Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

On Fractile Transformation of Covariates in Regression

Bodhisattva Sen [a] & Probal Chaudhuri [b]

[a] Department of Statistics , Columbia University , New York , NY , 10027 , USA

[b] Theoretical Statistics and Mathematics Unit , Indian Statistical Institute , 203 B.T. Road, Kolkata , India , 700 108

Accepted author version posted online: 31 Jan 2012.Published online: 11 Jun 2012.

To cite this article: Bodhisattva Sen & Probal Chaudhuri (2012) On Fractile Transformation of Covariates in Regression, Journal of the American Statistical Association, 107:497, 349-361, DOI: 10.1080/01621459.2011.646916

To link to this article: http://dx.doi.org/10.1080/01621459.2011.646916

On Fractile Transformation of Covariates in Regression

Bodhisattva SEN and Probal CHAUDHURI

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 distributions and the ranges 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 article, we describe a method of standardizing the covariates and estimating the transformed regression function, which then become comparable. We develop smooth estimates of the 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 the real examples leads to new and useful statistical conclusions that are missed by comparison of the usual regression functions.

KEY WORDS: Asymptotic consistency; Fractile regression; Groups of transformations; Invariance and equivariance; Kernel smoothing; 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.

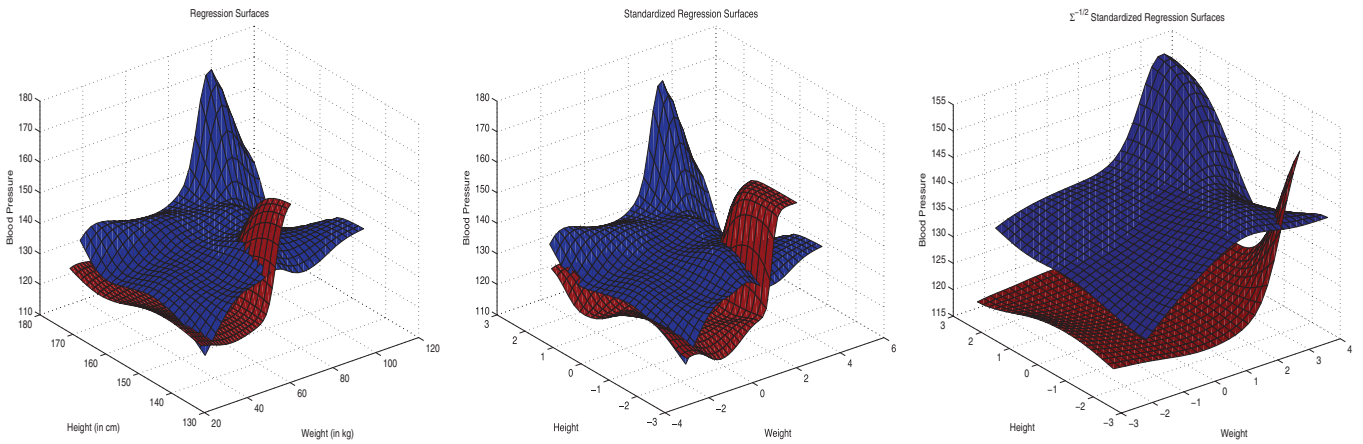
Example 1. Data were collected on 258 individuals from the Bhutia tribe and 305 individuals from the Toto tribe in India on blood pressure, height, and weight by the scientists of the Human Genetics Unit at Indian Statistical Institute, Kolkata. It is of interest to compare the relationship of blood 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 shown in Figure 1(a). 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. Such coordinate-wise location and scale-adjusted regression surfaces are shown in Figure 1(b), whereas Figure 1(c) shows the regression surfaces, where we standardize the covariate vector by subtracting the sample mean vector and multiplying by the inverse of the square root of the sample dispersion matrix. But the surfaces are still not quite in comparable forms—the supports 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 smoothing bandwidths, we have 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. This convention is followed in computing all the bivariate regression surfaces illustrated in the article. 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 (overweight) person. As will be shown later, these two features in the regression surfaces are indeed spurious and lead 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 nongovernment, nonfinancial 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 such as 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 2(a) shows the usual regression surfaces for

Bodhisattva Sen (E-mail: *bodhi@stat.columbia.edu*) is Assistant Professor, Department of Statistics, Columbia University, New York, NY 10027, USA. Probal Chaudhuri (E-mail: *probal@isical.ac.in*) is Professor at the Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 108, India. The authors are grateful to three referees and an associate editor for their insightful comments that helped improve the article. They would also like to thank Partha P. Majumder (Human Genetics Unit, Indian Statistical Institute, Kolkata, India) for providing them with the blood pressure data for the Bhutia and the Toto tribes and the Reserve Bank of India for supplying the data on sales and profits of companies in India.

Journal of the American Statistical Association
March 2012, Vol. 107, No. 497, Theory and Methods
DOI: 10.1080/01621459.2011.646916



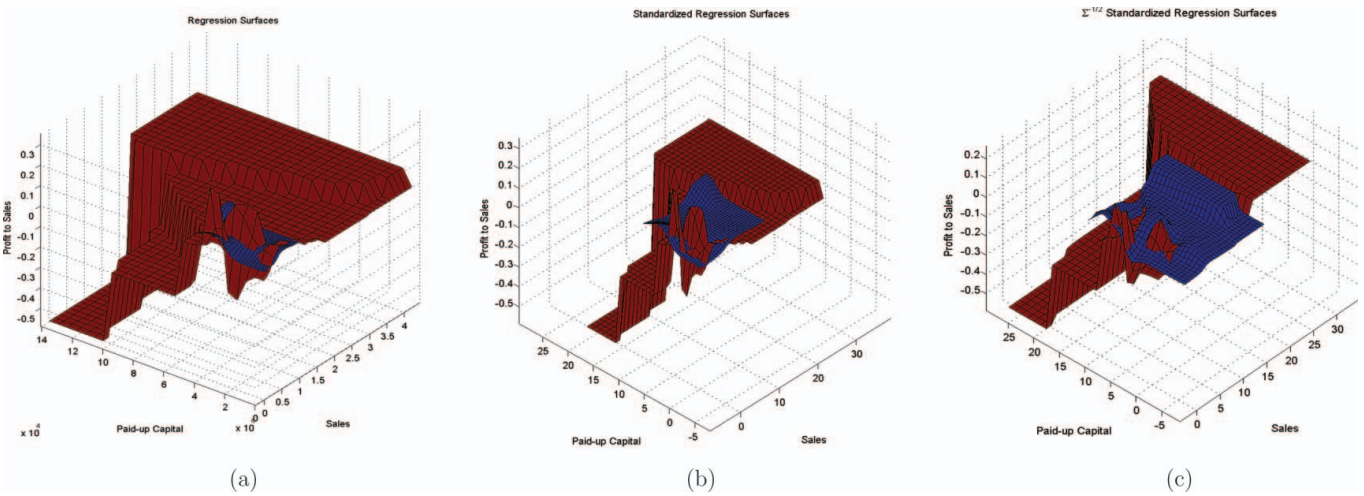
(a) (b) (c)

Figure 1. (a) Usual regression surfaces, (b) coordinate-wise 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 blood pressure on weight and height for the *Bhutia* (red) and *Toto* (blue) tribes.

the year 1997 (blue surface) and 2003 (red surface) with 944 and 1243 data points, respectively. Figures 2(b) and 2(c) 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.

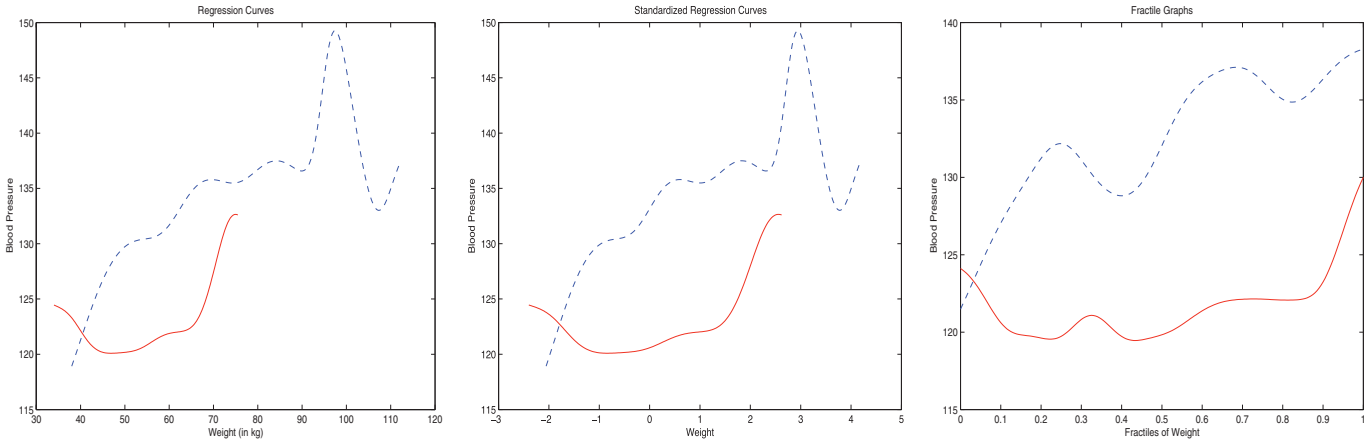
Both the preceding examples demonstrate the need for a methodology to appropriately standardize the covariate vectors before comparing the corresponding regression functions, when the distributions and supports of the covariates are very different in the two populations. Note that the usual location and coordinate-wise scale transformation standardizes the means and variances of the covariates for each population, whereas the standardization by subtracting the means and multiplying by the inverse of the square root of the dispersion matrix only normalizes the means and the dispersion matrix of the covariate vectors. This raises two related questions of interest that we address in this article: (1) How do we standardize the *distribution* of the covariates that will enable a more meaningful comparison of the regression functions? (2) Suppose that we have two random vectors $(\mathbf{X}_1, Y)$ and $(\mathbf{X}_2, Y)$ in $\mathbb{R}^{d+1}$ for $d \geq 1$, having continuous distributions, where the predictor $\mathbf{X}_2 = \mathbf{g}(\mathbf{X}_1)$ and $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ is an (unknown) invertible function. Can we find a standardization of the covariates $\mathbf{X}_1$ and $\mathbf{X}_2$ that will enable us to compare and conclude that the two regression functions for the populations are essentially the same?

In this article, we propose a method of standardizing the covariates using a multivariate transformation, which is derived from their multivariate distribution, that achieves the desired joint distributional standardization. The useful properties of the



(a) (b) (c)

Figure 2. (a) Usual regression surfaces, (b) coordinate-wise 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 profit (as a fraction of sales) on sales and paid-up capital for the years *1997* (blue) and *2003* (red).



(a) (b) (c)

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.

proposed multivariate standardization in enabling proper comparability of the regression functions, expressed in terms of invariance/equivariance properties of the standardized regression functions under groups of transformations acting on the covariates, are investigated. We also study the estimation of the corresponding regression functions based on the standardized covariates.

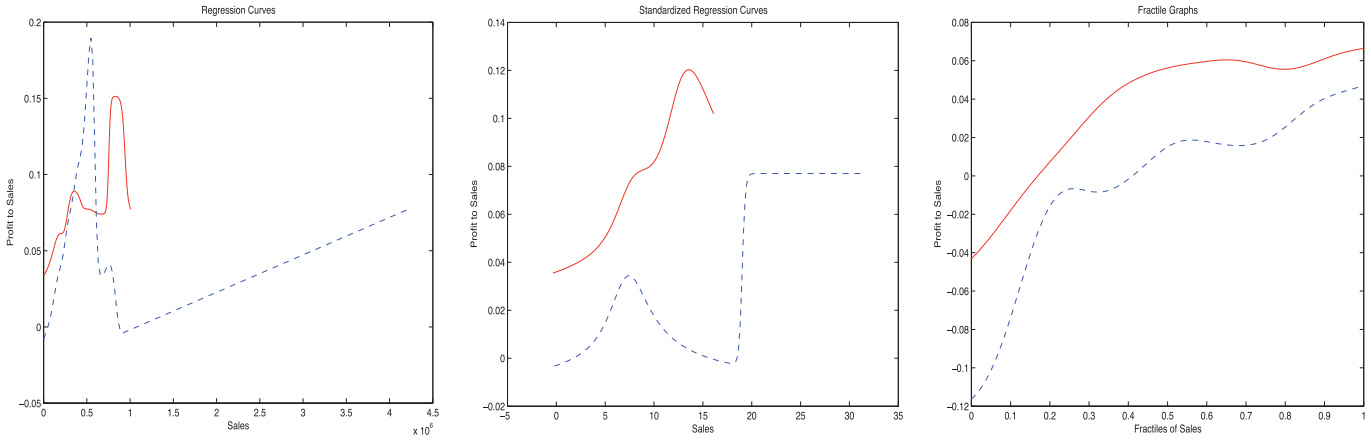
To start with, let us first look at the simpler problem when $d = 1$. Consider two bivariate random vectors (X_1, Y_1) and (X_2, Y_2) and the associated regression functions μ_1 and μ_2 where $\mu_1(x) = E(Y_1|X_1 = x)$ and $\mu_2(x) = E(Y_2|X_2 = x)$. Then the *fractile regression* functions are defined as

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

for $t \in (0, 1)$, where F_1 and F_2 are the distribution functions of X_1 and X_2, respectively (see Mahalanobis 1960). 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 tth 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, suppose that (X_1, Y) is a continuous bivariate random vector and if $X_2 = \phi(X_1)$, where ϕ is any strictly increasing transformation, then $E\{Y|F_1(X_1)\} = E\{Y|F_2(X_2)\}$, where F_1 and F_2 are the distribution functions of X_1 and X_2, respectively. This is a crucial property and can be interpreted in the following way: The fractile transformation makes the regression functions comparable even when the covariate in the second population is any increasing transformation of that of the first population. Fractile regression has been considered earlier in Mahalanobis (1960), Parthasarathy and Bhattacharya (1961), Sethuraman (1961), Bhattacharya and Müller (1993), and Sen (2005).

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 dataset 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 4(a) and 4(b) are 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 article, 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 Figures 3(a), 3(b), 4(a), and 4(b), there is a serious lack of comparability between the two regression curves, which is adequately resolved in Figures 3(c) and 4(c).

Fractile regression techniques with one covariate have been applied in diverse settings. Hertz-Picciotto and Din-Dzietham (1998) compared the infant mortality of African Americans and European Americans with gestational age using a "percentile based method" of standardization. They encountered a very similar problem as in Example 1, the two usual regression functions cross suggesting that African American infants do better than European-Americans about half the time, whereas the fractile regression functions remove this spurious visual impression. Nordhaus (2006) used fractile plots to study the dependence of log of "output density" on key geographic variables (temperature, precipitation, latitude, etc.). This application illustrates another usefulness of fractile regression: it enables us to simultaneously compare the effect of different, but possibly related, covariates (as opposed to the same variable in different populations) on one response variable by overplotting the



(a) (b) (c)

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

different fractile regression functions, which are now all defined on the same space $(0, 1)$.

The article is organized as follows. In Section 2, we consider a suitable notion of an invertible multivariate distribution function based on successive conditioning of the covariates and use it to define the fractile standardization and the corresponding regression function. The comparability of different fractile regression functions is investigated using invariance/equivariance properties of the transformed regression functions under different groups of transformations acting on the space of covariates. We also briefly discuss another notion of standardization along with the associated regression function. Section 3 discusses nonparametric smooth estimation of the fractile regression function from a random sample. 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 another application of fractile regression techniques on real data. We end this section with a comparison and discussion of the two standardizations proposed in the article on the basis of the real data examples. In Section 5, the concluding section, we discuss some extensions. Section 6, the Appendix, gives the proofs of the main results.

2. STANDARDIZATION OF COVARIATES USING TRANSFORMATIONS

Let $(Y, \mathbf{X})$ be a random vector having a continuous distribution on $\mathbb{R}^{d+1}$, $d \geq 1$, where $\mathbf{X} = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ is the covariate/predictor and $Y \in \mathbb{R}$ is the response variable. We want to study the effect of different kinds of "standardizations" of the predictor that would aid comparison of the regression functions, as discussed in the Introduction.

Let $\mathbb{P}$ be a class of covariate distributions on $\mathbb{R}^d$. Formally, a *"standardization"* can be defined as a function $\mathbf{T} : \mathbb{P} \times \mathbb{R}^d \to E \subset \mathbb{R}^d$, which is used to transform the covariate, such that $\mathbf{x} \mapsto \mathbf{T}(\mathbf{P}, \mathbf{x}) \equiv \mathbf{T}(\mathbf{X}, \mathbf{x})$ is an invertible map from $\mathcal{X}_\mathbf{P}$, the support of $\mathbf{P}$, onto E, for every $\mathbf{X} \sim \mathbf{P} \in \mathbb{P}$. The *standardized* regression function is then defined as

$$m_\mathbf{X}(\mathbf{t}) = E\{Y | \mathbf{T}(\mathbf{P}, \mathbf{X}) = \mathbf{t}\} \quad \text{for } \mathbf{t} \in E. \tag{1}$$

To study the effect of the standardization $\mathbf{T}$, we need to consider a group $\mathcal{G}$ of one-one transformations acting on the space of all d-variate predictors belonging to $\mathbb{P}$. Let $\mathbf{g}$ be one such transformation and denote by $\mathbf{g}(\mathbf{X})$ the random vector that takes on the value $\mathbf{g}(\mathbf{x})$ when $\mathbf{X} = \mathbf{x}$. We say that $\mathbf{T}$ is *invariant* under $\mathcal{G}$ if $\mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{x})) = \mathbf{T}(\mathbf{X}, \mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{g} \in \mathcal{G}$. We say that $\mathbf{T}$ is *equivariant* under $\mathcal{G}$ if $\mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{x})) = \mathbf{g}(\mathbf{T}(\mathbf{X}, \mathbf{x}))$, for all $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{g} \in \mathcal{G}$. The standardized regression function is *invariant* under the group action $\mathcal{G}$ if $m_{g(\mathbf{X})}(\mathbf{t}) := E\{Y | \mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{X})) = t\} = m_\mathbf{X}(\mathbf{t})$, for all $\mathbf{t} \in E$ and $\mathbf{g} \in \mathcal{G}$.

2.1 Standardization by the Fractile Transformation

In this section, 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 straightforward notion of an invertible distribution function [referred to as a *centered rank function*, see Serfling (2010)] and/or multivariate quantiles, because of the lack of natural ordering of points in $\mathbb{R}^d$ for $d > 1$. We restrict our attention to $\mathcal{P}$, the class of all covariate distributions on $\mathbb{R}^d$ having a continuously differentiable density (with respect to the Lebesgue measure) on its support. This ensures that the various univariate marginal and conditional distribution functions associated with the covariate distribution will be strictly increasing and invertible on their supports. For a d-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d) \sim \mathbf{P} \in \mathcal{P}$, we define the *fractile standardization* (transformation) $\mathbf{R}_\mathbf{P} : \mathbb{R}^d \mapsto [0, 1]^d$ (or $\mathbf{R}_\mathbf{X}$), as

$$\mathbf{R}_\mathbf{P}(\mathbf{x}) \equiv \mathbf{R}_\mathbf{X}(x_1, x_2, \ldots, x_d) = \big(F_1(x_1), F_{2|1}(x_2|x_1), \ldots, F_{d|1,2,\ldots,d-1}(x_d|x_1, x_2, \ldots, x_{d-1})\big),$$

where $F_1(x_1) = P(X_1 \leq x_1)$, $F_{2|1}(x_2) = P(X_2 \leq x_2 | X_1 = x_1)$, $\ldots$, $F_{d|1,2,\ldots,d-1}(x_d) = P(X_d \leq x_d | X_1 = x_1, X_2 = x_2, \ldots, X_{d-1} = x_{d-1})$. This is a multivariate analog of the univariate distribution transform [i.e., $x \mapsto F_X(x)$] and has a number of

desirable properties. The *fractile regression function*, that is, the $\mathbf{R_P}$-standardized regression function, can now be defined through (1) with $\mathbf{T}(\mathbf{P}, \cdot)$ replaced by $\mathbf{R_P}(\cdot)$.

Recently, a similar idea on successive conditioning of the coordinate variables of a random vector was used by Wei (2007) in the quantile regression setup on bivariate growth curves and Chesher (2003) and Ma and Koenker (2006) in quantile regression for structural econometric models. Salibián-Barrera and Wei (2008) used it to standardize the regressors to screen out leverage points.

It can be easily shown that $\mathbf{R_P}(\mathbf{X}) \sim$ Uniform$(0, 1)^d$, if $\mathbf{X} \sim \mathbf{P} \in \mathcal{P}$. This is a generalization of the invariance property shared by any continuous univariate distribution function F, that is, $F(X) \sim$ Uniform$(0, 1)$, where $X \sim F$. Thus, $\mathbf{R_P}$ achieves distributional standardization, making the transformed covariate vector have the same distribution for any $\mathbf{X} \sim \mathbf{P} \in \mathcal{P}$.

We next show that the fractile standardization has important invariance properties under statistically relevant groups of transformations on the covariates. If we consider the group of coordinate-wise scale and shift transformations $\mathcal{G} = \{\mathbf{g} : \mathbb{R}^d \mapsto \mathbb{R}^d \mid \mathbf{g}(\mathbf{x}) = \mathbf{D}\mathbf{x} + \mathbf{c}\}$, where $\mathbf{D}$ is a $d \times d$ diagonal matrix with positive diagonal entries and $\mathbf{c} \in \mathbb{R}^d$, acting on the space of covariates, then it is easy to see that the simple coordinate-wise location and scale standardization $\mathbf{T}_{ls}$ defined by $\mathbf{T}_{ls}(\mathbf{X}, \mathbf{x}) = \Gamma(\mathbf{X})^{-1/2}(\mathbf{x} - E(\mathbf{X}))$, where $\Gamma(\mathbf{X}) = \text{diag}(\text{var}(X_1), \text{var}(X_2), \ldots, \text{var}(X_d))$, is invariant under $\mathcal{G}$. As a consequence, $E\{Y|\mathbf{T}_{ls}(\mathbf{X}, \mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{T}_{ls}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{X})) = \mathbf{t}\}$ for all $\mathbf{t} \in \mathbb{R}^d$, for all $\mathbf{g} \in \mathcal{G}$, and thus the standardized regression function is invariant under $\mathcal{G}$.

But from the examples in the Introduction, we see that the coordinate-wise location and scale standardization of the covariates does not yield satisfactory results. However, it is intuitively obvious that monotonically increasing transformations of the covariates are meaningful and relevant for all the examples considered in Section 1, and hence it is reasonable to consider appropriate nonlinear monotonic transformations to standardize the covariates in such examples. For $\mathbf{x} \in \mathbb{R}^d$, let us write $\mathbf{x}_i = (x_1, x_2, \ldots, x_i)$, for $i = 1, 2, \ldots, d$, and consider the transformation $\mathbf{x} \mapsto \mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}_1), g_2(\mathbf{x}_2), \ldots, g_d(\mathbf{x}_d))$, where $g_i : \mathbb{R}^i \to \mathbb{R}$, is a strictly increasing transformation in x_i (the last coordinate) for every fixed $(x_1, x_2, \ldots, x_{i-1})$, and $(g_1, g_2, \ldots, g_i) : \mathbb{R}^i \to \mathbb{R}^i$ is invertible for every i, for $i = 1, 2 \ldots, d$. Consider the group $\mathcal{F}$ of all such transformations on the covariates. The following result shows that the standardization $\mathbf{R_P}$ is invariant under $\mathcal{F}$, and this 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 $\mathbf{X} = (X_1, X_2, \ldots, X_d) \sim \mathbf{P} \in \mathcal{P}$. Then, for $\mathbf{g} \in \mathcal{F}$, $\mathbf{R_X}(\mathbf{x}) = \mathbf{R}_{\mathbf{g}(\mathbf{X})}(\mathbf{g}(\mathbf{x}))$ for all $\mathbf{x} \in \mathbb{R}^d$, and, in particular, $E\{Y|\mathbf{R_X}(\mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{R}_{\mathbf{g}(\mathbf{X})}(\mathbf{g}(\mathbf{X})) = \mathbf{t}\}$ for all $\mathbf{t} \in (0, 1)^d$.

A special case of the previous theorem occurs when we consider the group $\mathcal{H} \subset \mathcal{F}$ of all coordinate-wise increasing transformations on the covariates, that is, $\mathbf{x} \mapsto \mathbf{g}(\mathbf{x}) = (g_1(x_1), g_2(x_2), \ldots, g_d(x_d))$, where $g_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2 \ldots, d$, is a strictly increasing function. The above theorem can then be interpreted as 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 in the two populations might be very different. Note that, $\mathbf{R_P}$ is invariant under the group of (linear) transformations $\mathcal{K} = \{\mathbf{g} : \mathbb{R}^d \mapsto \mathbb{R}^d \mid \mathbf{g}(\mathbf{x}) = \mathbf{D}\mathbf{x} + \mathbf{c}\}$, where $\mathbf{D}$ is a lower triangular nonsingular matrix with positive diagonal elements.

The next result shows that if we want the standardized regression function to be invariant under the group action $\mathcal{F}$, then the standardization $\mathbf{T}(\mathbf{X}, \cdot)$ has to be a function of $\mathbf{R_P}$. In addition, if we also want $\mathbf{T}(\mathbf{X}, \cdot)$ to achieve distributional standardization and belong to $\mathcal{F}$, then $\mathbf{R_P}$ is the only choice.

Theorem 2.2 Let $(\mathbf{X}, Y)$ be as in Theorem 2.1, and suppose that there exists a standardization $\mathbf{T} : \mathcal{P} \times \mathbb{R}^d \to E \subset \mathbb{R}^d$ such that $E\{Y|\mathbf{T}(\mathbf{X}, \mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{X})) = \mathbf{t}\}$ for all $\mathbf{t} \in E$ and $\mathbf{g} \in \mathcal{F}$, and equality holds for all joint distributions of $(\mathbf{X}, Y)$, with $\mathbf{X} \sim \mathbf{P} \in \mathcal{P}$. Then $\mathbf{T}(\mathbf{P}, \mathbf{x})$ must be a function of $\mathbf{R_P}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{P} \in \mathcal{P}$. Furthermore, if we assume that $\mathbf{T}(\mathbf{X}, \mathbf{X}) \sim$ Uniform$(0, 1)^d$ and $\mathbf{T}(\mathbf{X}, \cdot) \in \mathcal{F}$, then $\mathbf{T}(\mathbf{X}, \mathbf{x}) = \mathbf{R_P}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{X} \sim \mathbf{P} \in \mathcal{P}$.

The computation of $\mathbf{R_P}$ from a sample of independent, identically distributed (iid) data points $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim \mathbf{P} \in \mathcal{P}$, where $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d})$ for $i = 1, 2, \ldots, n$, requires the estimation of the conditional distribution functions. In order to estimate the conditional distributions, we may begin by estimating the multivariate density f of $\mathbf{X}_1$ using the kernel density estimator:

$$f_{n;1,2,\ldots,d}(\mathbf{x}) = \frac{1}{n(h_{1,n}h_{2,n}\ldots h_{d,n})} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i,1}}{h_{1,n}}\right) \times K\left(\frac{x_2 - X_{i,2}}{h_{2,n}}\right) \ldots K\left(\frac{x_d - X_{i,d}}{h_{d,n}}\right),$$

where K is a kernel function defined on $\mathbb{R}$, $h_{j,n}$ is the bandwidth parameter for the jth coordinate of the random vector at stage n, for $j = 1, 2, \ldots, d$, and $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$. This joint density estimate can then be used to compute estimates of the conditional densities. The conditional densities are computed as $f_{n;j|1,2,\ldots,j-1}(x_j|x_1, x_2, \ldots, x_{j-1}) = \frac{f_{n;1,2,\ldots,j}(x_1,x_2,\ldots,x_j)}{f_{n;1,2,\ldots,j-1}(x_1,x_2,\ldots,x_{j-1})}$, and then used to estimate the conditional distribution functions in a natural way as $F_{P_n,j|1,2,\ldots,j-1}(x_j|x_1, x_2, \ldots, x_{j-1}) = \int_{-\infty}^{x_j} f_{n;j|1,2,\ldots,j-1}(t_j|x_1, x_2, \ldots, x_{j-1})dt_j$. To compute the conditional densities in the subsequent sections, we have used the Gaussian kernel with bandwidths chosen by cross-validation and the computation was done using the "sm" package in R developed by Adrain Bowman and Adelchi Azzalini.

Let $\mathbf{R}_n$ be the estimated fractile standardization obtained from the sample. Under appropriate conditions on the kernel and the smoothing parameter(s) (e.g., when K is a bounded symmetric density on $\mathbb{R}$, $\|\mathbf{h}_n\| \to 0$ and $nh_{1,n}h_{2,n}, \ldots, h_{d,n} \to \infty$), it can be shown that $\mathbf{R}_n$ is a uniformly consistent estimator of $\mathbf{R_P}$, that is, $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{R}_n(\mathbf{x}) - \mathbf{R_P}(\mathbf{x})\| \xrightarrow{P} 0$. The result follows from noting that under the regularity conditions, the estimated conditional densities converge in probability to the their population counterparts and an application of Scheffe's theorem yields the desired result. Note that for $d = 1$, as we directly estimate $\mathbf{R}_n$ by the empirical distribution function of the covariate, the Glivenko–Cantelli theorem gives the desired result.

Though the multivariate transform $\mathbf{R_P}$ has nice invariance properties and simple probabilistic interpretations, in high dimensions, 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. For d greater than 4 or 5, the implementation of fractile regression using the $\mathbf{R_P}$ standardization becomes almost infeasible.

The fractile standardization $\mathbf{R_P}$ is not equivariant under reordering of the coordinates of the covariate vector which implies that it lacks general affine equivariance. But as pointed out by Van Keilegom and Hettmansperger (2002) and Serfling (2004) that when the covariates of interest have a special physical interpretation, as is the case in all our applications, there is no interest in affinely transforming them. In such an application, if there is a natural ordering of the importance of the covariates, we advocate the use of the $\mathbf{R_P}$ transform, conditioning successively on the less important predictor. For example, in Example 1 (in the Introduction), as body weight is known to affect blood pressure (the response) more acutely than height, the weight variable can be taken to be X_1. In Example 2, because such an ordering of the covariates is not obvious, we might like to use both orderings, (X_1, X_2) and (X_2, X_1), to construct two fractile regression functions and compare them. Alternatively, we can use a different procedure for standardization of the covariate in such a situation, which is discussed in the next section.

2.2 An Alternative to $\mathbf{R_P}$: The Marginal Rank Transform

Consider Example 2 in the Introduction. An economist might want to compare the mean profitability for companies with "median" sales and "median" paid-up capital for the two populations. For d-dimensional covariates, such comparisons involving the marginal quantiles of the covariate vector can be accomplished by considering the function $m(\mathbf{t}) = E\{Y|X_1 = F_1^{-1}(t_1), \ldots, X_d = F_d^{-1}(t_d)\}$ for the two populations, where F_i is the marginal distribution function of X_i, $i = 1, 2, \ldots, d$, and $\mathbf{t} = (t_1, t_2, \ldots, t_d) \in (0, 1)^d$. This leads to another nonparametric standardization of the covariate vector based on the marginal rank transformation $\mathbf{M_P}$: $(x_1, x_2, \ldots x_d) \mapsto (F_1(x_1), F_2(x_2), \ldots, F_d(x_d))$. This standardization retains the property of invariance possessed by $\mathbf{R_P}$ under the group of arbitrary coordinate-wise increasing transformations $\mathcal{H}$ (introduced in Section 2.1) of the covariates, that is, $\mathbf{M_{g(X)}}(\mathbf{g}(\mathbf{x})) = \mathbf{M_X}(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{g} \in \mathcal{H}$. In particular, $\mathbf{M_P}$ is invariant under the group of marginal scale and location changes of the covariates. Further, $\mathbf{M_P}$ is equivariant under the relabeling (i.e., permutation) of the covariate variables, which is not true for $\mathbf{R_P}$.

The standardized regression function is then defined as $m(\mathbf{t}) = E\{Y|\mathbf{M_P}(\mathbf{X}) = \mathbf{t}\}$ for $\mathbf{t} \in (0, 1)^d$, and is related to the usual regression function $\mu(\mathbf{x}) = E\{Y|\mathbf{X} = \mathbf{x}\}$ through $m(\mathbf{t}) = \mu(F_1^{-1}(t_1), F_2^{-1}(t_2), \ldots, F_d^{-1}(t_d))$, where $\mathbf{t} = (F_1(x_1), F_2(x_2), \ldots, F_d(x_d))$. We can estimate the usual regression function in any way we like, using nonparametric or parametric techniques, and then estimate the marginal quantile functions to estimate m.

The $\mathbf{M_P}$ standardization is computationally much simpler than the $\mathbf{R_P}$ transform as it only requires the estimation of univariate distribution functions. To compute $\mathbf{M}_n$, the sample analog of $\mathbf{M_P}$, from n iid data points, we replace F_i by the corresponding empirical distribution function of the ith covariate. Using this estimator, it can be easily shown, by an application of the Glivenko–Cantelli theorem, that $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{M}_n(\mathbf{x}) - \mathbf{M_P}(\mathbf{x})\| \to 0$ almost surely (a.s.) as $n \to \infty$.

3. SMOOTH ESTIMATION OF FRACTILE REGRESSION

In this section, we define smooth estimates of the standardized 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 $\mathbf{X}_i \sim \mathbf{P} \in \mathcal{P}$. Let $(\mathbf{X}, Y)$ be a generic random vector having the same joint distribution. For broader applicability, we describe the methodology for any standardization $\mathbf{H} : \mathbb{R}^d \to E \subset \mathbb{R}^d$ (which may or may not be $\mathbf{R_P}$ or $\mathbf{M_P}$). For notational convenience, we do not emphasize the dependence of $\mathbf{P}$ on $\mathbf{H}$, as $\mathbf{P}$ is fixed. We want to estimate the standardized regression function $m(\mathbf{t}) = E\{Y|\mathbf{H}(\mathbf{X}) = \mathbf{t}\}$ for $\mathbf{t} \in E$. We define the smooth estimated standardized regression function as

$$\widehat{m}_n(\mathbf{t}) = \sum_{i=1}^n Y_i W_{n,i}(\mathbf{t}) \quad \text{for } \mathbf{t} \in E, \tag{2}$$

where $W_{n,i}(\mathbf{t})$ is the weight function, which might depend on $\mathbf{H}_n$, the empirical or estimated value of $\mathbf{H}$. 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 $W_{n,i}(\mathbf{t}) = \frac{\mathbf{K}(\frac{\mathbf{t}-\mathbf{H}_n(\mathbf{X}_i)}{\mathbf{h}_n})}{\sum_{j=1}^n \mathbf{K}(\frac{\mathbf{t}-\mathbf{H}_n(\mathbf{X}_j)}{\mathbf{h}_n})}$, where $\mathbf{K}$ is a kernel function defined on $\mathbb{R}^d$, $\mathbf{t} = (t_1, t_2, \ldots, t_d) \in E$, $\frac{\mathbf{t}-\mathbf{H}_n(\mathbf{X}_i)}{\mathbf{h}_n} := (\frac{t_1-H_{n,1}(\mathbf{X}_i)}{h_{n,1}}, \frac{t_2-H_{n,2}(\mathbf{X}_i)}{h_{n,2}}, \ldots, \frac{t_d-H_{n,d}(\mathbf{X}_i)}{h_{n,d}})$, and $h_{n,1}, h_{n,2}, \ldots, h_{n,d}$ are the smoothing bandwidths.

3.1 Some Asymptotic Results

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, again stated for a general standardization.

Theorem 3.1. Fix $\mathbf{t} \in E$. Suppose that $m(\mathbf{t}) = E\{Y|\mathbf{H}(\mathbf{X}) = \mathbf{t}\}$ is continuous on E and $|m(\mathbf{u})| \le M$ for all $\mathbf{u} \in E$; the conditional variance of Y_i given $\mathbf{H}(\mathbf{X}_i)$ is bounded, that is, $v(\mathbf{u}) = \text{var}\{Y_i|\mathbf{H}(\mathbf{X}_i) = \mathbf{u}\} \le K_0$ for all $\mathbf{u} \in E$; and $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{H}_n(\mathbf{x}) - \mathbf{H}(\mathbf{x})\| \xrightarrow{P} 0$. Also assume the following conditions on the weight functions:

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

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

(W3) the weights are asymptotically localized, that is, there exists a sequence $\{\delta_n\}_{n=1}^\infty$, $\delta_n \to 0$ such that $\sum_{i=1}^n |W_{n,i}(\mathbf{t})| \mathbf{1}_{\{\|\mathbf{t}-\mathbf{H}_n(\mathbf{X}_i)\| > \delta_n\}} \xrightarrow{P} 0$ as $n \to \infty$;

Table 1. Ratio of the estimated IMSE for smoothed estimates of usual regression and fractile regression functions when the data-generating model is (a) $Y = \exp(-X) + \epsilon$ (left panel) and (b) $Y = X + \epsilon$ (right panel)

X	ϵ	IMSE ratio	X	ϵ	IMSE ratio
$\lvert N(0,1)\rvert$	$N(0, 0.2)$	1.10	$\lvert N(0,1)\rvert$	$N(0,1)$	0.90
$\lvert N(0,1)\rvert$	$N(0, 0.2X)$	2.25	$\lvert N(0,1)\rvert$	$N(0, X)$	2.18
$\lvert t_4\rvert$	$N(0, 0.2)$	3.43			
$\lvert t_4\rvert$	$N(0, 0.2X)$	4.81			

(W4) there exists $D \geq 1$ such that $P(\sum_{i=1}^n |W_{n,i}(\mathbf{t})| \leq D) = 1$ for all $n \geq 1$. Then, the conditional mean squared error of $\hat{m}_n(\mathbf{t})$ approaches 0 in probability. As a consequence, $\hat{m}_n(\mathbf{t}) \xrightarrow{P} m(\mathbf{t})$. Suppose now that (C1) for every $\eta > 0$, $\sum_{i=1}^n \frac{1}{s_n^2} \int_{e_i^2 > \eta^2 \frac{s_n^2}{W_{n,i}^2(\mathbf{t})}} W_{n,i}^2(\mathbf{t}) e_i^2 \, dP \to 0$ a.s., where $e_i = Y_i - E(Y_i|X_i)$, for $i = 1, 2, \ldots, n$. Letting $s_n^2 = \sum_{i=1}^n v(\mathbf{H}(\mathbf{X}_i)) W_{n,i}^2(\mathbf{t})$, we have

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

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

Note that conditions (W1)–(W4) are similar to those used by Stone (1977). Condition (C1) is essentially a version of the well-known Lindeberg–Feller condition. In the following, we briefly discuss the regularity conditions on the weight functions. Recall that the estimated multivariate centered rank functions $\mathbf{R}_n$ and $\mathbf{M}_n$ are uniformly consistent estimators of $\mathbf{R_P}$ and $\mathbf{M_P}$, 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 nonvanishing density of $\mathbf{H}(\mathbf{X})$ 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 $\mathbf{R_P}$ and $\mathbf{M_P}$. It must be noted that conditions (W1)–(W4) and (C1) are also satisfied for other weight functions, but we do not discuss them here.

Next we discuss an interesting optimality property of the uniform distribution on $[0, 1]^d$, the resulting distribution of the $\mathbf{R_P}$-transformed covariates. Suppose that the covariate vector is standardized by the transformation $\mathbf{H} : \mathbb{R}^d \to [0, 1]^d$. One plausible criterion to choose the optimal transformation might be to minimize the integrated asymptotic variance of the nonparametric function estimator [defined as in (2)], that is, $\text{IAV} = \lim_{n\to\infty} nh_{n,1}, \ldots, h_{n,d} \int_{(0,1)^d} E\{\hat{m}_n(\mathbf{t}) - E(\hat{m}_n(\mathbf{t}))\}^2 \, d\mathbf{t}$. Note that IAV is related to the scientific issue of reproducibility of the results obtained from a regression function estimate, when data are replicated, for instance, in repeated trials of an experiment. For our kernel regression estimate, we have the following result.

Theorem 3.2 Suppose that we are in a homoscedastic error model, and conditions (i)–(iii) (described above) hold along with $\|\mathbf{h}_n\| \to 0$. Then IAV is minimized when $\mathbf{H}(\mathbf{X})$ has uniform distribution on $[0, 1]^d$.

3.2 Finite Sample Performance of Fractile Regression

Consider a trivariate normal random vector $(X_1, X_2, Y) \sim N(\mathbf{0}, \Sigma)$ with $\Sigma = (\sigma_{i,j})_{3\times 3}$ such that $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0.5$ for

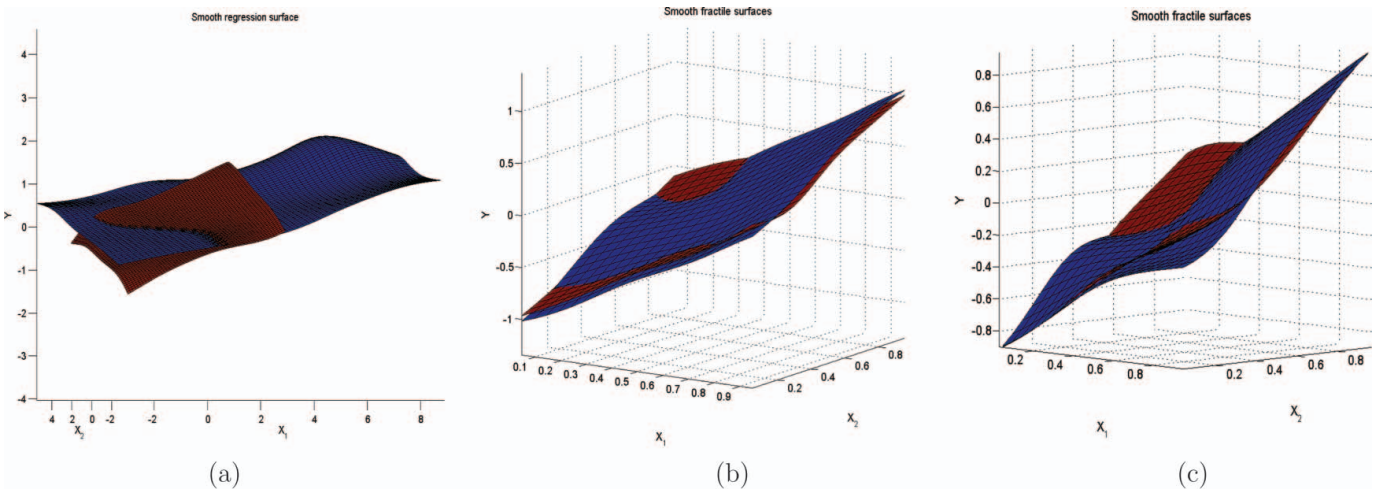


Figure 5. (a) Usual mean regression surfaces for $Y|X_1, X_2$, (b) estimated fractile regression surfaces, and (c) estimated $\mathbf{M_P}$-standardized regression surfaces for Gaussian data.

$i \neq j$. To illustrate the usefulness of fractile regression, we construct a second random vector (X_1', X_2', Y) where $X_1' = X_1^2$ and $X_2' = X_2^{1.75}$ for positive values of X_1 and X_2, and $X_1' = X_1$ and $X_2' = X_2$ for negative values of X_1 and X_2. We draw two samples of size 400 each from the two distributions mentioned above. Figure 5(a) shows the smoothed regression surfaces, whereas Figure 5(b) shows the smoothed fractile regression surfaces, using the fractile standardization $\mathbf{R_P}$, 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. Figure 5(c) shows the $\mathbf{M_P}$-standardized regression surfaces, and again we see that the regression surfaces are very similar and comparable.

In some situations, the use of fractile regression can provide better [e.g., in terms of integrated mean squared error (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 a valid comparison (and also for computational simplicity), 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 $\mathbf{R_P}$ standardization, 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, thereby producing more stable results.

Notice that in models (a) and (c) where the regression functions are bounded, 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 examples). Note that the uneven distribution of covariates in the simulation study is caused by the extreme observations generated from the heavy-tailed distributions. In the case of normal linear models (b) and (d), the performance of fractile regression is slightly inferior to that of the usual regression functions. Note that the true regression function in these models is linear, whereas the true fractile regression function has curvature, and this 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.

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) $Y = \exp(-X_1X_2) + \epsilon$ and (d) $Y = (X_1 + X_2)/2 + \epsilon$

Model	(X_1, X_2)	ϵ	IMSE ratio
(c)	$\lvert N(0,1)\rvert \times \lvert N(0,1)\rvert$	$N(0, 0.2)$	1.35
(c)	$\lvert N(0,1)\rvert \times \lvert N(0,1)\rvert$	$N(0, 0.2X_1)$	1.77
(c)	$\lvert t_4\rvert \times \lvert t_4\rvert$	$N(0, 0.2)$	2.62
(c)	$\lvert t_4\rvert \times \lvert t_4\rvert$	$N(0, 0.2X_1)$	3.39
(d)	$N_2((0,0), I_2)$	$N(0, 1)$	0.70
(d)	$N_2((0,0), I_2)$	$N(0, \lvert X_1\rvert)$	1.14

4. FRACTILE REGRESSION IN REAL APPLICATIONS

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 along with the $\mathbf{M}_n$-standardized regression function. 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 fractile standardization exactly achieves this purpose. In Figure 1, the surfaces were plotted with matched covariate values, but the matching covariates may belong to different 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. We thus see that fractile regression surfaces are robust to extreme values of covariates.

Example 2. In this example, we regress Y = ratio of profit to 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 standardized regression surfaces (using both notions of standardization $\mathbf{R_P}$ and $\mathbf{M_P}$) for the year 2003 lie almost completely below those 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 for this might be an increase in the competitiveness among the companies, which is due to growth in the number of companies as well as the emergence of foreign multinational companies over time. 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 difference in the distributions of the covariates corresponding to the two time points.

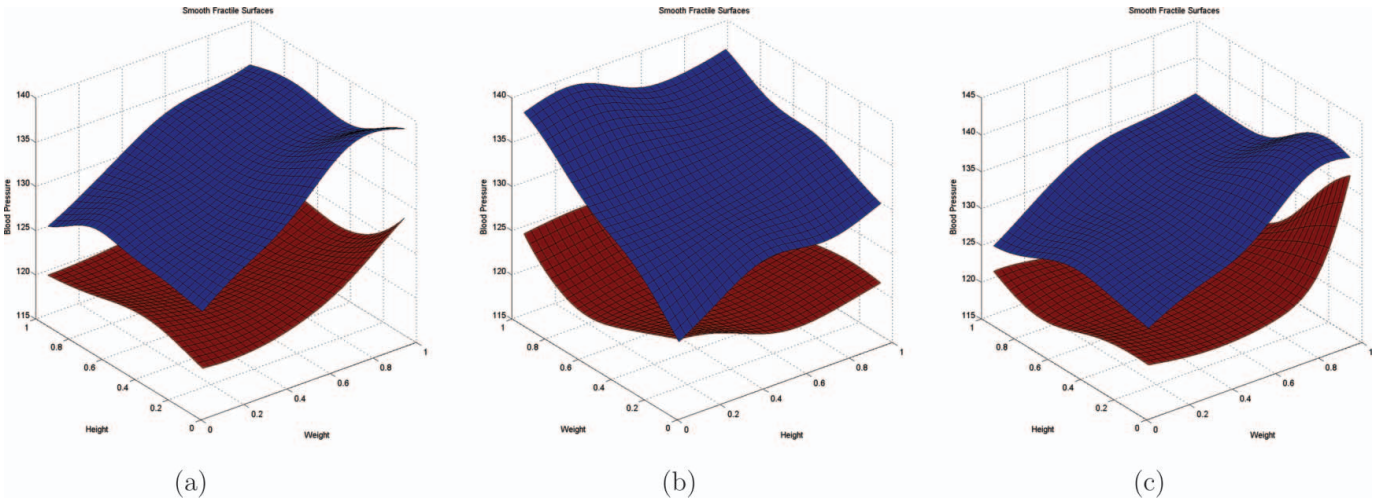


Figure 6. Smooth standardized regression surfaces for comparing Y = blood pressure in the *Bhutia* (red) and *Toto* (blue) tribes in Example 1, (a) using the standardization $\mathbf{R}_n$ with X_1 = weight and X_2 = height, (b) using the standardization $\mathbf{R}_n$ with the order of the covariates reversed, and (c) using the standardization $\mathbf{M}_n$.

4.1 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 information). It was created as part of a project analyzing poverty and existing social assistance programs in the transitional economies. What immediately arrests attention is the startling drop in income and increase in inequality accompanying the transition of these countries to market economies. We investigate this inequality in income and compare the economic condition of the transitional economies. A simple measure of the economic well-being of a population can be taken as the proportion of expenditure on food as a fraction of total expenditure per capita per household (in USD). This proportion would be quite small for rich and wealthy people, but for the poor it would be close to one. 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 datasets for two countries from the HEIDE database, namely, Poland (with 16,051 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 8(a) 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 (January to June 1993 for Poland and January to June 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

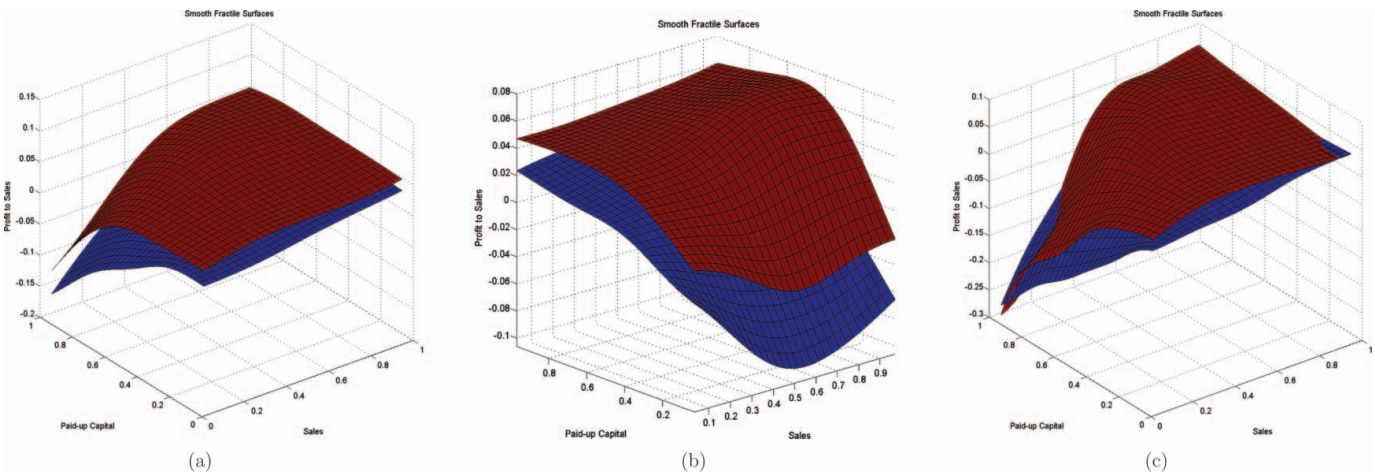
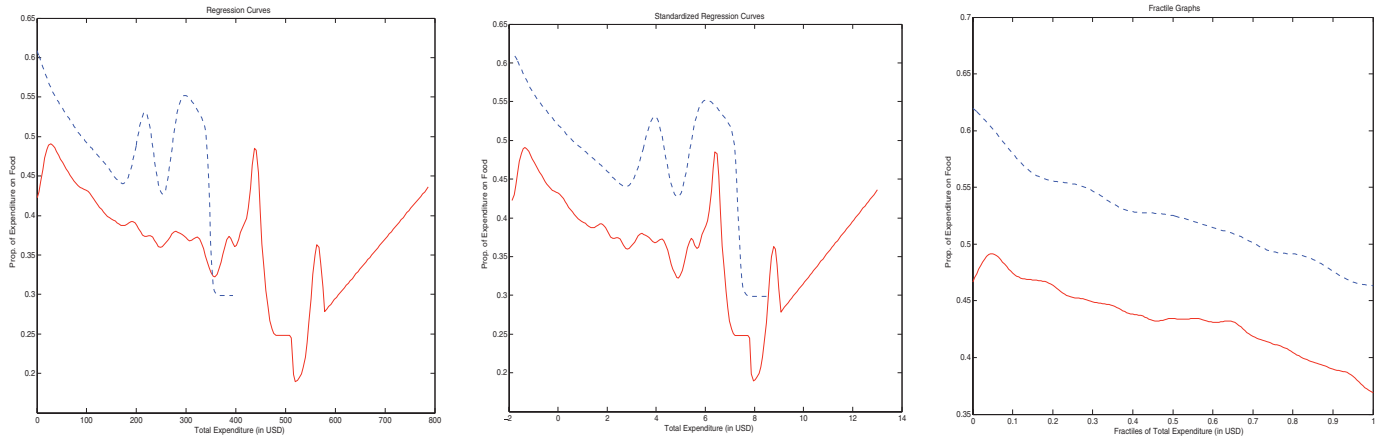


Figure 7. Smooth standardized regression surfaces for comparing Y = ratio of profit to sales for the years *1997* (red) and *2003* (blue) in Example 2, (a) using the standardization $\mathbf{R_P}$ with X_1 = sales and X_2 = paid-up capital, (b) using the standardization $\mathbf{R_P}$ with the order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



(a) (b) (c)

Figure 8. (a) Usual regression curves, (b) location and scale-adjusted regression curves, and (c) fractile curves, for proportion of expenditure on food on total expenditure for *Poland* (in red, solid line) and *Bulgaria* (in blue, dashed line) as discussed in Section 4.1.

Figure 8(b), the two curves are more aligned, but still the wiggliness for higher total expenditure values is disturbing. To make the regression curves comparable, we need some standardization of the covariates.

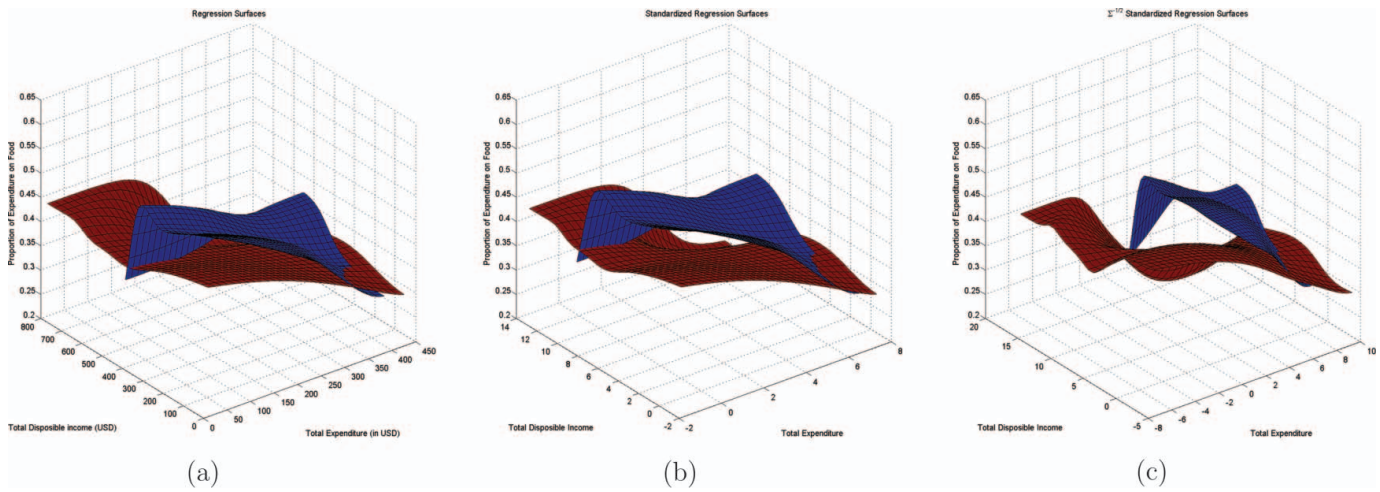
We would really like to compare the mean proportion of food expenditure for the poor (or the rich) in one population with that of the poor (or the rich) in the other population. The fractile curves accomplish exactly this, enabling us to compare the mean response values for fixed percentiles of total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

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 9(a) shows the usual regression surfaces, while Figure 9(b) shows the coordinate-wise location and scale-adjusted regression surfaces. Figure 9(c) 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 in Bulgaria during the time of the survey.

4.2 Choice of the Standardization: $\mathbf{R_P}$ Versus $\mathbf{M_P}$

In the preceding examples, we implemented both the multivariate standardizations $\mathbf{R_P}$ and $\mathbf{M_P}$. In the following, we discuss some of the advantages and disadvantages of the two methods.

Both $\mathbf{R_P}$ and $\mathbf{M_P}$ transform the covariate space to $(0, 1)^d$ and achieve invariance under component-wise increasing functions.



(a) (b) (c)

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).

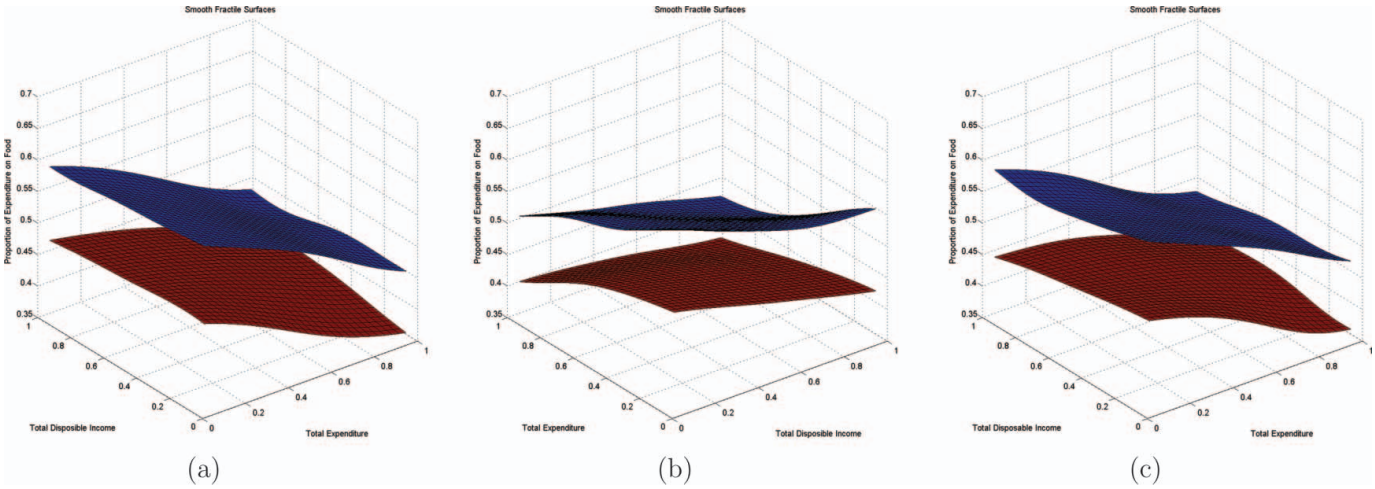


Figure 10. Smooth standardized regression surfaces for comparing Y = proportion of expenditure on food for the countries *Poland* (red) and *Bulgaria* (blue) (a) using the standardization $\mathbf{R}_n$ with X_1 = total expenditure and X_2 = total disposable income, (b) using the standardization $\mathbf{R}_n$ and the order of the covariates reversed, and (c) using the standardization $\mathbf{M}_n$.

But $\mathbf{R_P}$ achieves invariance under the group $\mathcal{F}$ (see Section 2.1), which is in fact much larger than the group of componentwise increasing transformations of the covariates. Recall from Section 2 that $\mathbf{R}_n$ and $\mathbf{M}_n$ are uniformly consistent estimators of $\mathbf{R_P}$ and $\mathbf{M_P}$, respectively. Further, $\mathbf{M}_n$ is a $n^{1/2}$-consistent estimator of $\mathbf{M_P}$, which is a consequence of the fact that the usual empirical process converges at $n^{1/2}$-rate. However, for $d \geq 2$, $\mathbf{R}_n$ converges to $\mathbf{R_P}$ at a slower rate, which depends on d, because it involves estimation of conditional distribution functions. As pointed out in Section 2.2, $\mathbf{M_P}$ is computationally simpler and does not depend on the ordering of the covariates.

But it is only $\mathbf{R_P}$ that achieves the required joint distributional standardization. $\mathbf{M_P}$, being a marginal standardization, does not take into account the multivariate distribution of $\mathbf{X}$ and, as a consequence, does not achieve joint distributional standardization. Note that all the marginal distributions of the standardized covariate vector $\mathbf{M_P}(\mathbf{X})$ are Uniform$(0, 1)$, but the joint distribution is not Uniform$(0, 1)^d$ unless the covariates are independent. The multivariate uniform distribution of the $\mathbf{R_P}$-transformed covariates also has a salutary effect on the estimation of the regression function as demonstrated in Theorem 3.2.

In situations, where the covariates are correlated among themselves, as in all the examples considered in this section, $\mathbf{M_P}$ leads to an inadequate standardization of the covariates for the two populations under comparison. On the other hand, $\mathbf{R_P}$ yields an adequate standardization of the covariate distributions by not only standardizing the marginal distributions but also the joint distributions. Figure 11 shows the scatterplots of the $\mathbf{M}_n$-transformed covariates for the three examples discussed in this section. We note that the transformed covariates have significant correlations between themselves, varying from 0.55 to 0.78. Further, the $\mathbf{M}_n$-transformed covariate space can still have regions of data sparsity within $(0, 1)^d$, as can be seen in Figure 11(b). Recall from Figure 7 that while comparing Y = ratio of profit to sales for the years 1997 and 2003 in Example 2 with X_1 = sales and X_2 = paid-up capital, the figure corresponding to the $\mathbf{M}_n$-transformed covariates still shows crossings of the two regression surfaces at the two extreme corners,

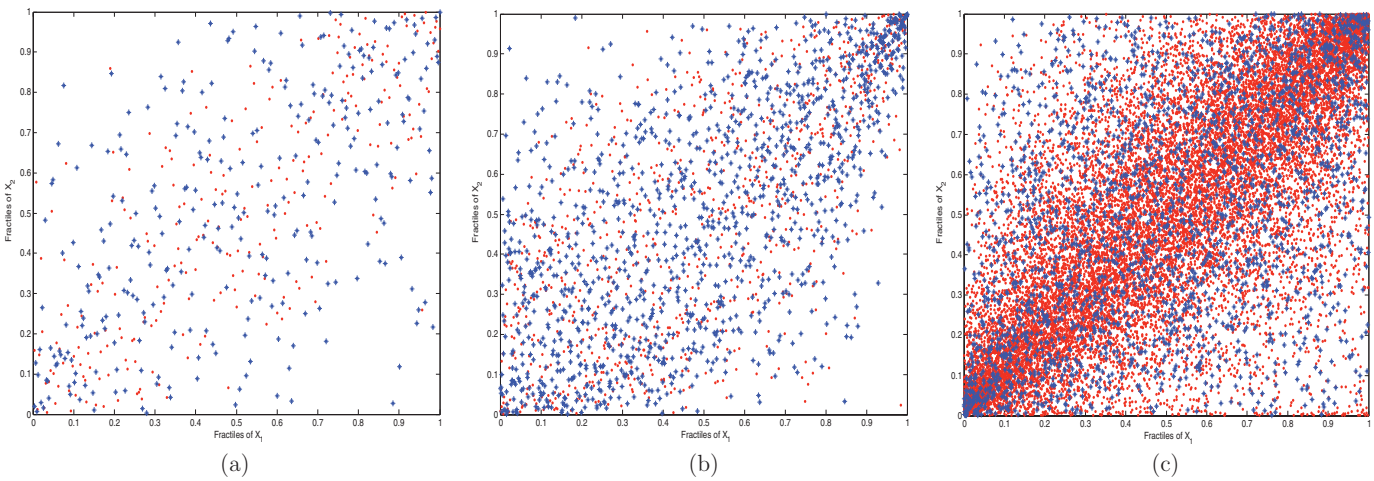


Figure 11. Scatterplot of the $\mathbf{M}_n$-transformed covariates in the three real examples: (a) for **Bhutia** (red) and **Toto** (blue) tribes with X_1 = weight and X_2 = height; (b) for the years **1997** (red) and **2003** (blue) with X_1 = sales and X_2 = paid-up capital; and (c) for **Poland** (red) and **Bulgaria** (blue) with X_1 = total expenditure and X_2 = total disposable income.

a feature that is not seen in the $\mathbf{R}_n$-transformed regression surfaces.

Heuristically, one can think of $\mathbf{M_P}$ as a nonparametric version of the usual coordinate-wise location and scale standardization, which has no effect on the correlation structure of the covariate distribution. $\mathbf{R_P}$, on the other hand, can be viewed as a nonparametric analog of the well-known standardization based on the subtraction of the mean vector and multiplication by the inverse of the square root of the dispersion matrix.

5. CONCLUDING REMARKS

In this article, we discuss standardization of covariates using the fractile transformation $\mathbf{R_P}$ in a regression setup to aid the comparison of two (or more) regression functions when the covariate vectors in the different populations have different distributions and supports. The $\mathbf{R_P}$ transform achieves distributional standardization, that is, the transformed covariates always have a common Uniform$(0, 1)^d$ distribution, and the corresponding fractile regression function possesses useful invariance properties under groups of transformations acting on the space of covariates. We also discuss the marginal transform $\mathbf{M_P}$ and study its equivariance/invariance properties. We develop smooth estimates of the transformed regression functions and illustrate their asymptotic properties.

Let us note that the $\mathbf{M_P}$ transform is particularly appealing when we assume an additive structure in the regression function, that is, $\mu(\mathbf{x}) = \theta_0 + \sum_{i=1}^d \theta_i(x_i)$ (see Stone 1985; Hastie and Tibshirani 1990). In this case, the fractile regression function using the $\mathbf{M_P}$ transform also has an additive structure and can be expressed as $m(\mathbf{t}) = \theta_0 + \sum_{i=1}^d \theta_i(F_i^{-1}(t_i))$, where $\mathbf{t} = (t_1, t_2, \ldots, t_d)$. This, in particular, facilitates estimation of $m(\mathbf{t})$ 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).

Though it is not very relevant in the examples considered in this article, sometimes affine or rotational equivariance of the chosen standardization may be an important requirement. For instance, in order to compare regression surfaces in a problem involving spatial covariates (e.g., rainfall data recorded in different locations in two geographical regions), orthogonal/affine equivariance is a natural requirement for the method of standardization for the covariates because an inference procedure should preferably not be affected by the choice of spatial coordinate systems. In such situations, we propose the use of the spatial rank transformation (discussed below), obtained by inverting the spatial quantile (or geometric quantile), introduced and studied by Chaudhuri (1996) and Koltchinskii (1997) (also see Breckling and Chambers 1988).

We define the spatial rank function (see Möttönen and Oja 1995) as $\mathbf{S_P}(\mathbf{x}) = E_{\mathbf{P}}(\frac{\mathbf{x}-\mathbf{X}}{\|\mathbf{x}-\mathbf{X}\|})$ for all $\mathbf{x} \in \mathbb{R}^d$, where $\mathbf{X} \sim \mathbf{P}$. Suppose that we have a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim \mathbf{P}$ in $\mathbb{R}^d$. The empirical spatial rank function is defined as $\mathbf{S}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \frac{\mathbf{x}-\mathbf{X}_i}{\|\mathbf{x}-\mathbf{X}_i\|}$. Computation of $\mathbf{S}_n(\mathbf{x})$ is simple, and its asymptotic properties are known. From Theorem (5.5) in Koltchinskii (1997), it follows that $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{S}_n(\mathbf{x}) - \mathbf{S_P}(\mathbf{x})\| \to 0$ a.s. as $n \to \infty$. See Serfling (2004) for a detailed account of the compelling strong points possessed by the spatial rank function.

Unlike $\mathbf{R_P}$, $\mathbf{S_P}$ is equivariant under the group of orthogonal linear transformations of $\mathbf{X}$, that is, for the group of transformations $\mathbf{x} \mapsto \mathbf{g}(\mathbf{x}) := A\mathbf{x}$, where A is orthogonal, we have $S_{\mathbf{g}(\mathbf{X})}(\mathbf{g}(\mathbf{x})) = AS_{\mathbf{X}}(\mathbf{x})$. This ensures the equivariance of the corresponding standardized regression function under such transformations, that is, $m_{\mathbf{g}(\mathbf{X})}(\mathbf{t}) := E\{Y|S_{\mathbf{g}(\mathbf{X})}(\mathbf{g}(\mathbf{X})) = \mathbf{t}\} = m_{\mathbf{X}}(A'\mathbf{t})$, and makes it a particularly useful standardization tool when dealing with spherically symmetric covariate distributions. As a consequence, $\mathbf{S_P}$, like $\mathbf{M_P}$, is equivariant under permutations of the coordinates of $\mathbf{X}$. However, unlike $\mathbf{R_P}$ or $\mathbf{M_P}$, $\mathbf{S_P}$ is not equivariant under arbitrary increasing transformations of the marginal variables. In fact, $\mathbf{S_P}$ is also not equivariant under general affine transformations. However, affine equivariant versions of the spatial multivariate quantile using the transformation–retransformation approach (see Chakraborty, Chaudhuri, and Oja 1998; Chakraborty 2001) can be used to extend the equivariance of the $\mathbf{S_P}$-standardized regression function under affine transformations, making it useful in problems involving covariates having elliptically symmetric distributions. Also, see Serfling (2010) for another related notion of centered rank function and its affine equivariant/invariance properties.

APPENDIX

Proof of Theorem 2.1. The ith coordinate of $\mathbf{R_X}(\mathbf{x})$ is $F_{\mathbf{X},i|1,2,\ldots,i-1}(x_i|x_1, \ldots, x_{i-1})$, where $F_{\mathbf{X},i|1,2,\ldots,i-1}$ is the conditional distribution function of X_i given $X_1, X_2, \ldots, X_{i-1}$, for $i = 1, 2, \ldots, d$. The result now follows from noting that the ith coordinate of $\mathbf{R}_{\mathbf{g}(\mathbf{X})}(\mathbf{g}(\mathbf{x}))$ is $F_{\mathbf{g}(\mathbf{X}),i|1,2,\ldots,i-1}(g_i(\mathbf{x}_i) \mid g_1(\mathbf{x}_1), \ldots, g_{i-1}(\mathbf{x}_{i-1}))$, which simplifies as

$$\begin{aligned} &P\left(g_i(\mathbf{X}_i) \le g_i(\mathbf{x}_i) \mid g_1(\mathbf{X}_1) = g_1(\mathbf{x}_1), \ldots, g_{i-1}(\mathbf{X}_{i-1}) = g_{i-1}(\mathbf{x}_{i-1})\right) \\ &\quad = P(X_i \le x_i \mid X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) \\ &\quad = F_{\mathbf{X},i|1,2,\ldots,i-1}(x_i|x_1, \ldots, x_{i-1}). \end{aligned}$$

■

Proof of Theorem 2.2. We first show that the following two statements are equivalent.

(i) $E\{Y|\mathbf{T}(\mathbf{X}, \mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{X})) = \mathbf{t}\}$ for all $\mathbf{t} \in E$, for all random vectors $(\mathbf{X}, Y)$ with $\mathbf{X} \sim \mathbf{P} \in \mathcal{P}$.

(ii) $\mathbf{T}(\mathbf{X}, \mathbf{x}) = \mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{x}))$ for all $\mathbf{x} \in \mathbb{R}^d$.

Note that $\mathbf{T}(\mathbf{X}, \mathbf{x}) = \mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{x}))$ for all $\mathbf{x} \in \mathbb{R}^d$ trivially implies

$$E\{Y|\mathbf{T}(\mathbf{X}, \mathbf{X}) = \mathbf{t}\} = E\{Y|\mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{X})) = \mathbf{t}\} \tag{A.1}$$

and hence $(ii) \Rightarrow (i)$. Also, choosing $Y = X_i$ and simplifying the conditional expectations on both sides of (3), for $i = 1, 2, \ldots, d$, it follows that $(i) \Rightarrow (ii)$. Therefore we have

$$\mathbf{T}(\mathbf{X}, \mathbf{x}) = \mathbf{T}(\mathbf{g}(\mathbf{X}), \mathbf{g}(\mathbf{x})) \tag{A.2}$$

for all $\mathbf{x} \in \mathbb{R}^d$, for all $\mathbf{g} \in \mathcal{F}$.

Fix $\mathbf{P} \in \mathcal{P}$, and take $\mathbf{g} = \mathbf{R_P}$. Note that the transformation $\mathbf{R_P} : \mathbb{R}^d \to \mathbb{R}^d$ belongs to $\mathcal{F}$. The result now follows immediately from (4) and observing that $\mathbf{g}(\mathbf{X}) \sim$ Uniform$(0, 1)^d$. Thus $\mathbf{T}(\mathbf{X}, \mathbf{x}) = \mathbf{h}(\mathbf{R_P}(\mathbf{x}))$, for some $\mathbf{h} : (0, 1)^d \to E$.

Now, if we assume that $\mathbf{T}(\mathbf{X}, \mathbf{X}) = \mathbf{h}(\mathbf{U}) \sim \mathbf{U} = (U_1, U_2, \ldots, U_d) \overset{d}{=}$ Uniform$(0, 1)^d$, we have $E = (0, 1)^d$. Also, if $\mathbf{h} = (h_1, h_2, \ldots, h_d) \in \mathcal{F}$, noting that $h_1(U_1) \sim$ Uniform$(0, 1)$, we have $P(h_1(U_1) \le u_1) = u_1$ for all $u_1 \in (0, 1)$ which implies $h_1(u_1) = u_1$ for all $u_1 \in (0, 1)$. Using this and the fact that $\mathbf{h} \in \mathcal{F}$, we can sequentially show that $h_i(u_1, u_2, \ldots, u_i) = u_i$ for $i = 1, 2, \ldots, d$. Thus, $\mathbf{h}$ is the identity function and this proves the result. ■

Proof of Theorem 3.1. 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 variance term, $E\{\widehat{m}_n(\mathbf{t}) - E(\widehat{m}_n(\mathbf{t}))\}^2$, can be simplified as

$$E\left[\sum_{i=1}^{n}\{Y_i - m(\mathbf{H}(\mathbf{X}_i))\}\, W_{n,i}(\mathbf{t})\right]^2 = \sum_{i=1}^{n} E\,\{Y_i - m(\mathbf{H}(\mathbf{X}_i))\}^2\, W_{n,i}^2(\mathbf{t}) \tag{A.3}$$

which is bounded by $K_0 \sum_{i=1}^{n} W_{n,i}^2(\mathbf{t}) = o_P(1)$, by assumption (W1) and the fact that $v(\cdot)$ is bounded. We decompose the conditional bias $\sum_{i=1}^{n} m(\mathbf{H}(\mathbf{X}_i))W_{n,i}(\mathbf{t}) - m(\mathbf{t})$ as

$$\sum_{i=1}^{n}\{m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})\}\, W_{n,i}(\mathbf{t}) + m(\mathbf{t})\left\{\sum_{i=1}^{n} W_{n,i}(\mathbf{t}) - 1\right\}. \tag{A.4}$$

Note that the second term in (6) 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 $\{\|\mathbf{t} - \mathbf{H}_n(\mathbf{X}_i)\| \le \delta_n\}$ as $E_{n,i}$. Therefore,

$$\begin{aligned} P\left(\left|\sum_{i=1}^{n} V_{n,i}\right| > \epsilon\right) &\le P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}}\right| > \epsilon/2\right) \\ &+ P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}^c}\right| > \epsilon/2\right) \le P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}}\right| > \epsilon/2\right) \\ &+ \eta/2 \quad \text{for all } n \ge N_1 \end{aligned} \tag{A.5}$$

as $P(|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}^c}| > \epsilon/2) \le P(2M \sum_{i=1}^{n} |W_{n,i}(\mathbf{t})|\mathbf{1}_{E_{n,i}^c} > \epsilon/2) \le \eta/2$ for all $n \ge N_1$ by (W3) and the fact that $m(\mathbf{t})$ is bounded.

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, $\|\mathbf{t} - \mathbf{H}_n(\mathbf{X}_i)\| \le \delta_n$ and $\|\mathbf{H}(\mathbf{X}_i) - \mathbf{H}_n(\mathbf{X}_i)\| \le B_n$ implies that $\|\mathbf{H}(\mathbf{X}_i) - \mathbf{t}\| \le \delta_n + B_n$ for all $i = 1, 2, \ldots, n$. Also notice that as $m(\cdot)$ is continuous at $\mathbf{t}$, there exists $\delta > 0$ such that $\|\mathbf{H}(\mathbf{X}_i) - \mathbf{t}\| \le \delta \Rightarrow |m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})| \le \frac{\epsilon}{2D}$. Now,

$$\begin{aligned} &P\left(\left|\sum_{i=1}^{n} V_{n,i}\mathbf{1}_{E_{n,i}}\right| > \epsilon/2\right) \le P\left(\sum_{i=1}^{n}\left|V_{n,i}\right|\mathbf{1}_{E_{n,i}} > \epsilon/2\right) \\ &\le P\left(\max_{1\le i\le n}|m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})|\,\mathbf{1}_{E_{n,i}}\sum_{i=1}^{n}|W_{n,i}(\mathbf{t})| > \epsilon/2\right) \\ &\le P\left(\max_{1\le i\le n}|m(\mathbf{H}(\mathbf{X}_i)) - m(\mathbf{t})|\,\mathbf{1}_{E_{n,i}} > \frac{\epsilon}{2D}\right) \le P\left(\delta_n + B_n > \delta\right) < \eta/2 \end{aligned} \tag{A.6}$$

for all $n \ge N_2$ 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 $\|\mathbf{t} - \mathbf{H}_n(\mathbf{X}_i)\| \le \delta_n$, which in turn implies that $\delta_n + B_n > \delta$.

Using (6), (7), and (8), 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 $n \ge \max\{N_1, N_2\}$. Thus, the conditional mean squared error of $\widehat{m}_n(\mathbf{t})$ approaches 0 in probability. An application of Chebyshev's inequality completes the proof of the weak consistency of $\widehat{m}_n(\mathbf{t})$.

Note that $\widehat{m}_n(\mathbf{t}) - E\{\widehat{m}_n(\mathbf{t})\} = \sum_{i=1}^{n} W_{n,i}(\mathbf{t})e_i$, where $e_i = Y_i - E(Y_i|X_i)$. 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 $i = 1, 2, \ldots, n$, and $S_n = \sum_{i=1}^{n} Z_{n,i}$. We use the Lindeberg–Feller central limit theorem to find the asymptotic distribution of S_n. Observe that $E(Z_{n,i}) = 0$ and $\sigma_{n,i}^2 = \mathrm{var}(Z_{n,i}) = v(\mathbf{H}(\mathbf{X}_i))W_{n,i}^2(\mathbf{t})$. Then $s_n^2 = \sum_{i=1}^{n}\sigma_{n,i}^2$. For any $\eta > 0$ and nonzero $W_{n,i}^2(\mathbf{t})$, the Lindeberg–Feller condition is exactly (C1), and thus the result follows. ■

Proof of Theorem 3.2. Under the conditions of the theorem, using (5) and noticing that $nh_{n,1}\ldots h_{n,d}\sum_{i=1}^{n} W_{n,i}^2(\mathbf{t}) \approx \sum_{i=1}^{n} n^{-1}(h_{n,1}\ldots h_{n,d})^{-1}\mathbf{K}_{\mathbf{h}}^2((\mathbf{t} - \mathbf{H}(\mathbf{X}_i))/\mathbf{h})/f^2(\mathbf{t})$ for the Nadaraya–Watson estimator, where f is the density of $\mathbf{H}(\mathbf{X})$, it can be shown that IAV $= \{\sigma^2\int_{\mathbb{R}^d}\mathbf{K}^2(\mathbf{u})d\mathbf{u}\}\times\int_{[0,1]^d}\{1/f(\mathbf{t})\}d\mathbf{t}$. The result now follows from the fact that the minimizer of $\int_{[0,1]^d}\{1/f(\mathbf{t})\}d\mathbf{t}$ with the constraint $\int_{[0,1]^d} f(\mathbf{t})d\mathbf{t} = 1$ is obtained when $f \equiv 1$ on $[0, 1]^d$. The result is also true for the local linear estimator [see Section 5.9, p. 140, of Wand and Jones (1995) for the key step in the proof] and many other linear smoothers. ■

[Received October 2009. Revised January 2011.]

REFERENCES

Bhattacharya, P. K., and Müller, H. G. (1993), "Asymptotics for Nonparametric Regression," *Sankhyā*, Series A, 53, 420–441. [351]

Breckling, J., and Chambers, R. (1988), "M-Quantiles," *Biometrika*, 75, 761–771. [360]

Chakraborty, B. (2001), "On Affine Equivariant Multivariate Quantiles," *Annals of the Institute of Statistical Mathematics*, 53, 380–403. [360]

Chaudhuri, P. (1996), "On a Geometric Notion of Quantiles for Multivariate Data," *Journal of the American Statistical Association*, 91, 862–872. [360]

Chesher, A. (2003), "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. [353]

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman & Hall/CRC. [360]

Hertz-Picciotto, I., and Din-Dzietham, R. (1998), "Comparisons of Infant Mortality Using a Percentile-Based Method of Standardization for Birthweight or Gestational Age," *Epidemiology*, 9, 61–67. [351]

Koltchinskii, V. I. (1997), "M-Estimation, Convexity and Quantiles," *The Annals of Statistics*, 25, 435–477. [360]

Ma, L., and Koenker, R. (2006), "Quantile Regression Methods for Recursive Structural Equation Models," *Journal of Econometrics*, 134, 471–506. [353]

Mahalanobis, P. C. (1960), "A Method for Fractile Graphical Analysis," *Econometrica*, 28, 325–351. [351]

Möttönen, J., and Oja, H. (1995), "Multivariate Spatial Sign and Rank Methods," *Journal of Nonparametric Statistics*, 5, 201–213. [360]

Nordhaus, W. D. (2006), "Geography and Macroeconomics: New Data and New Findings," *Proceedings of the National Academy of Sciences*, 103, 3510–3517. [351]

Parthasarathy, K. R., and Bhattacharya, P. K. (1961), "Some Limit Theorems in Regression Theory," *Sankhyā*, Series A, 23, 91–102. [351]

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270. [351]

Salibián-Barrera, M., and Wei, Y. (2008), "Weighted Quantile Regression With Nonelliptically Structured Covariates," *The Canadian Journal of Statistics*, 36, 595–611. [353]

Sen, B. (2005), "Estimation and Comparison of Fractile Graphs Using Kernel Smoothing Techniques," *Sankhyā*, 67, 305–334. [351,360]

Serfling, R. (2004), "Nonparametric Multivariate Descriptive Measures Based on Spatial Quantiles," *Journal of Statistical Planning and Inference*, 123, 259–278. [354,360]

Serfling, R. (2010), "Equivariance and Invariance Properties of Multivariate Quantile and Related Functions, and the Role of Standardization," *Journal of Nonparametric Statistics*, 7, 915–936. [352,360]

Sethuraman, J. (1961), "Some Limit Distributions Connected With Fractile Graphical Analysis," *Sankhyā*, Series A, 23, 79–90. [351]

Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals of Statistics*, 5, 595–620. [354,355]

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [360]

Van Keilegom, I., and Hettmansperger, T. (2002), "Inference Based on Multivariate *M*-Estimtors Based on Bivariate Consored Data," *Journal of the American Statistical Association*, 97, 328–336. [354]

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall. [349,361]

Wei, Y. (2007), "An Approach to Multivariate Covariate-Dependent Quantile Contours With Application to Bivariate Conditional Growth Charts," *Journal of the American Statistical Association*, 103, 397–409. [353]