Reference curves estimation via Sliced Inverse Regression

Ali Gannoun, Stephane Girard, Christiane Guinot, Jerôme Saracco

Reference curves estimation via Sliced Inverse Regression

Ali Gannoun, Stéphane Girard, Christiane Guinot and Jérôme Saracco

Abstract

In order to obtain 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 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.

1 Introduction

Reference intervals are a tool of some importance in clinical medecine. They provide a guideline to clinicians 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 and for typical individuals. The conventional definition of a reference interval is a pair of numbers that bind, for example, the central 90% of a set of values obtained from a specified group of subjects (the reference subjects).

The need for reference curves, rather than a simple reference interval, 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 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 [10].

Let $\alpha \in (0,1)$, the conditional quantile of Y given $X = x$, denoted by $q_\alpha(x)$, is naturally defined as 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 $(2\alpha - 1)\%$ reference curves are defined, when x varies, by

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

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

In the last decade a nonparametric theory has been developed 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, $\hat{F}_n(y|x)$, can be affected to $F(y|x)$:

$$\hat{F}_n(y|x) = \sum_{i=1}^{n} K\{(x - X_i)/h_n\} I_{\{Y_i \leq y\}} \Big/ \sum_{i=1}^{n} K\{(x - X_i)/h_n\}\ , \qquad (2)$$

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

$$\hat{F}_n(y|x) = \alpha. \qquad (3)$$

Many authors are interested in this estimator, see for example [1, 12]. Note that 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.

Although, theoretically, the extension of conditional quantiles to higher dimension p of X is obvious, its practical success, while depending on the number of observations, 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 rendered difficult making it less directly useful for exploratory purposes (unlike the one-dimensional case). When $p > 2$, 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 a sufficient summary plot that contains all 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 dimension-reduction, via Sliced Inverse Regression (SIR) method, in order to get an efficient estimator of conditional quantiles from which we can then deduce reference curves. In Section 2, we present the dimension-reduction context and we derive the corresponding semiparametric estimator of conditional quantiles. We also give an asymptotic result. Simulations are conducted in Section 3 to assess the performance of this estimator in finite-sample situation. Numerical example involving real data application is reported in Section 4.

2 Dimension-reduction and estimation procedure

2.1 Dimension-reduction context

We suppose that there exists a $p \times q$ matrix $(q \leq p)$ B such that

$$Y \perp X \mid \beta^T X, \tag{4}$$

where the columns of the $p \times d$ matrix β $(d \leq p)$ are linearly independent. Consequently, in the current study, the statement (4) is equivalent to

$$F(y|x) = F(y|\beta^T x),$$

for all values of x in the sample space. In a straightforward manner, it follows that

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

The SIR method can be used to estimated a basis of the subspace $S(\beta)$ spanned by the columns of β. More details and comments on the SIR estimation procedure can be found in [6, 11].

2.2 Estimation procedure

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, $i = 1, \ldots, n$.

• *Step 1: SIR estimation step.* With SIR method, we get $\{\hat{b}_k\}_{k=1}^d$, an estimated basis of $S(\beta)$. In practice, the dimension d is replaced with an estimate $\hat{d}$ (see for instance [8]).

• *Step 2: Conditional quantile estimation step.* For the sake of convenience, we assume that $d = 1$ and we use the notation $\hat{b} = \hat{b}_1$. Using the SIR estimates and following (2), an kernel estimator of $F(y|x)$ is defined, from the data $\{(Y_i, \hat{b}^T X_i)\}_{i=1}^n$, by

$$F_n\left(y \left| \hat{b}^T x\right.\right) = \frac{\sum_{i=1}^n K\{(\hat{b}^T x - \hat{b}^T X_i)/h_n\} I_{\{Y_i \leq y\}}}{\sum_{i=1}^n K\{(\hat{b}^T x - \hat{b}^T X_i)/h_n\}}. \tag{5}$$

Then, as in (3), we derive from (5) an estimator of $q_\alpha(x)$ by

$$q_{n,\alpha}\left(\hat{b}^T x\right) = F_n^{-1}(\alpha \mid \hat{b}^T x). \tag{6}$$

As a consequence of the above result, for $\alpha > 0.5$, the corresponding estimated $(2\alpha - 1)\%$ reference curves are given by the following

$$I_{n,\alpha}(x) = [q_{n,1-\alpha}(\hat{b}^T x), q_{n,\alpha}(\hat{b}^T x)], \quad \text{as } x \text{ varies.}$$

Under usual assumptions, we obtain the consistency of $q_{n,\alpha}(\hat{b}^T x)$ (see [9]) for the proof): for a fixed x in $\mathbb{R}^p$,

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

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 $\{\hat{b}_k\}_{k=1}^d$ and $\{(Y_i, \hat{b}_1^T X_i, \ldots, \hat{b}_d^T X_i)\}_{i=1}^n$. Then we use the classical multi-kernel estimation to get $q_{n,\alpha}(\hat{b}_1^T x, \ldots, \hat{b}_d^T x)$ as in (6).

3 Simulation study

We study the numerical performances of the proposed method on simulated data. In particular, we compare our method with the classical nonparametric estimation method. Let us introduce the following estimators of $q_\alpha(x)$:

(a) $q^{(a)}_{n,\alpha}(x) := q_{n,\alpha}(\widehat{b}^T x)$ is the estimator defined in (6).

(b) $q^{(b)}_{n,\alpha}(x) := q_{n,\alpha}(\beta^T x)$ 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) $q^{(c)}_{n,\alpha}(x) := q_{n,\alpha}(x)$ is the classical conditional nonparametric quantile estimator.

The kernels are the densities of the standard normal or multivariate normal distribution, and the bandwidth is chosen by a cross-validation technique. The estimated conditional quantiles are computed by numerically inversing the corresponding conditional distribution function.

3.1 Simulated models

We consider the following regression model

$$Y = f(\beta^T X) + \varepsilon, \tag{7}$$

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. 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 $q_\alpha(x) = f(\beta^T x) + N_\alpha$, where N_α is the α-quantile of the standard normal distribution.

3.2 Evaluation of the results

Our goal is to compare successively the three estimators **(a)**, **(b)** and **(c)** to the true quantile in the situations **(M1)**, **(M2)** and **(M3)**. To this end, the $N = 100$ 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:

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

The boxplots 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 are represented on Figure 1. Figure 1.1 shows no difference between the distribution of $E_{n,\alpha}^{(a)}$ and $E_{n,\alpha}^{(b)}$. The estimation of the direction β by $\widehat{b}$ has no significant consequence on the accuracy of the estimation of the reference curves. 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 1.3). 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 situations. Note that the quality of the estimation of β is not severely affected by the covariates number. Finally, in view of Figure 1.2, the nature of the link function f does not seem to have any influence on the relative behaviors of the three estimators.

4 Application to real data

4.1 Data

When studying skin biophysical properties of healthy women, knowledge about the reference "curves" of certain parameters is lacking. The aim is to establish 90% reference "curves" for some of the biophysical properties of the skin (here the conductance, which reflects the hydratation status of the skin) of healthy Caucasian women, on two facial areas and one forearm area, using the age and a set of covariates. The data collection was conducted from November 1998 to March 1999 on $n = 322$ Caucasian women between 20 and 80 years old with apparently healthy skin, and living in the Ile de France (in around Paris) area. The volunteers were preselected by a subcontractor company. Each volunteer was examined at CE.R.I.E.S (Private Research Center in Human Skin founded by Chanel) in a controlled environment. This evaluation included self-administered questionnaire on skin-related habits, a medical examination and a biophysical evaluation. The age of the volunteer, the temperature and relative humidity of the controlled environment occur in each study as covariates. The other available covariates included are some biophysical properties of the skin (as the skin surface temperature or the skin surface pH).

4.2 Results

We only give here the results for the forearm area. In step 1, the SIR method gives $\hat{d} = 1$ and the corresponding vector $\hat{b}$. Then in step 2, after a simplification of the index $\hat{b}^T X$ (see [9] for details), we construct the 90% reference curves for the conductance of the skin (variable named KBRAS) using this estimated index, see Figure 2. The results of the analysis on the forearm index 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

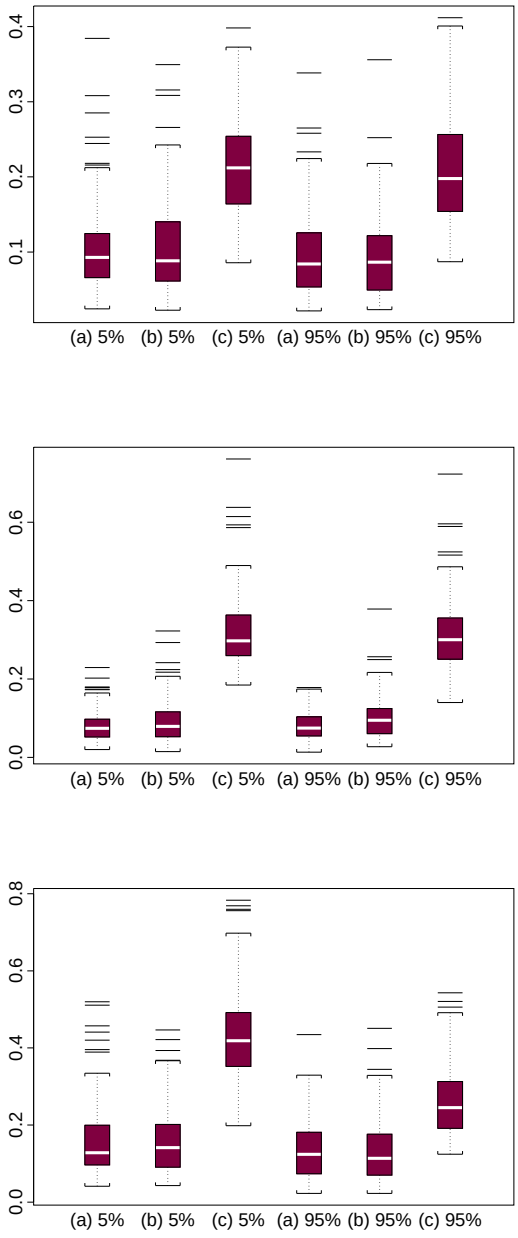

Figure 1: Boxplots obtained on the three different models **(M1)**, **(M2)** and **(M3)** (from top to bottom) with the three different estimates.

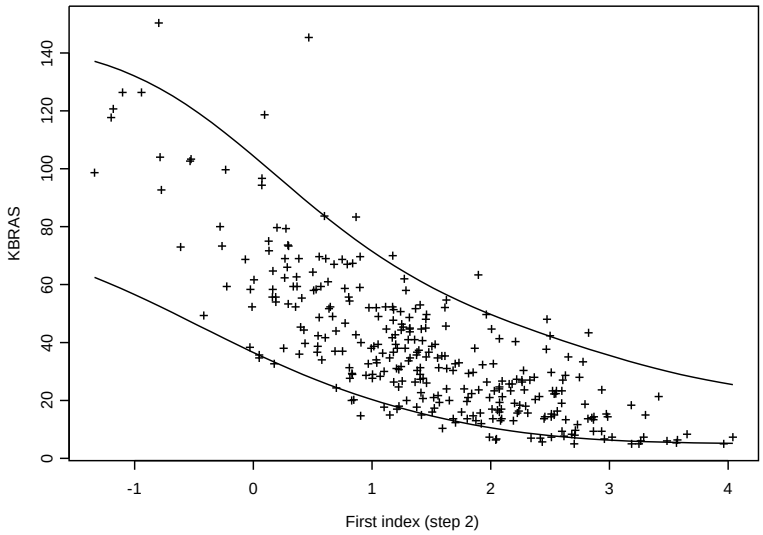

Figure 2: Estimated 90%-reference curves for the forearm area.

clinically-related with skin hydration: skin surface pH, capacitance and transepidermal water loss.

References

[1] Berlinet, A., Gannoun, A., and Matzner-Lober, E. (2001). Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, **35**, 139 – 169.

[2] C. Bernard-Michel, S. Dout, M. Fauvel, L. Gardes and S. Girard. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression, *Journal of Geophysical Research - Planets*, **114**, E06005.

[3] C. Bernard-Michel, L. Gardes and S. Girard. (2008). A Note on Sliced Inverse Regression with Regularizations, *Biometrics*, **64**, 982–986.

[4] C. Bernard-Michel, L. Gardes and S. Girard. (2009). Gaussian Regularized Sliced Inverse Regression, *Statistics and Computing*, **19**, 85–98.

[5] M. Chavent, S. Girard, V. Kuentz, B. Liquet, T.M.N. Nguyen and J. Saracco. (2014). A sliced inverse regression approach for data stream, *Computational Statistics*, to appear.

[6] Chen, C. H., and Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, **8**, 289 – 316.

[7] R. Coudret, S. Girard and J. Saracco. (2014). A new sliced inverse regression method for multivariate response, *Computational Statistics and Data Analysis*, to appear.

[8] Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, **93**, 132 – 140.

[9] Gannoun, A., Girard, S., Guinot, C. and Saracco, J. (2004). Sliced inverse regression in reference curves estimation. *Computational Statistics and Data Analysis*, **46**, 103–122.

[10] Healy, M. J. R., Rasbash, J., and Yang M. (1998). Distribution-free estimation of age-related centiles. *Annals of Human Biology*, **15**, 17 – 22.

[11] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316 – 342.

[12] Samanta, T. (1989). Non-parametric estimation of conditional quantiles. *Statistics and Probability Letters*, **7**, 407 – 412.

Ali Gannoun, Jérôme Saracco: Institut de Mathématiques et de Modélisation de Montpellier, Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier cedex 5, France.

Stéphane Girard: SMS/LMC/IMAG, Université Grenoble 1, BP 53, 38041 Grenoble cedex 9, France.

Christiane Guinot: CE.R.I.E.S, Biometrics and Epidemiology Department, 20, Rue Victor Noir, 92 521 Neuilly sur Seine cedex, France.