Mahalanobis's Fractile Graphs: Some History and New Developments

Bodhisattva Sen
Columbia University
New York, NY 10027

and

Probal Chaudhuri
Indian Statistical Institute
203 B.T. Road, Kolkata 700108

October 26, 2010

Abstract

Mahalanobis's famous paper on Fractile Graphical Analysis introduced a plethora of new statistical concepts and techniques [see Mahalanobis (1960)]. The method was originally proposed to compare two regression functions. We discuss and re-interpret some of his work, highlighting his contributions and some of the difficulties encountered. We develop a bootstrap based hypothesis test to compare the fractile regression curves based on their isotonized estimators. The proposed procedure does not depend on the choice of any tuning parameter and is computationally simple. Through an extensive simulation study, we illustrate the finite sample performance of our procedure. We also discuss three real data applications that illustrate the scope of the methodology.

Keywords: Bootstrap, comparison of regression functions, fractile regression, isotonic estimators, standardization of covariate.

1 Introduction

Fractile Graphical Analysis (FGA) was proposed and investigated by Prasanta Chandra Mahalanobis in a series of papers and seminars during the period 1950-70 as a method to compare two regression functions. The procedure is graphical in nature and can be thought of as the regression of the response variable on the ranks of the predictor using non-parametric techniques; and thus the name FGA.

1.1 A Brief History of Fractile Graphical Analysis

Mahalanobis's famous paper "A method of fractile graphical analysis" first appeared in a special issue of *Econometrica* [see pages 325-351 in volume 35 of the journal, published in the year 1960] that was brought out in honour of Ragnar Frisch, joint winner of the first Nobel Prize in Economics, in the year of his sixty-fifth birthday. The stage was indeed grand, and the issue had contributions from other real stars in economics including Nobel laureates like Jan Tinbergen, Paul Samuelson, Kenneth Joseph Arrow and Robert Merton Solow. Mahalanobis in his paper provided "some examples of the use of fractile graphical analysis", which he described as "a new method for the comparison of economic data relating to the same population

over time or to any two populations that differ as to geographical region or in any other way". The paper instantly created a sense of excitement among statisticians in India and abroad. Sankhyā in 1961 reprinted Mahalanobis's original article along with a series of other papers on this topic by Takeuchi (1961), Sethuraman (1961), Parthasarathy and Bhattacharya (1961). The paper also generated some controversy – Subramanian Swamy [see Swamy (1963)] wrote a paper in *Econometrica* criticizing Mahalanobis's approach, and later Iyengar and Bhattacharya (1965) published an article responding to Swamy's paper. A nice summary of the developments on FGA can be found in a collection of articles by Mahalanobis that was edited by P. K. Bose and published in 1988 [Mahalanobis (1988)].

Although a somewhat forgotten statistical tool now, the developments made by Mahalanobis in this area remain extremely relevant as it can be said to have lead to the inception of modern nonparametrics. We mention here some of his contributions in nonparametrics and allied areas. (a) FGA is one of the earliest works on nonparametric regression. In fact Watson's (1964) seminal paper on nonparametric kernel regression cited and discussed fractile graphs. (b) Mahalanobis (1960) also provided some guidelines for formally testing the equality of two regression curves nonparametrically. (c) FGA used one of the earliest forms of resampling (subsampling) techniques for carrying out statistical tests [see Hall (2003)]. (d) Mahalanobis (1988) provides one of the earliest versions of a multivariate fractile/quantile transformations. But the emphasis and "the aim of the method (FGA) is rather to produce distribution-free tests" to compare the fractile graphs than to accurately (nonparametrically) estimate the regression curves, as pointed out by Watson (1964) in his seminal paper. We maintain the same spirit in our discussion, and investigate nonparametric tests for the formal comparison of the fractile graphs.

Mahalanobis initially developed and used FGA to study the economic condition of rural India on the basis of data collected on household consumption and expenditure over two different time periods: the 7th (October 1953 to March 1954) and the 9th (May to November 1955) rounds of the National Sample Survey of India. It is obviously of great importance to policy makers of a country like India to understand the economic condition of the rural community. They would also like to ascertain whether their policies have been able to improve the economic condition of the rural population over a period of time. As a measure of the economic well-being of the rural community, one may consider the fraction of total expenditure that is spent on food articles to the total expenditure incurred. It is expected that lower this proportion, the greater is the possibility of the rural community being better off.

Let X be the total expenditure per capita per 30 days in a household and Y be the fraction of total expenditure on food articles per capita per 30 days in the household. Mahalanobis wanted to perform a regression analysis of Y on X and was interested in comparing the regression functions at two different time points. But due to inflation, the total expenditure (per capita per 30 days) for the two time points are not comparable. Just comparing the regression functions for the two populations did not make much sense. In fact, Mahalanobis was aware that the comparative

study that he was interested in will be inadequate even if one uses inflation adjusted figures for expenditures at the two different time points. So, he chose to compare the means of the Y-variable in different fractile (rank) groups corresponding to the X-variable. This approach leads to a novel way of standardizing the covariate X so that comparison of the two regression functions over two different time periods can be done in a more meaningful way. More precisely, FGA does this required standardization by considering $F(X)$ instead of X as the regressor, where F is the distribution function of X.

Recently there has been a renewed interest in FGA, and several papers have been written highlighting the usefulness and applicability of FGA in diverse settings; see e.g., Nordhaus (2006), Hertz-Picciotto and Din-Dzietham (1998), Montes-Rojas (2010), Bera and Ghosh (2006), Sen (2005) and Sen and Chaudhuri (2010). Nordhaus (2006) shows fractile plots of key geographic variables (temperature, precipitation, latitude, etc.) against the fractiles of log of "output density" while trying to explore the linkage between economic activity and geography. Hertz-Picciotto and Din-Dzietham (1998) compare the infant mortality using a "percentile based method" of standardization for birthweight or gestational age. Their motivation underlying the percentile-based method of standardization is that comparable health for two population groups will be expressed as equal rates of disease or mortality at equal percentiles in the distributions of either birthweight or gestational age. Montes-Rojas (2010) considers nonparametric estimators of average and quantile treatment effects, in applications arising in econometrics, using ideas from FGA. Bera and Ghosh (2006) discuss FGA with some historical perspectives and consider some relevant applications in Economics and Finance. The estimation and formal testing of fractile graphs using smooth nonparametric estimators is considered in Sen (2005). Sen and Chaudhuri (2010) investigate FGA when the covariate can be multi-dimensional.

Our contribution goes beyond the aforementioned papers in the following ways. We develop tests of hypotheses for the comparison of the fractile regression curves using isotonized nonparametric estimators. As mentioned before, this comparison is a central issue in FGA, and Mahalanobis himself did not have very clear results in this direction. In the process, we review and re-interpret some of the main ideas of FGA. Although Sen (2005) addressed this comparison, the paper was restricted to the use of smooth regression estimators. His approach, although conceptually simple, made the formal comparison of the fractile regressions very ad-hoc as it depended on the choice of many tuning parameters. In this paper, we propose tests of hypotheses that do not depend on the choice of tuning parameters, and illustrate, through an extensive simulation study, the remarkable finite sample properties of the method. The proposed procedure can be implemented easily and is computationally fast.

In Section 2, we define and study some properties of the population version of fractile regression. In Section 3, we introduce the nonparametric estimators of fractile regression to be considered in the paper. In particular, we motivate the define the isotonized estimators. The formal comparison of the fractile regression curves is con-

sidered in Section 4. The finite sample performance of our method is investigated in Section 5. Section 6 discusses application of FGA in three different examples. We end with some concluding remarks, mentioning some of the open research problems in this area and discuss possible extensions in Section 7.

2 Fractile Regression

In precise mathematical terms, FGA can be described as follows: consider two bivariate random vectors (X_1, Y_1) and (X_2, Y_2) and the associated regression functions μ_1 and μ_2 where $\mu_1(x) = E(Y_1|X_1 = x)$ and $\mu_2(x) = E(Y_2|X_2 = x)$. Then the *fractile regression* functions are defined as

$$m_1(t) = E\{Y_1|F_1(X_1) = t\} \quad \text{and} \quad m_2(t) = E\{Y_2|F_2(X_2) = t\}$$

for $t \in (0, 1)$, where F_1 and F_2 are the distribution functions of X_1 and X_2 respectively [see Mahalanobis (1960)]. Mahalanobis's goal was to test the equality of the fractile regression curves m_1 and m_2 nonparametrically. Note that the comparison of $m_1(t)$ and $m_2(t)$ amounts to comparing the means of the responses Y_1 and Y_2 at the t-th quantile of the covariates rather than the same value of the covariates, as is done in usual regression.

Notice that the transformed covariates $F_1(X_1)$ and $F_2(X_2)$ both have a uniform distribution on $(0, 1)$. This achieves a distribution-free nonparametric standardization of the covariates. The uniform distribution of the transformed covariate also yields optimality properties of the estimated fractile regression function; see Sen and Chaudhuri (2010). This standardization makes the fractile regression functions invariant under all strictly increasing transformations of the covariate. In other words, suppose that (X_1, Y) is a continuous bivariate random vector and if $X_2 = \phi(X_1)$, where ϕ is any strictly increasing transformation, then $E\{Y|F_1(X_1)\} = E\{Y|F_2(X_2)\}$, where F_1 and F_2 are the distribution functions of X_1 and X_2 respectively. This is a crucial property and can be interpreted in the following way: the fractile transformation makes the regression functions comparable even when the covariate in the second population is *any* increasing transformation of that of the first population.

In fact, it can be shown that if any transformation of the covariate achieves *invariance* of the regression functions under the group of all strictly increasing functions of the covariate then it must be a function of the fractile transform $X_1 \mapsto F_1(X_1)$. Furthermore, if we impose the additional assumption of uniformity of the transformed covariate distribution, then the fractile transformation is the *only choice*; see Sen and Chaudhuri (2010) for a discussion and a proof of the results. This property is very important while comparing regression functions, where the distributions of the covariates are very different for the two populations under comparison.

3 Estimation of Fractile Regression

Before we describe our procedure, let us first introduce Mahalanobis's ideas. Consider a random sample $\{(X_i, Y_i)\}_{i=1}^n$ from a bivariate population, where $X_1 \sim F$. Suppose that the data points are ranked in the ascending order of X. The n data points are divided into g *fractile* groups each of equal size $n' = n/g$. On the x-axis, g equidistant points $1, 2, \ldots, g$ are marked to represent the g fractile groups, and the corresponding means of the y-variable, labeled as $y'_1, y'_2, \ldots, y'_g$, are plotted. Each pair of adjoining points y'_i and y'_{i+1} for $i = 1, 2, \ldots, g-1$ are joined by straight lines to get a polygonal curve called the *fractile graph*.

In this paper, we consider two kinds of nonparametric estimation techniques for the fractile regression function

$$m(t) = E\{Y_1 | F(X_1) = t\} \qquad \text{for } t \in [0, 1].$$

The first method uses smoothing techniques, whereas the second method relies heavily on known "shape" constraints on the regression function (e.g., increasing, decreasing). Sen (2005) considered the estimation of fractile regression using kernel smoothing methods in great detail, and we just summarize the main ideas below.

A general class of smoothed nonparametric estimators of fractile regression, called linear smoothers [as proposed by Stone (1977)], can be expressed in the form:

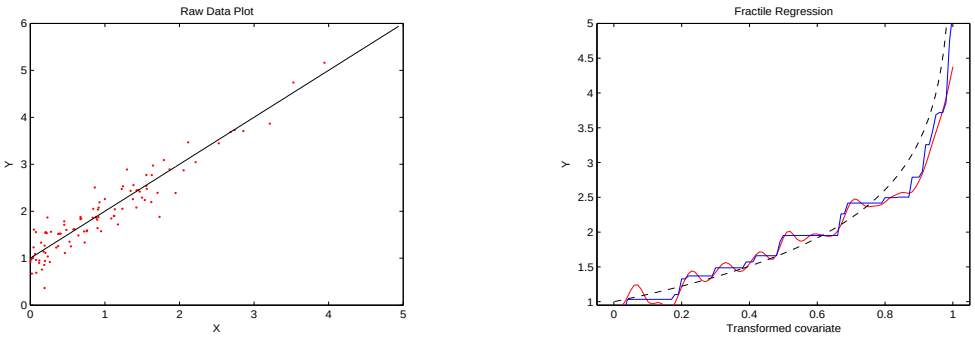
$$\widehat{m}_n^S(t) = \sum_{i=1}^n Y_i W_{n,i}(t),$$

where $W_{n,i}(t)$ is a proper weight function depending on the input data, preferably satisfying $\sum_{i=1}^n W_{n,i}(t) = 1$, for all $t \in [0, 1]$. If we use nonparametric kernel regression procedures [see e.g., Muller (1988), Härdle (1990), Wand and Jones (1995), Fan and Gijbels (1996)], the weight function can depend on the smoothing bandwidth and can take the form $W_{n,i}(t) = W_{n,i}(t, h_n, F_n(X_1), F_n(X_2), \ldots, F_n(X_n)), 1 \le i \le n$, where F_n is the empirical distribution function of $\{X_i\}_{i=1}^n$ (i.e., $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \le x\}}$ for $x \in \mathbb{R}$), and h_n is the smoothing bandwidth based on a sample of size n. Note that as we are regressing Y on the ranks of X, in a sense, we can pretend that our observations are $\{(F(X_i), Y_i)\}_{i=1}^n$, where F is the distribution function of X_1. But as the distribution function F is not known, we work with the empirical distribution function F_n, and it is used in the weight functions for the fractile regression estimators.

Fractile regression can also be obtained as a transformation of the usual regression function [see Bhattacharya and Muller (1993)] by observing that

$$m_i(t) = \mu_i \circ F_i^{-1}(t), \quad \text{for } i = 1, 2.$$

This provides an alternative way of estimating fractile regression: first estimate the usual regression function and then estimate the distribution function of the covariate to obtain the final estimator. But this requires a smooth estimation of the distribution function of the covariate, making it more difficult to implement; and in fact,



(a) (b)

Figure 1: *(a) The scatter plot with the true regression function, (b) the fractile regression function (in dashed black) along the smoothed (in solid red) and the shape constrained increasing (in solid blue) nonparametric estimators.*

our simulations showed that it had worse finite sample properties. Thus, we do not explore this further.

The choice of the smoothing bandwidth h_n is of crucial importance in the above smoothing procedures. Although there are several methods proposed in the literature [e.g., see Rice (1984), Härdle (1990), Wand and Jones (1995), Fan and Gijbels (1996)] for choosing the optimal bandwidth, in practice, the performance of most bandwidth selectors is far from satisfactory. This motivates the use of "shape" constrained nonparametric estimators that are completely automated, and do not need the choice of any tuning parameter.

Suppose that the usual regression function is known to be increasing (or decreasing). In that case the fractile regression function $m(\cdot)$ will also be increasing (or decreasing). As we will see, in most of the real applications considered in this paper, such shape restrictions arise naturally. Representing the transformed data set as $\{(i/n, Y_{[i:n]})\}_{i=1}^n$, where $Y_{[i:n]}$ denotes the concomitant of the i-th order statistic of X, we can now define the *isotonic* (increasing) estimator of m as

$$\widehat{m}_n^I = \arg\min_{f \in \mathcal{C}} \sum_{i=1}^{n} \left\{ Y_{[i:n]} - f\left(i/n\right) \right\}^2 \tag{1}$$

where $\mathcal{C}$ denotes the class of all increasing real valued functions on $[0,1]$; see Brunk (1970) and Robertson et al. (1988). Thus, the isotonic estimator is obtained by minimizing the least squares criterion over all increasing functions. A unique solution to problem (1) exists and can be expressed as [see Robertson et al. (1988), page 24]

$$\widehat{m}_n^I\left(k/n\right) = \max_{i \le k} \min_{j \ge k} \frac{Y_{[i:n]} + \ldots + Y_{[j:n]}}{j - i + 1},$$

for $k = 1, \ldots, n$. The estimator can be easily computed using the pool adjacent violators algorithm (PAVA); see Barlow et al. (1972). If skillfully implemented, PAVA has a computational complexity of $O(n)$ [see Grotzinger and Witzgall (1984)]. There is quite a large literature on isotonic regression. Barlow et al. (1972) is a classic reference along with Robertson et al. (1988). A recent paper by de Leeuw at al. (2009) gives an overview of the problem's history and computational aspects.

Note that $\widehat{m}_n^I$ is only defined at the x-values i/n, but it can be defined on the entire interval $[0, 1]$ using a piece-wise constant (or linear) interpolation. The main advantage of the isotonized estimator $\widehat{m}_n^I$ is that it avoids the specification of any tuning parameter, and as we will see later, drastically simplifies the testing problem considered in Section 4.

As an example of fractile regression, we demonstrate the smooth kernel based estimator, the shape constrained (increasing) estimator in Figure 1 along with the true fractile regression curve. We generated a random samples of size $n = 100$ from the population $Y = 1.0 + X + \epsilon$ where $\epsilon \sim N(0, 0.09)$ and $X \sim Exp(1)$. Although the two estimators are quite similar, the kernel based method produces a very wiggly function (a consequence of the chosen optimal bandwidth being too small). For our smooth estimator, we have used the Nadaraya-Watson type weight function [see Sen (2005)] with the standard normal kernel and optimal bandwidth obtained by the method of least squares cross validation.

4 Comparison of Estimated Fractile Regression Functions

Suppose that we have data $\{(X_{1i}, Y_{1i})\}_{i=1}^{n_1}$ and $\{(X_{2i}, Y_{2i})\}_{i=1}^{n_2}$ from two populations, and we want to test the hypotheses

$$H_0 : m_1 = m_2 \qquad vs. \qquad H_A : m_1 \neq m_2, \tag{2}$$

where $m_1(t) = E(Y_{1i}|F_1(X_{1i}) = t)$ and $m_2(t) = E(Y_{2i}|F_2(X_{2i}) = t)$, and F_1 and F_2 are the continuous strictly increasing distribution functions of X_{1i} and X_{2i} respectively.

Much effort has been devoted to the problem of comparing nonparametric regression curves in the recent literature [e.g., see Delgado (1993); Munk and Dette (1998); Dette and Neumeyer (2003)]. These authors considered the testing problem

$$H_0 : g_1 = g_2 \qquad vs. \qquad H_A : g_1 \neq g_2, \tag{3}$$

where g_1 and g_2 are the usual regression curves corresponding to two different populations. Most authors concentrated on equal design points to develop tests for (3). Kulasekera (1995) proposed a test for the hypotheses in (3) using quasi-residuals which is applicable under the assumption of different design points for both the samples. Kulasekera and Wang (1997) considered the selection of smoothing parameters to obtain optimal power in tests of regression curves. Munk and Dette (1998),

Neumeyer and Dette (2003) considered the problem of the comparison of nonparametric regression curves under a very general set-up. Delgado (1993), Kulasekera (1995) and Kulasekera and Wang (1997) considered marked empirical processes to develop tests for the hypotheses in (3).

All the above procedures use nonparametric smoothing techniques that involve the choice of a number of tuning parameters and an optimal choice in finite samples is indeed very difficult. In this section, we outline a resampling (bootstrap) based hypothesis testing procedure that does not involve the choice of any tuning parameter and is completely automated. Our method is applicable in situations when the fractile regression function is known to obey "shape" restrictions like monotonicity (decreasing/increasing). Also, none of the above mentioned authors address the problem of possible effects of transformations on the covariate for the two populations. Further, some of the usual methods for comparison of the regression curves do not generalize in a straight forward manner in our setup as in fractile regression the covariate X_i is replaced by $F_n(X_i)$, and the $F_n(X_i)$'s are not independent even if the X_i's are so.

4.1 Mahalanobis's Idea for Comparing two Fractile Graphs

The first sample of size n_1 is obtained from the first bivariate population by drawing two independent ("interpenetrating") random half-samples each of size $n_1/2$. The first half-sample is then considered, and the fractile graph $G(1)$ is constructed from it [see Section 3 for Mahalanobis's construction of the fractile graphs]. The second half-sample is used to get the second fractile graph $G(2)$. Clearly, the two half-sample fractile graphs $G(1)$ and $G(2)$ have identical statistical distributions.

Mahalanobis's idea was to mix the two half-samples to form the combined sample of size n_1 from the first population. The combined sample is again ranked according to the X-values and divided into g fractile groups each containing n_1' ($n_1' = n_1/g$) units. The y-averages of the corresponding fractile groups are plotted to get the combined fractile graph $G(1,2)$. The "error area" $a(1,2)$ associated with the combined sample is defined as the area bounded between the two half-sample fractile graphs $G(1)$ and $G(2)$ (i.e., $a(1,2) = \int |G(1) - G(2)|$).

The second bivariate population is considered next from which a pair of independent ("interpenetrating") half-samples are drawn. The second set of fractile graphs $G'(1), G'(2)$ and $G'(1,2)$ are computed from the half-samples obtained from the second population. The area bounded between $G'(1)$ and $G'(2)$ is called the second "error area" associated with the second population and is denoted by $a'(1,2)$ (i.e., $a'(1,2) = \int |G'(1) - G'(2)|$). The area between the two combined fractile graphs $G(1,2)$ and $G'(1,2)$ is called the "separation area" and is denoted by $S(1,2)$ (i.e., $S(1,2) = \int |G(1,2) - G'(1,2)|$). The statistical error E to be associated with the "separation area" $S(1,2)$ is defined by the formula $E = \sqrt{a^2(1,2) + a'^2(1,2)}$. The significance of the observed value of $S(1,2)$ is tested by considering the test-statistic $S^2(1,2)/E^2$, which Mahalanobis thought would be distributed approximately like a chi-square random variable. However, there does not appear to be any mathematical

validity for this result. Interested readers are referred to Takeuchi (1961), Mitrofanova (1961) and Mahalanobis (1988) for some of the statistical properties of the "error area" in FGA. Sethuraman (1961) introduced other measures of divergence between fractile graphs and investigated their limit distributions.

The distribution of the test-statistic $S^2(1,2)/E^2$ is not in general chi-square. The implementation of the method also requires the choice of the "group" size g, a tuning parameter. Mahalanobis's idea of using the subsamples to approximate the null distribution can, in light of modern developments in bootstrap and other resampling techniques, be improved. In the following we propose a bootstrap based test that uses the essential ideas of Mahalanobis and discuss its implementation.

4.2 A Resampling Based Test

After obtaining the estimates of fractile regression, $\widehat{m}_{1,n_1}$ and $\widehat{m}_{2,n_2}$, for the two populations, to test the hypotheses in (2), we might use the test statistic

$$T^{(p)}_{n_1,n_2} = \int_0^1 |\widehat{m}_{1,n_1}(t) - \widehat{m}_{2,n_2}(t)|^p dt \tag{4}$$

where $p \geq 1$. When $p = 1$ this gives us the "separation area" between the two fractile regression functions. Under the null hypothesis, we expect the test statistics to be small, whereas large values of the test statistics would support the alternative hypothesis. For mathematical and computational tractability, we recommend taking $p = 2$, and use it in our data analysis. A major technical barrier in using the test statistic $T^{(p)}_{n_1,n_2}$ is that its sampling distribution is analytically intractable. But recently there has been progress in this direction, and Durot (2007) shows that under appropriate conditions, when we use the shape restricted nonparametric estimators,

$$n_1^{1/6}\left\{n_1^{p/3}\int_0^1 |\widehat{m}_{1,n_1}(t) - m_1(t)|^p dt - m_p\right\} \to_d N(0,\sigma_p^2),$$

where m_p and σ_p are unknown constants that depend in a complicated way on the regression function and the error distribution. A similar result would also hold for the estimator for the second sample. The use of the asymptotic normal distribution is difficult as that involves estimating the nuisance parameters m_p and σ_p. But this suggests that an appropriate bootstrap [see e.g., Hall (1992) and Efron and Tibshirani (1993)] based test may provide good approximation for the P-value of testing the hypotheses in (2). In light of this, we propose a resampling based procedure to test the hypotheses in (2). A complete theoretical justification of our method is beyond the scope of the present paper, and will be the topic of future research.

We describe below the steps involved in computing the bootstrap estimates of the P-values when $T_{n_1,n_2} \equiv T^{(2)}_{n_1,n_2}$ is used as the test statistic.

- We transform the covariate X into its quantiles, i.e., we transform the data set into $\{(i/n_1, Y_{1[i:n_1]})\}_{i=1}^{n_1}$ and $\{(i/n_2, Y_{2[i:n_2]})\}_{i=1}^{n_2}$, where $Y_{1[i:n_1]}$ denotes the concomitant of the i-th order statistic of X for the first sample [see e.g., David and Nagaraja (2003)].

- After transforming the covariate, we obtain the estimated isotonized fractile regression functions $\widehat{m}_{1,n_1}$ and $\widehat{m}_{2,n_2}$ for the two samples as explained in Section 3. We compute T_{n_1,n_2} from the data. Note that we use the piece-wise constant extension of $\widehat{m}_{1,n_1}$ and $\widehat{m}_{2,n_2}$ on the interval [0,1] which drastically simplifies the computation of T_{n_1,n_2}.

- To test the significance of the observed value of T_{n_1,n_2}, we first compute the pooled estimate of fractile regression $\widehat{m}(\cdot)$ combining the two data sets. This is accomplished by defining

$$\widehat{m} = \arg\min_{f \in \mathcal{C}} \sum_{i=1}^{n_1+n_2} \left\{ y_{[i:n_1+n_2]} - f\left(i/(n_1+n_2)\right) \right\}^2 w_i \tag{5}$$

  where $y_{[i:n_1+n_2]}$ is the concomitant of the i-th order statistic of X for the pooled sample, and $w_i = 1/n_1$ or $1/n_2$ depending on whether the i-th data point is from the first or second sample. Note that (5) is the weighted version of isotonic regression, and slightly different from (1). But invoking the PAVA algorithm on the $n_1 + n_2$ response values in the pooled sample gives the solution. From the definition of the weights w_i, it is clear that the sum of the weights for each sample is 1, and this ensures that both the samples get equal weightage in calculating the pooled estimator, even though their sample sizes might be different. Under the null hypothesis, $\widehat{m}$ acts as the surrogate for the true fractile regression function.

- We compute the residual at each i/n_j for $1 \le i \le n_j$ and $j = 1, 2$ using the pooled estimate $\widehat{m}(\cdot)$, i.e.,

$$\hat{\epsilon}_{ji} = Y_{ji} - \widehat{m}(i/n_j).$$

  To construct the bootstrap samples of sizes n_1 and n_2, we draw from the distribution of the residuals, to construct the bootstrap response values. If we assume homoscedastic errors, this could be achieved by drawing a simple random samples with replacement $\{\epsilon^*_{ji} : i = 1, 2, \ldots, n_j; j = 1, 2\}$ from the residuals $\{\hat{\epsilon}_{ji}\}$, and then defining

$$Y^*_{j[i:n_j]} = \widehat{m}(i/n_j) + \epsilon^*_{ji}, \quad \text{for } i = 1, 2, \ldots, n_j; j = 1, 2.$$

  To take care of heteroscedasticity, we use wild bootstrap to generate $\epsilon^*_{ji} = \hat{\epsilon}_{ji} V_{ji}$, where $V_{11}, \ldots, V_{1n_1}, V_{21}, \ldots, V_{2n_2}$ are zero mean i.i.d. random variables that are independent from the two samples [see e.g., page 257 of Mammen (1993)]. In this paper we consider the V_{ji}'s as i.i.d. random variables with masses $(\sqrt{5}+1)/(2\sqrt{5})$ and $(\sqrt{5}-1)/(2\sqrt{5})$ at the points $(1-\sqrt{5})/2$ and $(1+\sqrt{5})/2$ (note that this distribution satisfies $E(V_{ji}) = 0; E[V^2_{ji}] = E[V^3_{ji}] = 1$).

- Let T^*, the bootstrap version of the test statistic, be defined as in (4), with $p = 2$, based on the bootstrapped fractile regression curves obtained from the bootstrapped samples $\{Y^*_{j[i:n_j]}\}$. These computations are repeated N times (we have used $N = 2000$ in our numerical studies) to yield $\{T^*_i\}_{i=1}^N$.

	n_2	25	50	100	25	50	100	25	50	100
n_1	Model	μ_1 vs. μ_2			μ_1 vs. μ_3			μ_1 vs. μ_4		
25	1	0.044	0.044	0.058	0.595	0.695	0.767	0.796	0.887	0.937
50	1	0.053	0.048	0.046	0.697	0.773	0.837	0.871	0.931	0.966
25	2	0.043	0.033	0.022	0.478	0.560	0.649	0.762	0.852	0.918
50	2	0.043	0.035	0.036	0.615	0.678	0.788	0.857	0.924	0.964
25	3	0.042	0.047	0.066	0.153	0.188	0.250	0.319	0.435	0.547
50	3	0.052	0.046	0.050	0.182	0.237	0.303	0.417	0.567	0.697
25	4	0.054	0.040	0.034	0.069	0.071	0.079	0.150	0.183	0.259
50	4	0.081	0.046	0.034	0.107	0.089	0.081	0.220	0.261	0.330
25	5	0.037	0.050	0.061	0.347	0.432	0.530	0.815	0.921	0.947
50	5	0.051	0.052	0.060	0.432	0.589	0.687	0.926	0.978	0.988
25	6	0.050	0.036	0.029	0.184	0.201	0.236	0.581	0.698	0.773
50	6	0.068	0.043	0.030	0.238	0.272	0.335	0.724	0.782	0.865

Table 1: Rejection probabilities of a wild bootstrap version of the test for various sample sizes and the regression functions when the nominal level is $\alpha = 0.05$.

- We compare the observed difference between the fractile regressions (i.e., T_{n_1,n_2}) with the empirical distribution of the test statistic T_i^*. The bootstrap estimate of the P-value is the proportion of times T_i^* exceeds the observed value of T_{n_1,n_2}.

5 Simulation Study

To investigate the finite sample properties of our testing procedure, we consider two samples obtained from the two regression models:

$$Y_{1i} = \mu_1(X_{1i}) + \epsilon_{1i}, \text{ and } Y_{2i} = \mu_j(X_{2i}) + \epsilon_{2i},$$

for $j = 2, 3, 4$. We study three kinds of increasing regression functions – one linear and two non-linear – with homoscedastic and heteroscedastic errors. We consider the following regression models with $X_{1i} \overset{iid}{\sim} Exp(1)$, $X_{2i} \overset{iid}{\sim} Exp(1)$, and $\epsilon_{1i} \overset{iid}{\sim} N(0, 0.09)$ for sample sizes $(n_1, n_2) = (25, 25), (25, 50), (25, 50), (50, 25), (50, 50), (50, 100)$:

1. $\mu_1(x) = \mu_2(x) = 1$; $\mu_3(x) = 1 + 0.5x$; $\mu_4(x) = 1 + 2.0x$; $\epsilon_{2i} \overset{iid}{\sim} N(0, 0.09)$.
2. $\mu_1(x) = \mu_2(x) = 1$; $\mu_3(x) = 1 + 0.5x$; $\mu_4(x) = 1 + 2.0x$; $\epsilon_{2i} \sim N(0, 0.09X_{2i})$.
3. $\mu_1(x) = \mu_2(x) = e^{-x}$; $\mu_3(x) = e^{-1.5x}$; $\mu_4(x) = e^{-2.0x}$; $\epsilon_{2i} \overset{iid}{\sim} N(0, 0.09)$.
4. $\mu_1(x) = \mu_2(x) = e^{-x}$; $\mu_3(x) = e^{-1.5x}$; $\mu_4(x) = e^{-2.0x}$; $\epsilon_{2i} \sim N(0, 0.09X_{2i})$.
5. $\mu_1(x) = \mu_2(x) = \sqrt{x + 1.0}$; $\mu_3(x) = \sqrt{x + 1.5}$; $\mu_4(x) = \sqrt{x + 2.0}$; $\epsilon_{2i} \overset{iid}{\sim} N(0, 0.09)$.
6. $\mu_1(x) = \mu_2(x) = \sqrt{x + 1.0}$; $\mu_3(x) = \sqrt{x + 1.5}$; $\mu_4(x) = \sqrt{x + 2.0}$; $\epsilon_{2i} \sim N(0, 0.09X_{2i})$.

	n_2	25	50	100	25	50	100	25	50	100
n_1	Model	μ_1 vs. μ_2			μ_1 vs. μ_3			μ_1 vs. μ_4		
25	1	0.005	0.008	0.010	0.345	0.472	0.624	0.645	0.781	0.8720
50	1	0.004	0.004	0.009	0.568	0.698	0.781	0.790	0.882	0.952
25	2	0.002	0.003	0.002	0.211	0.286	0.391	0.558	0.698	0.831
50	2	0.004	0.004	0.004	0.391	0.517	0.644	0.720	0.847	0.933
25	3	0.006	0.008	0.012	0.037	0.068	0.084	0.120	0.204	0.298
50	3	0.010	0.005	0.009	0.060	0.096	0.149	0.188	0.357	0.505
25	4	0.006	0.006	0.007	0.011	0.007	0.022	0.035	0.068	0.117
50	4	0.011	0.006	0.007	0.024	0.019	0.022	0.078	0.090	0.145
25	5	0.005	0.010	0.014	0.126	0.191	0.303	0.601	0.786	0.882
50	5	0.005	0.008	0.011	0.199	0.330	0.492	0.808	0.930	0.975
25	6	0.004	0.004	0.003	0.054	0.067	0.082	0.365	0.461	0.586
50	6	0.012	0.006	0.005	0.070	0.105	0.136	0.503	0.620	0.737

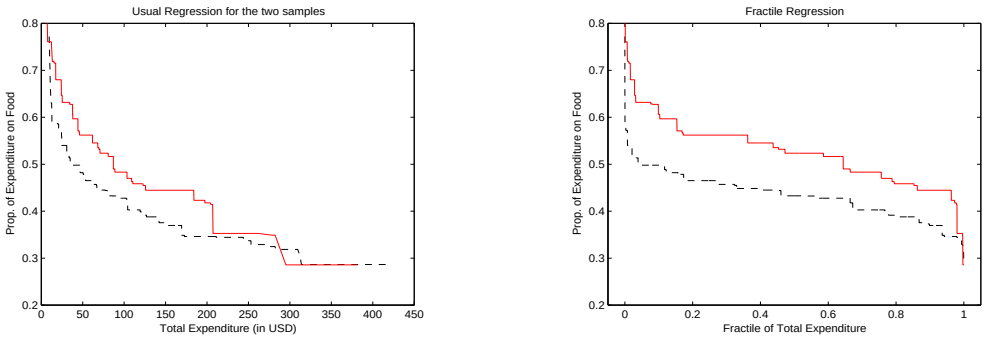
Table 2: Rejection probabilities of a wild bootstrap version of the test for various sample sizes and the regression functions when the nominal level is $\alpha = 0.01$.

We use the wild bootstrap methodology to generate the bootstrap samples (as discussed in Section 4). Table 1 shows the rejection probabilities of the wild bootstrap version of the test (i.e., the power of the statistical test) for various sample sizes and the three different regression functions when the nominal level of the test (α) is fixed at 0.05. We see that for testing μ_1 versus μ_2, the estimated rejection probabilities are quite close to 0.05, and as we move further away from the null hypothesis, the probabilities very rapidly move towards 1. Note that as the regression functions in models 3 and 4 are quite similar the power of the test is comparatively low. Table 2 shows the rejection probabilities of the test for the same models when $\alpha = 0.01$. The test seems to be slightly conservative as most of the rejection probabilities are less than 0.01 when the null hypothesis is true, but this could also be an artifact of the relatively small number of replications consider. Note that to construct both the tables we generate 2000 bootstrap samples per data set (to compute the cut-off value for the test statistic) and repeat the analysis for 2000 data sets to compute the rejection probabilities.

6 Some Real Data Illustrations

In this section, we describe three examples involving real data that illustrate the scope and usefulness of FGA. We analyze the data sets as par the methodology developed in the paper using isotonic methods to estimate the fractile regression functions. See Sen and Chaudhuri (2010) for more details and another analysis of these examples using smooth estimates of fractile regression.

Example 1: The Household Expenditure and Income Data for Transitional Economies (HEIDE) database contains data from household survey maintained by the World Bank Group; and it includes four countries in Eastern Europe and the former So-



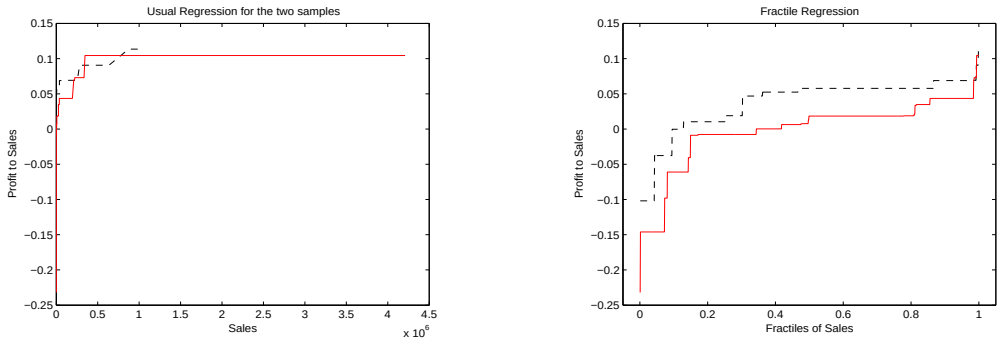
(a) (b)

Figure 2: *(a) Usual regression curves, and (b) the estimated fractile regression functions, for proportion of expenditure on food on total expenditure for Poland (in red, solid line) and Bulgaria (in black, dashed line).*

viet Union (see http://www.worldbank.org/ for more information). It was created as part of a project analyzing poverty and existing social assistance programs in the transitional economies. What immediately arrests attention is the startling drop in income and increase in inequality accompanying the transition of these countries to market economies. We investigate this inequality in income and compare the economic condition of the transitional economies.

A simple measure of the economic well-being of a population can be taken as the proportion of expenditure on food as a fraction of total expenditure per capita per household (in USD). This proportion would be quite small for rich and wealthy people, but for the poor it would be close to one. Thus it is known a priori that the regression functions are decreasing. And by regressing this proportion on the total expenditure, we can get a fair idea of the inequality in income and the economic condition of the populations.

We consider data sets for two countries from the HEIDE database, namely Poland (with 16051 data points) and Bulgaria (with 2466 data points), and estimate the regression functions. Figure 2 shows the estimates of the usual regression functions and that of the fractile regression curves with proportion of expenditure on food as the response and total expenditure per capita per household (in USD) as the predictor. Both the regression curves in Figure 2a clearly show the decreasing trend as expected. The ranges of the covariates are somewhat different in the two populations even though both of them are measured in USD. This might be partly because the data for the two populations were collected at different time points (Jan-Jun 1993 for Poland and Jan-Jun 1995 for Bulgaria). It might also be partly due to the disparity in purchasing powers of 1 USD in the two countries at two different time points. The crossing of the two regression functions for large covariate values is also disturbing. To make the regression curves comparable, we need some standardization of the covariates.

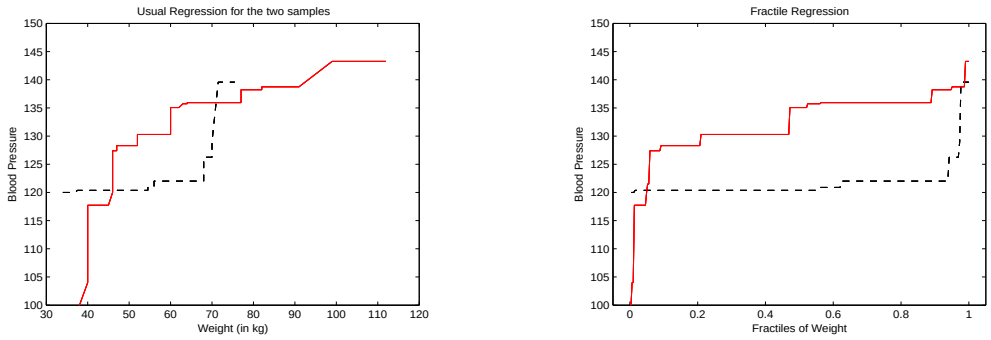


(a) (b)

Figure 3: *(a) Usual regression curves, and (b) the estimated fractile regression functions, for profit (as a fraction of sales) against sales for the years 1997 (in dashed black) and 2003 (in solid red).*

We would really like to compare the mean proportion of food expenditure for the poor (or the rich) in one population with that of the poor (or the rich) in the other population. The fractile curves accomplish exactly this, enabling us to compare the mean response values for fixed percentiles of total expenditure. The transformed covariate values close to 0 correspond to the very poor people and values close to 1 correspond to the richest people in the populations if we take total expenditure as a measure of economic condition. In Figure 2b, the two fractile regression functions are properly aligned and it appears that the condition of households in Poland is uniformly economically better than those of Bulgaria. A formal test of hypothesis using the resampling procedure outlined in Section 4 yields a P-value very close to 0.

Example 2: The Reserve Bank of India keeps data on the sales (in Indian rupees) and profit (as a fraction of sales) for non-government, non-financial public limited companies in India over different years. The Reserve Bank of India is interested in comparing the profitability of the companies against sales, at two time points. This gives rise to a regression problem where one regresses profit (as a fraction of sales) against sales. One would like to compare the two regression functions corresponding to two time points. But the comparison of usual regression functions is not meaningful as, due to inflation and other economic changes over time, the covariate values at two different time points happen to differ by several orders of magnitude. Figure 3a shows the usual regression functions for the year 1997 (dashed black) and 2003 (red solid) with 944 and 1243 data points respectively while Figure 3b shows the corresponding fractile regression functions. The uneven covariate distribution leads to data sparsity in certain regions of the covariate space. Besides, the large difference in the covariate values for the years 1997 and 2003 makes the two regression functions virtually incomparable in Figure 3a. The estimated fractile regression functions clearly show that the two functions are not equal and the P-value obtained is 0.001.



(a) (b)

Figure 4: *(a) Usual regression curves, and (b) the estimated fractile regression functions, for blood pressure against weight for the Bhutia (in dashed black) and Toto (in solid red) tribes.*

Example 3: The usefulness of FGA is not only restricted to financial/economic data as is illustrated in this example. Data were collected on 258 individuals from the Bhutia tribe and 305 individuals from the Toto tribe in India on blood pressure and weight by the scientists of the Human Genetics Unit at Indian Statistical Institute, Kolkata. It is of interest to compare the relationship of blood pressure with the weight of an individual for the two populations. For example, a biologist might want to compare the mean blood pressure for the two tribes with median weight for the two populations. Such comparisons involving the notion of quantiles can be accomplished by studying the fractile regression functions. We know from various scientific considerations that the regression functions will be increasing in this case, and the shape restricted function estimators for the two populations are plotted in Figure 4a. The two usual regression functions are not comparable as the covariates have very different distributions in the two populations. In fact, the ranges of the covariates are quite different. The crossing of the two regression functions is also disturbing. But the fractile regression functions in Figure 4b adequately resolve these comparability issues. From the fractile regression function, it can be easily seen that for the Bhutia tribe, blood pressure remains almost constant over the entire domain of the weight variable, a feature not very apparent in Figure 4a. A formal comparison of the two fractile regression functions yields a P-value very close to 0.

7 Concluding Remarks

The comparison of two regression functions, when the distribution of the covariates in the two populations are different, arises quite often in statistics and is the central issue of FGA. In this paper, we have developed resampling based hypothesis testing procedures to compare the fractile regression curves using their isotonized

nonparametric estimators. Our procedure does not depend on the choice of any tuning parameter, a major disadvantage of most of the earlier methods available in the literature. In course of our research, we revisit some of Mahalanobis's results and provide a brief history of FGA. Our approach can also be extended to compare k fractile regression curves, for $k \geq 2$. We end with a brief discussion on some open problems in this area.

The proposed procedure is computationally simple and has satisfactory finite sample performance. But the theoretical validity of the method is not investigated adequately in the paper, although some heuristics are provided in Section 4. Indeed, it is mathematically challenging and beyond the scope of the present article. It will be a topic of future research. Throughout this paper we assumed that there is prior knowledge on the shape (increasing/decreasing) of the regression function. This helped us to use the isotonized estimators that are free from tuning parameters. In certain applications, such restrictions might not be known, and might not be very appropriate. In such situations, the comparison of the fractile curves is problematic and requires further investigation. Sen (2005) discusses some of the issues in this setup, and proposes hypothesis testing procedures that depend on certain tuning parameters. A thorough theoretical study of the performance of these proposed tests is unknown and would be an interesting problem for future research.

A natural extension of Mahalanobis's ideas is to investigate FGA when the dimension of the covariate can be greater than one. Such an extension is not immediate, as in multi-dimension, there is no unique concept of rank or quantile. Mahalanobis (1988) provided some ideas in this direction. Sen and Chaudhuri (2010) consider this problem in greater detail and discuss examples that arise in diverse applications. But they focus mainly only on the estimation of fractile regression. A formal comparison of the fractile functions in such a setup is an open problem and deserves attention.

References

[1] BARLOW, R.E., BARTHOLOMEW, D., BREMNER, J.M. and BRUNK, H.D. (1972). *Statistical inference under order restrictions; the theory and application of isotonic regression*, Wiley, New York.

[2] BERA, A.K. and GHOSH, A. (2006). Fractile regression and its applications. Technical report. University of Illinois at Urbana-Champaign.

[3] BHATTACHARYA, P.K. and MULLER, H.G. (1993). Asymptotics for Nonparametric Regression. *Sankhyā, Ser.A*, **53**, 420-441.

[4] BRUNK, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.) 177195. Cambridge Univ. Press.

[5] DAVID, H.A. and NAGARAJA, H.N. (2003). *Order Statistics, 3rd edn.* Wiley, New York.

[6] de LEEUW, J., HORNIK, K. and MAIR, P. (2009). Isotone optimization in R: Pool-adjacent-violators (PAVA) and active set methods, *J. Statist. Software*, **32**, 1-24.

[7] DUROT, C. (2007). On the L_p-error of monotonicity constrained estimators. *Ann. Statist.*, **35**, 1080-1104.

[8] DELGADO, M.A. (1993). Testing the Equality of Nonparametric Regression Curves. *Statist. Probab. Lett.*, **17**, 199-204.

[9] EFRON, B. and TIBSHIRANI, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

[10] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Application*. Chapman and Hall, London.

[11] GROTZINGER, S. J. and WITZGALL, C. (1984). Projections onto simplices. *Appl. Math. Optimiz.*, **12**, 247-270.

[12] HALL, P. and HART, J.D. (1990). Bootstrap Test for Difference between Means in Nonparametric Regression. *J. Amer. Statist. Assoc.*, **85**, 1039-1049.

[13] HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

[14] HALL, P. (2003). A Short Prehistory of the Bootstrap. *Statist. Sc.*, **18**, 158-167.

[15] HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

[16] HERTZ-PICCIOTTO, I. and DIN-DZIETHAM, R. (1998). Comparisons of infant mortality using a percentile-based method of standardization for birthweight or gestational age. *Epidemiology*, **9**, 61–67.

[17] IYENGAR, N.S. and BHATTACHARYA, N. (1965). Some Observations on Fractile Graphical Analysis. *Econometrica*, **33**, 644-645.

[18] KULASEKERA, K.B. (1995). Comparison of Regresion Curves Using Quasi Residuals. *J. Amer. Statist. Assoc.*, **90**, 1085-1093.

[19] KULASEKERA, K.B. and WANG, J. (1997). Smoothing Parameter Selection for Power Optimality in Testing of Regression Curves. *J. Amer. Statist. Assoc.*, **92**, 500-511.

[20] MAHALANOBIS, P.C. (1960). A Method for Fractile Graphical Analysis. *Econometrica*, **28**, 325-351.

[21] MAHALANOBIS, P.C. (1988). *Fractile Graphical Analysis*, (Editor: P.K. Bose). Statistical Publishing Society, Calcutta.

[22] MAMMEN, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *Ann. Statist.*, **21**, 255-285.

[23] MITROFANOVA, N.M. (1961). On Some Problems of Fractile Graphical Analysis. *Sankhyā, Ser.A*, **23**, 145-154.

[24] MONTES-ROJAS, G.V. (2010). Nonparametric Estimation of ATE and QTE: An Application of Fractile Graphical Analysis. Technical report (Available at http://www.staff.city.ac.uk/ sbbc685/FGAATEQTE.pdf).

[25] MULLER, H.G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer-Verlag, Berlin.

[26] MUNK, A. and DETTE, H. (1998). Nonparametric Comparison of Several Regression Functions: Exact and Asymptotic Theory. *Ann. Statist.*, **26**, 2339-2368.

[27] NEUMEYER, N. and DETTE, H. (2003). Nonparametric Comparison of Regression Curves: An Empirical Process Approach. *Ann. Statist.*, **31**, 880-920.

[28] NORDHAUS, W.D. (2006). Geography and Macroeconomics: New Data and New Findings. *Proc. Natl. Acad. Sci.*, **103**, 3510–3517.

[29] PARTHASARATHY, K.R. and BHATTACHARTYA, P.K.(1961). Some Limit Theorems in Regression Theory. *Sankhyā, Ser.A*, **23**, 91-102.

[30] RICE, J.A. (1984). Bandwidth Choice for Nonparametric Regression. *Ann. Statist.*, **12**, 1215-1230.

[31] ROBERTSON, T., WRIGHT, F.T. and DYKSTRA, R.L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

[32] SEN, B. (2005). Estimation and Comparison of Fractile Graphs Using Kernel Smoothing Techniques. *Sankhyā*, **67**, 305–334.

[33] SEN, B. and CHAUDHURI, P. (2010). On Fractile Transformation of Covariates in Regression. (submitted).

[34] SETHURAMAN, J. (1961). Some Limit Distributions Connected with Fractile Graphical Analysis. *Sankhyā, Ser.A*, **23**, 79-90.

[35] STONE, C.J. (1977). Consistent Nonparametric Regression. *Ann. Statist.*, **5**, 505-545.

[36] SWAMY, S. (1963). Notes on Fractile Graphical Analysis. *Econometrica*, **31**, 551-554.

[37] TAKEUCHI, K. (1961). On Some Properties of Error Area in the Fractile Graph Method. *Sankhyā, Ser.A*, **23**, 65-78.

[38] WAND, M.P. and JONES, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

[39] WATSON, C.S. (1964). Smooth Regression Analysis. *Sankhyā, Ser.A*, **26**, 359-372.