Sliced Inverse Regression In Reference Curves Estimation

Ali Gannoun, Stéphane Girard, Christiane Guinot, Jerôme Saracco

Sliced Inverse Regression In Reference Curves Estimation

Ali Gannoun[1,2], Stéphane Girard[1,3], Christiane Guinot[4], Jérôme Saracco[1]

Abstract

In order to obtain reference curves for data sets when the covariate is multidimensional, we propose in this paper a new procedure 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.

Keywords: conditional quantiles; dimension reduction; kernel estimation; semiparametric method; reference curves; sliced inverse regression (SIR).

[1] Laboratoire de Probabilités et Statistique, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France. e-mail: {gannoun, saracco}@stat.math.univ-montp2.fr

[2] Statistical Genetics and Bioinformatics Unit, National Human Genome Center, Howard University, Washington D.C. 20059, USA. e-mail: agannoun@howard.edu

[3] SMS/LMC, Université Grenoble I, 38041 Grenoble Cedex 9, France. e-mail: Stephane.Girard@imag.fr

[4] CE.R.I.E.S, 20, Rue Victor Noir, 92 521 Neuilly sur Seine Cedex, France. e-mail: christiane.guinot@ceries-lab.com

1 Introduction

The reference intervals are an important tool in clinical and medical practice. They provide a guideline to clinicians or clinical chemists seeking to interpret a measurement obtained from a new patient. Many experiments, in particular in 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 interval is the range of values bounded by a pair of quantiles (the reference limits), such as the 5th and 95th centiles for a 90% reference interval, 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, circumferences and skinfold) against an appropriate covariate (age, time). For details, the readers may refer, for example, to the work of Healy et al. (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 et al. (1991), Hendricks and Koenker (1992).

Mathematically speaking, let $\alpha \in (0, 1)$. The αth-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 $\alpha > 0.5$, the $100 \times (2\alpha - 1)\%$ reference curves are defined, when x varies, by

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

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

When n observations $\{(x_i, y_i)\}_{i=1}^n$ of (X, Y) are available, basic parametric or nonparametric estimation methods can be considered. The choice of the method typically depends on the sample size. When n is small, parametric assumptions are usually added to reduce the number of parameters that need to be estimated. For instance, $F(y|x)$ is often assumed to be Gaussian, so that an estimate of the α-th quantile $q_\alpha(x)$ is $\hat{\mu}(x)+N_\alpha\hat{\sigma}(x)$ where $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ are estimates of $\mu(x) = E[Y|X = x]$ and $\sigma^2(x) = Var(Y|X = x)$ respectively, and where N_α denotes the αth-quantile of the standard normal distribution $\mathcal{N}(0, 1)$. For large samples, one nonparametric approach proposed by Goldstein and Pan (1992) is to group data along the X-axis into bins, estimate the αth-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 scatterplot.

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 (1964) and Watson (1964) estimator, $\hat{F}_{NW}(y|x)$, can be affected to $F(y|x)$. If we write $Y^\star = \mathbb{I}_{\{Y\leq y\}}$ where $\mathbb{I}$ denotes the indicator function, 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 type of 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 esti-

mation methods, the reader may also refer to 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 (2002).

Although, theoretically, the extension of conditional quantiles to higher dimension p of the covariate (denoted by $\mathbf{X}$ in this multidimensional context) is obvious, its practical success, while depending on the number of observations, suffers from the so-called *curse of dimensionality*. The spareness of high dimensional data is a general problem in nonparametric estimation. Further, because reference curves are, in this case, a pair of p-dimensional hyper-surfaces, their visual display is rendered difficult making it less directly useful for exploratory purposes (unlike the one-dimensional case). On the other hand, when $p \leq 2$, two- and three-dimensional plots can provide useful information on such changes, as discussed by Cook and Weisberg (1994). Though it is now easy to view three-dimensional plots with widely available software, 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. Note also that, from theoretical point of view, the convergence rate of non parametric estimators depends on the dimension p of predictors space. It decreases when p increases. So, reducing the dimension p increases automatically the rate of convergence. For details and deep information on this topic, one can see Stone (1982) and references therein.

Motivated by this, the key is then to reduce the dimension of the predictor vector $\mathbf{X}$ without loss of information on the conditional distribution of Y given $\mathbf{X}$ and without requiring a prespecified parametric model. Sufficient dimension-reduction leads naturally to the idea of a sufficient summary plot that contains all information on the regression available from the sample. Methods to reduce the dimension exist in the literature. Stone (1985, 1986) used additive regression models to cope with curse of dimensionality in nonparametric function estimation. Chaudhuri (1991) used this technique in order to estimate conditional quantiles.

In this paper, we focus on linear projection method of reducing the dimensionality of the covariates in order to construct a more efficient estimator of conditional quantiles and consequently reference curves. The specific dimension reduction method used

is based on Li's well known Sliced inverse regression (SIR), see Li (1991). This method is used as pre-step of the main analysis of the data. It is fairly robust, especially against some outliers in the regressor observations. The rest of the paper is organized as follows. In Section 2, we present the dimension reduction context and the SIR method. We derive in Section 3 a semiparametric estimator of conditional quantiles based on this 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. Finally, 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 an associated basis by the SIR method. We also emphasize the consequence of the dimension-reduction on the definition of the conditional quantile.

2.1 Dimension reduction subspaces

A convenient data reduction formulation is to assume there exists a $p \times r$ matrix $(r \times p)$ B such that

$$F(y|\mathbf{x}) = F(y|B^T\mathbf{x}), \tag{5}$$

where $F(.|.)$ is the conditional distribution function of the response Y given the second argument. Such matrix always exists because (5) is trivially true when $B = I_p$ the $p \times p$ identity matrix. The assumption (5) implies that the $p \times 1$ predictor vector $\mathbf{X}$ can be replaced by the $r \times 1$ predictor $B^T\mathbf{X}$) without loss of regression information. Most importantly, if $r < p$, then sufficient reduction in the dimension of the regression is achieved. The linear subspace $S(B)$ spanned by the columns of B is a dimension reduction subspace, see Li (1991), and its dimension denotes the number of linear components of $\mathbf{X}$ needed to model Y. When (5) holds, then it also

holds with B replaced by any matrix whose columns form a basis for $S(B)$. Clearly, knowledge of the smallest dimension reduction subspace would provide the most parcimonious characterization of Y given $\mathbf{X}$, as it provides the greatest dimension reduction in the predictor vector. Let $S_{Y|\mathbf{X}}$ denote the unique smallest dimension reduction subspace, referred to the central dimension reduction subspace, see Cook (1994,1996,1998). Let $d = dim(S_{Y|\mathbf{X}})$, $d \leq r$, the dimension of this subspace, and β the $p \times d$ matrix whose columns form a basis of $S_{Y|\mathbf{X}}$. Then, from (5), we have

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

2.2 Characterization

Let $S_{E(\mathbf{X}|Y)}$ denote the subspace spanned by $\{E(\mathbf{X}|Y) - E(\mathbf{X}) : \ Y \in \Omega_Y\}$ where $\Omega_Y \in \mathbb{R}$ is the sample space of Y. Given (5), assume that the marginal distribution of the predictors $\mathbf{X}$ satisfies the following *linearity condition*:

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

Let Σ be the variance matrix of $\mathbf{X}$, supposed positive-definite. Under **(LC)**, Li (1991) showed that the centered inverse regression curve $E(\mathbf{X}|Y) - E(\mathbf{X}) \in S(\Sigma\beta)$. Thus,

$$S_{E(\mathbf{X}|Y)} \subseteq S(\Sigma\beta) = \Sigma S_{Y|\mathbf{X}}. \tag{6}$$

Let $\mathbf{Z}$ denote the standardized version of the predictor $\mathbf{X}$ defined by $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - E(\mathbf{X}))$, where $\Sigma^{-1/2}$ is the unique symmetric positive-definite square root of Σ^{-1}. Cook (1998) showed that there is no loss of generality working on the Z-scale, because any basis for $S_{Y|\mathbf{Z}}$ can be back-transformed to a basis for $S_{Y|\mathbf{X}}$ since $S_{Y|\mathbf{X}} = \Sigma^{-1/2} S_{Y|\mathbf{Z}}$. Therefore, in view of (6), we have

$$S_{E(\mathbf{Z}|Y)} \subseteq S(\eta) = S_{Y|\mathbf{Z}}, \tag{7}$$

where $\eta = \Sigma^{1/2}\beta$. This does not guarantee equality between $S_{E(\mathbf{Z}|Y)}$ and $S_{Y|\mathbf{Z}}$ and, thus, inference about $S_{E(\mathbf{Z}|Y)}$ possibly covers only part of $S_{Y|\mathbf{Z}}$. Moreover, it is clear that

$$S\{Var(E(\mathbf{Z}|Y))\} = S_{E(\mathbf{Z}|Y)}, \tag{8}$$

except on a set of measure zero, see Cook (1998).

Using results (7) and (8), the estimation of the inverse regression curve $E(\mathbf{Z}|Y)$ serves to estimate the central dimension-reduction subspace by estimating the matrix $Var(E(\mathbf{Z}|Y))$. 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 $S_{Y|\mathbf{Z}}$.

Remark. The linearity condition **(LC)** 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 $\mathbf{X}$, see Eaton (1986). This condition holds for instance when $\mathbf{X}$ 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 for instance, 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.

2.3 Sliced 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 $\mathcal{S}_1, \ldots, \mathcal{S}_H$. Then, the p components of $\mathbf{Z}$ 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 $\mathbf{Z}$. Let M denote the matrix $Var(E(\mathbf{Z}|\widetilde{Y}))$. Using (7) and (8), it is clear that

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

The last inclusion in (9) holds because $\widetilde{Y}$ is a function of Y, which implies that $S_{Y|\mathbf{Z}}$ is a dimension-reduction subspace for the regression of $\widetilde{Y}$ on $\mathbf{Z}$. Note that, using the slicing $\mathcal{S}_1, \ldots, \mathcal{S}_H$, M is written

$$M = \sum_{h=1}^{H} p_h m_h m_h^T, \tag{10}$$

where $p_h = \mathrm{P}(Y \in \mathcal{S}_h)$ and $m_h = E[\mathbf{Z}|Y \in \mathcal{S}_h]$. Let $s_1 \geq \ldots \geq s_p$ denote the singular values of M, and $u_1, \ldots, u_p$ denote the corresponding left singular vectors. Assuming that $d = \dim(S(M))$, $S(M) = S(u_1, \ldots, u_d)$. Transforming back to the $\mathbf{X}$ scale, $\{b_k = \Sigma^{-1/2} u_k\}_{k=1}^{d}$ form a basis of $S(\beta)$. Following SIR vocabulary, the dimension reduction subspace $S(\beta)$ is called the *effective dimension-reduction* (EDR) space, and the vectors b_k are named EDR directions. As we focus our dimension reduction approach on the SIR method, we will use this terminology from now on.

Pathological case for SIR. Li (1991, 1992) and Cook and Weisberg (1991) mention that SIR can miss EDR directions even if the **(LC)** condition is valid. The reason is that it is "blind" for symmetric dependencies. In this case, the inverse regression curve does not contain any information about the EDR directions. For handling such cases, in order to recover the EDR directions, a natural extension is to consider higher moments of the conditional distribution of $\mathbf{X}$ given Y. Various methods based on second moments for estimating the EDR space have been developped: for example, SIR-II and SIR$_\alpha$ (Li, 1991), SAVE (Cook and Weisberg, 1991), the pooled slicing version of these methods and the choice of α (Saracco, 2001, and Gannoun and Saracco, 2000, 2002) and Principal Hessian directions (Li, 1992). These methods may help, at least for completeness, when SIR fails to capture all the 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, see for example, Hsing and Carroll (1992), Kötter (1996), Zhu and Fang (1996), Saracco (1997, 1999) among others. The estimation of the dimension of the EDR space has been studied for instance by Schott (1994) and Ferré (1998). Carroll and Li (1992) used 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) used a multivariate linear model for the inverse regression curve. The case of censored regression data is considered by Li et al. (1999).

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 $\mathbf{x}_i$ denote the corresponding $p \times 1$ vector of observed covariate values, $i = 1, \ldots, n$. The $1 \times (p+1)$ data $(\mathbf{x}_i^T, y_i)$ are assumed to be independent and identically distributed observations from the multivariate random vector $(\mathbf{X}^T, Y)$ with finite moment.

3.1 SIR estimation step

Let $\overline{\mathbf{x}}$ and $\widehat{\Sigma}$ be the sample mean and the sample variance matrix of the $\mathbf{x}_i$'s. Let $\widehat{\mathbf{z}}_i$ be the estimated standardized predictor defined by $\widehat{\mathbf{z}}_i = \widehat{\Sigma}^{-1/2}(\mathbf{x}_i - \overline{\mathbf{x}}),\ i = 1, \ldots, n$. Then the SIR estimate of M defined in (10) 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 $\widehat{\mathbf{z}}$ within slice h. Let $\hat{s}_1 \geq \ldots \geq \hat{s}_p$ denote the singular values of $\widehat{M}$, and $\hat{u}_1, \ldots, \hat{u}_p$ denote the corresponding left singular vectors.

Assuming that the dimension d of $S(M)$ is known, $S\left(\widehat{M}\right) = S(\hat{u}_1, \ldots, \hat{u}_d)$ 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 for example, Li (1991), 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, $\{\hat{b}_k = \widehat{\Sigma}^{-1/2}\hat{u}_k\}_{k=1}^{d}$ forms 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.

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

$$F(y|\mathbf{x}) = F(y|\beta^T\mathbf{x}) = F(y|b^T\mathbf{x}), \tag{11}$$

and

$$q_\alpha(\mathbf{x}) = q_\alpha(\beta^T\mathbf{x}) = q_\alpha(b^T\mathbf{x}).$$

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

$$\{\widehat{v}_i = \hat{b}^T\mathbf{x}_i\}_{i=1}^n \quad \text{and} \quad \hat{v} = \hat{b}^T\mathbf{x}.$$

Following (3), from the data $\{(y_i, \hat{v}_i)\}_{i=1}^n$, we define a kernel estimator of $F(y|\mathbf{x})$ by

$$F_n\left(y\left|\hat{b}^T\mathbf{x}\right.\right) = F_n(y|\hat{v}) = \sum_{i=1}^n K\{(\hat{v} - \widehat{v}_i)/h_n\}\mathbb{I}_{\{y_i \le y\}} \Big/ \sum_{i=1}^n K\{(\hat{v} - \widehat{v}_i)/h_n\}\ . \tag{12}$$

Then, as in (4), we derive from (12) an estimator of $q_\alpha(\mathbf{x})$ by

$$q_{n,\alpha}\left(\hat{b}^T\mathbf{x}\right) = q_{n,\alpha}(\hat{v}) = F_n^{-1}(\alpha \mid \hat{v}). \tag{13}$$

As a consequence of the above result, for $\alpha > 0.5$, the corresponding estimated $100 \times (2\alpha - 1)\%$ reference curves are given, as $\mathbf{x}$ varies, by

$$I_{n,\alpha}(\mathbf{x}) = [q_{n,1-\alpha}(\hat{v}), q_{n,\alpha}(\hat{v})] = [q_{n,1-\alpha}(\hat{b}^T\mathbf{x}), q_{n,\alpha}(\hat{b}^T\mathbf{x})]. \tag{14}$$

Remark. The above definitions have been presented in the context of single index. A natural extension is to consider the general multiple indices ($d > 1$) and to work with $\{\hat{b}_j = \widehat{\Sigma}^{-1/2}\hat{u}_j\}_{j=1}^d$ and $\{\hat{v}_i = (\hat{b}_1^T\mathbf{x}_i, \ldots, \hat{b}_d^T\mathbf{x}_i)\}_{i=1}^n$. Then we follow the classical multi-kernel estimation to get $q_{n,\alpha}(\hat{b}_1^T\mathbf{x}, \ldots, \hat{b}_d^T\mathbf{x})$ as in (13): for instance, the corresponding kernel K used in (12) can be the d-dimensional normal density and the bandwidth h_n can be chosen by a cross-validation technique, see for instance Schimek (2000).

4 Asymptotic properties

In this section, we study the consistency of $q_{n,\alpha}(\hat{b}^T\mathbf{x})$. The additional assumptions under which the results in this paper are derived are gathered together below for easy reference.

Assumptions

(A1) The random vectors $(\mathbf{X}_i, Y_i), i \geq 1$, 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 $\int v^2K(v)dv < \infty$.

(A3) The sequence of bandwidth h_n tends to zero such that $nh_n/\ln n \longrightarrow \infty$.

(A4) The variable $\mathbf{X}$ admits a continuous marginal density.

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

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

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

With this in mind, we have the following results:

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

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

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

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

Comments on Theorem 2. Theorem 2 gives the weak convergence of the estimator. This convergence is enough to make application. To get the uniform convergence, one can suppose that $\mathbf{X}$ is defined on compact set of $\mathbb{R}^p$, and proceed by the same manner as in Berlinet et al. (2001).

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 αth-conditional quantile at point $\mathbf{x}$:

- **(a)** $q_{n,\alpha}^{(a)}(\mathbf{x}) \stackrel{\text{def}}{=} q_{n,\alpha}(\hat{b}^T\mathbf{x})$, the estimator defined in (13). 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)** $q_{n,\alpha}^{(b)}(\mathbf{x}) \stackrel{\text{def}}{=} q_{n,\alpha}(\beta^T\mathbf{x})$ 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)** $q_{n,\alpha}^{(c)}(\mathbf{x}) \stackrel{\text{def}}{=} q_{n,\alpha}(\mathbf{x})$ 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)**.

First, we consider the following regression model

(M1) $Y = f(\beta^T \mathbf{X}) + \varepsilon,$

with the nonlinear link function $f(t) = 1+\exp(2t/3)$. The random variable $\mathbf{X}$ 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 $\mathbf{X}$. The motivation for introducing this model is to investigate the behaviour of the estimation methods when the dimension increases: $p \in \{3, 5, \ldots, 13\}$. Introducing $c = [1, \ldots, 1]$ the line vector of length $(p-1)/2$, the vector β is chosen such as $\beta^T = (p-1)^{-1/2}[c, -c, 0]$. Let us note that the true conditional αth-quantile can be written: $q_\alpha(\mathbf{x}) = f(\beta^T \mathbf{x}) + N_\alpha$, where N_α is the αth-quantile of the standard normal distribution.

Second, we introduce the mixture model

(M2) $Y = (1-\theta)g(\beta^T \mathbf{X}) + \theta h(\mathbf{X}) + \varepsilon,$

in dimension $p = 3$ and with linear link function $f(t) = 1 + 2t/3$. The parameter θ tunes the importance of the contamination of the regression model by the function $h(x, y, z) = 2xyz/3$. The motivation for introducing this model is to investigate the robustness of the estimation methods when the contamination increases: $\theta \in \{0, 0.2, \ldots, 1\}$, and thus when the linearity condition (**LC**) is less and less satisfied. Similarly to the previous situation, $\beta^T = 2^{-1/2}[1, -1, 0]$, $\mathbf{X}$ follows the standard multinormal distribution $\mathcal{N}_3(0, I_3)$ and where ε is normally distributed $\varepsilon \sim \mathcal{N}(0, 1)$ and is independent from $\mathbf{X}$. Note that $\text{var}(g(\beta^T \mathbf{X})) = \text{var}(h(\mathbf{X})) = 1$. The true conditional αth-quantile can be written: $q_\alpha(\mathbf{x}) = (1-\theta)g(\beta^T \mathbf{x}) + \theta h(\mathbf{x}) + N_\alpha$.

Note that both link functions f and g are chosen such that there is no symmetric dependence (pathological model), thus SIR can recover the (EDR) direction β.

Our goal is to compare successively the three estimators **(a)**, **(b)** and **(c)** to the true quantile in the situations **(M1)** and **(M2)**. 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

$$E_{n,\alpha}^{(\Theta)} = \frac{1}{125} \sum_{\ell=1}^{125} \left(q_{n,\alpha}^{(\Theta)}(z_\ell) - q_\alpha(z_\ell) \right)^2, \quad \text{where } \Theta \in \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}.$$

5.4 Results

– All the experiments concerning the behaviour of the estimation methods with respect to the dimension have been conducted on the model **(M1)** and are summarized in Figure 1. First of all, an example of graphical representation of the estimated quantiles for the **(M1)** model is presented. The superimposition in Figure 1.1 of the theoretical quantile 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 $\Theta \in \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ and $\alpha \in \{0.05, 0.95\}$. The mean values (over the $N = 100$) samples are plotted as a function of the dimension p on Figure 1.3 (the plot is represented as a continuous graph for the sake of visual clarity). It appears that there is no difference between the means of $E_{n,\alpha}^{(\mathbf{a})}$ and $E_{n,\alpha}^{(\mathbf{b})}$. The estimation of the direction β by $\widehat{b}$ has no significant consequence on the accuracy of the estimation of the reference curves whatever the dimension can be. At the opposite, the mean value of $E_{n,\alpha}^{(\mathbf{c})}$ increases with the dimension. The corresponding boxplots are represented on Figures 1.4–1.6. They confirm that there is no difference between the distribution of $E_{n,\alpha}^{(\mathbf{a})}$ and $E_{n,\alpha}^{(\mathbf{b})}$ for $d \in \{3, 9, 13\}$. 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. In this case, the curse of dimensionality becomes an essential limitation to the use of estimator **(c)**, and thus estimator **(a)** is particularly useful in such a situation.

– The experiments concerning the robustness of the estimation methods with respect to a contamination have been conducted on model **(M2)** and are reported on Figure 2. It can be observed on Figure 2.1 that estimators **(a)** and **(b)** involving a dimension-reduction step outperform estimator **(c)** provided the contamination ratio is less that 40%. This result is due to the robustness of the SIR method which yields correct estimations of the dimension-reduction direction for $\theta \leq 0.4$, see the squared cosines boxplots on Figure 2.2 and the first eigenvalue percent boxplots on Figure 2.3. Other experiments, which are not reported there, showed that the SIR estimate of the dimension-reduction direction is more robust that other ones, such as the MLR (Maximum Likelihood Regression). Back to the estimated quantiles, the error boxplots presented in Figures 2.3–2.6 illustrate that estimator **(a)** remains acceptable provided that $\theta \leq 0.6$. These experiments indicate that estimator **(a)** may be very useful for estimating conditional quantiles in high dimensions, even though the linearity condition **(LC)** is violated.

6 Application to real data

When studying biophysical skin properties of healthy women, knowledge about the reference "curves" of certain parameters is lacking. Information concerning these parameters refers to only few studies. The aim is to establish 90%-reference "curves" for some of the 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 apparently healthy skin (i.e. without any sign of ongoing skin disease or general disease with proven cutaneous manifestations), and living in the Ile de France (in around Paris) 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 were made on each area. In this paper, we will only exhibit the results concerning the forearm area. The results corresponding to the two facial areas are described in Gannoun et al. (2001).

In this study, we are interested by the estimation of the 90%-reference "curves" for the variable of interest KBRAS which is the conductance of the skin, using the corresponding set of covariates described below. 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. The other available covariates included are some biophysical properties of the skin: the skin temperature denoted by TBRAS, the transepidermal water loss denoted by BRAS1, the skin pH denoted by PBRAS, the skin hydration by capacitance denoted by C2BRAS, 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) measured by the variables ABRAS, BBRAS and LBRAS.

6.2 Procedure

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

– **Step 1.** We apply the SIR method using the variable of interest Y and the covariates of $\mathcal{X}$. From the eigenvalues scree plot, we determine the number $\hat{d}$ 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 $\hat{d}$ is then the number of the eigenvalues located before this jump. Note that if no jump is detected, no dimension reduction is possible. The eigenvalues scree plot approach used here is a useful explonatory tool in determining d. Of course testing procedure could be also used to identify d, see for instance, Schott (1994) or Ferré (1998). For simplicity of notation, we continue to write d for $\hat{d}$ in the following.

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

  We visualize the structure of the "reduced" data $\left\{(y_i, \hat{b}_1^T\mathbf{x}_i, \ldots, \hat{b}_d^T\mathbf{x}_i)\right\}_{i=1}^n$.
– **Step 2.** The aim here is to "simplify" the indices $\hat{b}_k^T\mathbf{x}$ in order to obtain an simpler interpretation. To this end, for each index, we make a forward-selected linear regression model of $\hat{b}_k^T\mathbf{x}$ on the covariates of $\mathcal{X}$ (based on the AIC criterion). 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 $\widetilde{b}_1, \ldots, \widetilde{b}_d$. Finally, we graphically check that each plot $\left\{(\hat{b}_k^T\mathbf{x}_i, \widetilde{b}_k^T\mathbf{x}_i)\right\}_{i=1}^n$ has a linear structure.
– **Step 3.** We are now able to estimate the reference "curves" (which are hypersurfaces when $d > 1$) on the sample $\left\{(y_i, \widetilde{b}_1^T\mathbf{x}_i, \ldots, \widetilde{b}_d^T\mathbf{x}_i)\right\}_{i=1}^n$, 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 $\mathcal{N}(0, I_d)$ and the bandwidth is selected by cross validation. The estimation of the reference curves are obtained using C codes. If $d = 1$, each reference curve is evaluated on 50 points equidistributed on the range of the estimated index, otherwise the reference hypersurfaces are evaluated on an appropriate grid depending on the dimension d.

We now apply the procedure developped in the previous subsection in order to get the 90%-reference "curves" for the variable KBRAS. Note that, in this kind of biophysical study, the presence of the covariate AGE is imposed in the model by the clinician. Following **Step 1** of the procedure, from the eigenvalues scree plot (see Figure 3), we select the dimension $d = 1$. Testing procedure could also be used in order to identify d. With graphic tools, we clearly observe in Figure 4 a structure between the first estimated index and KBRAS. Note that no structure has been detected with the second estimated index. These observations confirm the exploratory 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 5 confirms that the index based on the remaining selected covariates is very similar to the one based on the whole covariates. In **Step 3**, we construct the 90%-reference curves for KBRAS using this estimated index, see Figure 6. The results of the analysis on the forearm show that, apart from age (imposed in the index), five covariates enter in the model. As expected, two of these covariates represent the environmental conditions of the measurements (temperature and relative humidity). The three other covariates are directly clinically-related with skin hydration: skin pH, capacitance and transepidermal water loss. Altogether the clinicians consider that the results provided by our approach do enable the construction of reference 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 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, and to Pr. Simon Thacker from Rochambau School whose remarks led to an important improvement of the presentation.

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 $\mathbf{x} \in \mathbb{R}^p$ we have $\hat{b}^T\mathbf{x}$ converges to $b^T\mathbf{x}$. Let us also recall that $F(y \mid \mathbf{x}) = F(y \mid b^T\mathbf{x})$.
Now $\forall y \in \mathbb{R}$,

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

Consequently,

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

Let us first show that **(P1)** converges to 0. Let $\gamma > 0$ and $J(\gamma) := [b^T\mathbf{x} - \gamma, b^T\mathbf{x} + \gamma]$, then

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

where $\bar{J}(\gamma)$ denotes the complement of $J(\gamma)$. Moreover, it is clear that

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

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

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

which converges to 0 under **(LC)** and because $\mathbb{I}_{\{\hat{b}^T\mathbf{x}\in \bar{J}(\gamma)\}} = 0$ for n which is 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 $\hat{b}^T\mathbf{x}$ converges to $b^T\mathbf{x}$, we have that

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

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

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

which proves that **(P2)** converges to 0. This concludes 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 $F(. \mid b^T\mathbf{x})$ is a distribution function with a unique quantile of order α (Assumption **(A6)**), for any $\varepsilon > 0$, there exists $\xi(\varepsilon) > 0$ defined by

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

such that

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

Now, the expansion

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

yields $\mid F(q_{n,\alpha}(\hat{b}^T\mathbf{x}) \mid b^T\mathbf{x}) - F(q_\alpha(b^T\mathbf{x}) \mid b^T\mathbf{x}) \mid \leq \sup_{y\in\mathbb{R}} \mid F(y \mid b^T\mathbf{x}) - F_n(y \mid \hat{b}^T\mathbf{x}) \mid$. We thus get $P(\mid q_{n,\alpha}(\hat{b}^T\mathbf{x}) - q_\alpha(b^T\mathbf{x}) \mid > \varepsilon) \leq P(\sup_{y\in\mathbb{R}} \mid F(y \mid b^T\mathbf{x}) - F_n(y \mid \hat{b}^T\mathbf{x}) \mid > \xi(\varepsilon))$, 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. (1997), "Dimension Reduction via Parametric Inverse Regression", L_1-*Statistical procedures and related topics*, MS Lecture Notes, 31,215-228.

Cai, Z. (2002), "Regression Quantiles for Time Series", *Econometric Theory*, 18, 169-192.

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.

Chaudhuri, P. (1991), "Global Nonparametric Estimation of Condtional Quantile Functions and their Derivative", *Journal of Multivariate Analysis*, 39, 246-269.

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. (1998), "Determining the Dimension in Sliced Inverse Regression and Related Methods", *Journal of the American Statistical Association*, 93, 132-140.

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., Girard, S., Guinot, C. and Saracco, J. (2001), "Dimension-reduction in reference curves estimation", *Technical Report ENSAM-INRA-UMII*, 01-06.

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. (2002), "An Asymptotic Theory for SIR_α Method", *To appear in Statistica Sinica*.

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

Hall, P., and Li, K. C (1993), "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. (1988), "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.

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. (1991), "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).

Nadaraya, E. A. (1964) "On Estimating Regression", *Theory of Probability and Its Applications*, 9, 141-142.

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

Saracco, J. (2001), "Pooled Slicing Methods Versus Slicing Methods", *Communications in Statistics - Simulations and Computations*, 30, 499-511.

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

Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression", *Journal of the American Statistical Association*, 89, 141-148.

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

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression", *The Annals of Statistics*, 10, 1040-1053.

Stone, C. J. (1985), "Additive Regression and other Nonparametric Models", *The Annals of Statistics*, 13, 689-705.

Stone, C. J. (1986), "The Dimensionality Reduction Principle for Generalized Additive Models", *The Annals of Statistics*, 14, 590-606.

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.

Watson, G. S. (1964) "Smooth regression analysis", *Sankhya, Series A*, 26, 359-372.

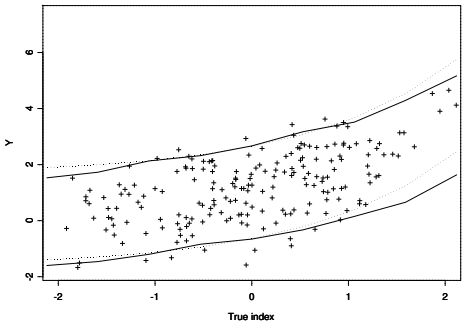
Yu, K. and Jones, M. C. (1998), "Local Linear Quantile Regression", *Journal of the American Statistical Association*, 93, 228-237.

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

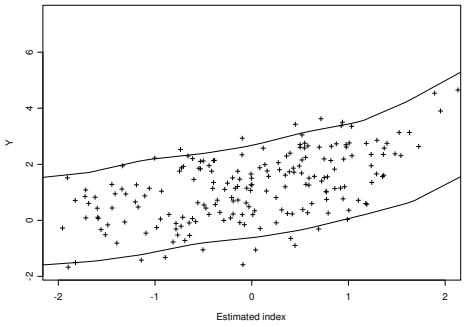
Covariate	AIC	EDR
AGE	312.93	0.011
C2BRAS	35.29	-0.126
PBRAS	23.00	0.438
HYGRO	17.86	-0.057
BRAS1	17.20	-0.066
TEMP	16.89	0.312

Table 1

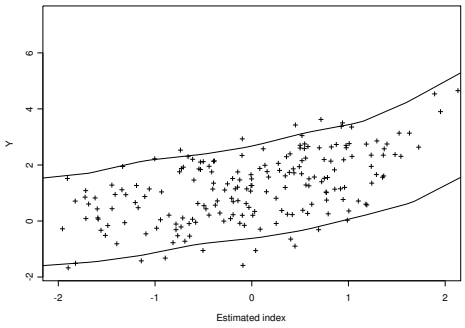
Results of the forward-selected linear regression model for the index concerning KBRAS. For each selected covariate (first column), we indicate its negative or positive contribution to the EDR (third column) as well as the value of the AIC criterion (second column).



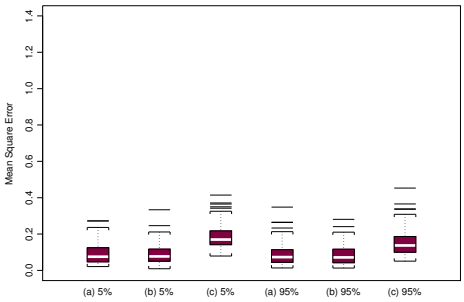
1. True reference curves (dotted line) and estimator **(b)** (solid line)



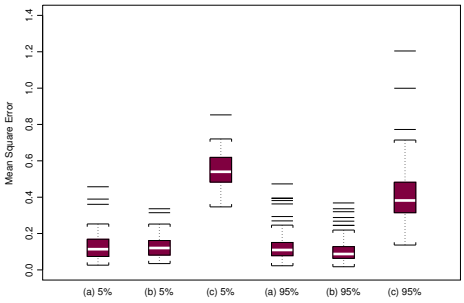
2. Estimator **(a)**



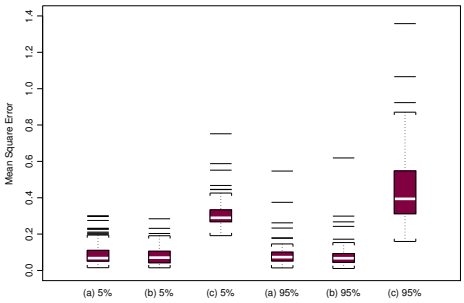
3. Mean errors. **(a)**: solid line **(b)**: dotted line, **(c)**: dashed line



4. Boxplot obtained with $d = 3$

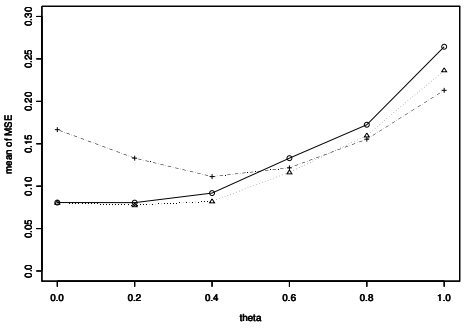


5. Boxplot obtained with $d = 9$

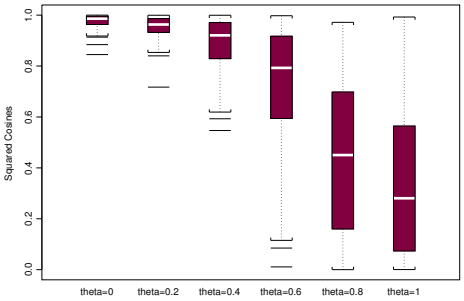


6. Boxplot obtained with $d = 13$

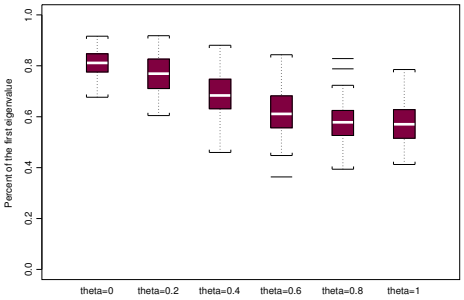
Fig. 1. *1–2: Comparison of the kernel estimated and true 90%-reference curves on the model* **(M1)**. *Estimator* **(b)** *of the reference curves is obtained using the true index whereas estimator* **(a)** *is obtained with the SIR estimated index. 3: Comparison of the mean errors obtained by the three estimates on the model* **(M1)** *as a function of the dimension. 4–6: Comparison of the error boxplot obtained on the model* **(M1)** *with the three different estimates for different values of the dimension.*



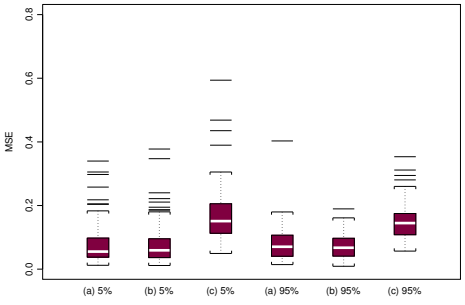
1. Mean errors. **(a)**: solid line **(b)**: dotted line, **(c)**: dashed line



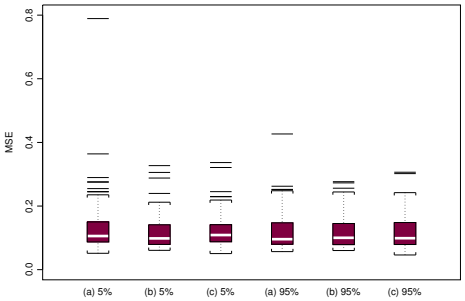
2. Squared cosines boxplot



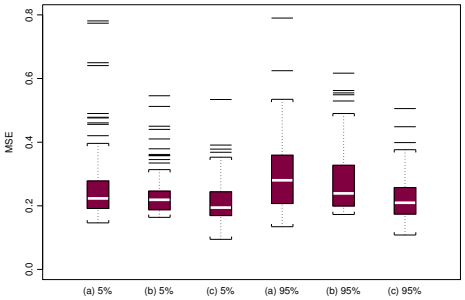
3. First eigenvalue boxplot



4. Boxplot obtained with $\theta = 0$



5. Boxplot obtained with $\theta = 0.6$



6. Boxplot obtained with $\theta = 1$

Fig. 2. *1: Comparison of the mean errors obtained by the three estimates on the model* **(M2)** *as a function of the contamination parameter θ. 2: Boxplots of the squared cosines between the true and estimated dimension reduction when θ increases. 3: Boxplots of the first eigenvalue when θ increases. 4–6: Comparison of the error boxplots obtained on the model* **(M2)** *with the three different estimates for different values of θ.*

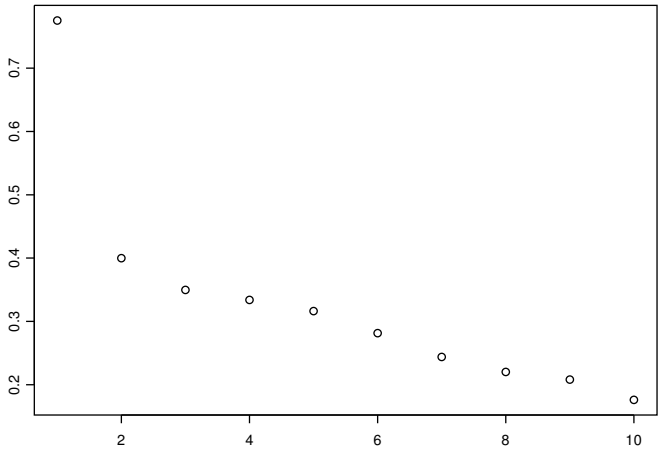


Fig. 3. *Scree plot of eigenvalues for variable KBRAS. The break in size of eigenvalues between the first and second SIR directions suggests that one dimension should be chosen.*

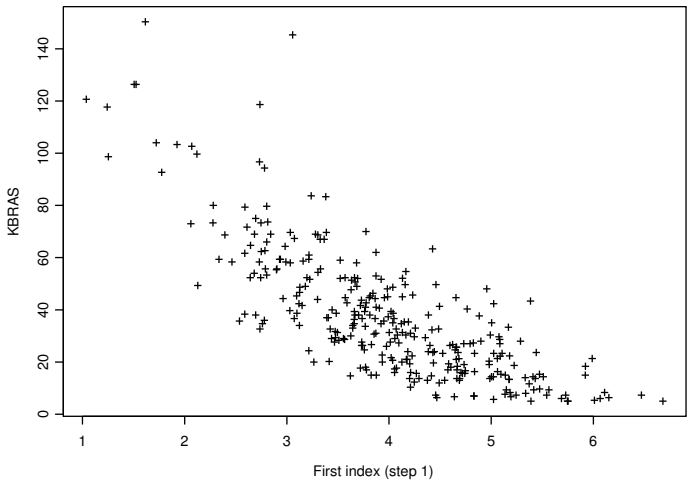


Fig. 4. *Scatterplot of the response variables versus the first SIR index computed with all the covariates (step 1). This first index reveals a strong structure in the scatterplot.*

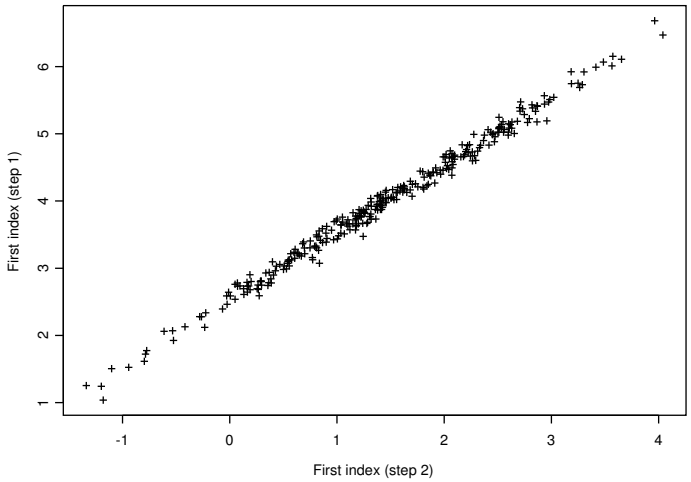


Fig. 5. *Graphical validation of the covariate selection step. The index computed at the first step with all covariates are plotted versus the index computed at the second step with the selected covariates. Since the plots reveal a linear structure ($R^2 = 0.990$), there is no loss of information working with only the subset of selected covariates.*

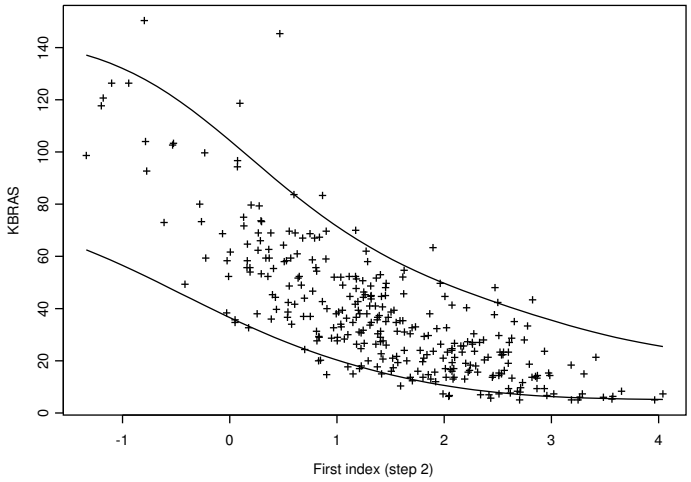


Fig. 6. *Estimated 90%-reference curves for variable KBRAS. These curves are estimated using the index computed with the selected covariates (step 2).*