On Fractile Regression

Bodhisattva Sen and Probal Chaudhuri

University of Michigan Indian Statistical Institute

Ann Arbor, MI 48109 203 B.T. Road, Kolkata 700108 *

Abstract

The need for comparing two or more regression functions arises frequently in statistical applications. Comparison of the usual regression functions is not very meaningful in situations where the distribution and the range of the covariates have changed for the populations. For instance, in econometric studies, the prices of commodities and people's incomes observed at different time points may not be on comparable scales due to inflation and other economic factors. In this paper we describe, motivated by an idea of Mahalanobis (1960), a method of standardizing the covariates and estimating the transformed regression function, which now become comparable. We develop smooth estimates of fractile regression function and study its statistical properties. We prove the consistency and asymptotic normality of the estimated fractile regression function defined through general weight functions. We illustrate our method through analysis of three data sets: blood pressure and related measurements of two tribes in India, profit and sales of

*Bodhisattva Sen is a PhD Student, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *bodhi@umich.edu*). Probal Chaudhuri is Professor at Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India (E-mail: *probal@isical.ac.in).*

private companies in India at two time points, and data on household income and expenditure of two East European transitional economies.

KEY WORDS: Consistency and asymptotic normality, geometric quantile, kernel smoothing, multivariate fractile, smooth estimates, transformation of covariates.

1 Introduction

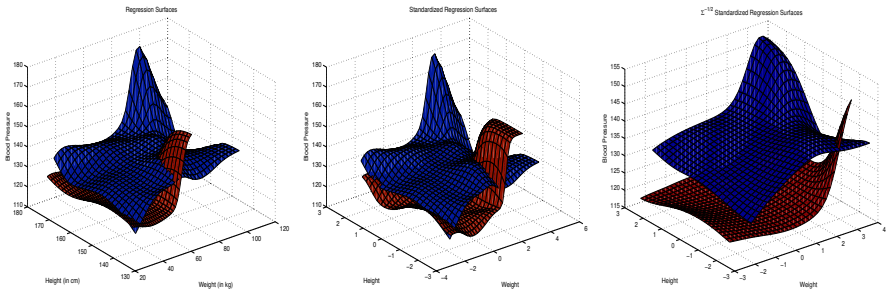
Comparison of two or more regression functions arise quite often. But the comparison of usual regression functions can be very difficult when the covariates for the two populations have different distributions. Let us consider a couple of examples to illustrate this point, where nonparametric estimates of regression functions are used.

Example 1: Data was collected on 258 individuals from the Bhutia tribe and 305 individuals from the Toto tribe in India on blood pressure, height 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 height and the weight of an individual for the two populations. A common approach would be to compare the two regression surfaces as is shown in Figure (1a). But the two regression surfaces are not comparable as the covariates have very different distributions in the two populations. In fact, the ranges of the covariates are quite different. Probably the simplest way to standardize the covariates in order to make the regression functions comparable would be to subtract the mean from each of the covariate values and divide

by the standard deviation. The location and scale shifted regression surfaces are shown in Figure (1b), whereas Figure (1c) shows the regression surfaces, where we standardize each covariate vector by subtracting the sample mean vector and multiplying by the inverse of the square-root of the observed dispersion matrix. But the surfaces are still not quite in comparable forms – the ranges of the standardized covariates still tend to differ quite a bit. We have used the Nadaraya-Watson smoother with the standard bivariate gaussian kernel to produce the regression surfaces. For choosing the optimal smoothing bandwidths, we have used the least squares cross validation method [see Wand and Jones (1995)] and computation was done by using the "sm" package in R developed by Adrain Bowman and Adelchi Azzalini.

A disturbing feature in the three figures is the crossing of the red surface (for the Bhutia tribe). The Toto population is usually believed to have higher blood pressure than the Bhutia population. An obvious question that arises is whether the crossing is a real feature of the Bhutia population or not. Another anomaly illustrated in the figures is the high peak of the blue surface (for the Toto tribe) at large values of height and weight. A tall and heavy person would not usually be expected to have a higher blood pressure than a short and heavy (over-weight) person. As will be shown later, these two features in the regression surfaces are indeed spurious and can be attributed to the skewness of the covariates and a misleading comparison of the two regression surfaces.

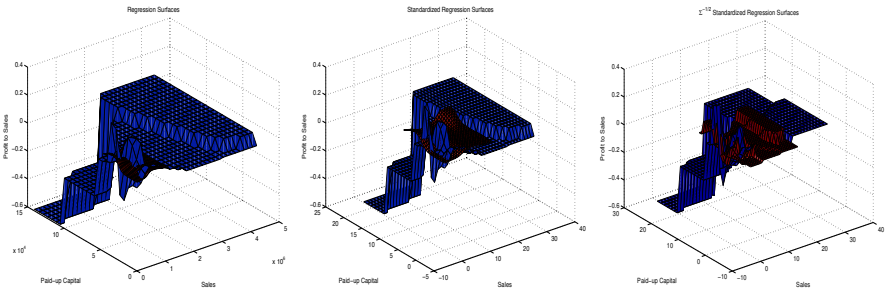
Example 2: The Reserve Bank of India keeps data on the sales (in Indian rupees), paid-up capital (in Indian rupees) and profit (as a fraction of sales) for



(a) (b) (c)

Figure 1: *(a) Usual regression surfaces, (b) location and scale shifted regression surfaces, and (c) regression surfaces when the covariates are standardized by the inverse of the square-root of the dispersion matrix of blood pressure with weight and height for the Bhutia (red) and Toto (blue) tribes.*

non-government, non-financial public limited companies in India over different years. They are interested in comparing the profitability of the companies against measures like the sales and the paid-up capital, at two time points. This gives rise to a regression problem where one regresses profit (as a fraction of sales) against sales and paid-up capital. One would like to compare the two regression surfaces for two time points. But the comparison of usual regression surfaces 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 (2a) shows the usual regression surfaces for the year 1997 (red surface) and 2003 (blue surface) with 944 and 1243 data points respectively. Figure (2b) and Figure (2c) show the regression surfaces with the covariate vector standardized by a simple coordinate-wise location and scale change, and by the inverse of the square-root of the dispersion matrix respectively. We use similar non-parametric techniques as in Example 1 to estimate the regression surfaces. The skewness of the covariates causes distortion of the estimated regression surfaces, and the choice of the



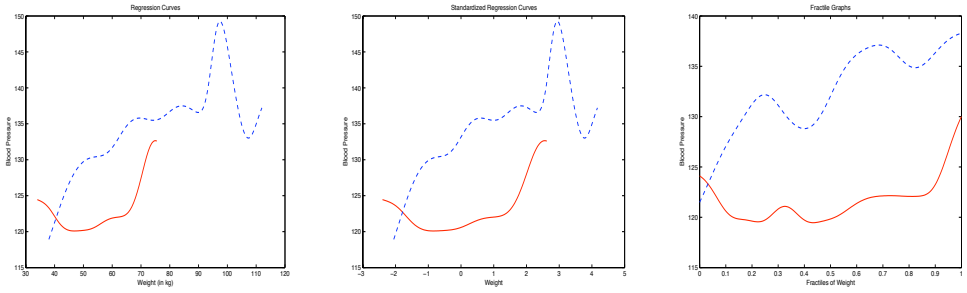
(a) (b) (c)

Figure 2: *(a) Usual regression surfaces, (b) location and scale shifted regression surfaces, and (c) regression surfaces when the covariates are standardized by the inverse of the square-root of the dispersion matrix of profit (as a fraction of sales) with sales and paid-up capital for the years 1997 (red) and 2003 (blue).*

smoothing bandwidth also becomes very difficult. Besides, the large difference in the covariate values for the years 1997 and 2003 makes the two regression surfaces virtually incomparable in the figures.

Both the preceding examples demonstrate the need for a methodology to appropriately standardize the covariates before comparing the regression functions. In this paper we propose a method for standardizing the covariates by using a multivariate transformation, which is derived from their multivariate distribution. We also discuss the estimation of the corresponding regression functions based on the transformed covariates. Let us first look at the problem when there is just one covariate. 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

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

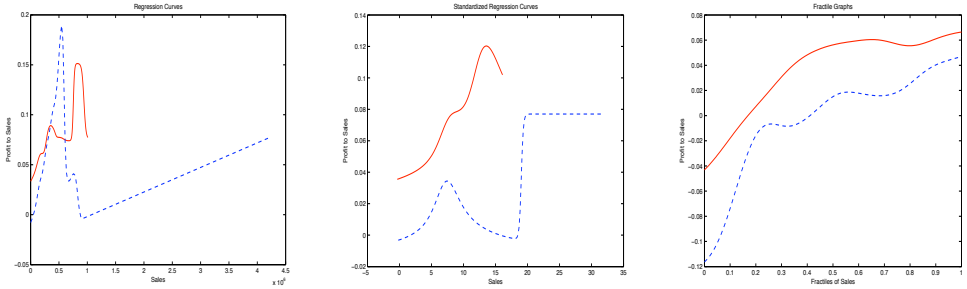


(a) (b) (c)

Figure 3: *(a) Usual regression curves, (b) location and scale shifted regression curves, and (c) fractile curves of blood pressure with weight for the Bhutia (in red, solid line) and Toto (in blue, dashed line) tribes.*

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)]. Note that the transformed covariates $F_1(X_1)$ and $F_2(X_2)$ both have a uniform distribution on $(0,1)$. This distribution-free nonparametric standardization of the covariates makes comparison of the regression functions meaningful even when the real valued covariates have very different distributions in the two populations. 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. Also, this standardization makes the fractile regression functions invariant under all strictly increasing transformations of the covariate. In other words, if $X_2 = \phi(X_1)$, where ϕ is any strictly increasing transformation, then $E\{Y_1|F_1(X_1)\} = E\{Y_2|F_2(X_2)\}$. Fractile regression has been considered earlier in Mahalanobis (1960), Sethuraman (1961), Parthasarathy and Bhattacharya (1961), Bhattacharya and Muller (1993) and Sen (2005).

In Figure (3), we have plotted the usual regression curves, regression curves



(a) (b) (c)

Figure 4: *(a) Usual regression curves, (b) location and scale shifted regression curves, and (c) fractile curves of profit (as a fraction of sales) against sales for the years 1997 (in red, solid line) and 2003 (in blue, dashed line).*

with covariates standardized for location and scale and the smooth estimates of fractile curves with blood pressure as the response and body weight as the predictor for the two populations discussed in Example 1. Figure (4) shows the corresponding three plots for the data set in Example 2 with profit on sales as the response and sales as the predictor. We used the Nadaraya-Watson smoother with the standard normal kernel to estimate the regression functions [see Sen (2005) for more details on estimating the fractile curves]. The highly irregular regression curves obtained in Figures (4a) and (4b) is due to the skewness of the data. The performance of data driven bandwidths for the regression curves in this example was very poor. We made a subjective choice of the smoothing parameter after observing several plots with different bandwidths. In all the other plots we used the direct plug-in bandwidth estimator developed by Ruppert, Sheather and Wand (1995). Bandwidth selection is a relatively simpler problem for fractile regression as the transformed covariate values are uniformly spaced over the interval $(0, 1)$. In each of the Figures (3a), (3b), (4a) and (4b) there is a serious lack of comparability between the two regression curves, which is adequately resolved in Figures (3c)

and (4c).

In this paper, we develop and investigate fractile regression when the dimension of covariates might be more than one. The first hurdle in defining fractile regression with multiple covariates is the absence of a straight-forward notion of multivariate quantiles, because of the lack of natural ordering of points in $\mathbb{R}^d$ for $d > 1$. In Section 2, we discuss two suitable notions of multivariate quantiles and use it to define the fractile regression function. Section 3 discusses nonparametric smooth estimation of the fractile regression function from a sample of data points. We investigate asymptotic properties of such estimates in Section 4. The fractile surfaces for Examples 1 and 2 are presented in Section 5 followed by a brief discussion. Section 6 provides another application of fractile regression techniques on real data, namely, Household Survey data for Transitional Economies. In Section 7, the Appendix, we give the proofs of the main results and state the necessary conditions on the weight functions required for Theorem 4.2.

2 Fractile regression function

The lack of objective basis for ordering multivariate observations is a major problem in extending the notion of quantiles to multi-dimensions. For a d-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ with probability distribution $\mathbf{P}$ on $\mathbb{R}^d$, we define a multivariate analogue $\mathbf{R_P} : \mathbb{R}^d \mapsto [0,1]^d$ of the univariate distribution function [i.e., $x \mapsto F_X(x)$] as

$$\mathbf{R_P}(x_1, x_2, \ldots, x_d) = \left(F_1(x_1), F_{2|1}(x_2|x_1), \ldots, F_{d|1,2,\ldots,d-1}(x_d|x_1, x_2, \ldots, x_{d-1})\right), \quad (1)$$

where $F_1(x_1) = P(X_1 \leq x_1), F_{2|1}(x_2) = P(X_2 \leq x_2 | X_1 = x_1), \ldots, F_{d|1,2,\ldots,d-1}(x_d)$ $= P(X_d \leq x_d | X_1 = x_1, X_2 = x_2, \ldots, X_{d-1} = x_{d-1})$. The quantile map obtained by inverting this transformation can be taken as a version of *multivariate fractile map*, and it is expressed as

$$\mathbf{G_P}(\mathbf{u}) = \left(F_1^{-1}(u_1), F_{2|1}^{-1}(u_2|u_1), \ldots, F_{d|1,2,\ldots,d-1}^{-1}(u_d|u_1, u_2, \ldots, u_{d-1})\right) \qquad (2)$$

for $\mathbf{u} = (u_1, u_2, \ldots, u_d) \in (0,1)^d$ [see Bhattacharya (1963)]. We index d-dimensional multivariate quantiles by points in the open unit cube $(0,1)^d$. Points close to $(0.5, 0.5, \ldots, 0.5)$ correspond to the central quantiles whereas points close to the boundary of the cube would correspond to extreme quantiles. Suppose that $\mathbf{X}$ and $\mathbf{Z}$ are two random vectors having distributions $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively on $\mathbb{R}^d$ with continuous density functions. If $\mathbf{R}_{\mathbf{P}_1}(\mathbf{x}) = \mathbf{R}_{\mathbf{P}_2}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$ then $\mathbf{P}_1 = \mathbf{P}_2$. Thus the above notion of multivariate distribution transform characterizes the distribution of the random vector.

An interesting invariance property shared by any continuous univariate distribution function F is that $F(X) \sim$ Uniform$(0,1)$, where X has distribution function F. A similar result holds for $\mathbf{R_P}$. Suppose that $\mathbf{X} = (X_1, X_2, \ldots, X_d) \sim \mathbf{P}$, then it can be easily shown that $\mathbf{R_P}(\mathbf{X}) \sim$ Uniform$(0,1)^d$, if $\mathbf{X}$ has a density on $\mathbb{R}^d$.

As we have discussed in the previous section, we need some nonparametric standardization of the covariates to make meaningful comparison of the two regression functions for the two populations. In the single covariate setup, the usual location and scale standardization of the covariate will make the regression curve invariant under any linear transformation of the covariate. For two

real populations, the covariates might be related by a complicated monotonic transformation. We would still like to compare the regression functions and if the distribution of the response remains the same, we may like to conclude that the transformed regression functions have not changed. The standardization $X \mapsto F(X)$, F being the distribution function of X, makes the transformed regression curves invariant under any increasing transformation of the covariate. In fact, it can be shown that any transformation of the covariate X that makes the regression curves invariant under all increasing transformations will necessarily be a function of F. So, in a sense, the fractile transformation is a very strong notion of standardization.

In the multiple covariate setup, we use $\mathbf{R_P}$, the multivariate distribution transform discussed above, to standardize the covariates and regress Y on $\mathbf{R_P}(\mathbf{X})$ (where $\mathbf{X} \sim \mathbf{P}$). In other words, we define the fractile regression function obtained by using the multivariate distribution transform $\mathbf{R_P}$ as

$$m(\mathbf{t}) = E\{Y|\mathbf{R_P}(\mathbf{X}) = \mathbf{t}\} \text{ for } \mathbf{t} \in (0,1)^d.$$

We now state a result on the invariance of this fractile regression function under any coordinate-wise strictly increasing transformation of the covariate vector that justifies the use of $\mathbf{R_P}$ as a nonparametric standardization tool for the covariates.

Theorem 2.1 *Let* $(\mathbf{X}, Y) \in \mathbb{R}^{d+1}$ *be a random vector such that* $\mathbf{X} = (X_1, X_2, \ldots, X_d) \sim \mathbf{P}_1$ *has a density on* $\mathbb{R}^d$. *Also, let* $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_d) \sim \mathbf{P}_2$, *where* $Z_j = \phi_j(X_j)\ \forall\ j = 1, 2, \ldots, d$, *and each* ϕ_j *is a strictly increasing function on* $\mathbb{R}$. *Then* $\mathbf{R}_{\mathbf{P}_1}(\mathbf{x}) = \mathbf{R}_{\mathbf{P}_2}(\mathbf{\Phi}(\mathbf{x}))$ *where* $\mathbf{\Phi}(\mathbf{x}) = (\phi_1(x_1), \phi_2(x_2), \ldots, \phi_d(x_d))$. *In*

particular,

$$E\{Y|\mathbf{R}_{\mathbf{P}_1}(\mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{R}_{\mathbf{P}_2}(\mathbf{Z}) = \mathbf{t}\} \text{ for all } \mathbf{t} \in (0,1)^d.$$

The above theorem says that even if each covariate gets transformed by an arbitrary strictly increasing transformation, the fractile regression function will not change. This property is quite desirable when we would like to standardize the covariates and compare two regression functions, where the distribution of the covariates might be very different. Note that the transformed covariates will always have the same Uniform$(0,1)^d$ distribution, making the fractile regression functions comparable.

The transformation $\mathbf{R}_\mathbf{P}$ standardizing the covariates is invertible and depends on the distribution $\mathbf{P}$ of the covariates. The transformed vector of covariates $\mathbf{R}_\mathbf{P}(\mathbf{X})$ always has a fixed distribution irrespective of $\mathbf{P}$. Let $\mathcal{P}$ be the class of all distributions on $\mathbb{R}^d$ having a density. Suppose now that $\mathbf{T} : \mathcal{P}\times\mathbb{R}^d \to E \subset \mathbb{R}^d$ is another transformation such that $\mathbf{x} \mapsto \mathbf{T}(\mathbf{P}, \mathbf{x})$ is an invertible map from $\mathbb{R}^d$ onto E for every $\mathbf{P} \in \mathcal{P}$. Then, the transformed regression function can be defined as $E\{Y|\mathbf{T}(\mathbf{P}, \mathbf{X}) = \mathbf{t}\}$ for $\mathbf{t} \in E$. The following theorem shows that if the transformed regression function does not change for any coordinate-wise increasing transformation of the covariates then $\mathbf{T}(\mathbf{P}, \mathbf{X})$ must necessarily have a fixed distribution, like $\mathbf{R}_\mathbf{P}(\mathbf{X})$, for all $\mathbf{P} \in \mathcal{P}$.

Theorem 2.2 *Let* $\mathbf{X}, Y, \mathbf{Z}, \mathbf{P}_1$ *and* $\mathbf{P}_2$ *be as in Theorem 2.1, and suppose that there exists a transformation* $\mathbf{T} : \mathcal{P}\times\mathbb{R}^d \to E \subset \mathbb{R}^d$ *as described above such that* $E\{Y|\mathbf{T}(\mathbf{P}_1, \mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{T}(\mathbf{P}_2, \mathbf{Z}) = \mathbf{t}\}$ *for all* $\mathbf{t} \in E$*, and equality holds for all joint distributions* $(\mathbf{X}, Y)$*, with* $\mathbf{X} \sim \mathbf{P}_1 \in \mathcal{P}$*. Then* $\mathbf{T}(\mathbf{P}, \mathbf{X})$

must have a fixed distribution, i.e., we must have $\mathbf{T}(\mathbf{P}, \mathbf{X}) \stackrel{d}{=} \mathbf{V}$*, for some fixed random vector* $\mathbf{V}$*, for all* $\mathbf{X} \sim \mathbf{P} \in \mathcal{P}$.

The computation of $\mathbf{R_P}$ from a sample of data points $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim \mathbf{P}$ in $\mathbb{R}^d$ requires the estimation of the conditional distribution functions as the exact distribution will be unknown for almost all practical problems. We might use a kernel estimate of the multivariate density of $\mathbf{X}_i$, and then use it to get various marginal and conditional densities. Let $\mathbf{R}_n$ be the estimated distribution transform obtained from the sample. For the examples used in this paper, we have used a gaussian kernel with smoothing bandwidths chosen by cross validation. Under appropriate conditions on the kernel and the smoothing parameter(s), it can be easily shown that $\mathbf{R}_n$ is a uniformly consistent estimator of $\mathbf{R_P}$, i.e.,

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{R}_n(\mathbf{x}) - \mathbf{R_P}(\mathbf{x})\| \xrightarrow{P} 0.$$

Mahalanobis (1988) (pages 68-71) and Bhattacharya (1963) suggested an alternative method for computing the conditional quantiles of the covariates which essentially consists of nested binning of the covariates.

2.1 An alternative to $\mathbf{R_P}$: geometric quantile and related distribution transform

Though the multivariate transform $\mathbf{R_P}$ has nice invariance properties and simple probabilistic interpretations, sometimes it can be difficult to estimate, as it requires estimation of the conditional distribution functions. As the dimension d increases, the density estimation becomes more difficult and the computational complexity increases at an exponential rate. Also note that $\mathbf{R_P}$ depends on the ordering of the coordinates of the covariate vector. Changing

the order of the coordinate variables would change the transformation. In this subsection, we discuss another notion of multivariate quantile, which is computationally simpler and does not depend on the ordering of the co-ordinate random variables. This concept of multivariate quantile, called the geometric quantile (or spatial quantile), was introduced and studied by Chaudhuri (1996) and Koltchinskii (1997).

Suppose that we have a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim \mathbf{P}$ in $\mathbb{R}^d$. We index d-dimensional multivariate quantiles by points in the open unit ball $B^{(d)} = \{\mathbf{u} : \mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\| < 1\}$. For any $\mathbf{u} \in B^{(d)}$ and $\mathbf{t} \in \mathbb{R}^d$, the geometric quantile $\widehat{\mathbf{Q}}_n(\mathbf{u})$ corresponding to $\mathbf{u}$ and based on the d-dimensional sample is defined as $\widehat{\mathbf{Q}}_n(\mathbf{u}) = \arg\min_{\mathbf{Q}\in\mathbb{R}^d} \sum_{i=1}^n \Phi(\mathbf{u}, \mathbf{X_i} - \mathbf{Q})$, where $\Phi(\mathbf{u}, \mathbf{t}) = \|\mathbf{t}\| + \langle \mathbf{u}, \mathbf{t} \rangle$. Here $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the usual Euclidean inner product and norm. A $\mathbf{u}$ for which $\|\mathbf{u}\|$ is close to 1 corresponds to an extreme quantile whereas a $\mathbf{u}$ for which $\|\mathbf{u}\|$ is close to zero corresponds to a central quantile. For a random vector $\mathbf{X} \sim \mathbf{P}$ on $\mathbb{R}^d$, we can define the geometric quantile $\mathbf{Q_P}(\mathbf{u})$, for $\mathbf{u} \in B^{(d)}$ as $\mathbf{Q_P}(\mathbf{u}) = \arg\min_{\mathbf{Q}\in\mathbb{R}^d} E\{\Phi(\mathbf{u}, \mathbf{X} - \mathbf{Q}) - \Phi(\mathbf{u}, \mathbf{X})\}$.

It is known that the solution $\mathbf{Q_P}(\mathbf{u})$ of the above minimization problem always exists for any $\mathbf{u}$. The geometric quantile function $\mathbf{Q_P}(\mathbf{u})$ characterizes the associated distribution in the sense that $\mathbf{Q}_{\mathbf{P}_1} = \mathbf{Q}_{\mathbf{P}_2} \Rightarrow \mathbf{P}_1 = \mathbf{P}_2$. Computation of the sample geometric quantile function is simple and the asymptotic properties are known [see Chaudhuri (1996) and Koltchinskii (1997) for more details]. But the geometric quantile is not equivariant under arbitrary increasing transformations, like $\mathbf{G_P}$.

We define the geometric distribution function $\mathbf{S_P}$ corresponding to the geometric quantile map $\mathbf{Q_P}$ as $\mathbf{S_P}(\mathbf{x}) = E_{\mathbf{P}}\left(\frac{\mathbf{x}-\mathbf{X}}{\|\mathbf{x}-\mathbf{X}\|}\right)\ \forall\ \mathbf{x} \in \mathbb{R}^d$, where $\mathbf{X} \sim \mathbf{P}$. Note that $\mathbf{S_P}$ and $\mathbf{Q_P}$ are invertible functions, and one is the inverse of the other. The geometric distribution function can also be viewed as an extension of the usual distribution function in the univariate case. Unlike $\mathbf{R_P}(\mathbf{X}), \mathbf{S_P}(\mathbf{X})$ does not have a fixed distribution for all $\mathbf{X} \in \mathcal{P}$. The empirical geometric distribution function is defined as

$$\mathbf{S}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x}-\mathbf{X}_i}{\|\mathbf{x}-\mathbf{X}_i\|}.$$

From Theorem (5.5) in Koltchinskii (1997), it follows that $D_n = \sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{S}_n(\mathbf{x}) - \mathbf{S_P}(\mathbf{x})\| \to 0$ a.s. as $n \to \infty$.

3 Smooth estimation of fractile regression

In this section, we define smooth estimates of the fractile regression function. As pointed out by Stone (1977), most nonparametric regression estimates can be expressed as a weighted sum of the response values. We develop a similar kind of theory by using general weight functions satisfying some regularity conditions. Suppose that we have a sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from a population in $\mathbb{R}^{d+1}$ with a continuous density function, where $\mathbf{X}_i \sim \mathbf{P}$. Let $(\mathbf{X}, Y)$ be a generic random vector having the same joint distribution. The methodology is described with a general version of multivariate distribution transform $\mathbf{H} : \mathbb{R}^d \to E \subset \mathbb{R}^d$ (which may or may not be $\mathbf{R_P}$ or $\mathbf{S_P}$). We want to estimate the fractile regression function

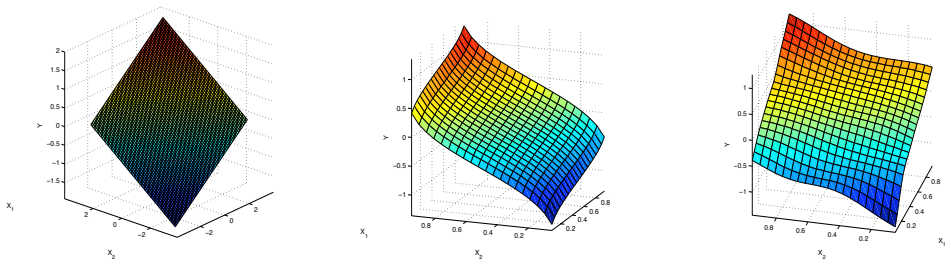
$$m(\mathbf{t}) = E\{Y|\mathbf{H}(\mathbf{X}) = \mathbf{t}\} \text{ for } \mathbf{t} \in E.$$

We define the smooth estimated fractile regression function as

$$\widehat{m}_n(\mathbf{t}) = \sum_{i=1}^{n} Y_i W_{n,i}(\mathbf{t}) \text{ for } \mathbf{t} \in E, \tag{3}$$

where $W_{n,i}(\mathbf{t})$ is the weight function, which might depend on $\mathbf{H}_n$, the empirical or estimated value of $\mathbf{H}$. If we use kernel based weights, one common choice of weight function is the Nadaraya-Watson type weight function, i.e., $W_{n,i}(\mathbf{t}) = \frac{\mathbf{K}\left(\frac{\mathbf{t}-\mathbf{H}_n(\mathbf{X}_i)}{\mathbf{h}_n}\right)}{\sum_{j=1}^{n}\mathbf{K}\left(\frac{\mathbf{t}-\mathbf{H}_n(\mathbf{X}_j)}{\mathbf{h}_n}\right)}$, where $\mathbf{K}$ is a density function defined on $\mathbb{R}^d$, $\mathbf{t} = (t_1, t_2, \ldots, t_d) \in E$, $\frac{\mathbf{t}-\mathbf{H}_n(\mathbf{X}_i)}{\mathbf{h}_n} = \left(\frac{t_1 - H_{n,1}(\mathbf{X}_i)}{h_{n,1}}, \frac{t_2 - H_{n,2}(\mathbf{X}_i)}{h_{n,2}}, \ldots, \frac{t_d - H_{n,d}(\mathbf{X}_i)}{h_{n,d}}\right)$, and $h_{n,1}, h_{n,2}, \ldots, h_{n,d}$ are the smoothing bandwidths.

As an example, we demonstrate our smooth estimate of fractile regression surface, along with the actual fractile regression surface, using the multivariate distribution transform $\mathbf{R_P}$ in Figure (5), where we have a tri-variate normal population $(X_1, X_2, Y) \sim N(\mathbf{0}, \Sigma)$ with $\Sigma = (\sigma_{i,j})_{3\times 3}$ such that $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0.5 \ \forall \ i \neq j$. We have generated a sample of size 200 to construct the smooth estimated fractile regression surface. For choosing the optimal bandwidths for the weight functions, we used the least squares cross validation method [see Wand and Jones (1995)] and computation was done using the "sm" package in R developed by Adrain Bowman and Adelchi Azzalini. The transformed vector of covariates $\mathbf{R_P}(\mathbf{X})$ has a uniform distribution on $(0,1)^d$, which makes the choice of bandwidths easier and more stable. We use the standard multivariate gaussian density as the kernel $\mathbf{K}$ and the Nadaraya-Watson type weight function in our examples.



(a) (b) (c)

Figure 5: *(a) Usual mean regression surface for $Y|X_1, X_2$, (b) actual fractile regression surface for $Y|X_1, X_2$, and (c) estimated fractile regression surface for $Y|X_1, X_2$ using smooth estimates for gaussian data.*

4 Asymptotic properties of estimated fractile regression

In this section, we prove the consistency and asymptotic normality of the smooth estimated fractile regression function. The following theorem gives the asymptotic distribution of $\widehat{m}_n(\mathbf{t})$ when we center it around its conditional mean.

Theorem 4.1 *Fix $\mathbf{t} \in E$. Suppose that $Y_i = g(\mathbf{X}_i) + e_i$, where the e_i's are i.i.d. with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$ and independent of the $\mathbf{X}_i$'s. Under the assumption $\frac{s_n^2}{\max_{1\leq i\leq n} W_{n,i}^2(\mathbf{t})} \longrightarrow \infty$ a.s., where $s_n^2 = \sigma^2 \sum_{i=1}^n W_{n,i}^2(\mathbf{t})$, we have*

$$\frac{\widehat{m}_n(\mathbf{t}) - E\{\widehat{m}_n(\mathbf{t})|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}}{s_n} \xrightarrow{d} N(0,1) \tag{4}$$

conditional on the $\mathbf{X}_i$'s, for almost all sequences $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$.

The following theorem shows that the estimated fractile regression function is a consistent estimator of $m(\mathbf{t})$.

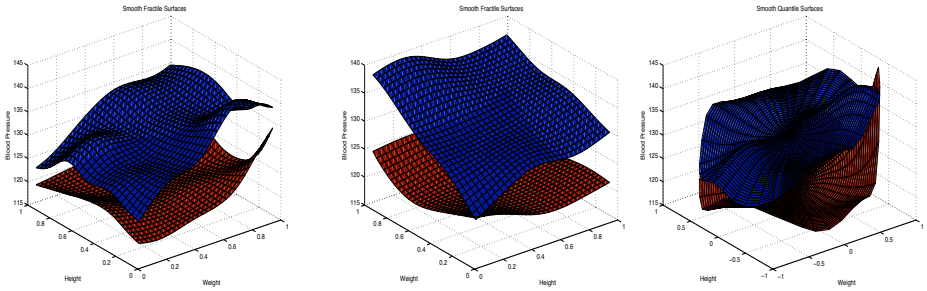
Theorem 4.2 *Fix* $\mathbf{t} \in E$. *Suppose that* $m(\mathbf{t}) = E\{Y|\mathbf{H}(\mathbf{X}) = \mathbf{t}\}$ *is continuous on* E *and* $|m(\mathbf{t})| \leq M \ \forall \ \mathbf{t} \in E$; *the conditional variance of* Y_i *given* $\mathbf{H}(\mathbf{X}_i)$ *is bounded, i.e.,* $v(\mathbf{t}) = Var\{Y_i|\mathbf{H}(\mathbf{X}_i) = \mathbf{t}\} \leq K_0 \ \forall \ \mathbf{t} \in E$; *and* $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{H}_n(\mathbf{x}) - \mathbf{H}(\mathbf{x})\| \xrightarrow{P} 0$. *Also assume conditions (W1)-(W4) on the weight functions, described in the appendix. Then, the conditional mean squared error of* $\widehat{m}_n(\mathbf{t})$ *approaches* 0 *in probability. As a consequence,*

$$\widehat{m}_n(\mathbf{t}) \xrightarrow{P} m(\mathbf{t}). \tag{5}$$

Recall that the estimated multivariate distribution transforms $\mathbf{R}_n$ and $\mathbf{S}_n$ are uniformly consistent estimators of $\mathbf{R_P}$ and $\mathbf{S_P}$ respectively. For the Nadaraya-Watson type weight function, conditions (W2) and (W4) are immediate. For compactly supported kernels, which are nonzero and bounded in a neighborhood of $\mathbf{0}$, and also the standard gaussian kernel, (W3) follows easily if $\|\mathbf{h}_n\| \longrightarrow 0$. Under the additional assumptions (i) $nh_{n,1}h_{n,2}\ldots h_{n,d} \longrightarrow \infty$, (ii) the uniform consistency of the estimated multivariate transform $\mathbf{H}_n$ and (iii) the existence of a non-vanishing density of $\mathbf{H}(\mathbf{X})$ in E, we can verify condition (W1). The condition on the weight function in Theorem 4.1 can also be verified under the above mentioned assumptions. Thus the conclusions of Theorems 4.1 and 4.2 hold for estimates based on the Nadaraya-Watson type weight function defined using the multivariate transforms $\mathbf{R_P}$ and $\mathbf{S_P}$.

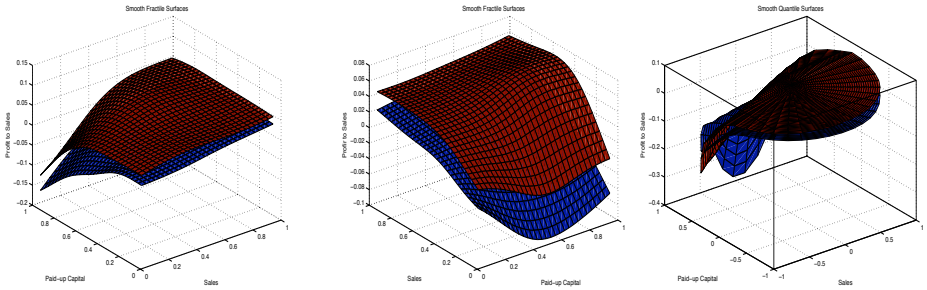
5 Fractile regression surfaces in examples

Example 1**:** On an average, individuals in the Toto tribe are heavier than those of the Bhutia tribe, and this makes the comparison of the usual regression surfaces difficult. Figure (6) shows the fractile regression surfaces for the



(a) (b) (c)

Figure 6: *Smooth fractile surfaces for comparing* Y = *blood pressure in the* ***Bhutia*** *(red) and* ***Toto*** *(blue) tribes in Example 1 (a) using the transformation* $\mathbf{R_P}$ *with* X_1 = *weight and* X_2 = *height, (b) using the transformation* $\mathbf{R_P}$ *with the order of the covariates reversed, and (c) using the transformation* $\mathbf{S_P}$.



(a) (b) (c)

Figure 7: *Smooth fractile surfaces for comparing* Y = *ratio of profit to sales for the years* ***1997*** *(red) and* ***2003*** *(blue) in Example 2 (a) using the transformation* $\mathbf{R_P}$ *with* X_1 = *sales and* X_2 = *paid-up capital, (b) using the transformation* $\mathbf{R_P}$ *with the order of the covariates reversed, and (c) using the transformation* $\mathbf{S_P}$.

Bhutia and Toto tribes using the multivariate distribution transforms $\mathbf{R_P}$ and $\mathbf{S_P}$. The two surfaces do not cross any longer because of the proper comparison of the regression surfaces. While comparing the regression surfaces, it is more meaningful to compare blood pressure of individuals in the same fractile group of height and weight for the two tribes rather than their actual covariate values. The multivariate distribution transform $\mathbf{R_P}$ exactly achieves this purpose.

An increase in weight increases blood pressure (on an average) for both the populations, though the relation is much more visible for the Toto tribe. The large peak in the blue surface [see Figure (1)] for large values of weight and height is absent in the fractile regression surfaces. On further investigation, we see that the spike was a result of covariate skewness and data sparsity for such large values of height and weight. We thus see that fractile regression surfaces are robust to extreme values of covariates. Also, the transformed covariates are now approximately independent whereas the original covariates – the height and the weight - are correlated. This is another interesting feature of the transformation $\mathbf{R_P}$ that can be exploited when dealing with correlated covariates.

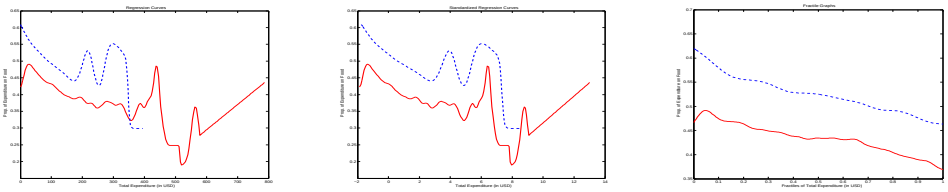
Example 2: In this example, we regress Y = ratio of profit to sales = $\frac{\text{profit}}{\text{sales}}$ against X_1 = sales and X_2 = paid-up capital. We study data for the years 1997 and 2003. The fractile regression surfaces for the two samples are shown in Figure (7). The estimated fractile surfaces (using both notions of multivariate distribution transform $\mathbf{R_P}$ and $\mathbf{S_P}$) for the year 2003 lie almost completely below that of 1997 indicating a fall in profit to sales ratio over the years. This decrease in profitability might be due to several reasons. One plausible reason

might be an increase in the number of companies (specially the emergence of foreign multinational companies) – the competitiveness among the companies has decreased their profitability. The analysis also indicates that larger companies (i.e., companies with large sales and paid-up capital) enjoy greater profitability whereas, on an average, those with low sales and high paid-up capital suffer the worst losses, as might be expected. These features are not at all prominent in the usual regression surfaces. It is very difficult to compare the usual regression surfaces as shown in Figure (2) because of the large changes in the distribution of the covariates over the two time points.

6 A Further Example

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 Soviet Union (see http://www.worldbank.org/research/inequality/data.htm 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.

A simple measure of the economic well-being of a population can be taken as the proportion of expenditure on food as a ratio of total expenditure per capita per household (in USD). This proportion would be quite small for rich



(a) (b) (c)

Figure 8: *(a) Usual regression curves, (b) location and scale shifted regression curves, and (c) fractile curves for proportion of expenditure on food with total expenditure for Poland (in red, solid line) and Bulgaria (in blue, dashed line) as discussed in Section 6.*

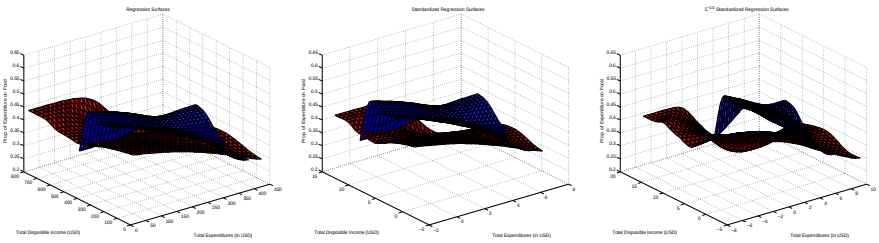
and wealthy people, but for the poor it would be close to one. By regressing this proportion on the total expenditure we can get a fair idea of the inequality in income and the economic condition of the populations.

To illustrate our point, 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 (8) shows the usual regression curves, regression curves with covariates standardized for location and scale and the smooth estimates of 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 (8a) show an initial decreasing trend but become very wiggly as total expenditure increases. Also the ranges of the covariates are quite 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. In Figure (8b), the two curves are more aligned, but still the wigglyness for higher total expenditure values is disturbing. To make the regression curves comparable, we need some standardization of the covariates.

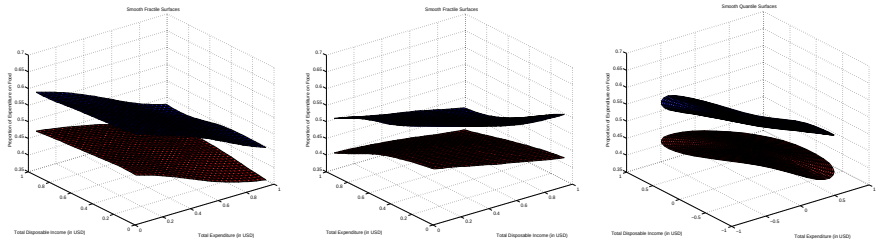
We would really like to compare the mean proportion of food expenditure for the poor (or the rich) in one population with that of the poor (or the rich) in the other population. The fractile curves accomplish exactly this, enabling us to compare the mean response values for fixed percentiles of total 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. From Figure (8c), it appears that the condition of households in Poland is uniformly economically better than those of Bulgaria. The standardization of the covariate also eliminates of the wigglyness of the earlier curves, which might be due to data sparseness for large covariate values.

As total disposable income is another financial indicator, our next step is to compare the proportion of expenditure on food to the total expenditure and the total disposable income. Figure (9a) shows the usual regression surfaces, while Figure (9b) shows the coordinate-wise location and scale shifted regression surfaces. Figure (9c) shows the regression surfaces when we standardize the covariate vector by subtracting its mean vector and multiplying by the inverse of the square-root of the dispersion matrix. It is important to know whether the crossing of the two surfaces at high covariate values is a real



(a) (b) (c)

Figure 9: *(a) Usual regression surfaces, (b) location and scale shifted regression surfaces, and (c) regression surfaces when the covariates are standardized by the inverse of the square-root of the dispersion matrix for proportion of expenditure on food (as a fraction of total expenditure) on total expenditure and total disposable income for the countries Poland (red) and Bulgaria (blue).*



(a) (b) (c)

Figure 10: *Smooth fractile surfaces for comparing Y = proportion of expenditure on food for the countries Poland (red) and Bulgaria (blue) (a) using the transformation $\mathbf{R_P}$ with X_1 = total expenditure and X_2 = total disposable income; (b) using the transformation $\mathbf{R_P}$ and the order of the covariates reversed; and (c) using the transformation $\mathbf{S_P}$.*

feature, as that would imply sharper economic inequality in Bulgaria (blue surface). But Figure (10) shows that the fractile surfaces do not cross; they rather share a very similar pattern over the entire domain of the covariates. This possibly reconfirms the fact that the households in Poland were better off than those of Bulgaria during the time of the survey.

Acknowledgement: We would like to thank Partha P. Majumder (Human Genetics Unit, Indian Statistical Institute, Kolkata) for providing us with the blood pressure data for the Bhutia and Toto tribes and the Reserve Bank of India for supplying the data on sales and profits of companies in India.

7 Appendix

Proof of Theorem 2.1 Suppose that $\mathbf{R}_{\mathbf{P}_1}(\mathbf{x}) = \mathbf{u}$, i.e., $F_{\mathbf{P}_1,i|1,2,\ldots,i-1}(x_i|x_1,\ldots,x_{i-1}) = u_i \; \forall \; i = 1,2,\ldots,d$, where $\mathbf{u} = (u_1, u_2, \ldots, u_d)$. Note that,

$$
\begin{aligned}
& F_{\mathbf{P}_1,i|1,2,\ldots,i-1}(x_i|x_1,\ldots,x_{i-1}) = u_i \\
\Rightarrow \quad & P(X_i \leq x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) = u_i \\
\Rightarrow \quad & P\left(\phi_i(X_i) \leq \phi_i(x_i) | \phi_1(X_1) = \phi_1(x_1), \ldots, \phi_{i-1}(X_{i-1}) = \phi_{i-1}(x_{i-1})\right) = u_i \\
\Rightarrow \quad & F_{\mathbf{P}_2,i|1,2,\ldots,i-1}(\phi_i(x_i)|\phi_1(x_1),\ldots,\phi_{i-1}(x_{i-1})) = u_i \; \forall \; i = 1,2,\ldots,d
\end{aligned}
$$

where $F_{\mathbf{P}_2,i|1,2,\ldots,i-1}$ is the conditional distribution function of Z_i given $Z_1, Z_2, \ldots, Z_{i-1}$, for $i = 1,2,\ldots,d$. Therefore, we have $\mathbf{R}_{\mathbf{P}_1}(\mathbf{x}) = \mathbf{u} = \mathbf{R}_{\mathbf{P}_2}(\mathbf{\Phi}(\mathbf{x}))$. Now, $E\{Y|\mathbf{R}_{\mathbf{P}_1}(\mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{R}_{\mathbf{P}_2}(\mathbf{\Phi}(\mathbf{X})) = \mathbf{t}\} = E\{Y|\mathbf{R}_{\mathbf{P}_2}(\mathbf{Z}) = \mathbf{t}\} \; \forall \; \mathbf{t} \in (0,1)^d$, which gives us the desired result. □

Proof of Theorem 2.2 We will show that $E\{Y|\mathbf{T}(\mathbf{P}_1, \mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{T}(\mathbf{P}_2, \mathbf{Z}) = \mathbf{t}\}$ for all $\mathbf{t} \in E$, for all random vectors $(\mathbf{X}, Y)$ with $\mathbf{X} \sim \mathbf{P}_1 \in \mathcal{P}$ is equivalent to $\mathbf{T}(\mathbf{P}_1, \mathbf{x}) = \mathbf{T}(\mathbf{P}_2, \mathbf{\Phi}(\mathbf{x}))\ \forall\ \mathbf{x} \in \mathbb{R}^d$. Given that $\mathbf{T}(\mathbf{P}_1, \mathbf{x}) = \mathbf{T}(\mathbf{P}_2, \mathbf{\Phi}(\mathbf{x}))\ \forall\ \mathbf{x} \in \mathbb{R}^d$ it is trivial to see that $E\{Y|\mathbf{T}(\mathbf{P}_1, \mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{T}(\mathbf{P}_2, \mathbf{Z}) = \mathbf{t}\}$. The other part follows from choosing $Y = X_i$ and simplifying the conditional expectations on both sides, for $i = 1, 2, \ldots, d$.

Note that $\mathbf{T}(\mathbf{P}_1, \mathbf{x}) = \mathbf{T}(\mathbf{P}_2, \mathbf{\Phi}(\mathbf{x}))\ \forall\ \mathbf{x} \in \mathbb{R}^d$ implies that $\mathbf{T}(\mathbf{P}_1, \mathbf{X}) \stackrel{d}{=} \mathbf{T}(\mathbf{P}_2, \mathbf{\Phi}(\mathbf{X}))$. Let $\mathbf{U} \sim \lambda$, where λ is the Uniform$(0,1)^d$ measure. Then $\mathbf{T}(\lambda, \mathbf{U}) \sim \mathbf{V}$ for some random vector $\mathbf{V}$. As any continuous random vector $\mathbf{X} \sim \mathbf{P}$ can be constructed from $\mathbf{U}$ using a coordinate-wise increasing transformation (which amounts to choosing suitable $\mathbf{\Phi}$), we conclude that $\mathbf{T}(\mathbf{P}, \mathbf{X}) \stackrel{d}{=} \mathbf{V}$ for all $\mathbf{X} \sim \mathbf{P} \in \mathcal{P}$. □

Proof of Theorem 4.1 As in the proof of Theorem 4.2, all expectations and variances denoted below are conditional on the $\mathbf{X}_i$'s, $i = 1, 2, \ldots, n$. Note that

$$\widehat{m}_n(\mathbf{t}) - E\{\widehat{m}_n(\mathbf{t})\} = \sum_{i=1}^{n} W_{n,i}(\mathbf{t}) e_i.$$

To find the conditional limiting distribution of $\sum_{i=1}^n W_{n,i}(\mathbf{t})e_i$ given the $\mathbf{X}_i$'s, let us define $Z_{n,i} = W_{n,i}(\mathbf{t})e_i$ for $i = 1, 2, \ldots, n$. Let $S_n = \sum_{i=1}^n Z_{n,i}$. We use the Lindeberg-Feller Central Limit Theorem to find the asymptotic distribution of S_n.

Observe that $E(Z_{n,i}) = 0$ and $\sigma_{n,i}^2 = Var(Z_{n,i}) = \sigma^2 W_{n,i}^2(\mathbf{t})$. Let $s_n^2 = \sum_{i=1}^n \sigma_{n,i}^2 = \sigma^2 \sum_{i=1}^n W_{n,i}^2(\mathbf{t})$. For any $\eta > 0$ and nonzero $W_{n,i}^2(\mathbf{t})$, the Lindeberg-

Feller condition can be simplified as

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{1}{s_n^2} \int_{|Z_{n,i}|>\eta s_n} Z_{n,i}^2 dP = \sum_{i=1}^{n} \frac{1}{s_n^2} \int_{e_i^2 > \eta^2 \frac{s_n^2}{W_{n,i}^2(\mathbf{t})}} W_{n,i}^2(\mathbf{t}) e_i^2 dP \\
\leq \; & \sum_{i=1}^{n} \frac{1}{s_n^2} \int_{e_i^2 > \eta^2 \frac{s_n^2}{\max_{1\leq i\leq n} W_{n,i}^2(\mathbf{t})}} W_{n,i}^2(\mathbf{t}) e_i^2 dP \\
\leq \; & \sigma^{-2} \int_{e_1^2 > \eta^2 \frac{s_n^2}{\max_{1\leq i\leq n} W_{n,i}^2(\mathbf{t})}} e_1^2 dP \longrightarrow 0 \text{ as } n \to \infty \qquad (6)
\end{aligned}$$

by the assumption of the theorem. The result now follows from the Lindeberg-Feller Central Limit Theorem, i.e., $\frac{\sum_{i=1}^n W_{n,i}(\mathbf{t}) e_i}{\sigma \sqrt{\sum_{i=1}^n W_{n,i}^2(\mathbf{t})}} \xrightarrow{d} N(0,1)$ conditional on the $\mathbf{X}_i$'s, for almost all sequences $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. □

The regularity conditions on the weight functions applicable for Theorem 4.2 are described below.

(W1) $\sum_{i=1}^n W_{n,i}^2(\mathbf{t}) \xrightarrow{P} 0$ as $n \to \infty$.

(W2) $\sum_{i=1}^n W_{n,i}(\mathbf{t}) \xrightarrow{P} 1$ as $n \to \infty$.

(W3) The weights are asymptotically localized, i.e., there exists a sequence $\{\delta_n\}_{n=1}^{\infty}$ with $\delta_n \to 0$ such that

$$\sum_{i=1}^{n} |W_{n,i}(\mathbf{t})| \mathbf{1}_{\{\|\mathbf{t}-\mathbf{H}_n(\mathbf{X}_i)\|>\delta_n\}} \xrightarrow{P} 0 \text{ as } n \to \infty.$$

(W4) There exists $D \geq 1$ such that $P(\sum_{i=1}^n |W_{n,i}(\mathbf{t})| \leq D) = 1 \; \forall \; n \geq 1$.

Note that conditions (W1)-(W4) are and similar to those used by Stone (1977).

Proof of Theorem 4.2 In the following theorem, all expectations are conditional expectations given the $\mathbf{X}_i$'s, $i = 1, 2, \ldots, n$. For $\mathbf{t} \in E$, the conditional

mean squared error can be decomposed as

$$E\left\{\widehat{m}_n(\mathbf{t}) - m(\mathbf{t})\right\}^2 = E\left\{\widehat{m}_n(\mathbf{t}) - E(\widehat{m}_n(\mathbf{t}))\right\}^2 + \left\{E(\widehat{m}_n(\mathbf{t})) - m(\mathbf{t})\right\}^2. \quad (7)$$

The conditional variance term, $E\{\widehat{m}_n(\mathbf{t}) - E(\widehat{m}_n(\mathbf{t}))\}^2$, can be simplified as

$$\begin{aligned}
& E\left[\sum_{i=1}^{n}\left\{Y_i - m(\mathbf{H}(\mathbf{X}_i))\right\} W_{n,i}(\mathbf{t})\right]^2 \\
= \; & E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{Y_i - m(\mathbf{H}(\mathbf{X}_i))\right\}\left\{Y_j - m(\mathbf{H}(\mathbf{X}_j))\right\} W_{n,i}(\mathbf{t}) W_{n,j}(\mathbf{t})\right] \\
= \; & \sum_{i=1}^{n} E\left\{Y_i - m(\mathbf{H}(\mathbf{X}_i))\right\}^2 W_{n,i}^2(\mathbf{t}) \\
\leq \; & K_0 \sum_{i=1}^{n} W_{n,i}^2(\mathbf{t}) \xrightarrow{P} 0 \text{ by assumption (W1) and the fact that } v(\mathbf{t}) \text{ is bounded.}
\end{aligned}$$

To show that the conditional bias goes to 0 in probability, we decompose it as

$$\begin{aligned}
& \sum_{i=1}^{n} m(\mathbf{H}(\mathbf{X}_i)) W_{n,i}(\mathbf{t}) - m(\mathbf{t}) \\
= \; & \sum_{i=1}^{n}\left\{m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})\right\} W_{n,i}(\mathbf{t}) + m(\mathbf{t})\left\{\sum_{i=1}^{n} W_{n,i}(\mathbf{t}) - 1\right\}. \quad (8)
\end{aligned}$$

Note that the second term in (8) goes to 0 in probability by assumption (W2). We will show that $\sum_{i=1}^{n} V_{n,i} \xrightarrow{P} 0$ as $n \to \infty$, where $V_{n,i} = \{m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})\}$ $W_{n,i}(\mathbf{t})$. Let $\epsilon > 0$ and $\eta > 0$ be given. To simplify writing, we denote the event $\{\|\mathbf{t} - \mathbf{H}_n(\mathbf{X}_i)\| \leq \delta_n\}$ as $E_{n,i}$. Therefore,

$$\begin{aligned}
& P\left(\left|\sum_{i=1}^{n} V_{n,i}\right| > \epsilon\right) \leq P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}}\right| > \epsilon/2\right) + P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}^c}\right| > \epsilon/2\right) \\
\leq \; & P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}}\right| > \epsilon/2\right) + \eta/2 \quad \text{for all } n \geq N_1 \text{ as} \quad (9)
\end{aligned}$$

$P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}^c}\right| > \epsilon/2\right) \leq P\left(\sum_{i=1}^{n} |V_{n,i}|\mathbf{1}_{E_{n,i}^c} > \epsilon/2\right)$
$\leq P\left(2M\sum_{i=1}^{n} |W_{n,i}(\mathbf{t})|\mathbf{1}_{E_{n,i}^c} > \epsilon/2\right) \leq \eta/2 \; \forall \; n \geq N_1$ by the fact that $m(\mathbf{t})$ is bounded and (W3).

Let $B_n = \sup_{\mathbf{x}\in\mathbb{R}} \|\mathbf{H}_n(\mathbf{x}) - \mathbf{H}(\mathbf{x})\|$. By assumption, we know that $B_n \xrightarrow{P} 0$. Observe that, $\|\mathbf{t} - \mathbf{H}_n(\mathbf{X}_i)\| \leq \delta_n$ and $\|\mathbf{H}(\mathbf{X}_i) - \mathbf{H}_n(\mathbf{X}_i)\| \leq B_n$ implies that $\|\mathbf{H}(\mathbf{X}_i) - \mathbf{t}\| \leq \|\mathbf{H}_n(\mathbf{X}_i) - \mathbf{t}\| + \|\mathbf{H}(\mathbf{X}_i) - \mathbf{H}_n(\mathbf{X}_i)\| \leq \delta_n + B_n$ for all $i = 1, 2, \ldots, n$.

Also notice that as $m(\cdot)$ is continuous at $\mathbf{t}$, there exists $\delta > 0$ such that $\|\mathbf{H}(\mathbf{X}_i) - \mathbf{t}\| \leq \delta \Rightarrow |m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})| \leq \frac{\epsilon}{2D}$. Now,

$$\begin{aligned}
& P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}}\right| > \epsilon/2\right) \leq P\left(\sum_{i=1}^{n} |V_{n,i}|\,\mathbf{1}_{E_{n,i}} > \epsilon/2\right) \\
\leq \; & P\left(\max_{1\leq i\leq n} |m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})|\,\mathbf{1}_{E_{n,i}} \sum_{i=1}^{n} |W_{n,i}(\mathbf{t})| > \epsilon/2\right) \\
\leq \; & P\left(\max_{1\leq i\leq n} |m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})|\,\mathbf{1}_{E_{n,i}} > \frac{\epsilon}{2D}\right) \\
\leq \; & P\left(\delta_n + B_n > \delta\right) < \eta/2 \; \forall \; n \geq N_2 \qquad (10)
\end{aligned}$$

as $\delta_n + B_n \xrightarrow{P} 0$. The last two inequalities follow because $|m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})|$ $\mathbf{1}_{E_{n,i}} > \frac{\epsilon}{2D}$ implies that $\|\mathbf{H}(\mathbf{X}_i) - \mathbf{t}\| > \delta$ and $\|\mathbf{t} - \mathbf{H}_n(\mathbf{X}_i)\| \leq \delta_n$, which in turn implies that $\delta_n + B_n > \delta$.

Using (8), (9) and (10), we conclude $P(|\sum_{i=1}^{n}\{m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})\}W_{n,i}(\mathbf{t})| > \epsilon) < \eta$ for all $n \geq \max\{N_1, N_2\}$. Thus, the conditional mean squared error of $\widehat{m}_n(\mathbf{t})$ approaches 0 in probability.

Now, using Chebyshev's inequality, we have

$$\begin{aligned} & P(|\widehat{m}_n(\mathbf{t}) - m(\mathbf{t})| \geq \epsilon | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) \\ \leq \; & E[\{\widehat{m}_n(\mathbf{t}) - m(\mathbf{t})\}^2 | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n]/\epsilon^2 \xrightarrow{P} 0 \text{ as } n \to \infty. \end{aligned}$$

An application of the dominated convergence theorem completes the proof of the Equation (5). □

References

[1] BHATTACHARYA,P.K. (1963). On an Analog of Regression Analysis. *Ann. Math. Statis.* **34**, 1459-1473.

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

[3] CHAUDHURI,P. (1996). On a Geometric Notion of Quantiles for Multivariate Data. *J. Amer. Statist. Assoc.*, **91**, 862-872.

[4] KOLTCHINSKII,V.I. (1997). M- Estimation, Convexity and Quantiles, *Ann. Statist.*, **25**, No. 2, 435- 477.

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

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

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

[8] RUPPERT, D., SHEATHER, S.J., and WAND, M.P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, **90**, 1257-1270.

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

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

[11] STONE,C.J. (1977). Consistent Nonparametric Regression. *Ann. Statist.*, **5**, 595-620.

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