On Fractile Transformation of Covariates in Regression

Bodhisattva Sen and Probal Chaudhuri
Columbia University and Indian Statistical Institute
New York, NY 10027 203 B.T. Road, Kolkata 700108*

Abstract

The need for comparing two 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 are different 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 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 analytically as well as numerically. We also provide a few real examples that illustrate the difficulty in comparing the usual regression functions and motivate the need for the fractile transformation. Our analysis of

*Bodhisattva Sen is Assistant Professor, Department of Statistics, Columbia University, New York, NY 10027 (E-mail: bodhi@stat.columbia.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).
the real examples leads to new and useful statistical conclusions that are missed by comparison of the usual regression functions.

KEY WORDS: Asymptotic normality, consistency, fractile graphical analysis, kernel smoothing, multivariate quantile, nonparametric regression.

1 Introduction

Comparison of two regression functions can be a difficult task 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 were 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 coordinate-wise location and scale adjusted 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. This convention is followed in computing all the bivariate regression surfaces illustrated in the paper. One may use other standard nonparametric regression tools, but it is our empirical experience that it does not change the main results and findings.

A disturbing feature in the three figures is the crossing of the two regression surfaces. 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 leads to 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 non-government, non-financial public limited companies in India over different years. Here paid-up capital refers to the total amount of shareholder capital that has been paid in full by shareholders. The Reserve Bank of India is
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. The uneven covariate distribution leads to data sparsity in certain regions of the covariate space and causes distortion of the estimated regression surfaces. The choice of the 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 \(\mu_1\) and \(\mu_2\) where \(\mu_1(x) = E(Y_1|X_1 = x)\) and \(\mu_2(x) = E(Y_2|X_2 = x)\). Then the fractile regression functions are defined as

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

for \(t \in (0,1)\), where \(F_1\) and \(F_2\) are the distribution functions of \(X_1\) and \(X_2\) respectively [see Mahalanobis (1960)]. Note that the transformed covariance
Figure 3: (a) Usual regression curves, (b) location and scale adjusted regression curves, and (c) fractile regression curves, for blood pressure against weight for the Bhutia (in red, solid line) and Toto (in blue, dashed line) tribes.

izes $F_1(X_1)$ 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 $\phi$ is any strictly increasing transformation, then $E\{Y_1|F_1(X_1)\} = E\{Y_2|F_2(X_2)\}$. This is a crucial property and can be interpreted in the following way: The usual location and scale standardization makes the regression functions comparable in situations where the covariates in the two populations are linear transformations of one another, whereas, the fractile transformation makes the regression functions comparable even when the covariate in the second population is any increasing transformation of that of the first population. Fractile regression has been considered earlier in Mahalanobis (1960), Sethuraman (1961),

In Figure (3), we have plotted the usual regression curves, regression curves with covariates standardized for location and scale and the smooth estimates of fractile regression 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. The highly irregular regression curves obtained in Figures (4a) and (4b) is due to the very uneven covariate distribution with data sparsity in some regions of the covariate space. 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 univariate regression plots shown in the paper 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).

Fractile regression techniques with one covariate have been applied in diverse settings. Nordhaus (2006) shows fractile plots of key geographic variables (temperature, precipitation, latitude, etc.) against the fractiles of log of “output density” while trying to explore the linkage between economic activity and geography. Hertz-Picciotto and Din-Dzietham (1998) compare the infant mortality using a “percentile based method” of standardization for birthweight or gestational age. Their motivation underlying the percentile-based method of standardization is that comparable health for two population groups will be expressed as equal rates of disease or mortality at equal percentiles in the distributions of either birthweight or gestational age.

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 define a suitable notion of multivariate quantile based on successive conditioning of the covariates and use it to define the fractile regression function. We also briefly discuss another notion of multivariate quantile, namely the spatial quantile described in Chaudhuri (1996) and Koltchinskii (1997), and the associated fractile regression function. Section 3 discusses nonparametric smooth estimation of the fractile regression function from a sample of data points. A simulation study shows the superiority of our method over usual regression analysis without proper standardization.
of the covariates. We also prove the consistency and asymptotic normality of the fractile regression estimates. The fractile surfaces for Examples 1 and 2 are presented in Section 4 followed by a brief discussion. Section 5 provides another application of fractile regression techniques on real data, namely, Household Survey data for Transitional Economies. In Section 6 we discuss some extensions and alternative approaches to our method. 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 3.1.

2 Fractile Transformation

For a $d$-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ with distribution $P$ on $\mathbb{R}^d$, we define the fractile transformation $R_P : \mathbb{R}^d \mapsto [0, 1]^d$, a multivariate analogue of the univariate distribution function [i.e., $x \mapsto F_X(x)$], as

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

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 is expressed as $G_P(u) = (F_1^{-1}(u_1), F_{2|1}^{-1}(u_2|u_1), \ldots, F_{d|1,2,\ldots,d-1}^{-1}(u_d|u_1, u_2, \ldots, u_{d-1}))$ for $u = (u_1, u_2, \ldots, u_d) \in (0, 1)^d$ [see Bhattacharyya (1963)]. Recently Wei (2007) used a similar idea on successive conditioning of the coordinate variables of a random vector in the quantile regression setup on bivariate growth curves. Chesher (2003) and Ma and Koenker (2006) use the idea of recursive conditioning in quantile regression for structural econometric models.

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.

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

As we have discussed in the previous section, we need a 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. 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 $R_P$, the multivariate distribution transform discussed above, to standardize the covariates and regress $Y$ on $R_P(X)$ (where $X \sim P$). In other words, we define the fractile regression function obtained by using the multivariate distribution transform $R_P$ as

$$m(t) = E\{Y|R_P(X) = t\} \text{ for } t \in (0, 1)^d.$$
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 $R_P$ as a nonparametric standardization tool for the covariates.

**Theorem 2.1** Let $(X, Y) \in \mathbb{R}^{d+1}$ be a random vector such that $X = (X_1, X_2, \ldots, X_d) \sim P_1$ has a density on $\mathbb{R}^d$. Also, let $Z = (Z_1, Z_2, \ldots, Z_d) \sim P_2$, where $Z_j = \phi_j(X_j)$ for all $j = 1, 2, \ldots, d$, and each $\phi_j$ is a strictly increasing function on $\mathbb{R}$. Then $R_{P_1}(x) = R_{P_2}(\Phi(x))$ where $\Phi(x) = (\phi_1(x_1), \phi_2(x_2), \ldots, \phi_d(x_d))$.

In particular,

$$E\{Y|R_{P_1}(X) = t\} = E\{Y|R_{P_2}(Z) = t\} \text{ for all } t \in (0, 1)^d.$$ 

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

**Theorem 2.2** Let \( X, Y, Z, P_1 \) and \( P_2 \) be as in Theorem 2.1, and suppose that there exists a transformation \( T: \mathcal{P} \times \mathbb{R}^d \rightarrow E \subset \mathbb{R}^d \) as described above such that \( E\{Y|T(P_1, X) = t\} = E\{Y|T(P_2, Z) = t\} \) for all \( t \in E \), and equality holds for all joint distributions \((X, Y)\), with \( X \sim P_1 \in \mathcal{P} \). Then \( T(P, X) \) must have a fixed distribution, i.e., we must have \( T(P, X) \overset{d}{=} V \), for some fixed random vector \( V \), for all \( X \sim P \in \mathcal{P} \).

The computation of \( R_P \) from a sample of data points \( X_1, X_2, \ldots, X_n \sim P \) 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 \( X_i \), and then use it to get various marginal and conditional densities. To compute the conditional densities in the subsequent sections we have used the Gaussian kernel with bandwidths chosen by cross validation and computation was done by using the “sm” package in R developed by Adrain Bowman and Adelchi Azzalini, which has already been mentioned in Section 1.

Let \( R_n \) be the estimated distribution transform obtained from the sample. Under appropriate conditions on the kernel and the smoothing parameter(s), it can be easily shown that \( R_n \) is a uniformly consistent estimator of \( R_P \), i.e.,

\[
\sup_{x \in \mathbb{R}^d} \|R_n(x) - R(P, x)\| \overset{P}{\rightarrow} 0.
\]

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.
Though the multivariate transform $R_P$ 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 $R_P$ depends on the ordering of the coordinates of the covariate vector. Changing the order of the coordinate variables would change the transformation. In applications where there is a natural ordering of the importance of the covariates, we advocate the use of that ordering, conditioning successively with the less important predictor. In other situations, different orderings can be tried out and the ordering for which the distribution transform is closest to $\text{Uniform}(0,1)^d$ (in some metric) can be used.

2.1 An alternative to $R_P$: spatial distribution transform

There are several multivariate versions of quantiles that can be used to define fractile regression function. 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 variables in the covariate vector. This concept of multivariate quantile, called the spatial quantile (or geometric quantile), was introduced and studied by Chaudhuri (1996) and Koltchinskii (1997) [also see Breckling and Chambers (1988)].

We define the spatial distribution function $S_P$ as

$$S_P(x) = E_P\left(\frac{x - X}{\|x - X\|}\right)$$

for all $x \in \mathbb{R}^d$, where $X \sim P$. Note that $S_P$ and the corresponding quantile map $Q_P$ are invertible functions, and one is the inverse of the other. Unlike
\( R_P(X), \ S_P(X) \) is equivariant under orthogonal linear transformations of \( X \), and this ensures equivariance of the corresponding fractile regression function under such transformations. This is very desirable especially in situations were the covariates have a spherically symmetric distribution. In particular, \( S_P(X) \) is equivariant under permutations of the co-ordinates of \( X \). However, the spatial quantile is not equivariant under arbitrary increasing transformations of the marginal variables. The spatial distribution function can also be viewed as an extension of the usual distribution function in the univariate case. Suppose that we have a sample \( X_1, X_2, \ldots, X_n \sim P \) in \( \mathbb{R}^d \). The empirical spatial distribution function is defined as

\[
S_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{x - X_i}{\|x - X_i\|}.
\]

Computation of the sample spatial distribution is simple and the asymptotic properties are known. From Theorem (5.5) in Koltchinskii (1997), it follows that \( D_n = \sup_{x \in \mathbb{R}^d} \|S_n(x) - S_P(x)\| \to 0 \) 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 \( \{(X_i, Y_i)\}_{i=1}^{n} \) from a population in \( \mathbb{R}^{d+1} \) with a continuous density function, where \( X_i \sim P \). Let \( (X, Y) \) be a generic random vector having the same joint distribution. The methodology is described with a general version of multivariate distribution (fractile) trans-
form $H : \mathbb{R}^d \rightarrow E \subset \mathbb{R}^d$ (which may or may not be $R_P$ or $S_P$). We want to estimate the fractile regression function

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

We define the smooth estimated fractile regression function as

$$\hat{m}_n(t) = \sum_{i=1}^{n} Y_i W_{n,i}(t) \text{ for } t \in E, \quad (1)$$

where $W_{n,i}(t)$ is the weight function, which might depend on $H_n$, the empirical or estimated value of $H$. Many standard nonparametric regression estimates (e.g., kernel, local polynomial, nearest neighbor, spline regressions) can be expressed in the form of such weighted averages with appropriate choices of weight functions. For instance, if kernel based Nadaraya-Watson type weight function is used, we have $W_{n,i}(t) = \frac{K(\frac{t-H_n(X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{t-H_n(X_j)}{h_n})}$, where $K$ is a kernel function defined on $\mathbb{R}^d$, $t = (t_1, t_2, \ldots, t_d) \in E$, $\frac{t-H_n(X_i)}{h_n} = \left(\frac{t_1-H_{n,1}(X_i)}{h_{n,1}}, \frac{t_2-H_{n,2}(X_i)}{h_{n,2}}, \ldots, \frac{t_d-H_{n,d}(X_i)}{h_{n,d}}\right)$, and $h_{n,1}, h_{n,2}, \ldots, h_{n,d}$ are the smoothing bandwidths.

### 3.1 Simulation study

**Fractile regression surfaces:** We consider two tri-variate normal samples of size 400 each drawn from the population $(X_1, X_2, Y) \sim N(0, \Sigma)$ with $

\Sigma = (\sigma_{i,j})_{3 \times 3}$ such that $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0.5$ for $i \neq j$. To illustrate the usefulness of fractile regression, we transform the covariates in the second sample nonlinearly as $X_1 \mapsto X_1^{2}$ and $X_2 \mapsto X_2^{1.75}$ for positive values of $X_1$ and $X_2$, while the negative values of $X_1$ and $X_2$ remain unchanged. Figure (5a) shows the smoothed regression surfaces, whereas Figure (5b) shows the smoothed
fractile regression surfaces, using the transformation $R_p$, for the two samples. It is easy to see that the usual regression surfaces are not comparable and look very different, whereas, the two estimated fractile regression surfaces are very similar, as should be the case.

**Comparison based on integrated mean squared error (IMSE):** In some situations, the use of fractile regression can provide better (e.g., in terms of IMSE) estimators of the underlying regression function. Tables 1 and 2 show the estimated IMSE in a simulation study, with one or two covariates (using sample size 400 and 500 monte carlo replications) for four models: (a) $Y = \exp(-X) + \epsilon$; (b) $Y = X + \epsilon$; (c) $Y = \exp(-X_1X_2) + \epsilon$; and (d) $Y = (X_1 + X_2)/2 + \epsilon$. For computational purpose, the IMSE was approximated by evaluating the squared difference of the estimator and the truth at all the data points and then taking a simple average. We see that the estimated fractile regression functions, using the transformation $R_p$, have considerably
lower IMSEs in most of the cases. The estimated regression functions perform poorly while estimating the mean response for extreme covariate values, because of data sparsity and/or high error variance. For extreme values of covariates, the averaging (smoothing) of the response involves only a few observations owing to the small number of data points present in the smoothing neighborhood, and this produces estimates with large variances. The fractile regression functions perform better as the transformed covariates are approximately uniformly distributed on $[0, 1]^d$, and smoothing over a fixed bandwidth involves averaging with similar number of observations.

Notice that in models (a) and (c), the regression function is bounded by 0 and 1. In general, for regression functions that are bounded (have bounded asymptotes), fractile regression works better than usual regression. This effect is more pronounced when we have unevenly distributed covariates and when the error distribution is heteroscedastic (both phenomena are observed in most of our applications). Note that the uneven distribution of covariates in the simulation study is caused by the extreme observations generated from the heavy-tailed distributions. In case of normal linear models (b) and (d), the performance of fractile regression is slightly inferior than that of the estimated regression functions. Note that the true regression function in these models is linear whereas the true fractile regression function has curvature, and this...
Table 2: Ratio of the estimated IMSE for smoothed estimates of usual regression and fractile regression functions with two covariates when the data generating model is (c) \( Y = \exp(-X_1 X_2) + \epsilon \) and (d) \( Y = (X_1 + X_2)/2 + \epsilon \).

<table>
<thead>
<tr>
<th>Model</th>
<th>((X_1, X_2))</th>
<th>(\epsilon)</th>
<th>IMSE ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c)</td>
<td>(</td>
<td>N(0, 1)</td>
<td>\times</td>
</tr>
<tr>
<td>(c)</td>
<td>(</td>
<td>N(0, 1)</td>
<td>\times</td>
</tr>
<tr>
<td>(c)</td>
<td>(</td>
<td>t_4</td>
<td>\times</td>
</tr>
<tr>
<td>(c)</td>
<td>(</td>
<td>t_4</td>
<td>\times</td>
</tr>
<tr>
<td>(d)</td>
<td>(N_2((0, 0), I_2))</td>
<td>(N(0, 1))</td>
<td>0.72</td>
</tr>
<tr>
<td>(d)</td>
<td>(N_2((0, 0), I_2))</td>
<td>(N(0,</td>
<td>X_1</td>
</tr>
</tbody>
</table>

makes its estimation more difficult, resulting in slightly larger IMSEs. As the distribution of the transformed covariates is Uniform(0,1)\(^d\), the choice of the smoothing bandwidth for estimating the fractile regression function is relatively simpler and more stable.

### 3.2 Some asymptotic results

The simulation study illustrates the superiority of fractile regression compared with usual regression in many settings. Since it is well known that the standard nonparametric regression estimators are consistent under very general conditions, one would expect similar asymptotic results to hold for fractile regression estimates. This is indeed the case as is illustrated in the following theorem.

**Theorem 3.1** Fix \( t \in E \). Suppose that \( m(t) = E\{Y|H(X) = t\} \) is continuous on \( E \) and \( |m(t)| \leq M \ \forall \ t \in E \); the conditional variance of \( Y_i \) given \( H(X_i) \) is bounded, i.e., \( v(t) = \text{Var} \{Y_i|H(X_i) = t\} \leq K_0 \ \forall \ t \in E \); and \( \sup_{x \in \mathbb{R}^d} \|H_n(x) - H(x)\| \overset{L^p}{\to} 0 \). Also assume conditions (W1)-(W4) on the weight functions, described in the Appendix. Then, the conditional mean
Figure 6: Smooth fractile surfaces for comparing $Y = \text{blood pressure}$ in the **Bhutia** (red) and **Toto** (blue) tribes in Example 1 (a) using the transformation $R_P$ with $X_1 = \text{weight}$ and $X_2 = \text{height}$, (b) using the transformation $R_P$ with the order of the covariates reversed, and (c) using the transformation $S_P$.

Squared error of $\hat{m}_n(t)$ approaches 0 in probability. As a consequence,

$$\hat{m}_n(t) \xrightarrow{P} m(t). \quad (2)$$

Suppose now that condition (C1) (stated in the Appendix) hold. Letting $s_n^2 = \sum_{i=1}^n v(H(X_i))W_{n,i}^2(t)$, we have

$$\frac{\hat{m}_n(t) - E\{\hat{m}_n(t)|X_1, X_2, \ldots, X_n\}}{s_n} \xrightarrow{d} N(0, 1) \quad (3)$$

conditional on the $X_i$’s, for almost all sequences $X_1, X_2, \ldots, X_n$.

4 Fractile regression surfaces in the two 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 Bhutia and Toto tribes using the multivariate distribution transforms $R_P$ and
The two surfaces do not cross any longer because of a more appropriate comparison of the regression surfaces. While comparing the regression surfaces, it is more meaningful to compare blood pressure of individuals in the same quantile group of height and weight for the two tribes rather than their actual covariate values. The multivariate distribution transform $R_P$ exactly achieves this purpose. In Figure (1), the surfaces were plotted with matched covariate values, but the matching covariates may belong to distinct fractile groups leading to improper comparison of the corresponding blood pressure values.

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 in Figure (1) corresponding to large values of weight and height is absent in the fractile regression surfaces. On a careful investigation, we saw that the spike was a result of uneven covariate distribution and data sparsity around such large values of height and weight. In such regions, the regression surfaces were essentially obtained as a weighted average of a few very large response values. After the fractile transformation the transformed covariates are uniformly distributed which resolves the earlier problem of data sparsity in certain regions of the covariate space. We thus see that fractile regression surfaces are robust to extreme values of covariates.

**Example 2:** In this example, we regress $Y = \text{ratio of profit to sales}$ against $X_1 = \text{sales}$ and $X_2 = \text{paid-up capital}$. 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 $R_P$ and $S_P$) for the year 2003 lie almost completely below that of 1997 indicating a fall in profit to sales ratio over the years. This
Figure 7: Smooth fractile surfaces for comparing $Y = \text{ratio of profit to sales}$ for the years 1997 (red) and 2003 (blue) in Example 2 (a) using the transformation $R_P$ with $X_1 = \text{sales}$ and $X_2 = \text{paid-up capital}$, (b) using the transformation $R_P$ with the order of the covariates reversed, and (c) using the transformation $S_P$.

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 change in the distribution of the covariates over the two time points.

5 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/ for more informa-
It was created as part of a project analyzing poverty and existing social assistance programs in the transitional economies. What immediately arrests attention is the startling drop in income and increase in inequality accompanying the transition of these countries to market economies. We investigate this inequality in income and compare the economic condition of the transitional economies.

A simple measure of the economic well-being of a population can be taken as the proportion of expenditure on food as a fraction of total expenditure per capita per household (in USD). This proportion would be quite small for rich and wealthy people, but for the poor it would be close to one. 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
Figure 9: (a) Usual regression surfaces, (b) location and scale adjusted 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).

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

As total disposable income is another financial indicator, our next step is to consider the regression problem with the fraction of expenditure on food as the response and total expenditure and total disposable income as the two covariates. We intend to compare the regression surfaces for the Bulgarian and the Polish populations. Figure (9a) shows the usual regression surfaces, while Figure (9b) shows the coordinate-wise location and scale adjusted 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 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.

6 Concluding remarks

There are many multivariate versions of quantiles that can be used to define the fractile regression function [see Hallin, Paindaveine and Siman (2009), Kong and Mizera (2008), Serfling (2002), Liu, Parelius, and Singh (1997), and Zuo and Serfling (2000) for various notions of multivariate quantiles]. In fact, the results in Section 3 are applicable for any version of multivariate fractile for which a uniformly consistent estimator exists (see Theorem 3.1). For the sake of illustration and in view of their appealing properties, we have discussed $R_\mathcal{P}$ and $S_\mathcal{P}$ in detail. Note that the distribution of $R_\mathcal{P}(X)$ is invariant to the distribution of $X \sim \mathcal{P}$, and the corresponding fractile regression function is also invariant under coordinate-wise increasing transformations of $X$ making $R_\mathcal{P}$ (and the associated quantile) very suitable for covariate standardization. The equivariance of $S_\mathcal{P}(X)$ under orthogonal linear transformations makes it a particularly useful standardization tool when dealing with spherically symmetric covariate distributions. Affine equivariant version of the spatial multivariate quantile [Chakraborty (2001)] can also be used to generalize the equivariance
of the fractile regression function to elliptically symmetric distributions. Also, the computation of \( S_p(X) \) is much simpler than some of the other versions of multivariate fractile.

As an alternative to modeling the joint distribution of the covariate, which involves dealing with multivariate quantiles, one can use a marginal quantile transformation to standardize the covariates. To motivate the notion of marginal modeling of covariates, consider the blood pressure example (Example 1 in Section 1). We would like to compare the mean blood pressures for the two tribes, Toto and Bhutia, for the same “quantile”-value of height and weight. A biologist might be interested in knowing the difference in mean blood pressure for individuals with “median” height and “median” weight for the two tribes. Such comparisons involving univariate notion of quantile can be accomplished by studying the function

\[
m(t) = E \{ Y | X_1 = F_1^{-1}(t_1), \ldots, X_d, 1 = F_d^{-1}(t_d) \}
\]

for the two populations, where \( X = (X_1, X_2, \ldots X_d) \), \( F_i \) is the marginal distribution function of \( X_i \), \( i = 1, 2, \ldots, d \), and \( t = (t_1, t_2, \ldots, t_d) \in (0, 1)^d \).

Thus, we can define another notion of nonparametric covariate standardization by considering the transformation \( M_P : (x_1, x_2, \ldots x_d) \mapsto (F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \). This standardization is conceptually and computationally much simpler as it only requires the estimation of univariate distribution functions. This transformation retains the property of invariance under arbitrary increasing transformations of the covariates (as discussed in Section 2). It also does not change with the relabeling of the covariates, unlike the \( \text{R}_P \) transform.

The transformed regression function is defined as \( m(t) = E \{ Y | M_P(X) = \)
for \( t \in (0,1)^d \), and is related to the the usual regression function \( \mu(x) = E\{Y|X = x\} \) through \( m(t) = \mu(F_1^{-1}(t_1), F_2^{-1}(t_2), \ldots, F_d^{-1}(t_d)) \). We can estimate the usual regression function in any way we like, using non-parametric or parametric techniques, and then estimate the marginal quantile functions (which involves one-dimensional smoothing) to estimate \( m \).

One obvious drawback of \( \text{MP} \), being a marginal transformation, is that it does not take into account the joint distribution of \( X \). In particular, it does not characterize the associated distribution of \( X \). Note that both the \( \text{RP} \) and the \( \text{SP} \) transforms uniquely determine the distribution of \( X \), i.e., if \( X_1 \) and \( X_2 \) are two random vectors having distributions \( P_1 \) and \( P_2 \) respectively on \( \mathbb{R}^d \) with continuous density functions and \( \text{RP}_1(x) = \text{RP}_2(x) \) for all \( x \in \mathbb{R}^d \) then \( P_1 = P_2 \). In this connection, also note that if \( X \sim P \) is absolutely continuous on \( \mathbb{R}^d \), \( \text{RP}(X) \sim \text{Uniform}(0,1)^d \); which, in particular, implies that the transformed covariates are independent. Such strong covariate standardization properties are not possessed by the \( \text{MP} \) transform. The \( \text{MP} \) transform is particularly useful when we assume an additive structure in the regression function, i.e., \( \mu(x) = \theta_0 + \sum_{i=1}^d \theta_i(x_i) \) [see Stone (1985) and Hastie and Tibshirani (1990)]. In this case the fractile regression function using the \( \text{MP} \) transform also has an additive structure and can be expressed as \( m(t) = \theta_0 + \sum_{i=1}^d \theta_i(F_i^{-1}(t_i)) \), where \( t = (t_1, t_2, \ldots, t_d) \). This facilitates computation of \( m(t) \) using the backfitting algorithm [see Hastie and Tibshirani (1990)], and the asymptotic properties of the estimator can be derived using techniques similar to that in Sen (2005).

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
Proof of Theorem 2.1  Suppose that $\mathbf{R}_P(x) = u$, i.e., $F_{P_1,i|1,2,...,i-1}(x_1|\mu_i,\ldots,x_{i-1}) = u_i \forall i = 1, 2, \ldots, d$, where $u = (u_1, u_2, \ldots, u_d)$. Note that,

$$F_{P_1,i|1,2,...,i-1}(x_1|\mu_1,\ldots,x_{i-1}) = u_i$$

$$\Rightarrow P(X_i \leq x_i|X_1 = \mu_1, X_2 = \mu_2, \ldots, X_{i-1} = \mu_{i-1}) = u_i$$

$$\Rightarrow P(\phi_i(X_i) \leq \phi_i(x_i)|\phi_i(X_1) = \phi_1(\mu_1), \ldots, \phi_i(X_{i-1}) = \phi_i(\mu_{i-1})) = u_i$$

$$\Rightarrow F_{P_2,i|1,2,...,i-1}(\phi_i(x_i)|\phi_i(\mu_1), \ldots, \phi_i(\mu_{i-1})) = u_i \forall i = 1, 2, \ldots, d$$

where $F_{P_2,i|1,2,...,i-1}$ 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 $\mathbf{R}_P(x) = u = \mathbf{R}_P(Z)$. Now, $E\{Y|\mathbf{R}_P(X) = t\} = E\{Y|\mathbf{R}_P(\Phi(X)) = t\} = E\{Y|\mathbf{R}_P(Z) = t\} \forall t \in (0,1)^d$, which gives us the desired result. \qed

Proof of Theorem 2.2  We will show that $E\{Y|\mathbf{T}(P_1, X) = t\} = E\{Y|\mathbf{T}(P_2, Z) = t\}$ for all $t \in E$, for all random vectors $(X,Y)$ with $X \sim P_1 \in \mathcal{P}$ is equivalent to $\mathbf{T}(P_1, x) = \mathbf{T}(P_2, \Phi(x)) \forall x \in \mathbb{R}^d$. Given that $\mathbf{T}(P_1, x) = \mathbf{T}(P_2, \Phi(x)) \forall x \in \mathbb{R}^d$ it is trivial to see that $E\{Y|\mathbf{T}(P_1, X) = t\} = E\{Y|\mathbf{T}(P_2, Z) = t\}$. 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 $\mathbf{T}(P_1, x) = \mathbf{T}(P_2, \Phi(x)) \forall x \in \mathbb{R}^d$ implies that $\mathbf{T}(P_1, X) \overset{d}{=} \mathbf{T}(P_2, \Phi(X))$. Let $U \sim \lambda$, where $\lambda$ is the Uniform$(0,1)^d$ measure. Then $\mathbf{T}(\lambda, U) \sim \mathbf{V}$ for some random vector $\mathbf{V}$. As any continuous random vector $X \sim P$ can be constructed from $U$ using a coordinate-wise increasing
transformation (which amounts to choosing suitable $\Phi$), we conclude that $T(P, X) \overset{d}{=} V$ for all $X \sim P \in \mathcal{P}$.

The regularity conditions on the weight functions required for Theorem 3.1 are described below along with a brief discussion.

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

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

(W3) The weights are asymptotically localized, i.e., there exists a sequence $\{\delta_n\}_{n=1}^{\infty}$, $\delta_n \to 0$ such that $\sum_{i=1}^{n} |W_{n,i}(t)| 1_{\{|t-H_n(X_i)||>\delta_n\}} \overset{P}{\to} 0$ as $n \to \infty$.

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

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

For the asymptotic normality of the fractile regression function we need the following condition:

(C1) For every $\eta > 0$, $\sum_{i=1}^{n} \frac{1}{s_n^2} \int_{e_i^2 > \eta^2 \frac{s_n^2}{W_{n,i}(t)}} W_{n,i}^2(t)e_i^2 dP \to 0$ a.s. where $e_i = Y_i - E(Y_i|X_i)$ for $i = 1, 2, \ldots, n$.

This is essentially a version of the well-known Lindeberg-Feller condition.

Recall that the estimated multivariate distribution transforms $R_n$ and $S_n$ are uniformly consistent estimators of $R_P$ and $S_P$ 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 $0$, and also the standard gaussian kernel, (W3) follows easily if $\|h_n\| \to 0$. Under the additional assumptions (i) $nh_{n,1}h_{n,2} \ldots h_{n,d} \to \infty$, (ii) the uniform consistency of the estimated multivariate transform $H_n$, and (iii) the existence of a non-vanishing density of $H(X)$ in $E$, we can verify condition (W1). Condition (C1) can also be verified easily under the above mentioned assumptions if the response is bounded. Thus for bounded response, as is
the case in most of our applications, the conclusions of Theorem 3.1 hold for estimates based on the Nadaraya-Watson type weight function defined using the multivariate transforms $R_P$ and $S_P$.

**Proof of Theorem 3.1** In the following theorem, all expectations are conditional expectations given the $X_i$’s, $i = 1, 2, \ldots, n$. For $t \in E$, the conditional mean squared error can be decomposed as

$$E \{\hat{m}_n(t) - m(t)\}^2 = E \{\hat{m}_n(t) - E(\hat{m}_n(t))\}^2 + \{E(\hat{m}_n(t)) - m(t)\}^2. \quad (4)$$

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

$$E \left[ \sum_{i=1}^n \{Y_i - m(H(X_i))\} W_{n,i}(t) \right]^2$$

$$= E \left[ \sum_{i=1}^n \sum_{j=1}^n \{Y_i - m(H(X_i))\} \{Y_j - m(H(X_j))\} W_{n,i}(t)W_{n,j}(t) \right]$$

$$= \sum_{i=1}^n E \{Y_i - m(H(X_i))\}^2 W_{n,i}^2(t)$$

$$\leq K_0 \sum_{i=1}^n W_{n,i}^2(t) \overset{P}{\to} 0 \text{ by assumption (W1) and the fact that } v(t) \text{ is bounded.}$$

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

$$\sum_{i=1}^n m(H(X_i))W_{n,i}(t) - m(t)$$

$$= \sum_{i=1}^n \{m(H(X_i)) - m(t)\} W_{n,i}(t) + m(t) \left\{ \sum_{i=1}^n W_{n,i}(t) - 1 \right\}. \quad (5)$$

Note that the second term in (5) goes to 0 in probability by assumption (W2).

We will show that $\sum_{i=1}^n V_{n,i} \overset{P}{\to} 0$ as $n \to \infty$, where $V_{n,i} = \{m(H(X_i)) - m(t)\} W_{n,i}(t)$. Let $\epsilon > 0$ and $\eta > 0$ be given. To simplify writing, we denote the
event \( \{ \| t - H_n(X_i) \| \leq \delta_n \} \) as \( E_{n,i} \). Therefore,

\[
P \left( \left\| \sum_{i=1}^{n} V_{n,i} \right\| > \epsilon \right) \leq P \left( \left\| \sum_{i=1}^{n} V_{n,i} 1_{E_{n,i}} \right\| > \epsilon/2 \right) + P \left( \left\| \sum_{i=1}^{n} V_{n,i} 1_{E_{n,i}}^c \right\| > \epsilon/2 \right)
\]

\[
\leq P \left( \left\| \sum_{i=1}^{n} V_{n,i} 1_{E_{n,i}} \right\| > \epsilon/2 \right) + \eta/2 \quad \text{for all } n \geq N_1 \quad \text{as} \quad (6)
\]

\[
P \left( \left\| \sum_{i=1}^{n} V_{n,i} 1_{E_{n,i}}^c \right\| > \epsilon/2 \right) \leq P \left( \sum_{i=1}^{n} \left| V_{n,i} \right| 1_{E_{n,i}}^c > \epsilon/2 \right)
\]

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

Let \( B_n = \sup_{x \in \mathbb{R}} \| H_n(x) - H(x) \|. \) By assumption, we know that \( B_n \to 0 \). Observe that, \( \| t - H_n(X_i) \| \leq \delta_n \) and \( \| H(X_i) - H_n(X_i) \| \leq B_n \) implies that \( \| H(X_i) - t \| \leq \| H_n(X_i) - t \| + \| H(X_i) - H_n(X_i) \| \leq \delta_n + B_n \) for all \( i = 1, 2, \ldots, n \).

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

\[
P \left( \left\| \sum_{i=1}^{n} V_{n,i} 1_{E_{n,i}} \right\| > \epsilon/2 \right) \leq P \left( \sum_{i=1}^{n} \left| V_{n,i} \right| 1_{E_{n,i}} > \epsilon/2 \right)
\]

\[
\leq P \left( \max_{1 \leq i \leq n} \left| m(H(X_i)) - m(t) \right| 1_{E_{n,i}} \sum_{i=1}^{n} \left| W_{n,i}(t) \right| > \epsilon/2 \right)
\]

\[
\leq P \left( \max_{1 \leq i \leq n} \left| m(H(X_i)) - m(t) \right| 1_{E_{n,i}} > \frac{\epsilon}{2D} \right)
\]

\[
\leq P \left( \delta_n + B_n > \delta \right) < \eta/2 \quad \forall \ n \geq N_2 \quad \text{as} \quad (7)
\]

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

Using (5), (6) and (7), we conclude \( P(\sum_{i=1}^{n} \{|m(H(X_i)) - m(t)| W_{n,i}(t)| > \epsilon\}) < \eta \) for all \( n \geq \max\{N_1, N_2\} \). Thus, the conditional mean squared error of
\( \hat{m}_n(t) \) approaches 0 in probability.

Now, using Chebyshev’s inequality, we have

\[
P(|\hat{m}_n(t) - m(t)| \geq \epsilon |X_1, X_2, \ldots, X_n) \leq \frac{E[(\hat{m}_n(t) - m(t))^2 |X_1, X_2, \ldots, X_n]}{\epsilon^2} \xrightarrow{P} 0 \text{ as } n \to \infty.
\]

An application of the dominated convergence theorem completes the proof of the Equation (2).

Note that \( \hat{m}_n(t) - E[\hat{m}_n(t)] = \sum_{i=1}^{\infty} W_{n,i}(t)e_i \) where \( e_i = Y_i - E(Y_1|X_i) \).

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

Observe that \( E(Z_{n,i}) = 0 \) and \( \sigma_{n,i}^2 = Var(Z_{n,i}) = v(H(X_i))W_{n,i}(t) \). Then \( s_n^2 = \sum_{i=1}^{\infty} \sigma_{n,i}^2 \). For any \( \eta > 0 \) and nonzero \( W_{n,i}^2(t) \), the Lindeberg-Feller condition is exactly (C1), and thus the result follows.

\[ \square \]

References


