1-Statistical procedures and related topics*, MS Lecture Notes, 31,215-228.

Cai, Z. (2001), "Regression Quantiles for Time Series", *To appear in Econometric Theory.*

Carroll, R. J., and Li, K. C. (1992), "Measurement Error Regression with Unknown Link: Dimension Reduction and Data Visualization", *Journal of the American Statistical Association*, 87, 1040-1050.

Chaudhuri, P. (1991), "Nonparametric Estimates of Regression Quantiles and their Local Bahadur Representation", *The Annals of Statistics*, 19, 760-777.

Chen, C. H., and Li, K. C. (1998), "Can SIR Be as Popular as Multiple Linear Regression?", *Statistica Sinica*, 8, 289-316.

Cole, T. J. (1988), "Fitting Smoothed Centile Curves to Reference Data", *Journal of the Royal Statistical Society, Series A*, 151, 385-418.

Cook, R. D., and Weisberg, S. (1991), "Discussion on Li", *Journal of the American Statistical Association*, 86, 328-332.

Cook, R. D. (1994), "On the Interpretation of the Regression Plots", *Journal of the American Statistical Association*, 89,177-189.

Cook, R. D., and Nachtsheim, C. J. (1994), "Reweighting to Acheive Elliptically Contoured Covariates in Regression", *Journal of the American Statistical Association*, 89, 592-599.

Cook, R. D., and Weisberg, S. (1994), *An Introduction to Regression Graphics*, New York, Wiley.

Cook, R. D. (1996), "Graphics for Regressions with a Binary Response", *Journal of the American Statistical Association*, 91, 983-992.

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying the Regressions Through Graphics*, New York, Wiley.

Diaconis, P., and Freedman, D. (1984), "Asymptotic of graphical projection pursuit", The annals of Statistics, 12, 793-815.

Eaton, M. L. (1986), "A Characterization of Spherical Distributions, *Journal of Multivariate Analysis*, 20, 272-276.

Fan, J., Hu, T., and Truong, Y. K. (1994), "Robust Non-parametric Function Estimation", *Scandinavian Journal of Statistics*, 21, 433-446.

Ferré, L. (1997), "Dimension choice for Sliced Inverse Regression Based on Ranks",*Journal of the American Statistical Association*, 93, 132-140.

Ferré, L. (1998), "Determining the Dimension in Sliced Inverse Regression and Related Methods", *Student*, 2, 95-108.

Gannoun, A., (1990), "Estimation Non Paramétrique de la Médiane Conditionnelle, médianogramme et Méthode du Noyau", *Revue de l'Institut de Statistique de Université de Paris*, 45, 11-22.

Gannoun, A. and Saracco, J. (2000), "A Cross Validation Criterion for SIR_α and $PSIR_\alpha$ Methods in View of Prediction", *submitted.*

Gannoun, A., and Saracco, J. (2001), "An Asymptotic Theory for SIR$_\alpha$ Method", *submitted.*

Goldstein, H., and Pan, H. (1992), "Percentile Smoothing using Piecewise Polynomials, with Covariates", *Biometrics*, 48, 1057-1068.

Hall, P., and Li, K. C (1992), "On Almost Linearity of Low-Dimensional Projections from High-Dimensional Data", *The annals of Statistics*, 21, 867-889.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Healy, M. J. R., Rasbash, J., and Yang M. (1998), "Distribution-Free Estimation of Age-Related Centiles", *Annals of Human Biology*, 15, 17-22.

Hendricks, W. and Koenker, R. (1992), "Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity", *Journal of the American Statistical Association*, 99, 58-68.

Hsing, T. and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression", *The Annals of Statistics*, 20, 1040-1061.

Jones, M. C. and Hall, P. (1990), "Mean Squared Error Properties of Kernel Estimates of Regression Quantiles". *Statistics and Probability Letters*, 10, 283-289.

Kötter, T. (1996), "An asymptotic result for Sliced Inverse Regression", *Computational Statistics*, **11**, 113-136.

Kötter, T. (2000), "An Asymptotic Result for Sliced Inverse Regression", *Computational Statistics*, 11, 113-136.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction" (with discussion), *Journal of the American Statistical Association*, 86, 316-342.

Li, K. C. (1992), "On Principal Hessian Direction for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma", *Journal of the American Statistical Association*, 87, 1025-1039.

Li, K. C., Wang J. L., and Chen, C. H. (1999), "Dimension Reduction for Censored Regression Data", *The annals of Statistics*, 27, 1-23.

Magee, L., Burbidge, J. B., and Robb, A. L. (1986), "Computing Kernel-Smoothed Conditional Quantiles from Many Observations", *Journal of the American Statistical Association*, 86, 673-677.

Mint el Mouvid, M. (2000), “Sur l’Estimateur Linéaire Local de la Fonction de Répartition Conditionnelle”, Ph.D. Dissertation, Montpellier II University (France).

Prakasa Rao, B. L. S. (1983), Nonparametric Functional Estimation, *Academic Press, London.*

Poiraud-Casanova, S. (2000), “Estimation Non Paramétrique des Quantiles Conditionnels”, Ph.D. Dissertation, Toulouse I University (France).

Royston, P., and Altman, D. G. (1992), “Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling” (with discussion). *Applied Statistics*, 43, 429-467.

Samanta, T. (1989), “Non-parametric Estimation of Conditional Quantiles”, *Statistics and Probability Letters*, 7, 407-412.

Saracco, J. (1997), “An asymptotic Theory for Sliced Inverse Regression”, *Communications in Statistics - Theory and methods*, 26, 2141-2171.

Saracco, J. (2001), “Pooled Slicing Methods Versus Slicing Methods”, *To appear in Communications in Statistics - Simulations and Computations*, 30(4).

Saracco, J. (1999), “Sliced Inverse Regression Under Linear Constraints”, *Communication in Statistics - Theory and methods*, 28, 2367-2393.

Schimek, M. G. (2000), Smoothing and Regression. Approaches, Computation, and Application, *John Wiley & sons, Inc., New York.*

Schott, J. R. (1994), “Determining the Dimensionality of in Sliced Inverse Regression”, *Journal of the Americain Statistical Associationb*, 89, 141-148.

Stone, C. J. (1977), “Consistent Nonparametric Regression” (with discussion), *The Annals of Statistics*, 5, 595-645.

Stute, W. (1986), “Conditional Empirical Processes”, *The Annals of Statistics*, 14, 638-647.

Tsybakov, A. B. (1986), “Robust Reconstruction of Functions by the Local Approximation Method”, *Problems of Information Transmission*, 22, 133-146.

Yu, K. and Jones, M. C. (1998), “Local Linear Quantile Regression”, *Journal of the American Statistical Association*, 93, 228-237.

Zhu, L. X., and Fang, K. T. (1996), “ Asymptotics for kernel estimate of Sliced Inverse Regression”, *The Annals of Statistics*, 24, 1053-1068.

Covariate	AIC	EDR
AGE	312.93	0.011
C2BRAS	35.29	-0.126
PBRAS	23.00	0.438
HYGRO	17.86	-0.057
BRAS1	17.20	-0.066
TEMP	16.89	0.312

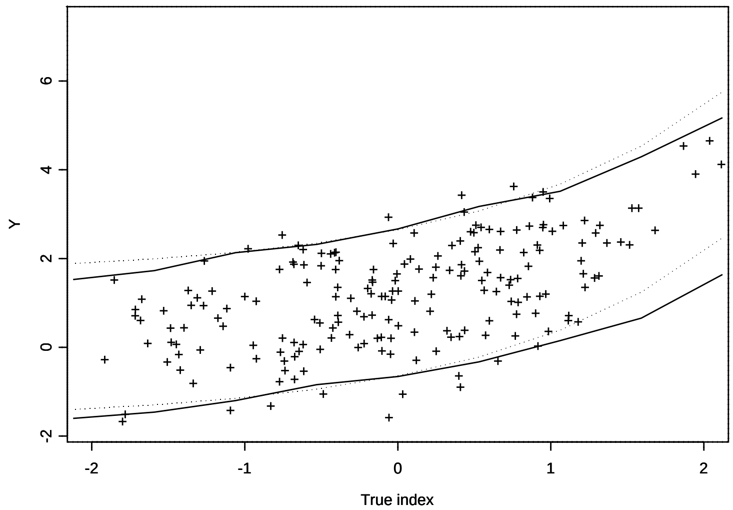
Table 1: *Results of the forward-selected linear regression model for the index concerning KBRAS. For each selected covariate (first column), we indicate its negative or positive contribution to the EDR (third column) as well as the value of the AIC criterion (second column).*

Covariate	AIC	EDR
AGE	301.85	-0.003
C2JOUE	29.56	-0.102
SJOUE	25.67	0.004
TEMP	23.51	0.439
JOUE1	20.07	-0.036
PJOUE	18.48	0.273

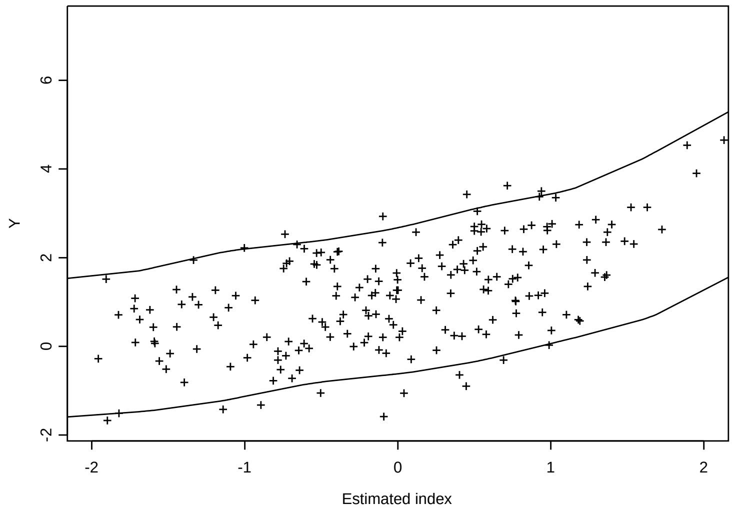
Table 2: *Results of the forward-selected linear regression model for the index concerning KJOUE.*

Covariate	AIC	EDR
AGE	314.20	-0.0001
C2FRONT	13.88	0.0954

Table 3: *Results of the forward-selected linear regression model for the index concerning KFRONT.*

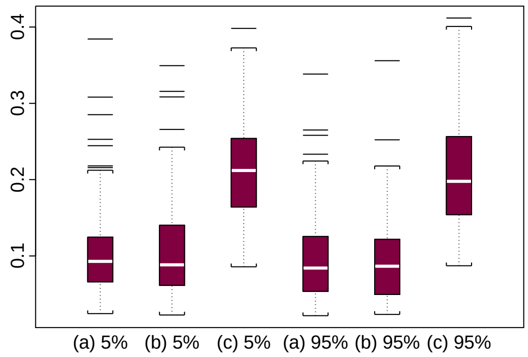


1. True reference curves (dotted line) and estimator **(b)** (solid line)

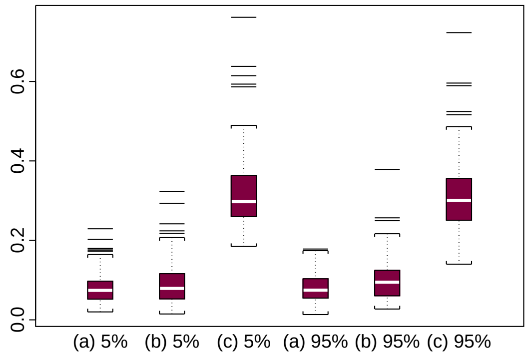


2. Estimator **(a)**

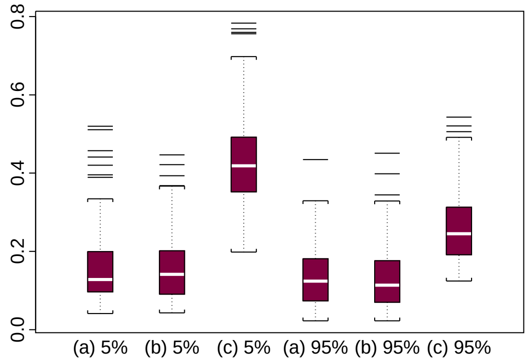
Figure 1: *Comparison of the kernel estimated and true 90%-reference curves on the model* **(M3)**. *Estimator* **(b)** *of the reference curves is obtained using the true index whereas estimator* **(a)** *is obtained with the SIR estimated index. All the resulting curves are visually similar to the theoretical ones.*



1. Model (**M1**)

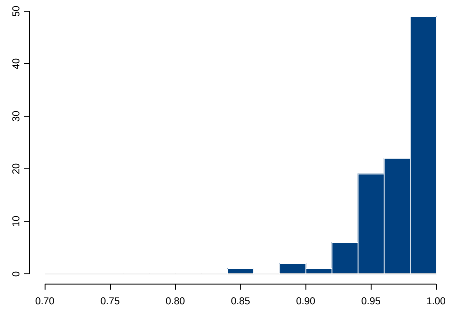


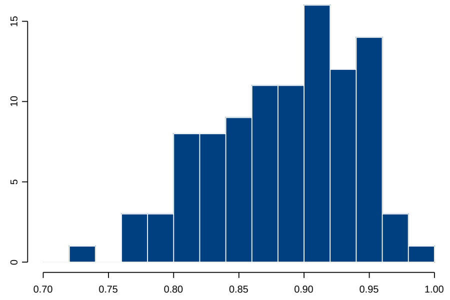
2. Model (**M2**)



3. Model (**M3**)

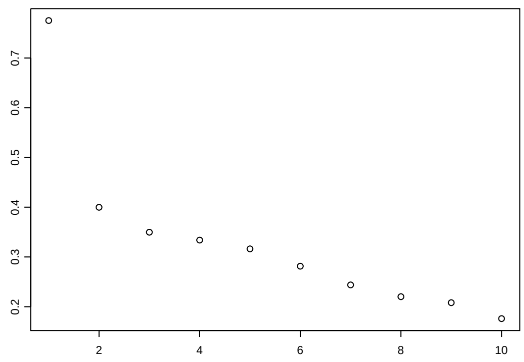
Figure 2: *Boxplots obtained on the three different models with the three different estimates. Errors obtained with the estimators* (**a**) *and* (**b**) *are similarly distributed. Moreover, they are smaller than those obtained without dimension-reduction (estimator* (**c**)*) whatever the nature of the model is.*



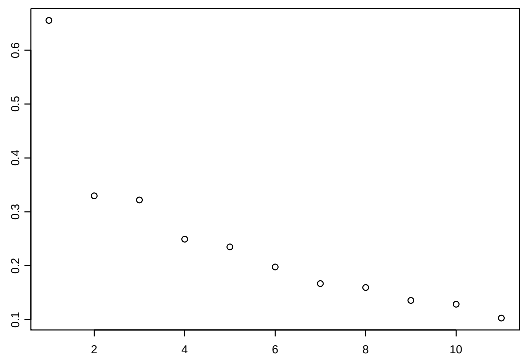


1. Model **(M1)**　　　　2. Model **(M2)**

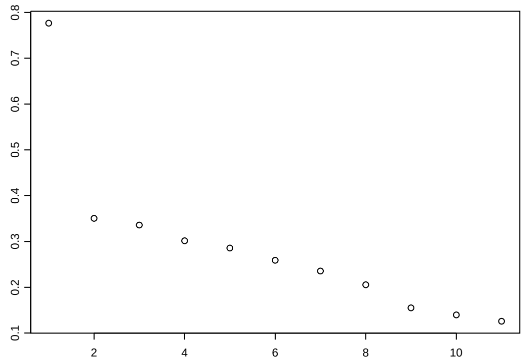
Figure 3: *Histograms of the square cosines of the angle between the true direction* β *and the SIR estimated direction* $\widehat{b}$*. These histograms explain the similarity of the boxplots obtained with estimators* **(a)** *and* **(b)** *in Figure 2. Because the distribution of square cosines is concentrated in the neighbourhood of 1, the direction is well estimated whatever the number of covariates is.*



1. KBRAS

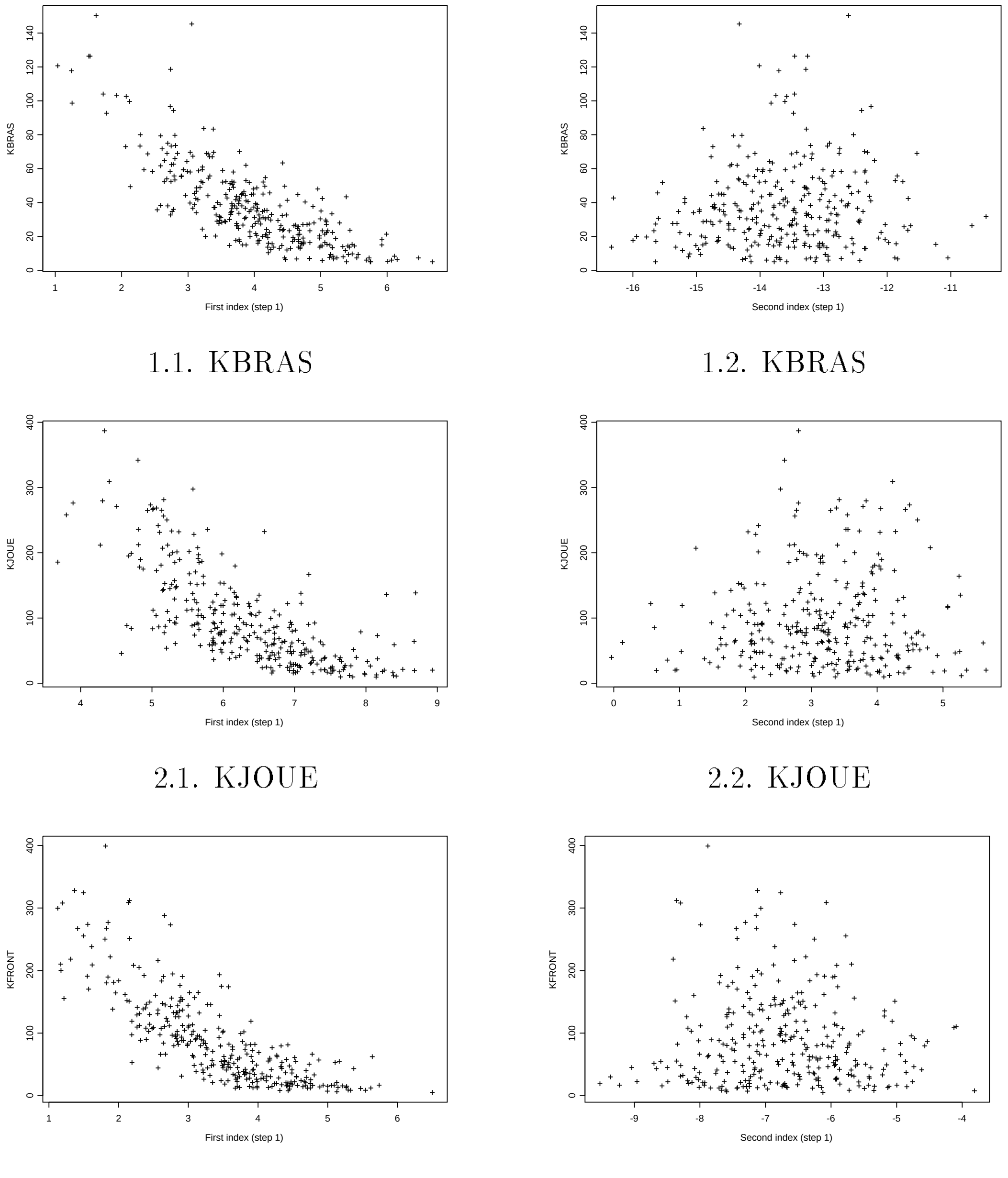


2. KJOUE



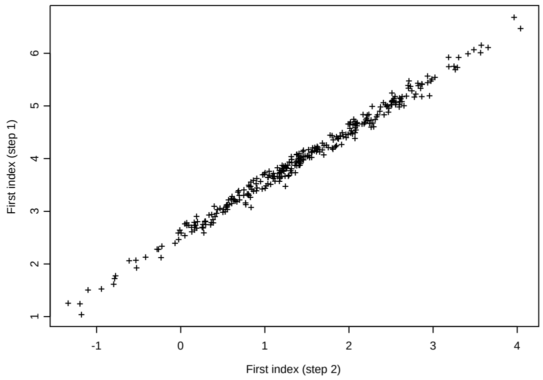
3. KFRONT

Figure 4: *Scree of eigenvalues for each response variable KBRAS, KJOUE and KFRONT. The break in size of eigenvalues between the first and second SIR directions suggests to choose a one dimensional dimension-reduction subspace.*

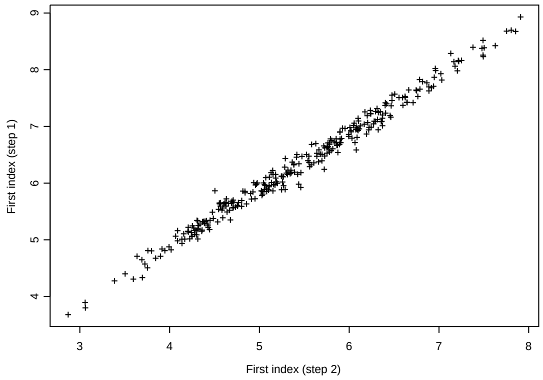


1.1. KBRAS
1.2. KBRAS
2.1. KJOUE
2.2. KJOUE
3.1. KFRONT
3.2. KFRONT

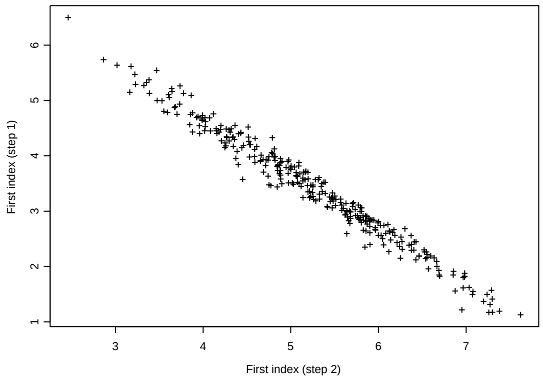
Figure 5: *Scatterplots of the response variables versus the SIR indices computed with all the covariates (step 1). Each variable KBRAS, KJOUE and KFRONT is plotted versus the first (resp. the second) corresponding index in the left (resp. the right) hand side column. As opposed to the second index, the first index reveals a strong structure of the scatterplot. This confirms the empirical choice of only one EDR direction.*



1. KBRAS ($R^2 = 0.990$)

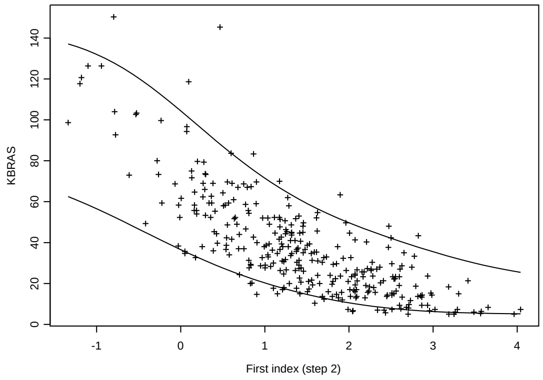


2. KJOUE ($R^2 = 0.984$)

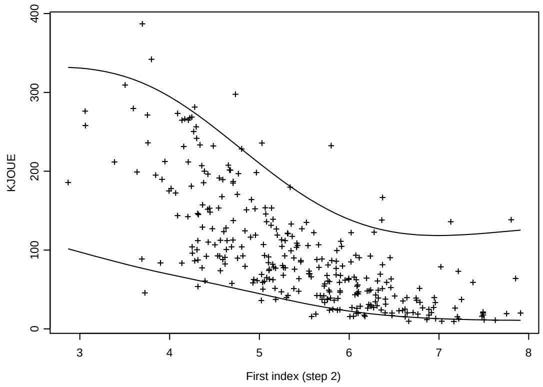


3. KFRONT ($R^2 = 0.975$)

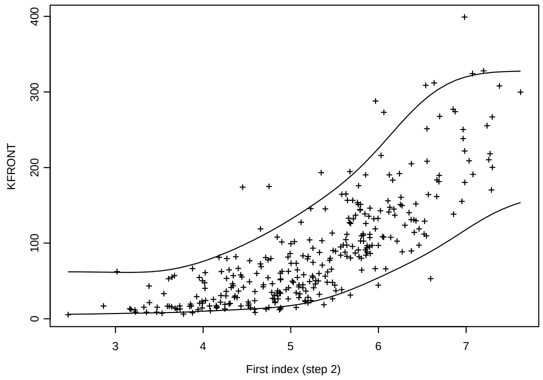
Figure 6: *Graphical validation of the covariate selection step. For each response variable KBRAS, KJOUE and KFRONT, the indices computed at the first step with all covariates are plotted versus the indices computed at the second step with the selected covariates. Since the plots reveal a linear structure (R^2 close to 1), there is no loss of information working with only the subset of selected covariates.*



1. KBRAS



2. KJOUE



3. KFRONT

Figure 7: *Estimated 90%-reference curves for each response variable KBRAS, KJOUE and KFRONT. These curves are estimated using the indices computed with the selected covariates (step 2).*