FRACTILES ON QUANTILE REGRESSION WITH APPLICATIONS

ANIL K. BERA, AUROBINDO GHOSH, AND ZHIJIE XIAO

Abstract. This year celebrates the 50th aniversary of Fractile Graphical Analysis proposed by Prashanta Chandra Mahalanobis (Mahalanobis, 1961) in a series of papers and seminars as a method for comparing two distributions controlling for the rank of a covariate through fractile groups.We revisit the technique of fractile graphical analysis with some historical perspectives. We propose a new non-parametric regression method called Fractile on Quantile Regression where we condition on the ranks of the covariate, and compare it with existing quantile regression techniques. We apply this method to compare mutual fund inflow distributions after conditioning on returns.

JEL Classification: C12, C14, C52

Date: January 31, 2011.

1991 *Mathematics Subject Classification.* Primary 62G08, 62G20, 62G30, Secondary 62E20, 62P20, 91B28.

Key words and phrases. Non-parametric regression, Fractile Graphical Analysis, rank regression, quantile regression, Gini coefficient, concentration curves.

Preliminary version, please do not quote without author's consent.

1. Motivation and Background

Fractile Graphical Analysis was proposed by Prashanta Chandra Mahalanobis (Mahalanobis, 1961) in a series of papers and seminars as a method to take into account the effect of a covariate while comparing two distributions. Unlike the parametric method of linear least squares regression analysis Mahalanobis proposed a more non-parametric way of controlling the covariates (possibly, more than one) using the ranks of "fractile" groups (possibly unequal). The method provides a graphical tool for comparing both complete distributions of the variable of interest (like income or expenditure) for all values of the covariate as well as specific fractiles. Mahalanobis used a visual method of approximating the standard error of the income at all the fractiles of the covariate for the same graph by taking two independently selected "interpenetrating subsamples" and obtaining a graph for each of the subsamples besides the combined sample. The method proposed by Mahalanobis for estimating the error area of a fractile graph was later hailed as a precursor to the genesis of latter day bootstrap methodology (Efron 1979a,b; Hall, 2003). FGA can used to test whether two distributions of the fractile graphs of two populations are different by looking at the "Area of Separation" between the two graphs.

It is worth mentioning that the fractile graphs are a more general version of the Lorenz concentration curve and more specific concentration curves where we look at the cumulative relative sums of the levels of the variable of interest (for example expenditure or income) in place of the actual values. Hence, FGA can be used to compare the error in estimating Lorenz curves or specific concentration curves. The main contribution of the Fractile Graphical Analysis were twofold, first it provided a method of using interpenetrating network of subsamples to estimate the error region and perform a simple graphical test of the whole or a range of values of the fractiles where the distributions are different (see the discussion in Swami, 1963 and Iyengar and Bhattacharya, 1965). The point raised in Swami(1963) that FGA was a novel way of looking at the age-old problem of concentration curves and Gini Coefficient is also misleading as FGA provides a method of comparison of same fractiles over different points of time or region, as well as, specific ranges of fractiles or the entire distribution.

Mahalanobis used Fractile Graphical Analysis as an instrument for evaluation of standard of living over different periods of time (for example, total consumption of households between the Eighth Round in July 1954-March 1955, and the Sixteenth Round, July 1960-June 1961 of National Sample Survey; see Srinivasan, 1996) that could be subsequently used for recommending policy variables. From a pure economic perspective, if we want to compare different groups of people with different levels of consumption of goods or services, we must assume that the relative prices of goods with respect to a numeraire are fixed. If the relative price changes so does the real income of individuals, the percentiles of individuals by income groups will be different for different relative prices. For the above mentioned example in the eighth round of NSS when the prices were low and the sixteenth round of NSS when the prices were high the fractile graphs were completely separated (that is, there is significant statistical difference between the real total consumption expenditure), with the fractile graph for the eighth round being closer to the line of equal distribution (the 45^0 line) However, a reverse thing happened when he looked at the specific concentration curve for a particular foodgrain consumption, with the 16th round fractile graph for the consumption of cereals or cereals being closer to the line of equal distribution. This can easily be explained using the fact that the relative price of cereals actually reduced, hence even though the price of cereals increased the poorer section of the population had a upward effect on their cereal consumption instead of the other commodities (substitution effect), this in turn increased their real income (income effect).

In a more current context, there has been quite a lot of discussion owing to the increase in life expectancy or the increase in the proportion of the aging population on whether raising the retirement age in some developed countries is a good idea? One major issue in this problem is that the age distribution of the population has been moving as well, so any solution to this issue must take into account a more robust measure of actual age groups like the ranks of age groups. Now, we can compare the different groups of income earners after controlling for the rank classes of age using a technique like Fractile Graphical Analysis.

The paper in this preliminary form is arranged in the following way. In the ensuing Section 2 we provide a brief historical background, and some perspectives about the development of this idea by Professor Mahalanobis. In Section 3, we introduce the original theory of Fractile Graphical Analysis as proposed by Mahalanobis with some propositions and conjectures. The relationship between FGA and Concentration curves is discussed with a proposition in Section 4. In Section

5 we re-introduce the concept of a nonparametric rank regression technique called *Fractile Regression* and discuss its relationship with existing non-parametric and semiparametric methods. Section 6 is devoted to finite sample and asymptotic properties of Fractile Regression estimates in the light of related work on *induced order statistics* or *concomitant variables* to the order statistics. We look at an illustrative empirical example in Section 7 on the inflow distribution of mutual funds conditional on returns. We conclude in Section 8.

2. Historical Background of Fractile Graphical Analysis

The genesis of Fractile Graphical Analysis was not as accidental as Mahalanobis' introduction to the field of applied statistics while answering the call of Professor Brajendra Nath Seal to do a statistical analysis of examination data for the University of Calcutta in 1917 ([26]). Professor Mahalanobis' study on anthropometric data on the racial inheritance of Anglo-Indians under the influence of Professor N. Annandale (then director of the Zoological and Anthropological Surveys of India and the Chairperson of Bangalore Session of the Indian Science Congress in 1924) led to the first serious statistical work of the Cambridge trained physicist in 1922. I think the initiation of Mahalanobis's thought on the decomposition of variation due to the natural statistical deviation and that due to measurement error came from his work with Sir Gilbert Walker (then Director-General of Observatories, also one of the coinventor of the Yule-Walker Normal Equations for AR(p) processes while studying factors affecting atmospheric phenomenon like the Southern Oscillations, later linked to El Niño) on atmospheric data on upper atmosphere. The first indication of seminal work on Mahalanobis D^2 came from anthropometric data analysis on the "Analysis of the Race Mixture of Bengal," presented as a part of his address in the Benaras Indian Science Congress in 1925. His work on the distribution of probable errors in agricultural experimental designs later known as the Fisherian methods of field experiments (after Professor R.A. Fisher, with whom Mahalanobis came in touch with owing to his research on field experiments) made him look deeper into the procedure of removing the effect of soil heterogeneity as a possible cause of variation in crop yields using non-linear "graduating curves" (a method used by Jerzy Neyman several years later in 1937 formulating the "Smooth test of Goodness-of-Fit"). As the elected president of the Indian Science Congress Association in Pune, 1950, his Presidential Address was titled "Why Statistics?" His pioneering effort to introduce statistical thinking in a Third World developing country just three years after gaining independence depicts how advanced Mahalanobis must have been for his time.

The foundation of the Indian Statistical Institute on December 17, 1931 was a culmination of the activities of Mahalanobis and his associates at the Statistical Laboratory in the University of Calcutta, and some leading academics (like Professor K.B. Madhava, Professor of Mathematical Economics and Statistics at Mysore, Minto Professor of Economics, Pramatha Nath Banerjee, Khaira Professor of Applied Mathematics, and P.C. Mahalanobis himself being a Professor of Physics, and other dignitaries) who felt the need for a society devoted to the advancement of statistics in India a discipline to solve both socioeconomic as well as fundamental scientific problems through the analysis of data. The institute took up research in prices of Indian commodities with respect to other economic factors under economists like Dr. H.C. Sinha. Professors Raj Chandra Bose and Samarendra Nath Roy worked on the derivation of the exact distribution of the generalized distance D^2 that measures the divergence between two populations developed by Mahalanobis in 1925 and later papers. One of the focus of this line of research was the identification, classification and discrimination in terms of variances and covariances of different populations.[1]

Mahalanobis had attracted a lot of renowned researchers like H. Wold of Upsala (Stationary Time Series) and Abraham Wald to conduct research collaborations in Indian Statistical Institute in late 1940's and early 1950's. After independence,

[1] In his appraisal of the role of the Indian Statistical Institute, Prof. R.A. Fisher noted, almost prophetically,

> "In regard to the future of statistical studies in India, at present it would seem that everything depends on the future of the Statistical Institute. This holds a key position, for the reason that India is rich in men capable of making a good showing on the basis of a university education, but poorly off in respect of those having a thorough technical grasp of what can in practice be done. Consequently, the mere creation of posts for statisticians followed by filling them with applicants possessing abundantly plausible credentials, would merely perpetuate an imitative adherence to the obsolete methods of older text books, and would stand in the way of real advances. There are, on the contrary, many young Indians capable of responding to the dual discipline of sound mathematical training, followed by practical, responsibly conducted research, in which sound judgment can be acquired, and the real meaning of mathematical studies brought to surface."

in 1949, Mahalanobis began to work in New Delhi as the Honorary Statistical Advisor to the Cabinet, Gverenment of India, and as Chairman of the Committee of Central Statisticians. Dr. Pitambar Pant who joined the Institute in 1946 as a scientific secretary to Professor Mahalanobis under the patronage of the first prime minister of independent India, Pandit Jawaharlal Nehru, was later instrumental in the creation of the Central Statistical Organization (CSO) in February 1951 which since its inception had very close links with the Indian Statistical Institute. As the chairman of the National Income Committee that was set up in 1949-50 (other members being Professors D.R. Gadgil and V.K.R.V. Rao), Professor Mahalanobis recommended to Pandit Nehru large scale sample surveys to fill in gaps in national statistics. This led to the creation of the National Sample Survey (NSS) in 1950 responsible for collecting comprehensive information annually pertaining to social, economic and demographic characteristics of both rural and urban sectors in two different rounds. Fully aware of the possibility of data corruption due to negligence and measurement errors Mahalanobis introduced the method of Interpenetrating Network of Subsamples (IPNS) at all stages of the processing of NSS data.

Independence in India brought with it the unique problem of staggeringly high unemployment and diminutive national income, Mahalanobis was given the arduous task of proposing a plan lowering the unemployment rate and at the same time doubling national income. The Planning Commission and the Department of Economic Affairs, Ministry of Finance, with the help of the Indian Statistical Institute and CSO designed the Second Five Year Plan on the basis of the draft Plan-frame of perspective planning proposed by Mahalanobis in March 1955. Economic and statistical research on underdeveloped economies were conducted in collaboration with researchers like Trevor Swan (Australia) and I.M.D. Little (U.K.) who came to India as part of research team from Centre for International Studies, MIT, USA, in the Planning Unit of the Indian Statistical Institute in Delhi. Several noted economists who have been affiliated with the planning unit in Delhi include Nobel Laureate Amartya Sen and the likes of Professors Jagdish Bhagwati, B. Minhas, A. Rudra and others. In the Calcutta centre of I.S.I. research was carried out in various fields in Applied Economics and Econometrics like growth models, input-output tables, estimation and use of expenditure elasticities for demand projection, setting up of a macro-econometric model of the Indian Economy, studies on national income and allied topics, trend and level of consumption in India, labor productivity and growth during the British period. In particular, there was sustained research on the preparation of a series on national income using survey of consumer expenditure under researchers like Ajit Biswas, H.K. Mazumdar, A. Rudra, A.K. Chakrabarti, I.B. Chatterjee, G.S. Chatterjee, S. Naqvi, N.K. Chatterjee, N. Bhattacharyya and N.S. Iyengar. A macro-econometric model was developed for the Indian economy under the supervision of Professor Gerhard Tintner and several studies were carried out on the time trend on the level and distribution of consumption in India and methods like Fractile Graphical Analysis was extensively used to control for covariates.[2]

It is indeed surprising that despite the presence of several noted economist, an applied statistician and physicist, Mahalanobis was entrusted with formulating the Draft Frame of the Second Five Year Plan with the main objective of eradicating poverty and unemployment in India based on a two-sector planning model. He created a number of study groups to examine specific economic and social problems like Jogobroto. Roy was in charge of the committee who investigated the impact of increase in income on consumer behavior. Dr. N. Bhattacharya, Dr. M. Mukherjee and Dr. J. Roy tirelessly worked along with others on the analysis of data from National Sample Survey on the sampling experiments for the first paper on Fractile Graphical Analysis (Mahalanobis 1958, 1962). In his role as the chairman of the Income Distribution Committee of the Government of India, even at an advanced age of 70, Professor Mahalanobis relentlessly worked through the night analyzing data (Bhattacharya in ISI newsletter, "Lekhon", 1997)

[2] Noted sociologist, Dr. Ramakrishna Mukherjee, reminisces an incident with Professor Mahalanobis in his tribute in the occasion of the 50th Anniversary of India's independence in the ISI newsletter "Lekhon" (1997, [27]),

> Professor Mahalanobis, a little morose over some agitation of workers at ISI, wondered,"Ramakrishna after my death what will happen to ISI? Would anything survive of what I created?" I (Ramakrishna Mukherjee) replied,"Professor, your Large Sample Survey will survive, D^2 will live, Fractile Graphical Analysis, though I don't quite understand it, will survive if its useful, and some students of yours will spread your message."
>
> Professor's eyes lit up, "And ISI?" I replied," What are you saying. It will become a University department."
>
> He stayed silent for a while contemplating, then replied, "Rabindranath used to say this is a riverine land, nothing survives in this climate for too long. To think of it, a country that assimilated Buddha and Rabindranath without a trace, there who am I to expect a legacy?"

J.K. Ghosh noted in his tribute to the contribution of Professor Mahalanobis on the occasion of the 50th Anniversary of India's independence in the ISI club newsletter ("Lekhon", 1997)

> In India's national life Professor Mahalanobis and ISI have three major contributions. First, to put India in the world map in the field of statistics-in no other branch of science is research in India so deep and far reaching. Second, to open horizons in scientific research in India in fields like Demography, Physical Anthropology, Paleantology and Sedimentary Geology, Computer and Computing Science etc. Third, concrete steps in Analytical Planning, construction of Statistical Databases, National Income Accounts measurement and distribution could be attributed to his efforts. Nowadays, although the methods of Planning has changed, but almost everyone agrees that Mahalanobis did not commit an error by emphasizing on the development of heavy industries in India. Economic development is not possible without these foundations.

Mahalanobis introduced the forward looking Harrod-Domar type two sector model for growth and development (later expanded it to a more realistic four sector model) of the Indian Economy where the state had to make direct investments to infrastructure building heavy industries, this investment was widely supported by practitioners and academicians alike. This need was particularly felt in a document circulated as the Bombay Plan published by leading industrialists who opined that "...the government of independent India should be in a better position to mobilize resources for investment on the large basic industries. as this would be beyond the capabilities of the private enterprise "(D.K. Bose, in Science Society and Planning).

3. Theory of Fractile Graphical Analysis

The exposition of this part of the paper is largely based on the work of Sethuraman (1961). Suppose, we have n pairs of observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) that are independently drawn from a population of the random variables (X, Y). Further, suppose we rank the observations with respect to the covariate x and define the series of indices $(i_1, i_2, ..., i_n)$ such that $x_{i_1} = x_{(1)}$, $x_{i_2} = x_{(2)}$, ..., $x_{i_n} = x_{(n)}$, hence $x_{i_1} \leq x_{i_2} \leq ... \leq x_{i_n}$. So we can write the data as $\left(x_{(1)}, y_{[1]}\right)$, $\left(x_{(2)}, y_{[2]}\right)$, ...,

$\left(x_{(n)}, y_{[n]}\right)$. We divide the data into m groups of size g each i.e. $n = mg$. Each of the group means of the variables ranked with respect to X is obtained, define

$$u_i = \frac{1}{m} \sum_{r=(i-1)m+1}^{im} x_{(r)}, i = 1, 2, ..., g \tag{3.1}$$

and

$$v_i = \frac{1}{m} \sum_{r=(i-1)m+1}^{im} y_{[r]}, i = 1, 2, ..., g. \tag{3.2}$$

We obtained two random samples (x_1^1, y_1^1) , (x_2^1, y_2^1) , ..., (x_n^1, y_n^1) and (x_1^2, y_1^2) , (x_2^2, y_2^2) , ..., (x_n^2, y_n^2) independently from the population P^{12}, hence the combined sample (x_1^{12}, y_1^{12}) , (x_2^{12}, y_2^{12}) , ..., $(x_{2n}^{12}, y_{2n}^{12})$ is also an independent sample of size $2n$ from the same population P^{12}. Similarly, we can obtain two random samples (x_1^3, y_1^3) , (x_2^3, y_2^3) , ..., (x_n^3, y_n^3) and (x_1^4, y_1^4) , (x_2^4, y_2^4) , ..., (x_n^4, y_n^4) independently from the population P^{34}, hence the combined sample (x_1^{34}, y_1^{34}) , (x_2^{34}, y_2^{34}) , ..., $(x_{2n}^{34}, y_{2n}^{34})$ is also an independent sample of size $2n$ from the same population P^{34}. We can define from equations (3.1) and (3.2) the group means $\left(v_1^1, v_2^1, ..., v_g^1\right)$, $\left(v_1^2, v_2^2, ..., v_g^2\right)$ of group size m and $\left(v_1^{12}, v_2^{12}, ..., v_g^{12}\right)$ of group size $2m$ from the samples drawn from population P^{12}. Similarly, define from equations (3.1) and (3.2) the group means $\left(v_1^3, v_2^3, ..., v_g^3\right)$, $\left(v_1^4, v_2^4, ..., v_g^4\right)$ of group size m and $\left(v_1^{34}, v_2^{34}, ..., v_g^{34}\right)$ of group size $2m$ from the samples drawn from population P^{34}. Let G^1,G^2 and G^{12} be the plots of the g group means $\left(v_1^1, v_2^1, ..., v_g^1\right)$, $\left(v_1^2, v_2^2, ..., v_g^2\right)$ and $\left(v_1^{12}, v_2^{12}, ..., v_g^{12}\right)$ against the group ranks 1 through g (See Figure FGAPlot.) Also define, for population P^{34}, G^3,G^4 and G^{34} be the plots of the group means $\left(v_1^3, v_2^3, ..., v_g^3\right)$, $\left(v_1^4, v_2^4, ..., v_g^4\right)$ and $\left(v_1^{34}, v_2^{34}, ..., v_g^{34}\right)$ against the covariate group ranks 1 through g. Continuing with some notations, define A_{12} be the *error area* bounded by fractile graphs G^1 and G^2 between the rank points of the covariate x, 1 and g; A_{34} be the *error area* bounded by graphs G^3 and G^4 between the rank points of the covariate x, 1 and g; and A_* be the *separation area* bounded between the combined graphs G^{12} and

G^{34}.

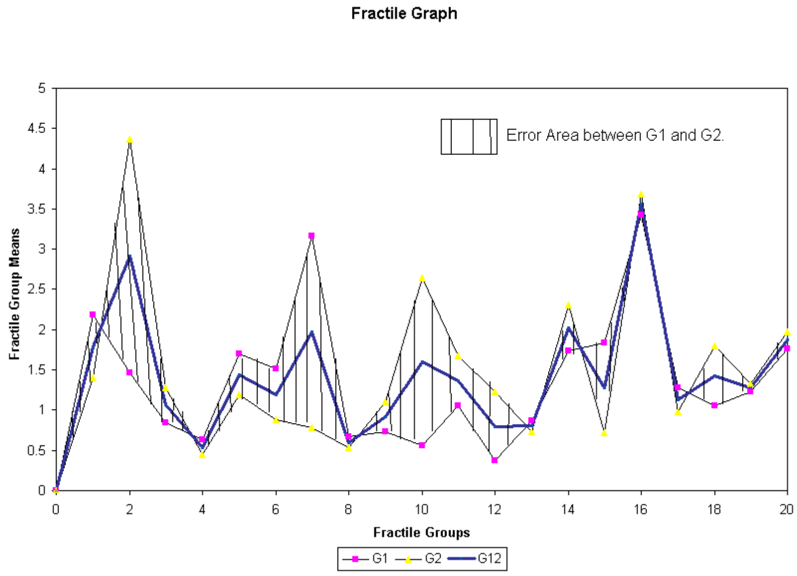


Figure 1: Fractile Graph

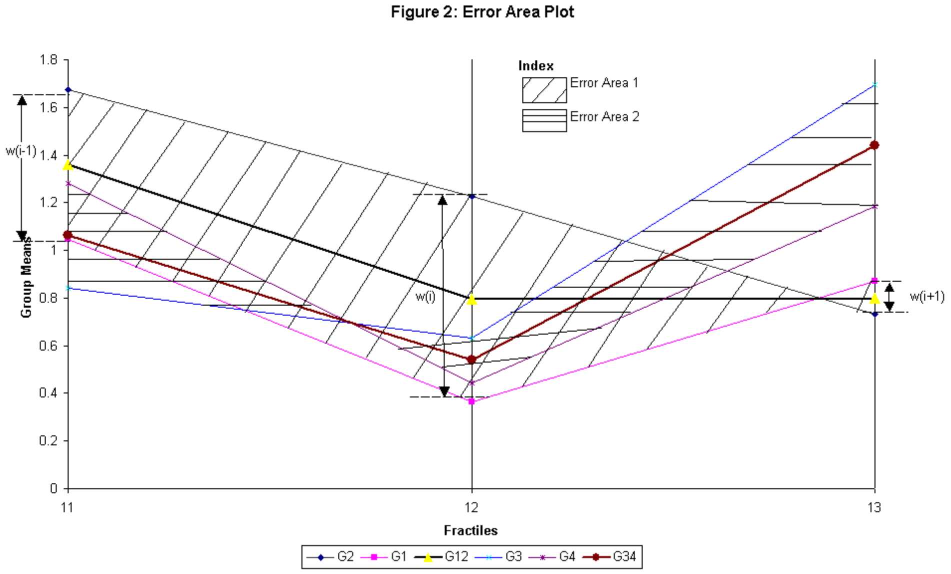


Figure 2: Error Area and Separation

Our first objective is to find out some analytical expressions for areas A_{12}, A_{34} and A_*. Noting that A_{12} and A_{34} would be similar, we focus on the area A_{12}, without loss of generality. Let us further define the following quantities of difference of means in the two groups.

$$\begin{aligned} v_i^1 - v_i^2 &= w_{i(12)}, \quad i = 1, 2, ..., g \\ v_i^3 - v_i^4 &= w_{i(34)}, \quad i = 1, 2, ..., g \\ v_i^{12} - v_i^{34} &= w_{i(*)}, \quad i = 1, 2, ..., g \end{aligned} \tag{3.3}$$

We can divide the area between G^1 and G^2 i.e. A_{12} into each constituent area between the ordinates i and $i+1$, say, $A_{12(i)}$. Let us summarize the construction of the area as the following Proposition 1.

Proposition 1. *(Takeuchi 1961) The* error area *bounded by graphs G^1 and G^2 is $A_{12} = \sum_{i=1}^{g-1} A_{12(i)}$ where*

$$\begin{aligned} A_{12(i)} &= \frac{1}{2}\left(|w_i| + |w_{i+1}|\right) - \partial\left(w_i, w_{i+1}\right)\frac{|w_i w_{i+1}|}{|w_i| + |w_{i+1}|} \\ \textit{where } \partial\left(a, b\right) &= \begin{cases} 0 & if \quad ab \geq 0 \\ 1 & if \quad ab < 0 \end{cases} \end{aligned} \tag{3.4}$$

Proof. **Case 1:** G^1 and G^2 does not cross between i and $i+1$, that means that $w_i w_{i+1} \geq 0$.

Area$\left(A_{i(12)}\right)$ is essentially that of a trapezoid between parallel vertical lines at i and $i+1$. Hence,

$$A_{i(12)} = \frac{1}{2}\left(|w_i| + |w_{i+1}|\right) \text{ if } w_i w_{i+1} \geq 0. \tag{3.5}$$

Note that, we can define $w_0 = 0$, so $A_{0(12)} = \frac{1}{2}\left(|w_1|\right)$. We can easily verify that (3.4) holds here.

Case 2: G^1 and G^2 does cross between i and $i+1$, that means that $w_i w_{i+1} < 0$.

Area$\left(A_{i(12)}\right)$ is essentially that of two equiangular triangles with their bases on the vertical lines on i and $i+1$, or in other words the bases are unit distance apart. Using the property of equiangular triangles, after construction, we can see that the altitude (or rise) of the triangles would be proportional to the base (or run). Since the bases of the triangles are $|w_i|$ and $|w_{i+1}|$ respectively, if x is the altitude of the triangle with base $|w_i|$ we observe

$$\frac{|w_i|}{|w_{i+1}|} = \frac{x}{1-x} \Rightarrow x = \frac{|w_i|}{|w_i| + |w_{i+1}|} \text{ or } 1 - x = \frac{|w_{i+1}|}{|w_i| + |w_{i+1}|}. \tag{3.6}$$

Using (3.6),

$$\begin{aligned}
A_{i(12)} &= \frac{1}{2}x\left|w_i\right| + \frac{1}{2}\left(1-x\right)\left|w_{i+1}\right| \\
&= \frac{1}{2}\frac{\left|w_i\right|^2 + \left|w_{i+1}\right|^2}{\left|w_i\right| + \left|w_{i+1}\right|} \\
&= \frac{1}{2}\frac{\left(\left|w_i\right| + \left|w_{i+1}\right|\right)^2 - 2\left|w_i\right|\left|w_{i+1}\right|}{\left(\left|w_i\right| + \left|w_{i+1}\right|\right)} \\
&= \frac{1}{2}\left(\left|w_i\right| + \left|w_{i+1}\right|\right) - \frac{\left|w_i w_{i+1}\right|}{\left|w_i\right| + \left|w_{i+1}\right|} \\
&= \frac{1}{2}\left(\left|w_i\right| + \left|w_{i+1}\right|\right) - \partial\left(w_i, w_{i+1}\right)\frac{\left|w_i w_{i+1}\right|}{\left|w_i\right| + \left|w_{i+1}\right|} \\
\text{where } w_i w_{i+1} &< 0 \qquad (3.7)
\end{aligned}$$

Note that, if $w_i = 0$, then $\left|w_i w_{i+1}\right| = 0$, so case 1 or case 2 would both work. □

One way of addressing the problem of the difference between two fractile graphs G^1 and G^2 is to look at a norm in a $g-$dimensional Euclidean space. The $\mathcal{L}_2-$norm can be defined as

$$\begin{aligned}
\triangle_{12} &= \left\|G^1 - G^2\right\| \\
&= \left\|\left(v_1^1, v_2^1, ..., v_g^1\right) - \left(v_1^2, v_2^2, ..., v_g^2\right)\right\| \\
&= \left\|\left(v_1^1 - v_1^2, v_2^1 - v_2^2, ..., v_g^1 - v_g^2\right)\right\| \\
&= \sqrt{w_{1(12)}^2 + w_{2(12)}^2 + ... + w_{g(12)}^2} \qquad (3.8)
\end{aligned}$$

Similarly, one can define $\triangle_{34} = \sqrt{w_{1(34)}^2 + w_{2(34)}^2 + ... + w_{g(34)}^2}$between G^3 and G^4, and finally, $\triangle_*$ between the combined graphs G^{12} and G^{34}. Suppose, $B = ((b_{ij}))$ is a positive definite matrix, then we can further define a more general class of distance measure as

$$\Gamma_{12}^2 = \sum_{i=1}^{g}\sum_{j=1}^{g} w_{i(12)}w_{j(12)}b_{ij} \qquad (3.9)$$

between the samples over the entire range of values. In particular the following Proposition is provided in Sethuraman (1961).

Proposition 2. *If (3.8) represents the distance between fractile graphs G^1 and G^2, and A_{12} represents the area between the two, then*

$$\frac{\triangle}{6} \leq A_{12} \leq \frac{\triangle}{\sqrt{g}}. \tag{3.10}$$

Proof. See Sethuraman (1961). □

3.1. **Asymptotic Distributions of the Dispersion Measures in FGA.**

(1) $m\Delta_{in}^2$ converges to a mixture of χ^2 variates, while $m\Gamma_{in}^2$ with a suitably chosen normalization matrix B converges to χ^2 with g degrees of freedom.

(2) For appropriate B, $E\,(\Gamma_{in}^2) \simeq g/m$, in general. Furthermore, $E\,(\Delta_{in}^2) \simeq$constant$(g/m)$ and $E\,(\varepsilon_{in}) \simeq$constant$(g/\sqrt{m})$ if (X, Y) is bivariate normal.

(3) $m\Delta_{in}^2$,$i = 1, 2$ and $2m\Delta_{*n}^2$ are asymptotically independent, so

$$\frac{2\Delta_{*n}^2}{(\Delta_{1n}^2 + \Delta_{2n}^2)} \rightarrow \text{Ratio of mixture of } \chi^2$$

Similarly, for a suitable normalization matrix B,

$$\frac{2\Gamma_{*n}^2}{(\Gamma_{1n}^2 + \Gamma_{2n}^2)} \rightarrow F_{g,2g}.$$

(4) The concentration ratios Σ_{in} are asymptotically normal.

4. Fractile Graphical Analysis and Concentration Curves

One aspect in which we can view Fractile Graphical Analysis is to look at it as a novel approach to the construction of Lorenz concentration curves (Swamy, 1963) of the variable of interest like income or expenditure (Y) or specific concentration curves of a particular item like expenditure on foodgrain. Let us define a variable z on the cumulative value of v, for each of the g fractile groups

$$z_0 = 0,\ z_i = \frac{v_1 + v_2 + ... + v_i}{v_1 + v_2 + ... + v_g},\ i = 1, 2, ..., g. \tag{4.1}$$

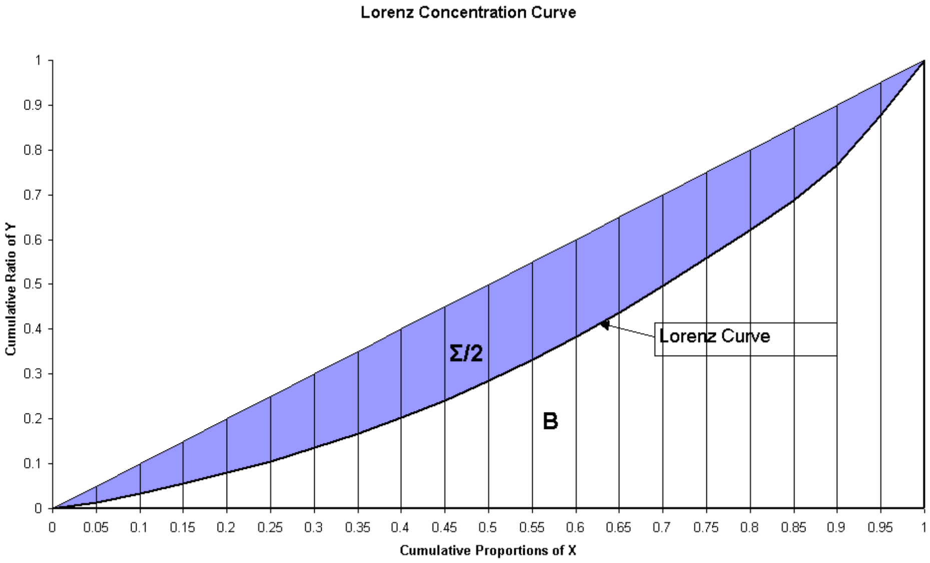


Figure 3: Concentration Curve

Proposition 3. *The concentration ratio,* $\sum$*, is twice the area between the concentration curve* C *and the line joining* $(0,0)$ *and* $(1,1)$ *, or*

$$\sum = 2\frac{\sum_{i=1}^{g} iv_i}{g\sum_{i=1}^{g} v_i} - \frac{g+1}{g}.$$

Proof. We can divide the area under the Lorenz curve (say, B) into g sub-areas, one for each of the fractiles, so $B = \sum_{k=1}^{g} B_k$. Further, let us define z_k as in (4.1), $k = 0, 1, 2, ..., g$. From Figure ConcCurve, we can see that the areas B_k are trapezoids between the vertical lines at $k-1$ and k on the x-axis, hence as $z_g = 1$,

$$\begin{aligned}
\sum_{k=1}^{g} B_k &= \sum_{k=1}^{g} \frac{1}{2}\left(z_{k-1}+z_k\right)\frac{1}{g} \\
&= \frac{1}{2g}\left[2\left(z_1+z_2+\ldots+z_{g-1}\right)+z_g\right] \\
&= \frac{1}{g}\sum_{k=1}^{g-1} z_k + \frac{1}{2g} \\
&= \frac{1}{g\sum_{i=1}^{g} v_i}\sum_{i=1}^{g-1}\sum_{k=i}^{g-1} v_i + \frac{1}{2g} \\
&= \frac{1}{g\sum_{i=1}^{g} v_i}\sum_{i=1}^{g}\left(g-i\right)v_i + \frac{1}{2g} \\
&= \frac{1}{g\sum_{i=1}^{g} v_i}\left(\sum_{i=1}^{g} g v_i - \sum_{i=1}^{g} i v_i\right) + \frac{1}{2g} \\
&= 1 - \frac{1}{g\sum_{i=1}^{g} v_i}\sum_{i=1}^{g} i v_i + \frac{1}{2g} \\
&= 1 + \frac{1}{2g} - \frac{\sum_{i=1}^{g} i v_i}{g\sum_{i=1}^{g} v_i} \qquad (4.2)
\end{aligned}$$

This means that the concentration ratio

$$\begin{aligned}
\sum &= 2\left[\frac{1}{2} - \sum_{k=1}^{g} B_k\right] \\
&= 1 - 2\left[1 + \frac{1}{2g} - \frac{\sum_{i=1}^{g} i v_i}{g\sum_{i=1}^{g} v_i}\right] \\
&= \frac{2\sum_{i=1}^{g} i v_i}{g\sum_{i=1}^{g} v_i} - \frac{g+1}{g}. \qquad (4.3)
\end{aligned}$$

□

Lorenz curves or more generally specific concentration curves are the most obvious generalizations of fractile graphical analysis method to measures of inequality or dispersion. This tool has been extensively used by authors like N. S. Iyengar to estimate Engel curves and Engel elasticities from survey data (Iyengar, 1960, 1964). We would also see later in how the area under the specific concentration curve weakly converges to the function of a standard Brownian motion process and a convolution of a independent Brownian Bridge process.

5. Fractile Regression

Our objective in this proposal is to look at the age-old problem of the effect of the covariates on distributions. Linear regression has always been the cornerstone of such an analysis where we investigate at the effects of the x-variables or covariates on the response variable y. A very simple example of that could be the effect of educational qualification measured in years of education on income or future income. It could be argued that educational qualification is a proxy for ability, hence higher educational qualification would lead to higher earning. However, performing simple linear regression on this somewhat naive model of "Returns to Education" misses some major parts of the story. First, the story of endogeneity, that is to say that it is very rare that education is randomly assigned, so individuals choose education based on their ability and opportunity cost. Hence it would be wrong to assign the credit of higher income solely to education, there could quite a few omitted variables. In fact, the error term ε in the population linear regression model, i.e.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where y is, say, log of income and x is the number of years of education, β_0 and β_1 are the partial regression coefficients, might be correlated with the independent variable x - problem often times referred to as "endogeneity" in Econometrics. However, the problem we are trying to address is not directly related to endogeneity, but the other aspect of the story missed by simple linear regression. It is very likely that people with high ability or high educational qualification might command a much higher salary for one extra year of education compared with someone with low ability or education. Linear regression fails to capture this "differential" treatment of the covariates or in particular "fractiles of the covariates." So instead of looking at regression of y on x we should be looking at the regression of Y grouped according to fractiles of X, i.e., we can answer the question for the bottom 10% of educational qualification in the society what is the effect of one more year of education all else remaining the same. This really brings us to the classical problem of non-parametric regression analysis. Let me briefly describe three very close neighbors in the field of regression analysis.

In non-parametric (Kernel-based) regression analysis we consider $Y_i \sim N\left(m\left(x_i\right), \sigma^2\right)$, $i = 1, 2, ..., n$, where conditional mean function $m\left(.\right)$ satisfies some regularity or smoothness conditions. Broadly, we can define the Nadaraya-Watson type location or regression estimator with the smoothing kernel $K\left(.\right)$ and bandwidth h as

$$\hat{m}_{NW}\left(x_o\right) = \arg\min_{\beta_o \in \mathcal{R}} \sum_{i=1}^{n} \left(y_i - \beta_o\right)^2 K\left(\frac{x - x_i}{h}\right) = \sum_{i=1}^{n} W_{in}^{NW}\left(x\right) y_i \qquad (5.1)$$

We can think of replacing x_i by a monotonic rank-score of x_i and use the weighted least squares type method as well. "Bandwidth" can be defined either in terms of actual width (kernel type) or the number of observations (nearest neighbor type). In Nearest Neighbor type regression estimate we replace x by the empirical distribution function $F_n\left(x\right)$ in Equation (5.1) to get

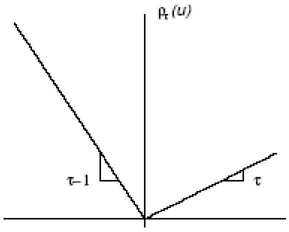
$$\hat{m}_{NN}\left(x_o\right) = \arg\min_{\beta_o \in \mathcal{R}} \sum_{i=1}^{n} \left(y_i - \beta_o\right)^2 K\left(\frac{F_n\left(x\right) - F_n\left(x_i\right)}{h_n}\right) = \sum_{i=1}^{n} W_{in}^{NN}\left(x\right) y_i. \qquad (5.2)$$

The major advantage that k-Nearest Neighbor type estimator has over the traditional kernel based estimator is that the former only depends on the ranks of $X_1, X_2, ..., X_n$. Hence, if $F\left(x\right)$ is continuous the problem gets transformed to much more tractable problem of estimating a regression function at $F\left(x_0\right)$ with the X-sample being uniformly distributed over $\left[0, 1\right]$. Its convergence properties in mean square has also been studied by Yang (1981). Stute (1984) showed that k-Nearest Neighbor type estimates are asymptotically normal if $E\left[Y^2\right] < \infty$, is much weaker than the conditions needed for existence of the Nadaraya-Watson type regression estimates like existence of the PDF $f\left(.\right)$ of X and that $E\left|Y\right|^3 < \infty$ (Schuster 1972).

In quantile regression, we look at the regression counterpart of univariate τ^{th} quantile of the dependent variable y is defined as

$$\hat{\alpha}\left(\tau\right) = \arg\min_{a \in \mathcal{R}} \sum_{i=1}^{n} \rho_\tau\left(y_i - a\right), \qquad (5.3)$$

where $\rho_t\left(u\right) = \left(\tau - I\left(u < 0\right)\right) u$ is often referred to as the check function.



τ^{th} Regression Quantile of Y on X (Koenker and Bassett, 1978)

$$\hat{\beta}(\tau) = \arg\min_{b\in\mathcal{R}^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^T b\right) \tag{5.4}$$

It should be noted that quantile regression controls for the quantiles of the y variable, and not of the original covariate i.e. the x variable.

To motivate for fractile regression let's think of a regression function of Y on $X = x$ as

$$m(x) = E[Y|X = x] \tag{5.5}$$

Let $F(x)$ is the marginal cumulative distribution function (CDF) of X with a density function (PDF) $f(x)$.

We can show that the regression function is invariant under a strictly monotonic transformation of the covariate X to its probability integral transform (PIT), $F(x)$,

$$\begin{aligned} r(u) &= E[Y|F(X) = u] \\ &\Rightarrow r(u) = E\left[Y|X = F^{-1}(u)\right] = m\left(F^{-1}(u)\right) \end{aligned} \tag{5.6}$$

If we want to find out the partial regression coefficients of $r(u)$ is given by

$$\frac{\partial r(u)}{\partial u} = \frac{\partial m}{\partial x}.\frac{\partial F^{-1}(u)}{\partial u} = m'(x)\frac{1}{f(x)}, \tag{5.7}$$

where we divide the non-parametric regression coefficients by the density function evaluated at x. One interpretation of that could be the regression coefficients are weighted less where the density of the covariate is low. As we can imagine now, that FGA is not just the "Prehistory of Bootstrap" (Hall 2003) but the "Prehistory of Inference on Non-parametric Regression" as well.

6. Properties of Fractile Regression Function and their ranks

Suppose X has a distribution function $\mathrm{F}(x)$, the conditional distribution of Y given $X = x$ is $G\left(y|X = x\right) = G_x$.

Lemma 1. *(Bhattacharya '74) If X has a continuous distribution function $F\left(.\right)$, the induced order statistics $Y_{[1]}, Y_{[2]}, ..., Y_{[n]}$ are conditionally independent given $X_1, X_2, ..., X_n$with conditional CDFs $G_{x_{(1)}}, G_{x_{(2)}}, ..., G_{x_{(n)}}$.*

Proof. Define $\mathbf{X} = \left(X_1, X_2, ..., X_n\right)'$, be a multivariate random variable and $\mathbf{x} = \left(x_1, x_2, ..., x_n\right)' \in \boldsymbol{R}^n$. Now consider a mapping $\lambda\left(k, \mathbf{X}\right) = j$ if $X_{(k)} = X_j \Rightarrow \lambda\left(k, \mathbf{X}\right) = j$ if $Y_{[j]} = Y_j$ (we can call $\lambda\left(k, \mathbf{X}\right)$ an order statistic index finder function.) Immediately, we can see that for $\mathbf{X}$ almost everywhere $\lambda\left(k, \mathbf{X}\right)$ is uniquely defined for all $k \in \left\{1, 2, 3, ..., n\right\}$.

$$\begin{aligned} & P\left(Y_{[1]} \leq y_{j_1}, Y_{[2]} \leq y_{j_2}, ..., Y_{[n]} \leq y_{j_n} |\mathbf{X} = \mathbf{x}\right) \\ = \ & P\left(Y_{\lambda(1,\mathbf{x})} \leq y_{j_1}, Y_{\lambda(2,\mathbf{x})} \leq y_{j_2}, ..., Y_{\lambda(n,\mathbf{x})} \leq y_{j_n} |\mathbf{X} = \mathbf{x}\right) \\ = \ & P\left(Y_{j_1} \leq y_{j_1}, Y_{j_2} \leq y_{j_2}, ..., Y_{j_2} \leq y_{j_n} |\mathbf{X} = \left(x_1, x_2, ..., x_n\right)'\right) \\ = \ & P\left(Y_{j_1} \leq y_{j_1} |X_{j_1} = x_{j_1}, Y_{j_2} \leq y_{j_2} |X_{j_2} = x_{j_2}, ..., Y_{j_2} \leq y_{j_n} |X_{j_n} = x_{j_n}\right) \\ = \ & P\left(Y_{j_1} \leq y_{j_1} |X_{j_1} = x_{j_1}\right) P\left(Y_{j_2} \leq y_{j_2} |X_{j_2} = x_{j_2}\right) ...P\left(Y_{j_2} \leq y_{j_n} |X_{j_n} = x_{j_n}\right) \\ = \ & G_{x_{j_1}}\left(y_{j_1}\right) G_{x_{j_2}}\left(y_{j_2}\right) ...G_{x_{j_n}}\left(y_{j_n}\right) \end{aligned}$$

since $\left(X_i, Y_i\right)', i = 1, 2...., n$ are independent and identically distributed random variables that have a continuos distribution function. In general, for $k = 1, 2, ..., n$,this argument will go through for any permutation $\lambda\left(k, \mathbf{X}\right)$ of $\left\{1, 2, ..., n\right\}$. □

Theorem 1. *Let X have a continuous distribution $F\left(x\right)$, and the distribution of Y given $X = x$ is also continuous and denoted by $G_x\left(y\right)$. Then given any $1 \leq r_1 < r_2 < ... < r_k \leq n$ for any fixed $k \leq n$,*

$$P\left[Y_{[r_i]} \leq y_i, 1 \leq i \leq k\right] = E\left[\prod_{i=1}^{k} G_{x_{(r_i)}}\left(y_i|X_{(r_i)} = x_{(r_i)}\right)\right]. \tag{6.1}$$

Proof. Applying lemma 1 we can immediately see that conditional on $\mathbf{X} = \mathbf{x}$ i.e.,$\left(X_1, X_2, ..., X_n\right)' = \left(x_1, x_2, ..., x_n\right)'$,

$$P\left[Y_{[r_i]} \leq y_i, 1 \leq i \leq k|\mathbf{X} = \mathbf{x}\right] = \prod_{i=1}^{k} G_{x_{(r_i)}}\left(y_i|X_{(r_i)} = x_{(r_i)}\right). \tag{6.2}$$

The result of the theorem follows from an application of the law of iterated expectation. □

Theorem 2. *(David, O'Connell and Yang '77) Under the above regularity conditions, where in $\theta_1 = P(X < x, Y < y)$, $\theta_2 = P(X < x, Y > y)$, $\theta_3 = P(X > x, Y < y)$ and $\theta_4 = P(X > x, Y > y)$, defining $R(.)$ as the rank function,*

$$\begin{aligned}
& P\left[R\left(Y_{[r]}\right) = s\right] \\
= \; & n\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_{k=0}^{\min(r-1,s-1)} \\
& \frac{(n-1)!}{k!\,(r-1-k)!\,(s-1-k)!\,(n-r-s+1+k)!} \\
& \times\theta_1^k\theta_2^{r-1-k}\theta_3^{s-1-k}\theta_4^{n-r-s+1+k}dG\left(y|x\right)dF\left(x\right). \qquad (6.3)
\end{aligned}$$

Proof. Our objective is to find out the probability distribution (or probability mass function) of the rank of the r^{th} induced order statistic (or r^{th} concomitant variable to the order statistics of X) when we have an *iid* sample (X_i, Y_i) $i = 1, 2, ..., n$ from a continuous bivariate distribution function $F(x, y)$ i.e. to find $P\left(Rank\left(Y_j\right) = s|Rank\left(X_j\right) = r\right) = P\left[R\left(Y_{[r]}\right) = s\right]$. We can illustrate the problem using the following table

	$X_{(r)} < x$	$X_{(r)} > x$	$Row\ Total$
$Y_{(s)} < y$	k	$s-1-k$	$s-1$
$Y_{(s)} > y$	$r-1-k$	$n-r-s+1+k$	$n-s$
$Col.\ Total$	$r-1$	$n-r$	$n-1$

where $k \in \{0, 1, 2, ..., \min(r-1, s-1)\}$.
Hence, if t=$\min(r-1, s-1)$ given $X = x$ and $Y = y$,

$$\begin{aligned}
P\left[R(Y_{[r]}) = s\right] \;=\; & \sum_{k=0}^{t}\frac{(n-1)!}{k!\,(s-1-k)!\,(r-1-k)!\,(n-r-s+1+k)!}\times \\
& \theta_1^k\theta_2^{r-1-k}\theta_3^{s-1-k}\theta_4^{n-r-s+1+k}.
\end{aligned}$$

Since there are n such variables Y, the unconditional probability for any value of $X = x$ and $Y = y$

$$P\left[R(Y_{[r]}) = s\right] = n\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_{k=0}^{t}\frac{(n-1)!}{k!\,(s-1-k)!\,(r-1-k)!\,(n-r-s+1+k)!}\times$$
$$\theta_1^k\theta_2^{r-1-k}\theta_3^{s-1-k}\theta_4^{n-r-s+1+k}dG_x\left(y\right)dF\left(x\right).$$

□

Theorem 3. *(Yang '77) Suppose the marginal distribution of X has a density function $f\left(x\right)$ that is bounded away from 0 in a neighborhood of $F^{-1}\left(\lambda_i\right), i = 1, 2, ..., k$. Also assume that the conditional probability distribution of $Y|X = x$, G_x at $y_1, y_2, ..., y_n$ are continuous at x.Then for $1 \leq r_1 < r_2 < ... < r_k \leq n$ such that $\frac{r_i}{n} \to \lambda_i \in (0,1)$ as $n \to \infty$,*

$$\lim_{n\to\infty} P\left[Y_{[r_i]} \leq y_i, 1 \leq i \leq k\right] = \prod_{i=1}^{k} G\left(y_i|F^{-1}\left(\lambda_i\right)\right). \tag{6.4}$$

Proof. This is an extension of the Theorem 1 in case the sample size diverges to ∞. Suppose for some fixed $k = 0, 1, 2, ..., n$ $\frac{r_i}{n} \to \lambda_i \in (0,1)$ as $n \to \infty$, then for an absolutely continuous distribution function $F\left(.\right)$, $F^{-1}\left(\frac{r_i}{n}\right) \to F^{-1}\left(\lambda_i\right)$ as $n \to \infty$. This follows from the convergence of continuous functions of quantiles of the Empirical Distribution Function (EDF) to the quantiles of the Cumulative Distribution Function (CDF) (see for example Rao(1973) page 423 6f.2(i).)
So as $n \to \infty$ and fixed k, using Theorem 1, and that $G_x\left(.\right)$ is a bounded continuous (conditional distribution) function

$$\lim_{n\to\to\infty} P\left[Y_{[r_i]} \leq y_i, 1 \leq i \leq k\right] = \lim_{n\to\infty}\prod_{i=1}^{k} G\left(y_i|F^{-1}\left(\frac{r_i}{n}\right)\right)$$
$$= \prod_{i=1}^{k} G\left(y_i|F^{-1}\left(\lambda_i\right)\right).$$

□

Theorem 4. *Suppose now that the joint distribution of (X, Y) say $p\left(x, y\right)$ is continuous and is bounded by some function $q\left(y\right)$ for all x in a neighborhood of $\xi_\lambda = F^{-1}\left(\lambda\right)$.Further suppose, the marginal density function $f\left(x\right)$ of X exists and is bounded away from 0 in a neighborhood of ξ_λ, $0 < \lambda < 1$. If $\frac{r}{n} \to \lambda$ as*

$n \to \infty$, and if we define $Z = G(Y)$ and $Z_\lambda = G(Y|F(X) = \lambda) = G_{F^{-1}(\lambda)}(Y)$ as the unconditional and the conditional probability integral transforms of Y, then

$$(i)\ \lim_{n\to\infty} E\left[\left(\frac{R\left(Y_{[r]}\right)}{n}\right)^k\right] = \int_{-\infty}^{\infty} G^k(y)\, dG\left(y|F^{-1}(\lambda)\right) = \int_{-\infty}^{\infty} Z^k dZ_\lambda \qquad (6.5)$$

$$(ii)\ \lim_{n\to\infty} P\left[R\left(Y_{[r]}\right) \leq na\right] = P\left[Z \leq a|F^{-1}(\lambda)\right] = P\left[Z_\lambda \leq a\right] \ \textit{for } 0 \leq a \leq 1. \qquad (6.6)$$

Proof. (i)This proof is based on results in David and Galambos (1974) and Yang (1977). For $I\{A\} = 1$, if A occurs, $I\{A\} = 0$ otherwise; and defining $\lambda(k, \mathbf{X}) = j$ if $Y_{[k]} = Y_j$, let's first consider an expression

$$\begin{aligned}\left[R\left(Y_{[r]}\right)\right]^k &= \left[\sum_{j=1}^{n} I\left\{Y_j \leq Y_{[r]}\right\}\right]^k \\ &= \sum_{j_1 \neq j_2 \neq \ldots \neq j_k \neq \lambda(r,\mathbf{X})} I\left\{Y_{j_1} \leq Y_{[r]}\right\} I\left\{Y_{j_2} \leq Y_{[r]}\right\} \ldots I\left\{Y_{j_k} \leq Y_{[r]}\right\} + O\left(n^{k-1}\right)\end{aligned}$$

which implies the k^{th} central moment of the rank of the r^{th} induced order statistic is

$$
\begin{aligned}
& E\left[\left(\frac{R\left(Y_{[r]}\right)}{n}\right)^{k}\right] \\
= & \frac{1}{n^{k}} \sum_{j_1 \neq j_2 \neq \ldots \neq j_k \neq \lambda(r,\mathbf{X})} P\left(Y_{j_1} \leq Y_{[r]}, Y_{j_2} \leq Y_{[r]}, \ldots, Y_{j_k} \leq Y_{[r]}\right) + O\left(n^{-1}\right) \\
= & \frac{1}{n^{k}} \sum_{l=1}^{n} \sum_{j_1 \neq j_2 \neq \ldots \neq j_k \neq l} P\left(Y_{j_1} \leq Y_{l}, Y_{j_2} \leq Y_{l}, \ldots, Y_{j_k} \leq Y_{l}, \operatorname{rank}\left(X_{l}\right) = r\right) + O\left(n^{-1}\right) \\
= & \frac{n\left(n-1\right)\ldots\left(n-k\right)}{n^{k}} P\left(Y_{1} \leq Y_{n}, Y_{2} \leq Y_{n}, \ldots, Y_{k} \leq Y_{n}, \operatorname{rank}\left(X_{n}\right) = r\right) + O\left(n^{-1}\right) \\
= & \frac{n\left(n-1\right)\ldots\left(n-k\right)}{n^{k}} \sum_{l=0}^{k} \binom{k}{l} \\
& \times P\left(\begin{array}{c} X_i \leq X_n, Y_i \leq Y_n, i = 1, 2, \ldots, l; \\ X_i > X_n, Y_i \leq Y_n, i = l+1, .., k; \\ \text{Exactly } \left(r-1-l\right) \ X_i \leq X_n, i = k+1, .., n \end{array}\right) + O\left(n^{-1}\right) \\
= & \frac{n\left(n-1\right)\ldots\left(n-k\right)}{n^{k}} \sum_{l=0}^{k} \binom{k}{l} \int_{-\infty}^{\infty} \left\{\int_{-\infty}^{\infty} p\left(x,y\right)^{l} \left(G\left(y\right) - p\left(x,y\right)\right)^{k-l} f\left(y|x\right) dy\right\} \times \\
& f_{r-l:n-k} dx + O\left(n^{-1}\right)
\end{aligned}
$$

where $f\left(y|x\right)$ is the density function of y given $X = x$, $f_{r-l;n-k}\left(.\right)$ is the density function of $X_{(r-l)}$ out of a possible $\left(n-k\right)$ $X's$. As n, r diverges to ∞, s.t. $\frac{r}{n} \rightarrow \lambda$, for fixed k, $X_{(r-l)}$ converges in probability to to λ. Hence, we get using binomial expansion

$$\begin{aligned}
\lim_{n\to\infty} E\left[\left(\frac{R\left(Y_{[r]}\right)}{n}\right)^k\right] &= \sum_{l=0}^{k}\binom{k}{l}\left\{\int_{-\infty}^{\infty} p\left(\lambda, y\right)^l\left(G\left(y\right)-p\left(\lambda, y\right)\right)^{k-l} f\left(y|\lambda\right) dy\right\} \\
&= \int_{-\infty}^{\infty}\left\{\sum_{l=0}^{k}\binom{k}{l}p\left(\lambda, y\right)^l\left(G\left(y\right)-p\left(\lambda, y\right)\right)^{k-l}\right\} f\left(y|\lambda\right) dy \\
&= \int_{-\infty}^{\infty}\left\{p\left(\lambda, y\right)+G\left(y\right)-p\left(\lambda, y\right)\right\}^k f\left(y|\lambda\right) dy \\
&= \int_{-\infty}^{\infty}\left\{G\left(y\right)\right\}^k dG\left(y|X=\lambda\right) \\
&= \int_{-\infty}^{\infty} Z^k dZ_\lambda
\end{aligned}$$

which proves result (i).

(ii) Part (i) essentially implies that the k^{th} moment of the normalized ranks of the r^{th} induced order statistics converges to the k^{th} moment of a uniformly distributed random variable conditional on the λ^{th} quantile. Hence, all the moments are continuous and bounded. This implies that the distribution of the normalized ranks are completely specified by its moments, hence for some $0 \leq a \leq 1$

$$\begin{aligned}
P\left[R\left(Y_{[r]}\right) \leq na\right] &= P\left[\left(\frac{R\left(Y_{[r]}\right)}{n}\right) \leq a|\ X_{(r)} = x_{(r)}\right] \\
&\Rightarrow \lim_{n\to\infty} P\left[R\left(Y_{[r]}\right) \leq na\right] = \lim_{n\to\infty} P\left[\left(\frac{R\left(Y_{[r]}\right)}{n}\right) \leq a|\ X_{(r)} = x_{(r)}\right] \\
&\Rightarrow \lim_{n\to\infty} P\left[R\left(Y_{[r]}\right) \leq na\right] = P\left[G\left(y\right) \leq a|X = F^{-1}\left(\lambda\right)\right] = P\left[Z_\lambda \leq a|\right].
\end{aligned}$$

□

This theorem simply implies that it is sufficient to work with the probability integral transforms of the Y variable after conditioning for the rank of the X variable.

Furthermore, the following corollary helps us to make inference on "fractile groups."

Corollary 1. *Suppose now there are two sequences* r *and* s *such that as* $n \to \infty$, $\frac{r}{n} \to \lambda_r$ *and* $\frac{s}{n} \to \lambda_s$ *where* $0 < \lambda_r < \lambda_s < 1$, *then for any* $0 \leq a \leq 1$ *and all any* i,

$$\begin{aligned}
&\lim_{n\to\infty} P\left[R\left(Y_i\right) \leq na|r \leq R(X_i) \leq s\right] \\
&= \frac{1}{\lambda_s - \lambda_r}\int_{\lambda_r}^{\lambda_s} P\left(Z \leq a|F^{-1}\left(\lambda\right)\right) d\lambda \\
&= \frac{1}{\lambda_s - \lambda_r}\int_{\lambda_r}^{\lambda_s} P\left(Z_\lambda \leq a\right) d\lambda. \qquad (6.7)
\end{aligned}$$

Proof. For any n,

$$\begin{aligned}
&P\left[R\left(Y_i\right) \leq na \mid r \leq R\left(X_i\right) \leq s\right] \\
&= \frac{n}{s-r+1}\sum_{j=r}^{s} P\left[R\left(X_i\right) = j\right] P\left[R\left(Y_i\right) \leq na \mid R\left(X_i\right) = j\right] \\
&= \frac{n}{s-r+1}\sum_{j=r}^{s} \frac{1}{n} P\left[R\left(Y_{[j]}\right) \leq na\right].
\end{aligned}$$

We also note that as $n \to \infty$, $\frac{r}{n} \to \lambda_r$ and $\frac{s}{n} \to \lambda_s$ where $0 < \lambda_r < \lambda_s < 1$, using Theorem 4

$$\begin{aligned}
\lim_{n\to\infty}\sum_{j=r}^{s}\frac{1}{n}P\left[R\left(Y_{[j]}\right) \leq na\right] &= \lim_{n\to\infty}\sum_{j=r}^{s}\frac{1}{n}P\left[Z \leq a \mid F^{-1}\left(\frac{j}{n}\right)\right] \\
&= \int_{\lambda_r}^{\lambda_s} P\left[Z \leq a \mid F^{-1}\left(\lambda\right)\right] d\lambda \\
&= \int_{\lambda_r}^{\lambda_s} P\left[Z_\lambda \leq a\ \right] d\lambda
\end{aligned}$$

using the Reimann Sum representation of an integral. □

6.1. **Asymptotics of Fractile Regression Analysis.** Let $R\left(t\right) = \int_0^t r\left(s\right) ds$, be the *Cumulative Fractile Regression function,* (Rao and Zhao, 1995) the

$$R_n\left(t\right) = \frac{1}{n}\sum_{j=1}^{[nt]} y_{[j]},. \qquad (6.8)$$

is an estimate of $R\left(t\right)$, where $[nt]$ is the largest integer less than or equal to nt. This term can be interpreted as a normalized partial (Reimann) sum that converges the area under the concentration curve (See Figure 3) upto point t, and is a measure of the *total variability or dispersion* in the induced order statistic Y among the lowest $100t\%$ of the population with respect to X.

Let the conditional variance of Y given $X = x$, be $\sigma^2(x) = Var(Y|X = x)$ .We can further define the *integrated volatility*

$$\psi(t) = \int_{-\infty}^{F^{-1}(t)} \sigma^2(x)\, dF(x) \tag{6.9}$$

with the sample counterpart as

$$\begin{aligned} \psi_n(t) &= \int_{-\infty}^{F_n^{-1}(t)} \sigma^2(x)\, dF_n(x) \text{ if } \frac{1}{n} \le t \le 1 \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned} \tag{6.10}$$

where $F_n(x)$ is the EDF of $X_1, X_2, ..., X_n$.

Lemma 2. *If $\sigma^2(x)$ is of bounded variation,*

$$\sup_{0 \le t \le 1} |\psi_n(t) - \psi(t)| \overset{a.s.}{\to} 0. \tag{6.11}$$

Proof. Using a change of variable of $U = F(X)$, and applying integration by parts on expression for integrated volatility,

$$\begin{aligned} \psi(t) &= \int_{-\infty}^{F^{-1}(t)} \sigma^2(x)\, dF(x) \\ &= \int_0^t \sigma^2\left(F^{-1}(u)\right) du \\ &= \sigma^2\left(F^{-1}(t)\right) t - \int_0^t u d\sigma^2. \end{aligned} \tag{6.12}$$

Now applying technique in equation (6.12) to equation (6.10),

$$\begin{aligned} \psi_n(t) &= \int_{-\infty}^{F_n^{-1}(t)} \sigma^2(x)\, dF_n(x) \\ &= \int_0^t \sigma^2\left(F_n^{-1}(u)\right) du \\ &= \sigma^2\left(F_n^{-1}(t)\right) t - \int_0^t u d\sigma^2. \end{aligned} \tag{6.13}$$

From equations (6.12) and (6.13),

$$\psi_n(t) - \psi(t) = \left(\sigma^2\left(F_n^{-1}(t)\right) - \sigma^2\left(F^{-1}(t)\right)\right) t. \tag{6.14}$$

Using result 6f.2(i) in Rao(1973, p. 423) that states if there exists a unique p^{th} quantile ξ_p i.e.$P\left[X \leq \xi_p\right] \geq p$ and $P\left[X \geq \xi_p\right] \geq (1-p) = q$, then if $\hat{\xi}_p$ is the p^{th} sample quantile,

$$P\left[\lim_{n\to\infty}\left|\hat{\xi}_p - \xi_p\right| = 0\right] = 1 \text{ or } \hat{\xi}_p \to \xi_p \text{ a.s.} \tag{6.15}$$

Using the Mean Value Theorem and bounded variation of $\sigma^2(.)$, defining $\xi_t = F^{-1}(t)$ and $\hat{\xi}_t = F_n^{-1}(t)$, for some ξ^* s.t. $\|\xi_t^* - \xi_t\| \leq \left\|\hat{\xi}_t - \xi_t\right\|$,

$$\begin{aligned}\sigma^2\left(\hat{\xi}_t\right) &= \sigma^2(\xi_t) + \left(\hat{\xi}_t - \xi_t\right)\left.\frac{d\sigma^2}{d\xi_t}\right|_{\xi_t=\xi_t^*} \\ &\Rightarrow \left|\sigma^2\left(\hat{\xi}_t\right) - \sigma^2(\xi_t)\right| \leq \left|\hat{\xi}_t - \xi_t\right| M, \qquad (6.16) \\ \text{where } \left.\frac{d\sigma^2}{d\xi_t}\right|_{\xi_t=\xi_t^*} &\leq M < \infty.\end{aligned}$$

Using equations (6.14) and (6.16),

$$\begin{aligned}\sup_{0\leq t\leq 1}\left|\psi_n(t) - \psi(t)\right| &= \sup_{0\leq t\leq 1}\left|\sigma^2\left(\hat{\xi}_t\right) - \sigma^2(\xi_t)\right| t \\ &\leq M \sup_{0\leq t\leq 1}\left|\hat{\xi}_t - \xi_t\right| \qquad (6.17)\end{aligned}$$

Using the continuity of $F(.)$, $F^{-1}(t) = \xi_t$ is uniquely defined, hence using equation (6.15) in (6.17) the result follows. □

Theorem 5. *(Bhattacharya 1977) Under some regularity conditions (viz., the continuity of* F*, bounded variability of* $\sigma^2(.)$*, and bounded fourth central moment of* Y*,* $\beta(x) = E\left[(Y - m(x))^4 | X = x\right] \leq B \in (0,\infty)$*),*

$$n^{\frac{1}{2}}\left[R_n(t) - R(t)\right] \Rightarrow \xi \circ \psi(t) + \int_0^t \eta(u)\, dr(u) \tag{6.18}$$

where ξ *and* η *are mutually independent standard Brownian motions and Brownian bridges respectively.*

Proof. Let us recall $m(x) = E[Y|X = x]$ is the conditional expectation or regression function of Y given X [see equation (5.5] and $r(u) = m \circ F^{-1}(u)$ is the fractile regression function at $F(X) = u$ [see equation (5.6)]. Let us first consider the case where F is unknown, so we use the empirical distribution function (EDF) of X.

$$\begin{aligned}
n^{\frac{1}{2}}\left[R_n(t)-R(t)\right] &= n^{-\frac{1}{2}}\sum_{i=1}^{[nt]}Y_{[i]}-n^{\frac{1}{2}}\int_0^t r(u)\,du \\
&= n^{-\frac{1}{2}}\sum_{i=1}^{[nt]}\left[Y_{[i]}-m\left(X_{(i)}\right)\right]+n^{-\frac{1}{2}}\sum_{i=1}^{[nt]}m\left(X_{(i)}\right)-n^{\frac{1}{2}}\int_0^t r(u)\,du \\
&= U_n-J_n, \\
\text{where } U_n &= n^{-\frac{1}{2}}\sum_{i=1}^{[nt]}\left[Y_{[i]}-m\left(X_{(i)}\right)\right] \\
\text{and } J_n &= n^{\frac{1}{2}}\left[\int_0^t r(u)\,du-n^{-1}\sum_{i=1}^{[nt]}m\left(X_{(i)}\right)\right] \qquad (6.19)
\end{aligned}$$

$$\begin{aligned}
n^{-1}\sum_{i=1}^{[nt]}m\left(X_{(i)}\right) &= n^{-1}\sum_{i=1}^{[nt]}m\left(F_n^{-1}\left(\frac{i}{n}\right)\right) \\
&= \int_0^t m\circ F^{-1}(u)\,d\left(F_n\circ F^{-1}(n)\right) \qquad (6.20) \\
&= \int_0^t r(u)\,dG_n(u), \\
\text{where } G_n &= F_n\circ F^{-1}.
\end{aligned}$$

Applying integration by parts on equations (6.19) and (6.20),

$$\begin{aligned}
\int_0^t r(u)\,dG_n(u) &= r(t)\,G_n(t)-\int_0^t G_n(u)\,dr(u), \\
\int_0^t r(u)\,du &= r(t)\,t-\int_0^t u\,dr(u). \qquad (6.21)
\end{aligned}$$

So,

$$\begin{aligned}
J_n &= n^{\frac{1}{2}} \left[\int_0^t r(u)\, du - n^{-1} \sum_{i=1}^{[nt]} m\left(X_{(i)}\right) \right] \\
&= n^{\frac{1}{2}} \left[\int_0^t \{G_n(u) - u\}\, dr(u) - r(t)\{G_n(t) - t\} \right] \\
&= n^{\frac{1}{2}} \left[\int_0^t V_n(u)\, dr(u) - r(t) V_n(t) \right], \text{ where} \\
V_n(u) &= \{G_n(u) - u\} \text{ and} \\
& \sup_{0 \le t \le 1} n^{\frac{1}{2}} V_n(t) \overset{a.s.}{\to} 0.
\end{aligned}$$

Under the given conditions, Bhattacharya (1974) showed that $U_n(t) \Rightarrow \xi \circ \psi$.Bhattacharya (1976) showed that under the same conditions with minor modifications like applying Skorohod's theorem conditional on $X_1, X_2, ..., X_n$ and using the triangle inequality on the integrated volatility

$$(U_n, V_n) \Rightarrow (\xi \circ \psi, \eta).$$

Since, $\phi(u,v)(t) = u(t) + \int_0^t v(s)\, dr(s)$ is a continuous function from the space differentiable continuous functions on $[0,1]$ of $\mathrm{D}^2[0,1]$ to continuous differentiable functions on $[a,b] \subset [0,1]$, we conclude the statement of the theorem on $[a,b]$. Similar result also holds when $F(.)$ is known. □

7. Illustrative Application of Fractile Regression on Comparing Distributions

There are several examples where we can use Fractile Graphical Analysis Techniques, and in particular, Fractile Regression methods. As discussed previously, male-female or younger-older workers wage gap with respect to returns to education; productivity gap between large and small firm productivity with respect to firm size;.difference on returns to equity with firm size; income distribution of different ethnic groups or countries with respect to age, etc.

For performing this test of comparison of distributions of we use the two sample version of Neyman (1937) smooth test procedure as proposed in Bera, Ghosh and Xiao (2004) (attached with this proposal for reference). Neyman's smooth test for $\mathbf{H}_0$: $\mathbf{F} = \mathbf{F}_0$. was for the one sample case with completely specified distribution under null hypothesis.$H_0 : f(x)$ is the true PDF (for review see Bera and Ghosh, 2001). This is equivalent to testing $H_0 : y = F(x) = \int\limits_{-\infty}^{x} f(u)\,du \sim U(0,1)$ .Neyman considered the following smooth alternative to the uniform density:

$$h(y) = C(\theta)\exp\left[\sum_{j=1}^{k}\theta_j\pi_j(y)\right] \tag{7.1}$$

$\pi_j(.)$ are orthogonal normalized Legendre polynomials. For $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k = 0$ has a test statistic

$$\Psi_k^2 = \sum_{j=1}^{k}\frac{1}{n}\left[\sum_{i=1}^{n}\pi_j(y_i)\right]^2 \sim \chi_k^2(0) \text{ under } H_0.$$

If we go problem of testing $H_0 : F = G$. We need to modify the original smooth test since both F and G are unknown. If $F(.)$ were known, we can construct a new random variable $Z_j = F(Y_j)$, $j = 1, 2, ..., m$.

The CDF of Z is given by

$$\begin{aligned} H(z) &= \Pr(Z \le z) = \Pr(F(Y) \le z) \\ &= G\left(F^{-1}(z)\right) = G(Q(z)) \end{aligned}$$

where $Q(z) = F^{-1}(z)$ is the quantile function of Z.

The PDF of Z is given by

$$\begin{aligned} h(z) &= \frac{d}{dz}H(z) = g\left(F^{-1}(z)\right)\frac{d}{dz}F^{-1}(z) \\ &= g\left(F^{-1}(z)\right)\frac{1}{f(F^{-1}(z))} \\ &= \frac{g(Q(z))}{f(Q(z))}, \quad 0 < z < 1. \end{aligned} \tag{7.2}$$

The main problem of comparing two distributions is to find a suitable measure of distance or norm between two distribution functions, i.e. to say, for any $x \in (-\infty, \infty)$,

$$\|G(x) - F(x)\|$$

If a density function exists over the support of F and G, then for any $t \in (0,1)$ this problem to be equivalent to the distance

$$\left|G \circ F^{-1}(t) - t\right|.$$

Under $H_0 : G = F$, $G \circ F^{-1}(t) = t$. In fact, the $h(z)$ in (7.2) is the corresponding PDF for the distribution function $G \circ F^{-1}$ defined over $(0,1)$. The PDF $h(z)$ is a ratio of two densities; and itself is a valid density function. Therefore, we will call it the *Ratio Density Function (RDF)* (Bera, Ghosh and Xiao, 2004).

When $H_0 : F = G$ is true (i.e. $f = g$) then from (7.2), $h(z) = \frac{g(Q(x))}{f(Q(x))} = 1, 0 < z < 1$. Z has the *Uniform* density in $(0,1)$.That means irrespective of what F and G are, the two-sample testing problems can be converted into testing only *one kind of hypothesis*; namely, *uniformity* of a transformed random variable.

For the two sample case with unknown F and G the Smooth test statistic is

$$\begin{aligned}
\Psi_k^2 &= \sum_{l=1}^{k} u_l^2,\ u_l = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \pi_l(z_j), l = 1,2,...,k \\
z_j &= F(y_j) = \int_{-\infty}^{y_j} f(\omega)\, d\omega,\ j = 1,2,...,m.
\end{aligned}$$

Under $H_0 : F = G, \Psi_k^2 \xrightarrow{D} \chi_k^2$.

The test has k components. Each component provides information regarding specific departures from $H_0 : F = G$.

However, in practice $F(.)$ is unknown. We use the Empirical Distribution Function,

$$\begin{aligned}
F_n(x) &= \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x), \qquad \hat{z}_j = F_n(y_j) \\
\hat{\Psi}_k^2 &= \sum_{l=1}^{k} \frac{1}{m} \left[\sum_{j=1}^{m} \pi_l(\hat{z}_j)\right]^2
\end{aligned}$$

The following two theorems [for proof and details see Bera, Ghosh and Xiao (2004)] provide some restrictions on relative sample sizes for consistent asymptotic χ^2 distribution of the test statistic, and also to minimize size distortion of the two sample smooth test of comparing two distributions.

Theorem 6. *If* $\frac{m \log\log n}{n} \to 0$ *as* $m, n \to \infty$ *then* $\hat{\Psi}_k^2 - \Psi_k^2 = o_p(1)$.

Theorem 7. *The optimal relative magnitude of m and n for minimum size distortion is given by* $\quad m = O\left(\sqrt{n}\right).$

For finite sample, for each fixed n_2, we may divide the index set $\mathcal{N} = \{1, \ldots, n\}$ into two mutually exclusive and exhaustive (large) sets $\mathcal{N}_1$ and $\mathcal{N}_2$ with cardinalities n_1 and n_2, where $n_1 + n_2 = n$, and define the **training set**

$$\mathcal{Z}_1 = \{(X_j), j \in \mathcal{N}_1\}$$

and the testing set

$$\mathcal{Z}_2 = \{(X_j), j \in \mathcal{N}_2\}.$$

Then we can estimate $F(\cdot)$ using data $\mathcal{Z}_1$ and construct

$$F_{n_1}(X_i) = \frac{1}{n_1} \sum_{j \in \mathcal{N}_1} I\left(X_j \leq X_i\right), \text{ for } i \in \mathcal{N}_2.$$

$\mathcal{Z}_1$ and $\mathcal{Z}_2$ are from the same distribution F, $F\left(X_i\right)$ $(i \in \mathcal{N}_2)$ are uniformly distributed and $F_{n_1}\left(X_i\right)$ provides an estimator for the uniform distribution, we may compare it with the CDF of standard uniform, say, using some criterion function

$$\frac{1}{n_2} \sum_{i \in \mathcal{N}_2} d(F_{n_1}\left(X_i\right), U[0,1])$$

and take average over R replications

$$\frac{1}{R} \sum_{r=1}^{R} \left[\frac{1}{n_2} \sum_{i \in \mathcal{N}_2} d(F_{n_1}^{r}\left(X_i\right), U[0,1]) \right]$$

For each value of n_2, we can calculate the above criterion function. We may choose n_2 that minimizes the above criterion.

Finally, we choose

$$m = \frac{n_2}{n_1} \times n.$$

The above method may have applications in more general settings. This is a cross-validation type procedure to select sample size.

One of the main problems we would investigate is the distributions of mutual fund inflows with before and after taxes with returns as covariate.(Poterba and Bergstressor, 2002). Means of mutual fund inflow distributions are different before and after taxes with past year returns as covariate.(Poterba and Bergstressor, 2002). "Return chasing" behavior among investors and excessive risk taking among fund managers (Chevalier and Ellison, 1997). Higher cash flow volatility and Fund performance might have a negative relationship (Edelen, 1999, Rakowski, 2002). Inflow distribution convey sentiments about stocks (Frazenni and Lamont, 2005). We want investigate how these distributions are different when we control for the fractiles of returns, hence we can predict the mutual fund inflow based on pre-tax or post-tax return information. This paper documented that mutual funds with heavily taxed returns have lower subsequent inflows compared to ones with lower tax burdens. Our objective is to see if there is evidence in the inflow distributions to show whether higher moments including volatility or skewness and kurtosis terms of inflow distributions are affected by tax exposure. Bergstresser and Poterba (2002) considered US domestic equity mutual funds data on January Releases from Morningstar Principia database with some conditions from 1993-1999.

For our current exposition we will only focus on the 1999 equity mutual fund returns and inflow data with similar characteristics. Bergstresser and Poterba (2002) found that after-tax returns do indeed have more influence on cash inflows on mutual funds, however they did not test whether higher order moments of the inflow distribution are affected by after tax returns.

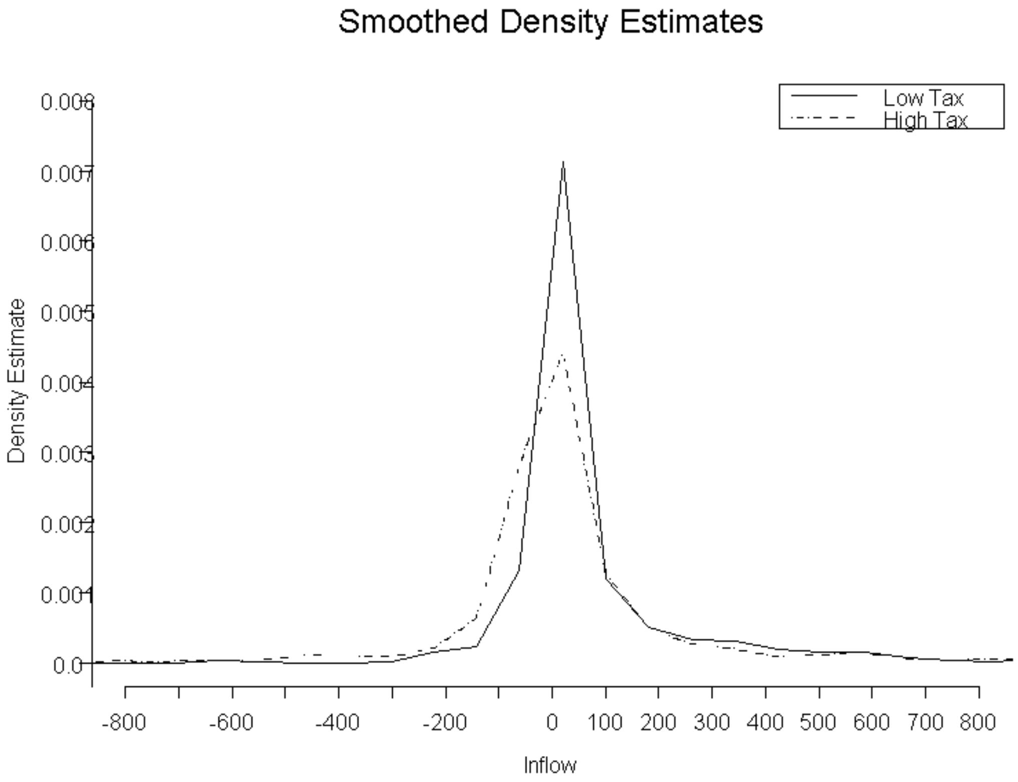


Figure: Inflow of Mutual Funds

Variable	Obs.	Mean	Median	Std. Dev.	Min.	1^{st} Q.	3^{rd} Q.	Max.	coef. skew	excs.kurt
Low Tax	864	187.43	12.15	782.77	-3686	-2.325	93.625	10454.5	7.019	69.15
High Tax	812	211.83	0.45	1427.04	-3961	-25.6	66.025	28651.3	11.888	204.738

Table 1. Summary Statistic for Fund Inflows

Test	KS	CvM
Statistic (T*)	5.3560346	9.371912
Critical Upper 0.1%	1.95	1.167

Table 2. Goodness-of-Fit Statistics based on the EDF.

We tend to reject $H_0 : F = G$ with the above tests but there is no indication of the nature of departure from H_0.

Ψ_4^2	u_1^2	u_2^2	u_3^2	u_4^2
192.33***	39.16***	111.19***	6.74**	2.01
(0.0)	(0.0)	(0.0)	(0.0094)	(0.1559)

Table 3. Smooth statistics and p-values (whole sample)

*** significant at 1%, ** significant at 5%, * significant at 10%.

To reduce the effect of the relative sample sizes, we took a random sample the inflow distribution from higher tax returns, and recomputed the smooth test statistics in Table 4 with the mutual inflows unadjusted for returns, then residuals from OLS, Median Regression, Fractile Regression and finally, a Median Regression on the fractiles of x.

Residuals with Returns	Ψ^2	u_1^2	u_2^2	u_3^2	u_4^2
Unadjusted	72.3725*** (7.11×10^{-15})	35.8095*** (2×10^{-9})	34.7677*** (4×10^{-9})	0.756 (0.3846)	1.0393 (0.308)
OLS	218.4541*** (0)	5.2483** (0.022)	117.8422*** (0)	13.1807*** (0.0003)	82.1828*** (0)
Median Regression	21.9341*** (0.0002)	0.0024 (0.9607)	6.7579*** (0.0093)	12.5554*** (0.0004)	2.6184 (0.1056)
Fractile Regression	170.7627*** (0)	1.4559 (0.2276)	114.9988*** (0)	8.0919*** (0.0044)	46.2161*** (0)
Median-Fractile	45.9366*** (2.54×10^{-9})	0.0074 (0.9317)	27.2038*** (1.8×10^{-7})	13.0462*** (0.0003)	5.6792** (0.0172)

Table 4. Smooth statistic and p-values (sample $m = 324$).

*** significant at 1%, ** significant at 5%, * significant at 10%

One obvious argument in this case is how to choose the mutual funds that have a comparatively high tax exposure, the only way to address this problem is to make fractile or rank groups of the returns. A detailed inspection of Table 4 reveals quite a few facets of the distribution of mutual fund inflows once adjusted for the covariate, in this case past years returns. We also see that the type of regression we use to adjust for the effect of mutual fund returns does indeed make a difference in the distribution of inflows with high and low tax exposure. From the smooth test technique discussed in Bera, Ghosh and Ziao (2004) (paper attached), we observe that the unadjusted inflow distribution for mutual funds with high and low tax exposure differs significantly in the first ($u_1^2 = 35.8095$) and second ($u_2^2 = 34.7677$) moment components. However, past year's mutual fund returns is the most important factor in determining mutual fund inflows (regression results not shown here, please refer to Bergstresser and Poterba, 2002). Hence, to compare the explanatory power of high and low tax exposure of the returns in explaining mutual fund inflows, we need to adjust for the variation in returns.

If we take ordinary least squares residuals (Begstresser and Poterba, 2002), the distribution of inflows adjusted for returns in the high and low tax exposure groups are distinctly different from each other in the direction of each of the first four moments (Table 4). This result could be due to the existence of extreme observations in the data. In order to reduce the effect of outliers we can use Median Regression (essentially Quantile Regression of the 50th percentile). We observe that the two adjusted distributions now only differ in the direction of the second and third moments (u_2^2 =6.7579 and u_3^2 =12.5554). This could be due to the difference in the risk preference and asymmetric loss function of the investors in those mutual funds. However, this result could also be an artifact of the possibility that the distributions of returns are distinctly different between the mutual funds with low tax exposure and those with high tax exposure.

So, in order to make the two groups comparable we have to standardize the covariates. Hence, we look at the residuals using the proposed fractile regression method without using any smoothing techniques. The returns adjusted inflow distribution differs in the directions of the second, third and fourth moments (u_2^2 =114.9988, u_3^2 =8.0919 and u_4^2 =46.2161), although the departure in the direction of the fourth moment is much reduced (u_4^2 =5.6792) and is only slightly significant if we combine quantile and fractile regressions.

This preliminary analysis reveals how we can adjust for a covariate that might not be comparable across two regimes using the linear rank transformation like the Empirical Distribution Function of the covariate, before comparing the distributions of the response variable across two regimes.

Method like this can easily be applied for determining the nature of departure of wages across genders or ethnic groups after adjusting for educational qualifications.Quantile Regression framework has been used subjectively to investigate shifts in location, scale and shapes of the wage distribution due to the effects of training on different quantiles across Europe. (Arulampalam, Booth and Bryan, 2004). Since training is not chosen endogenously by employees, endogeneity problems do not exacerbate the inference problem.To account for the endogeneity in schooling, GMM technique has been used in panel data to investigate how OLS regression might overestimate the gender gap.(Hansen and Wahlberg, 2005). This however doesn't answer the fact that the gap might be different controlling of percentiles of schooling.

8. Conclusions and Future Research

We look at a historical perspective of Mahalanobis' Fractile Graphical Analysis, particularly in the light of the real economic problem he was trying to deal with.

We reevaluate his contribution to the statistics and econometrics literature, as a precursor to k-Nearest Neighbor regression techniques. One of our main objectives in this paper is to define a new form of non-parametric regression namely Fractile Regression. We look at the different regression methods like kernel based non-parametric regression, quantile regression and fractile regression through an illustrative examples in inflow distribution of mutual funds We look at some asymptotic properties of Fractile Regression. We look at some empirical examples like the distribution of mutual fund inflow with pre-tax and after tax returns (Poterba and Bergstresser, 2002).

References

[1] Altman, N.S. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," The American Statistician, Volume 46:3,. pp. 175-185, 1992.

[2] Abdel-Ghany, M., Gehlken, A. and J. L. Silver. " Estimation of income elasticities from Lorenz concentration curves: Application to Canadian micro-data," International Journal of Consumer Studies, Volume 26:4, pp.278, 2002.

[3] Arulampalam, W., A. L. Booth and M. L. Bryan. "Are there Asymmetries in the Effects of Training on Conditional Male Wage Distribution?" IZA Discussion Paper No. 984, Mimeo, 2004.

[4] Bera, A. K. and A. Ghosh. "Neyman's Smooth Test and Its Applications in Econometrics," In Handbook of Applied Econometrics and Statistical Inference, Eds. A. Ullah, A. Wan and A. Chaturvedi, Marcel Dekker: New York, pp. 177-249, 2001.

[5] Bera, A. K., A. Ghosh and Z. Xiao. "Smooth Test for Copmaring Equality of Two Distributions," Working Paper, Singapore Management University, 2004.

[6] Bergstresser, D. and J. Poterba. "Do after-tax returns affect mutual fund inflows?" Journal of Financial Economics 63, pp.381-414, 2002.

[7] Bhattacharya, P. K. "On an analog of Regression Analysis," The Annals of Mathematical Statistics, Vol. 34:4, pp. 1459-1473, 1963.

[8] Bhattacharya, P. K. "Convergence of sample paths of normalized sums of induced order statistics," Annals of Statistics 2, pp. 1034-1039, 1974.

[9] Bhattacharya, P. K. "An invariance principle in regression analysis," Annals of Statistics 4, pp. 621-624, 1976.

[10] Bhattacharya, P. K. "Induced Order Statistics: Theory and Applications," In Handbook of Statistics, Vol. 4, P.R. Krishnaiah and P.K. Sen, Eds. Elsevier: New York, pp.383-403, 1984.

[11] Bhattacharya, P. K. and A. K. Gangopadhyay. "Kernel and Nearest-Neighbor Estimation of a Conditional Quantile," The Annals of Statistics, Vol. 18:3, pp. 1400-1415, 1990.

[12] Bhattacharya, P. K. and H.-G. Müller. "Asymptotics for Nonparametric Regression," Sankhyā, Series A, Volume dedicated to the memory of P. C. Mahalanobis 55:3, pp. 420-441, 1993.

[13] Boos, D.D. "Minimum Distance estimators for Location and Goodness of Fit," Journal of the American Statistical Association, Vol. 76:375, pp. 663-670, 1981.

[14] David, H. A. "Concomitants of order statistics," Bulletin of the International Statistical Institute 45, pp. 295-300, 1973.

[15] David, H. A. and J. Galambos. "The asymptotic theory of concomitants of order statistics," Journal of Applied Probability 11, pp. 762-770, 1974.

[16] David, H. A. and H.N. Nagaraja. Order Statistics, Third Edition, John Wiley and Sons: New Jersey, 2003.

[17] David, H. A., M.J. O'Connell and S.S. Yang. "Distribution and expected value of the rank of a concomitant of an order statistics," Annals of Statistics 5, pp. 216-223, 1977.

[18] Dutta, J., J.A. Sefton and M.R. Weale. " Income Distribution and Income Dynamics in the United Kingdom," Journal of Applied Econometrics, Vol. 16, pp. 599-617, 2001.

[19] Efron, B. "Computers and the theory of statistics: Thinking the Unthinkable," SIAM Rev. 21, pp. 460-480, 1979.

[20] Efron, B. "Bootstrap Methods: Another look at the Jackknife. Annals of Statistics, Volume 7, pp. 1-26, 1979.

[21] Efron, B. *The Jackknife, the Boostrap and Other Resamnpling Plans.* SIAM, Philadelphia, 1982.

[22] Ghosh, J.K. "Mahalanobis and the art and science of statistics: The Early Days," Indian Journal of History of Science, 29:1, pp. 89-98, 1994.

[23] Ghosh, J.K., P. Maiti, T.J. Rao and B.K. Sinha. "Evolution of Statistics in India," monograph, Indian Statistical Institute, Calcutta, India, pp.1-26.1998.

[24] Hall, P. "A Short Prehistory of the Bootstrap," Statistical Science, Volume 18:2, pp. 158-167, 2003.

[25] Hansen, J. and R. Wahlberg. "Endogenous schooling and the distribution of the gender wage gap," Empirical Economics 30, pp. 1-22, 2005.

[26] "History and Activities of the Indian Statistical Institute: 1931-1963", Courtesy, Prashanta Chandra Mahalanobis Memorial Museum and Archives, Indian Statistical Institute, Calcutta, pp. 1-41.

[27] "Lekhon: The Mouthpiece of Indian Statistical Institute Club," Ed. P. K. Chatterjee, 1997.

[28] Iyengar, N.S. "On a Method of Computing Engel Elasticities from Concentration Curves," Econometrica, Vol. 28:4, pp. 882-891, 1960.

[29] Iyengar, N.S. "A consistent method of estimating the Engel Curve from grouped survey data," Econometrica, Vol. 32:4, pp. 591-618, 1964.

[30] Iyengar, N.S. and N. Bhattacharya. "Some Observations on Fractile Graphical Analysis," Econometrica, Volume 33:3, pp. 644-645, 1965.

[31] Kawada, Y. "Some remarks concerning the expectation of the error area in fractile analysis," Sankhyā, Series A, Volume 23:1, pp. 155-160, 1961.

[32] Linder, A. and P. Czegledy. "Normality Test by Fractile Graphical Analysis," Sankhyā, Series B, Volume 35:1, pp.1-14, 1973.

[33] Lo, A.W. and A. C. MacKinlay. "Data-Snooping Biases in Tests of Financial Asset Pricing Models," Review of Financial Studies, Vol. 3:3, pp. 431-467, 1990.

[34] Mahalanobis, P.C. "A Method of Fractile Graphical Analysis," Econometrica, Volume 28:2, pp.325-351, 1961.

[35] Mahalanobis, P.C. " A Method of Fractile Graphical Analysis with Some Surmises of Results," Transactions of the Bose Research Institute, 1958.

[36] Mahalanobis, P.C. "Extensions of Fractile Graphical Analysis," Proceedings of the International Conference on Quality Control, Tokyo, 1969.

[37] Mahalanobis, P.C. "Extensions of Fractile Graphical Analysis to Higher Dimensional Data," In *Essays in Probability and Statistics*, Eds., R.C. Bose et. al. University of North Carolina Press, Chapel Hill, 1970.

[38] Mitrofanova, N.M. "On some problems of Fractile Graphical Analysis," Sankhyā, Series A, Volume 23:1, pp. 145-154, 1961.

[39] Neyman, J. “Smooth test” for goodness of fit. Skandinaviske Aktuarietidskrift 20:150-199, 1937.

[40] Parthasarathy, K. R. and P.K. Bhattacharya. " Some Limit Theorems in Regression Theory," Sankhyā, Series A, Volume 23:1, pp. 91-102, 1961.

[41] Rakowski, D. "Fund Flow Volatility and Performance," Manuscript, Georgia State University, 2002.

[42] Rao, C.R. "Prashanta Chandra Mahalanobis: June 29,1893- June 28, 1972," The IMS Bulletin, Volume 22:6, pp.593-597, 1993.

[43] Rao, C. R. "Statistics must have a purpose: The Mahalanobis Dictum," Sankhyā, Series A, Special Volume dedicated to the memory of P. C. Mahalanobis, 55:3, pp. 331-349, 1993.

[44] Rao, C. R. and L. C. Zhao. "Converegence Theorems for Empirical Cumulative Quantile Regression Functions," Mathematical Methods of Statistics, Volume 4:1, pp. 81-91, 1995.

[45] Rao, C. R. and L. C. Zhao. "Law of Iterated Logarithm for Empirical Cumulative Quantile Regression Functions," Statistica Sinica 6, pp. 693-702, 1996.

[46] Schuster, E.F. "Joint asymptotic distribution of the estimated regression function at a finite number of distinct points," Annals of Mathematical Statistics, v. 43, pp. 84-88, 1972.

[47] Sen, B. "Estimation and Comparison of Fractile Graphs Using Kernel Smoothing Techniques,"Sankhyā, Special Issue on Quantile Regression and Related Methods,Volume 67:2, pp. 305-334, 2005.

[48] Sen, P.K. "A Note on Invariance Principles for Induced Order Statistics,"Annals of Probability 4, pp.474-479, 1976.

[49] Sethuraman, J. "Limit Distributions connected with Fractile Graphical Analysis," Sankhyā, Series A, Volume 23:1, pp. 79-90, 1961.

[50] Srinivasan, T. N. "Professor Mahalanobis and Economics" in Chapter 11 pp. 224-252, In.*Prashanta Chandra Mahalanobis: A Biography*, Ed. Ashok Rudra. Oxford University Press.New Delhi, India,1996.

[51] Stone, C. J. "Consistent Nonparametric Regression," The Annals of Statistics, Vol. 5:4, pp. 595-620, 1977.

[52] Stute, W. "The Oscillation Behavior of Empirical Processes," The Annals of Probability, Vol 10:1, pp. 86-107, 1982.

[53] Stute, W. "Asymptotic Normality of Nearest Neighbor Regression Function Estimates," The Annals of Statistics, Vol 12:3, pp. 917-926, 1984.

[54] Stute, W. "The Oscillation Behavior of Empirical Processes: The Multivariate Case," The Annals of Probability, Vol 12:2, pp. 361-379, 1984.

[55] Swamy, S. " Notes on Fractile Graphical Analysis," Econometrica, Volume 31:3, pp. 551-554, 1963.

[56] Takeuchi, K. "On some properties of error area in the Fractile Graph Method," Sankhyā, Series A, Volume 23:1, pp. 65-78, 1961.

[57] Yang, S.S. "General distribution theory of the concomitants of order statistics," Annals of Statistics 5, pp. 996-1002, 1977.

[58] Yang, S.S. "Linear functions of concomitants of order statistics with applications to nonparametric estimation of a regression function," Journal of the American Statististical Association 76, pp. 658-662, 1981.

[59] Yang, S.S. "Linear combinations of concomitants of order statistics with applications to Testing and Estimation,"Annals of the Institute of Statistical Mathematics 33:A, pp. 463-470, 1981.

Dept. of Economics, University of Illinois,1206 S. Sixth Street, Champaign, IL 61820. USA. Phone: 217-333-4596. Fax: 217-244-6678.

E-mail address: abera@uiuc.edu

School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903.Phone: +65 6828 0863. Fax: +65 6828 0833.

E-mail address: aurobindo@smu.edu.sg

Department of Economics, Boston College

E-mail address: xiaoz@bc.edu