SFB
823

Discussion Paper

A study on the least square estimator of multiple isotonic regression function

Pramita Bagchi, Subhra Sankar Dhar

Nr. 5/2018

A study on the least square estimator of multiple isotonic regression function

Pramita Bagchi [a] and Subhra Sankar Dhar [b]

[a] *Institute of Statistics, Ruhr Universitat, Germany*
[b] *Department of Mathematics and Statistics, IIT Kanpur, India*

Abstract

Consider the problem of pointwise estimation of f in a multiple isotonic regression model $Z = f(X_1, \ldots, X_d) + \epsilon$, where Z is the response variable, f is an unknown non-parametric regression function, which is isotonic with respect to each component, and ϵ is the error term. In this article, we investigate the behaviour of the least square estimator of f and establish its asymptotic properties. We generalize the greatest convex minorant characterization of isotonic regression estimator for the multivariate case and use it to establish the asymptotic distribution of properly normalized version of the estimator. Moreover, we test whether the multiple isotonic regression function at a fixed point is larger (or smaller) than a specified value or not based on this estimator, and the consistency of the test is established. The practicability of the estimator and the test are shown on simulated and real data as well.

Keywords: Consistency, Convex function, Cumulative sum diagram, Non-standard asymptotic distribution, Rate of convergence

1 Introduction

Monotonicity perhaps is the most basic shape constraint for a real valued function on $\mathbb{R}$, and for various applications, monotonicity of the unknown regression function is assumed. For example, in environmental science, the number of days until freezing of Lake Mendota has been modelled as an isotonic function or in medical science, monotone relationship has been extensively used in growth curves (see, e.g., Barlow (1972)). Since then several attempts had been made to incorporate the monotonicity (i.e., isotonic) constraint on the unknown regression function of one variable. The use of isotonic methods, i.e., the least squares estimation under a monotonicity constraint for such estimation problem is motivated by its attractive properties as well as the marked advantage of not having to specify a user-specified bandwidth for estimation. The study of isotonic regression models for one variable dates back to Brunk (1958) and since then it has been studied by several authors under various assumptions (see, e.g., Brunk (1970), Groeneboom and Wellner (1992), Banerjee and Wellner (2001), Wang and Huang (2002), Abrevaya and Huang (2005), Bagchi et al. (2016) and Dhar (2016)).

Although, to the best of our knowledge, the investigation on consistency property of estimated isotonic regression function with more than one covariate (i.e., multivariate isotonic regression function) started in 1970's (see, e.g., Robertson and Wright (1973) and Makowski (1977)), the asymptotic distribution of the estimated multiple isotonic regression function has not been paid attention in the literature. Recently, Han et al. (2017) studied the risk bounds of the multiple isotonic least square estimator, and a sort of similar work was carried out by Chatterjee et al. (2018) for bivariate isotonic least square estimator. However, none of them derived the asymptitoic distribution of the least square estimator, which is essential for important statistical problems like testing and constructing confidence interval of the regression function.

Regarding real application of multiple isotonic regression model, we would like to mention that a natural monotone relationship between the response and several co-variates is quite common in practice. For instance, it is an accepted fact in medical science that the blood pressure is monotonically associated with the use of tobacco and the body weight (see, e.g., Moolchan et al. (2004)). Motivated by such type of real life examples, the multiple isotonic regression model is considered in this article, and to estimate the unknown multiple isotonic regression function, the methodology of the least square is adopted following the idea of Makowski (1977). In this article, the multivariate isotonic regression model is studied in two steps. At first, we establish a characterization of the isotonic least square estimator in terms of the cumulative sum process of the data, similar to greatest convex minorant characterization for univariate case. This characterization enables us to establish geometric properties as well as derive explicit asymptotic distribution of the estimator after an appropriate normalization. The limit distribution we obtain is a generalization of well-known Chernoff's distribution (see Brunk (1970)). Next, although the theoretical properties of this multivariate version requires extending the seminal works of Groeneboom (1989) and Groeneboom and Wellner (2001) for multivariate case and beyond the scope of this article, we propose simulation methods to compute the quantiles numerically. This quantiles are then used to implement this methodology such as in testing of hypothesis problem, which will be discussed in the subsequent paragraph.

As it is mentioned in the last paragraph, the level of blood pressure can be modelled by multiple isotonic regression model with the covariates like the amount of using tobacco and the body weight. In this example, one may have interest to know whether the diastolic blood pressure will be more than 90 or not when the body weight of an individual $= 80$ kilogram, and the individual consumes 5 cigarettes per day. To test such type of assertion, a formal testing of hypothesis problem is formulated, and a test statistic based on the least square estimator is proposed. Moreover, the consistency property of the test based on that test statistic is investigated under any fixed alternative. We also thoroughly explore the performance of the test through extensive simulation study when the sample size is finite, and the test is implemented on two benchmark data set as well.

The rest of the article is organized as follows. In Section 2, the theoretical properties of the least square estimator of the unknown isotonic regression function is developed. The geometric characterization and the asymptotic distribution of the estimator are explored in Sections 2.1 and 2.2, respectively. In Section 3, we test whether the multiple isotonic regression function at a fixed point is larger (or smaller) than a specified value or not based on the least square estimator, and the consistency of the test is established. The procedures of computing the quantile and estimating all necessary parameters associated with partial derivative of the non-parametric regression function are thoroughly discussed in Sections 4.1 and 4.2. Section 4.3 investigates the finite sample performance of the estimator and the test for various examples. Section 5 implements the test on two well-known real data sets, and Section 6 contains a few concluding remarks. Some technical details are provided in the Appendix.

2 Isotonic Regression Estimator and Its Asymptotic Properties

The notion of monotonicity and linearity can be extended to multivariate case in a number of ways. Before we introduce our regression model, we formally define the notion of monotonicity and linearity in $\mathbb{R}^d$ that we are going to use.

Definition 2.1. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be monotone (or linear) if for every coordinate $i \in \{1, 2, \ldots, d\}$ and every choice of $x_1, \ldots x_d \in \mathbb{R}$, the function $y \mapsto f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d)$ is monotone (or linear).*

We consider the isotonic regression problem with data $Z_{i_1 \ldots i_d}; i_k = 1, \cdots, n_k$ for $k = 1, \ldots, d$ from the regression model

$$Z_{i_1 \ldots i_d} = f(x_{1,i_1}, \ldots, x_{d,i_d}) + \epsilon_{i_1 \ldots i_d}, \tag{2.1}$$

where $x_{k,1} \leq x_{k,2} \leq \cdots \leq x_{k,n_k}$ are the fixed design points for each coordinate $k = 1, 2, \ldots, d$, $f : [0, 1]^d \to \mathbb{R}$ is continuous and non-decreasing at every coordinate, and $\epsilon_{i_1 \ldots i_d}$ are independent and identically distributed random variables with mean zero and finite variance σ^2. We here denote the total sample size to be $n := \prod_{i=1}^{d} n_i$ and define a d-dimensional monotone cone L as follows: let $y = (y_1, \ldots, y_d) \in L$, if $y' \in \mathbb{R}^d$ is such that $y'_k \geq y_k$ with strict equality for at least one coordinate, then $y' \in L$. Let $\mathcal{L}$ be the collection of all such monotone cones.

The isotonic regression estimator (IRE) for f in (2.1) is obtained by minimizing the squared error loss over monotone cones, which can be written as the following optimization problem

$$\arg\min_{f} \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} (Z_{i_1 \ldots i_d} - f_{i_1 \ldots i_d})^2, \qquad \text{s.t. } \{f_{i_1 \ldots i_d}\} \subset L \text{ for some } L \in \mathcal{L}. \tag{2.2}$$

The solution of the above optimization problem is given by the following max-min representation.

$$\hat{f}_n(u_1, \ldots, u_d) = \max_{\substack{L:(u_1,\ldots,u_d)\in L \\ L\in\mathcal{L}}} \min_{\substack{K:(u_1,\ldots,u_d)\in K, \\ K\in\mathcal{L}^c}} \sum_{(x_{1,i_1},\ldots,x_{d,i_d})\in K\cap L} \frac{Z_{i_1 \ldots i_d}}{|K \cap L|}, \tag{2.3}$$

where $\mathcal{L}^c := \{L^c : L \in \mathcal{L}\}$. The estimator in (2.3) can be numerically computed using multivariate pooled adjacent violators algorithm (PAVA) as described in Hoffmann (2009). However, this max-min representation is not particularly useful to derive the asymptotic properties of the estimator, due to the complex geometry of multivariate cones and their complements. The geometric characterization of the estimator will be investigated in the following subsection.

2.1 Geometric Characterization of Isotonic Regression Estimator

A very popular and useful characterization of classical isotonic regression estimator is as left-derivative of greatest convex minorant of the cumulative sum process of the data. We here show a similar geometric representation for the estimator defined in (2.3). The key element here is a generalization of the notion of greatest convex minorant for $\mathbb{R}^d$. To this end, for a real-valued function G defined on $\mathbb{R}^d$, we introduce d-dimensional left-slope $\partial_\ell G(u_1, \dots, u_d)$ as simply left partial derivative with respect to $u_1, \dots, u_d$. For example, if $d = 2$, we have

$$\partial_\ell G(u_1, u_2) := \lim_{h \to 0} \frac{G(u_1, u_2) - G(u_1 - h, u_2) - G(u_1, u_2 - h) + G(u_1 - h, u_2 - h)}{h^2}.$$

The right-slope $\partial_r G$ can be defined similarly. With this notations, we define the class of d-convex functions on $I \subset \mathbb{R}^d$ to be

$$\mathcal{C}_I := \{G : I \mapsto \mathbb{R}, \partial_\ell G \text{ is coordinate-wise monotone}\}. \tag{2.4}$$

Note that if $G \in \mathcal{C}_I$, G is convex on I then; however, the converse is not necessarily true.

Next, for any real-valued function S defined on $I \subset \mathbb{R}^d$, we define d-GCM $T_I(S)$ of S as the point-wise supremum of all d-convex function below S, i.e.,

$$T_I(S)(u_1, \dots, u_d) = \sup_{G \in \mathcal{C}_I; G \leq S} G(u_1, \dots, u_d). \tag{2.5}$$

Note here that Equation (2.5) implies that d-GCM is a d-convex function itself (see Lemma A.4). In the sequel, for sake of notational simplicity, if $I = \mathbb{R}^d$, we drop the subscript and write $T_{\mathbb{R}^d}(S)$ as $T(S)$. In this context, it should be mentioned that for $d = 1$, $T_I(S)$ is the regular GCM (greatest convex minorant) of S. In fact, it inherits some useful properties of regular GCM such as piecewise linearity, and indeed the following property stated in the lemma (similar to regular GCM), which is essential to study the geometry of d-GCM.

Lemma 2.1. *Suppose $S : I \mapsto \mathbb{R}$ is a continuous function on an interval $I \subseteq \mathbb{R}^2$. If $T_I(S)$ and S do not have any touch point in an interval $J \subseteq I$, then $T_I(S)$ is linear on J.*

Note that the collection of all touch points, i.e., the points where the functions S and $T_I(S)$ coincide are union of rectangles $I_1 \times \cdots \times I_d$ where each I_k is either an interval or a singleton set. Lemma 2.1 ensures in between these rectangles (which may be a point) the d-GCM is linear.

With the definition of d-GCM, we are ready to establish the relation between the IRE and the cumulative sum process. To this end, let S_n be the cumulative sum diagram of the data, i.e., technically speaking, we define S_n on $[0,1]^d$ as follows

$$S_n(x_{1,i_1},\ldots,x_{d,i_d}) = \frac{1}{n_1\ldots n_d}\sum_{l_1\le i_1}\cdots\sum_{l_d\le i_d} Z_{l_1\ldots l_d},\ \text{for } i_k=0,\ldots,n_k; k=1,\ldots,d$$

with the notations $x_{k,0}=0$ and $Z_{i_1\ldots i_d}=0$ if $i_k=0$ for any k and S_n is interpolated linearly at each coordinate in between the design points. Finally, we are now ready to state first main result which gives explicit characterization of Isotonic regression estimator.

Theorem 2.2. *Let $G := T_{[0,1]^d}(S_n)$ and $g(u_1,\ldots,u_d) := \partial_\ell G(u_1,\ldots,u_d)$ be the left-slope of G. Then g is the unique solution to the isotonic regression problem described in* (2.2).

The proof is a generalization of the argument given in the proof of Theorem 1.2.1 from Robertson et al. (1988) using induction type argument and is deferred to the Section A.2. This result also provides some insight to the geometry of the estimator. The following corollary is a direct consequence of Theorem 2.2 and Lemma 2.1.

Corollary 2.3. *The isotonic regression estimator is piecewise constant and right continuous.*

Corollary 2.3 provides us the idea about the feature of the isotonic regression estimator for a given data. It also opens a new research problem that how to construct a smooth (i.e., differentiable) estimator of the multivariate isotonic regression function.

2.2 Asymptotic Properties of the Isotonic Regression Estimator

With the geometric characterization explained in the earlier subsection, we now focus on the asymptotics of $\hat{f}_n$, the solution of (2.2). To be more precise, we are interested in obtaining a limit distribution of $d_n(\hat{f}_n(u_1,\ldots,u_d) - f(u_1,\ldots,u_d))$ for some fixed $(u_1,\ldots,u_d)\in(0,1)^d$ and appropriately chosen d_n such that $d_n\to\infty$ as $n\to\infty$, with $n=n_1\ldots n_d$. In the sequel, we establish such a distributional convergence result and derive the appropriate rate of convergence d_n in the process.

In order to study aforementioned issues, we proceed in two steps. First we write $d_n(\hat{f}_n(u_1,\ldots,u_d) - f(u_1,\ldots,u_d))$ as left-slope of d-GCM of a normalized and localized version of the partial sum process. We then establish a distributional convergence result of the partial sum process to a functional of regular Brownian Motion, where the convergence is shown on the space of continuous functions on $\mathbb{R}^d$. The distributional convergence of normalized version of the estimator is finally proved in view of arguments like localization and continuous mapping.

To this end, for any real-valued S defined on $\mathbb{R}^d$, we define

$$\Delta S(u_1, \dots, u_d, h_1, \dots, h_d) = \int_{u_1}^{u_1+h_1} \cdots \int_{u_d}^{u_d+h_d} dS(v_1, \dots, v_d).$$

With this notation, we have the following result:

Proposition 2.4. *The normalized isotonic regression estimator can be written as*

$$\begin{aligned} & d_n(\hat{f}_n(u_1, \dots, u_d) - f(u_1, \dots, u_d)) \\ & \quad = \left. \frac{d_n^{d+1} \partial^d T \left(\Delta S_n \left(u_1, \dots, u_d, \frac{s_1}{d_n}, \dots, \frac{s_d}{d_n} \right) - \frac{f(u_1, \dots, u_d) s_1 \dots s_d}{d_n^d} \right)}{\partial s_1 \dots \partial s_d} \right|_{s_1=0, \dots, s_d=0}. \end{aligned}$$

The assertion in Proposition 2.4 implies that to establish the asymptotic distribution of $d_n(\hat{f}_n(u_1, \dots, u_d) - f(u_1, \dots, u_d))$, one needs to have the convergence of the process $\mathbb{W}_n$ defined as

$$\mathbb{W}_n(s_1, \dots, s_d) := d_n^{d+1} \left(\Delta S_n \left(u_1, \dots, u_d, \frac{s_1}{d_n}, \dots, \frac{s_d}{d_n} \right) - \frac{f(u_1, \dots, u_d) s_1 \dots s_d}{d_n^d} \right). \qquad (2.6)$$

The following result describes the convergence of $\mathbb{W}_n$ to the process defined as

$$\mathbb{W}(s_1, \dots, s_d) = \mathbb{B}(s_1, \dots, s_d) + \frac{|s_1 \dots s_d|}{2} \big(|s_1| f_1(u_1, \dots, u_d) + \dots + |s_d| f_d(u_1, \dots, u_d) \big), \qquad (2.7)$$

where $\mathbb{B}(s_1, \dots, s_d)$ is a Gaussian process with zero mean and covariance kernel

$$K\big((s_1, \dots, s_d), (t_1, \dots, t_d)\big) = (|s_1| \wedge |t_1|) \times \dots \times (|s_d| \wedge |t_d|), \qquad (2.8)$$

and f_k is the partial derivative of f with respect to the k-th co-ordinate.

Theorem 2.5. *Let $d_n = n^{1/d+2}$, the partial sum process $\mathbb{W}_n$ converges in distribution to $\mathbb{W}$ uniformly on all compact sets of $\mathbb{R}^d$. In other words, for $I_c := [-c, c]^d$, the processes $\{\mathbb{W}_n(s_1, \dots, s_d)\}_{(s_1, \dots, s_d) \in I_c}$ converges in distribution to $\{\mathbb{W}(s_1, \dots, s_d)\}_{(s_1, \dots, s_d) \in I_c}$.*

Proof. Introduce the notations

$$\begin{aligned} F_n(u_1, \dots, u_d) = & \frac{1}{n} \sum_{x_{1,i_1} \le u_1} \cdots \sum_{x_{d,i_d} \le u_d} f(x_{1,i_1}, \dots, x_{d,i_d}) \\ F(u_1, \dots, u_d) = & \int_0^{u_1} \cdots \int_0^{u_d} f(x_1, \dots, x_d) dx_1 \dots dx_d. \end{aligned}$$

We decompose the partial sum process as follows:

$$\mathbb{W}_n(s_1,\ldots,s_d)=\mathbb{B}_n(s_1,\ldots,s_d)+R_n(s_1,\ldots,s_d)+M_n(s_1,\ldots,s_d)+E_n(s_1,\ldots,s_d),$$

with $\mathbb{B}_n(s_1,\ldots,s_d)=\frac{d_n^{d+1}}{n}\sum_{(i_1,\ldots,i_d)\in D_n}\epsilon_{i_1\ldots i_d}$, where the sum is over the rectangle $D_n:=[n_1u_1,n_1u_1+n_1s_1/d_n]\times\cdots\times[n_du_d,n_du_d+n_ds_d/d_n]$;

$$\begin{aligned}R_n(s_1,\ldots,s_d):=&d_n^{d+1}\left[\Delta F_n(u_1,\ldots,u_d,s_1/d_n,\ldots,s_d/d_n)-\Delta F(u_1,\ldots,u_d,s_1/d_n,\ldots,s_d/d_n)\right];\\ M_n(s_1,\ldots,s_d)=&d_n^{d+1}\left[\Delta F(u_1,\ldots,u_d,s_1/d_n,\ldots,s_d/d_n)-f(u_1,\ldots,u_d)\frac{s_1\ldots s_d}{d_n^d}\right],\end{aligned}$$

and E_n is the approximation error due to linear interpolation with $\sup_{s_1,\ldots,s_d}|E_n(s_1,\ldots,s_d)|=O(d_n^{d+1}/n)$. As f is bounded on $[0,1]^d$, we have $\sup_{[0,1]^d}|F_n-F|=O(1/n)$, and therefore, $\sup_{s_1,\ldots,s_d}|R_n(s_1,s_2)|=O(d_n^{d+1}/n)$. For, $s_1,\ldots,s_d>0$, using Taylor's expansion, we have

$$\begin{aligned}&M_n(s_1,\ldots,s_d)=d_n^{d+1}\left[\int_{u_1}^{u_1+s_1/d_n}\cdots\int_{u_d}^{u_d+s_d/d_n}f(x_1,\ldots,x_d)dx_1\ldots dx_d-f(u_1,\ldots,u_d)\frac{s_1\ldots s_d}{d_n^d}\right]\\ &=d_n^{d+1}\left[\int_{u_1}^{u_1+s_1/d_n}\cdots\int_{u_d}^{u_d+s_d/d_n}\left(f(u_1,\ldots,u_d)+\sum_{k=1}^{d}(x_k-u_k)f_k(u_1,\ldots,u_d)+O(d_n^{-2})\right)dx_1\ldots dx_d\right.\\ &\qquad\left.-f(u_1,\ldots,u_d)\frac{s_1\ldots s_d}{d_n^d}\right]\\ &=\frac{s_1^2}{2}s_2\ldots s_df_1(u_1,\ldots,u_d)+\cdots+s_1\ldots s_{d-1}\frac{s_d^2}{2}f_d(u_1,\ldots,u_d)+O(d_n^{-1}).\end{aligned}$$

So, for $(s_1,\ldots,s_d)\in\mathbb{R}^d$, we have

$$M_n(s_1,\ldots,s_d)=\frac{|s_1\ldots s_d|}{2}\big(|s_1|f_1(u_1,\ldots,u_d)+\cdots+|s_d|f_d(u_1,\ldots,u_d)\big)+O(d_n^{-1}).$$

Finally, for $s_1,\ldots,s_d>0$,

$$\begin{aligned}\mathbb{B}_n(s_1,\ldots,s_d)=&\frac{d_n^{d+1}}{n}\sum_{i_1=\lfloor n_1u_1\rfloor}^{\lfloor n_1u_1+s_1n_1/d_n\rfloor}\cdots\sum_{i_d=\lfloor n_du_d\rfloor}^{\lfloor n_du_d+s_dn_d/d_n\rfloor}\epsilon_{i_1\ldots i_d}\\ \overset{d}{=}&\frac{d_n^{d+1}}{n}\sum_{i_1=1}^{\lfloor\frac{s_1n_1}{d_n}\rfloor}\cdots\sum_{i_d=1}^{\lfloor\frac{s_dn_d}{d_n}\rfloor}\epsilon_{i_1\ldots i_d}+o(1).\end{aligned}$$

It is clear from the above expression that the sum will converge in distribution when $\frac{n}{d_n^{d+1}}=\sqrt{\frac{n_1\ldots n_d}{d_n^d}}\Rightarrow d_n^{d+2}=n$. As symmetric arguments can be made for $(s_1,\ldots,s_d)$ in all quadrants of $\mathbb{R}^d$ by general Central Limit Theorem, with $d_n=n^{1/d+2}$, on $[-c,c]^2$, the processes $\mathbb{B}_n(s_1,\ldots,s_d)$ converge in distribution to a zero mean Gaussian process with covariance kernel given by (2.8). Moreover, when $d_n=n^{1/d+2}$, the order of the residual terms $O(d_n^{d+1}/n) = O(n^{-1/(d+2)})$, this completes the proof. □

For any function $S : I \subset \mathbb{R}^d \to \mathbb{R}$, introduce the notation

$$T_I(S)'(u_1, \dots, u_d) := \frac{\partial^d T_I(S)(x_1, \dots, x_d)}{\partial x_1 \dots \partial x_d}\Big|_{x_1=u_1,\dots,x_d=u_d},$$

and we drop the subscript if $I = \mathbb{R}^d$. Note that by Proposition 2.4, the normalized estimator is $T(\mathbb{W}_n)'(0, \dots, 0)$, and we intend to use argument like continuous mapping theorem to show this converges in distribution to $T(\mathbb{W})'(0, \dots, 0)$. However, the map $f \mapsto T'(f)$ is not continuous. Though, one can show that the map $f \mapsto T'_K(f)$ is indeed continuous for any compact set K. Therefore, one needs the following localization result in order to effectively use continuous mapping argument.

Proposition 2.6. *Let $I_n = [-s_1 d_n, (1-s_1)d_n] \times \cdots \times [-s_d d_n, (1-s_d)d_n]$. Then for any compact interval $J \subset I_n$ and $C = [-c, c]^d$, given $\epsilon > 0$, we have*

$$\lim_{c\to\infty} \mathbb{P}\left(\sup_J |T_C(\mathbb{W})'(.) - T(\mathbb{W})'(.)| > \epsilon\right) = 0$$

and

$$\lim_{c\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sup_J |T_C(\mathbb{W}_n)'(.) - T_{I_n}(\mathbb{W}_n)'(.)| > \epsilon\right) = 0.$$

The proof is closely related to the proof of Theorem A.1 of Anevski and Hössjer (2006) and deferred to Section A.3.

Finally, we now state the main result:

Theorem 2.7. *If $\hat{f}_n$ is the solution of the optimization problem in* (2.2)*, we then have*

$$n^{1/(d+2)}\left(\hat{f}_n(u_1, \dots, u_d) - f(u_1, \dots, u_d)\right) \xrightarrow{d} T(\mathbb{W})'(0, \dots, 0),$$

where $\mathbb{W}$ is defined as in (2.7)

Proof. It follows from Proposition 2.4 that

$$n^{1/(d+2)}\left(\hat{f}_n(u_1, \dots, u_d) - f(u_1, \dots, u_d)\right) = T_{I_n}(\mathbb{W}_n)'(0, \dots, 0).$$

Now, using Theorem 2.5, one can claim that $\mathbb{W}_n$ converges to $\mathbb{W}$ as processes in $C(\mathbb{R}^d)$. By Lemma A.6, we have the map $T_K : C(K) \mapsto C(\mathbb{R}^d)$ is continuous for all compact $K \subset \mathbb{R}^d$. This along with the application of continuous mapping theorem implies that for all $c > 0$, we have $T_{[-c,c]^d}(\mathbb{W}_n)$ converges in distribution to $T_{[-c,c]^d}(\mathbb{W})$ as processes in $C(\mathbb{R}^d)$. Also, by Corollary A.10, $(0, \dots, 0)$ is not a touch point of $\mathbb{W}$ and $T_{[-c,c]^d}(\mathbb{W})$ almost surely, and therefore, Lemma 2.1 ensures that with probability 1, $(0, \dots, 0)$ is a point of differentiability of $T_{[-c,c]^d}(\mathbb{W})$. Hence, using continuity of left derivatives of d-convex functions, (Lemma A.2), we have $T_{[-c,c]^d}(\mathbb{W}_n)'(0,\ldots,0) \overset{d}{\to} T_{[-c,c]^d}(\mathbb{W})'(0,\ldots,0)$. Finally, Proposition 2.6 along with an application of converging together lemma (Theorem 8.6.2 in Resnick (1999)) gives the desired result. □

Remark 2.8. *In our model, we assume that the design points are fixed. In case the design points x_{ij}'s are random (and independent of the error distribution) the limit distribution of $\mathbb{W}_n$ defined in* (2.6)*, and consequently, the limit distribution of the isotonic regression estimator depend on the distribution of the design points. For the special case, x_{ij}'s are generated independently as n_i points from uniform* $[0,1]$ *random variable, for each $i \in \{1,\ldots,d\}$, the limit distribution of isotonic regression estimator is the same as Theorem 2.7.*

3 Applications : Testing of Hypothesis and Confidence Interval

In the last section, we established that the least square estimator of unknown multiple isotonic regression function converges weakly to a random variable associated with a certain functional of the multivariate Brownian Motion. In addition to the estimation of the unknown function, as it is indicated in Section 1, one may have interest to know whether the unknown multiple isotonic regression function at a fixed point is larger (or smaller) than any fixed value or not. Technically speaking, we want to test $H_0 : f(x_{1,0},\ldots,x_{d,0}) > c$ against $H_1 : f(x_{1,0},\ldots,x_{d,0}) \leq c$, where $\{x_{1,0},\ldots,x_{d,0}\}$ and c are specified. To test H_0 against H_1, for arbitrary c_0 and c_1 are such that $c_1 < c < c_0$, one may consider simple null and alternative hypotheses as $H_0^* : f(x_{1,0},\ldots,x_{d,0}) = c_0$ against $H_1^* : f(x_{1,0},\ldots,x_{d,0}) = c_1$, for technical simplicity.

In order to test H_0^* against H_1^*, the test statistic $T_n = n^{\frac{1}{d+2}}\{\hat{f}_n(x_{1,0},\ldots,x_{d,0}) - c_0\}$ can be formulated, which is essentially the difference between estimated regression function and the specified value of the function with appropriate normalization when the null hypothesis is true. The following theorem states the consistency property of the test based on T_n.

Theorem 3.1. *Let c_α be the $(1-\alpha)$-th quantile of the distribution of $T(\mathbb{W})'(0,\ldots,0)$. The test, which rejects H_0^* if $T_n > c_\alpha$, will have asymptotic size $= \alpha$. Moreover, $P_{H_1^*}[T_n > c_\alpha] \to 1$ as $n \to \infty$, i.e., the test will be a consistent test.*

Proof. First note that since under H_0^*, T_n converges weakly to $T(\mathbb{W})'(0,\ldots,0)$, the asymptotic level of the test based on T_n will be α.

The asymptotic power of the test will be

$$P_{H_1^*}\left[n^{\frac{1}{d+2}}\{\hat{f}_n(x_{1,0},\ldots,x_{d,0}) - c_0\} > c_\alpha\right] = P_{H_1^*}\left[n^{\frac{1}{d+2}}\{\hat{f}_n(x_{1,0},\ldots,x_{d,0}) - c_1 + c_1 - c_0\} > c_\alpha\right]$$
$$P_{H_1^*}\left[n^{\frac{1}{d+2}}\{\hat{f}_n(x_{1,0},\ldots,x_{d,0}) - c_1\} > c_\alpha - n^{\frac{1}{d+2}}(c_1 - c_0)\right]$$
$$\to 1 \text{ as } n \to \infty.$$

The implication follows form the facts that $c_\alpha - n^{\frac{1}{d+2}}(c_1 - c_0) \to -\infty$ as $c_1 > c_0$, and $n^{\frac{1}{d+2}}\{\hat{f}_n(x_{1,0}, \ldots, x_{d,0}) - c_1\}$ is bounded in probability under H_1^*. This completes the proof. □

To implement the test based on T_n, one needs to compute $\hat{f}_n(x_{1,0}, \ldots, x_{d,0})$ for a given data, and that can be done using the geometric property discussed in Section 2.1. Besides, in order to compute the the specified quantile of the distribution of $T(\mathbb{W})'(0, \ldots, 0)$, one may adopt the methodology that will be discussed in Section 4.1. In this discussion, we would like to emphasize that the assertion in Theorem 3.1 implies that the test based on T_n poses good power when the sample size is large enough.

Further, note that one can also construct the pointwise confidence interval of the multiple isotonic regression function based on the result stated in Theorem 2.7. For instance, $(1-\alpha)\%$ ($\alpha \in (0,1)$) confidence interval of $f(u_1, \ldots, u_d)$ at the point $(u_1, \ldots, u_d)$ based on our proposed least square estimator is $\left(\hat{f}_n(u_1, \ldots, u_d) - \frac{c_{1-\alpha}}{n^{\frac{1}{d+2}}}, \hat{f}_n(u_1, \ldots, u_d) + \frac{c_\alpha}{n^{\frac{1}{d+2}}}\right)$, where c_α and $c_{1-\alpha}$ are α and $(1-\alpha)$-th quantiles of the distribution of $T_{I_n}(\mathbb{W}_n)'(0, \ldots, 0)$, respectively. That is, in other words, the aforementioned $(1-\alpha)\%$ asymptotic confidence interval is the acceptance region of the test H_0 against H_1 when the level of significance $= \alpha$.

4 Finite Sample Simulation Study

In the earlier section, the asymptotic distribution of $\hat{f}_n(x_1, \ldots, x_n)$ has been established after appropriate normalization; that however does not address how the estimator behave for finite sample size. To study this issue, we explore the performance of $\hat{f}_n(x_1, \ldots, x_n)$ in this section when the sample size is finite. In the course of this study, one needs to know how to compute the quantile of the limiting distribution associated with finding the critical value and to estimate the partial derivatives involved in the limiting distributions. These two issues are discussed in Subsections 4.1 and 4.2, and the models of simulation study and results are discussed in Subsection 4.3.

4.1 Computing Quantiles of the Limiting Distribution

The random variable $T(\mathbb{W})'(0, \ldots, 0)$ appearing in the limit as described in Theorem 2.7 is indeed a generalization of the well-known Chernoff's distribution (See Groeneboom (1985)). The theoretical properties of this random variable will require a deep study of Gaussian process and its d-convex minorant similar to the works of Groeneboom (1989), Groeneboom and Wellner (2001) etc, and it is beyond the scope of this article. However, one can simulate the data from the distribution associated with $T(\mathbb{W})'(0, \ldots, 0)$ and compute the empirical version of a certain quantile based in the simulated data. As it is discussed in Section 3, that quantile can be used in estimating critical value and to formulate the point-wise confidence interval.

In the course of analyzing the random variable $T(\mathbb{W})'(0,\ldots,0)$, one can note that this random variable involves $(d+1)$ many parameters, namely, the error variance $=\sigma$ and the d-dimensional gradient of f defined as $\nabla f := (f_1(u_1,\ldots,u_d),\ldots,f_d(u_1,\ldots,u_d))^T$, where f_j is the partial derivative of f with respect to j-th component, $j=1,\ldots,d$. At first, we generate the data from the model $y_{i_1\ldots i_d} = (\nabla f)_1 i_1/n_1 + \cdots + (\nabla f)_d i_d/n_d + \epsilon_{i_1\ldots i_d}$ for $n_1 = \ldots = n_d = 10^3$, where $(\nabla f)_j$ is the j-th component of the d-dimensional vector (∇f), and $\epsilon_{i_1\ldots i_d}$ are i.i.d. with variance $=\sigma$. The estimation of the partial derivative of the unknown regression function is described in the next subsection. We employ then the multivariate PAVA to calculate the isotonic regression estimator $\hat{f}_n$ from this data at $u_1 = \cdots = u_d = 0$ and its normalized version. The same procedure is repeated $M = 10^4$ times, and $\alpha\%$ empirical quantile of the distribution associated with $T(\mathbb{W})'(0,\ldots,0)$ can be obtained from the $\alpha\%$ quantile of the M many values of $T(\mathbb{W})'(0,\ldots,0)$.

4.2 Estimation of the Partial Derivatives

As in the case of single covariate, estimation of the partial derivatives (i.e., f_j, $j=1,\ldots,d$) is one of the most challenging part for implementing this method. We here use a kernel based estimate from Banerjee and Wellner (2005) defined as

$$\hat{f}_j(u_1,\ldots,u_d) = \frac{1}{h_j}\int K\left(\frac{u_j - x}{h_j}\right) d\hat{f}_n(u_1,\ldots,u_{j-1},x,u_{j+1},\ldots,u_d) \tag{2.9}$$

for $j=1,\ldots,d$, where h_j is the bandwidth, and K is a Gaussian kernel (see Silverman (1986)). However, to implement the aforementioned methodology, one needs to choose the bandwidth h_j in an appropriate way. In the numerical study reported in this article, the bandwidth is chosen by the method of cross -validation, which is described as follows. To implement the cross validation technique, we divide the dataset into two parts randomly, and each data-point is assigned to one of the two sets with probability $=0.5$ using an auxiliary Bernoulli random variable having success probability $=0.5$. Let D_i denote the set of indices of the i-th data set ($i=1$ and 2), and we then estimate $\hat{f}_{k,D_i,h_j}$ as (2.9) using the data having the indices D_i. Next, $\hat{f}_{D_i,h_j}$ is calculated by numerically integrating $\hat{f}_{k,D_i,h_j}$ with respect to the k-th coordinate, and we finally calculate

$$CV_k(h_j) = \sum_{(i_1,\ldots,i_d)\in D_1} (Z_{i_1\ldots i_d} - \hat{f}_{D_2,h_j}(x_{1,i_1},\ldots,x_{d,i_d}))^2 + \sum_{(i_1,\ldots,i_d)\in D_2} (Z_{i_1\ldots i_d} - \hat{f}_{D_1,h_j}(x_{1,i_1},\ldots,x_{d,i_d}))^2. \tag{2.10}$$

The optimal bandwidth is obtained by minimizing $CV_k(h)$ with repsect to h.

4.3 Numerical Studies

We here investigate the performance of the proposed estimator $\hat{f}_n(\mathbf{x})$ for finite n, where $\mathbf{x} = (x_1,\ldots,x_d)$, and for that reason, the empirical mean sqaure error (EMSE) is defined here. For a given $\mathbf{x}$ and the model

$Y = f(\mathbf{x}) + \epsilon$, the EMSE of $\hat{f}_n$ is $\frac{1}{M}\sum\limits_{i=1}^{M}\left\{\hat{f}_{n,i}(\mathbf{x}) - f(\mathbf{x})\right\}^2$, where M is the number of replications, and $\hat{f}_{n,i}(\mathbf{x})$ is the value of $\hat{f}_n(\mathbf{x})$ for the i-th replication. In the numerical study, we consider the following examples.

Example 1: $f(\mathbf{x}) = x_1^2 + x_2^2$, where $\mathbf{x} := (x_1, x_2) = \left(\frac{i}{n_1+1}, \frac{j}{n_2+1}\right)$, i.e., $(x_1, x_2) \in (0,1) \times (0,1)$. Here $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$.

Example 2: $f(\mathbf{x}) = \exp(x_1 + x_2)$, where $\mathbf{x} := (x_1, x_2) = \left(\frac{i}{n_1+1}, \frac{j}{n_2+1}\right)$, i.e., $(x_1, x_2) \in (0,1) \times (0,1)$. Here also, $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$.

We would like to mention that here $d = 2$ is considered only because of concise presentation. In principle, one may study the behaviour of $\hat{f}_n$ for any dimension d. We here investigate the performance of $\hat{f}_n(\mathbf{x})$ at $\mathbf{x} = (0.5, 0.5)$ and choose $M = 1000$, and $n = 100$ ($n_1 = n_2 = 10$), 200 ($n_1 = 20$ and $n_2 = 10$), 300 ($n_1 = 30$ and $n_2 = 10$), 400 ($n_1 = 40$ and $n_2 = 10$) and 500 ($n_1 = 50$ and $n_2 = 10$). The different forms of error distribution, namely, standard normal (i.e., the form of the density function: $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$) and Laplace distributions (i.e., the form of the density function : $f(x) = \frac{1}{2}e^{-|x|}$, $x \in \mathbb{R}$) are considered here. The results are summerized in Table 1. The values in Table 1 indicates that the EMSE of $\hat{f}_n$ decreases as the sample size n increases when the errors are generated from Gaussian and Laplace distributions. Moreover, it performs better when the errors are generated from the Gaussian distribution, which is expected since the least sqaure estimator performs well for data following Gaussian distribution.

n	100	200	300	400	500
Error distribution : standard normal distribution					
Example 1 : EMSE of $\hat{f}_n(\mathbf{x})$	0.123	0.111	0.099	0.076	0.063
Example 2: EMSE of $\hat{f}_n(\mathbf{x})$	0.244	0.208	0.176	0.162	0.131
Error distribution : standard Laplace distribution					
Example 1 : EMSE of $\hat{f}_n(\mathbf{x})$	0.637	0.558	0.500	0.482	0.336
Example 2: EMSE of $\hat{f}_n(\mathbf{x})$	1.221	1.099	0.899	0.828	0.799

Table 1: The EMSE of $\hat{f}_n(\mathbf{x})$ different values of n.

We also want to test that whether $f(\mathbf{x})$ at $\mathbf{x} = (0.5, 0.5)$ is larger than 0.35 or not in Example 1, and in Example 2, the null hypothesis is $f(\mathbf{x})$ at $\mathbf{x} = (0.5, 0.5)$ is larger than 2 or not. To carry out the aforementioned testing of hypothesis problem, the permutation tests are done based on the permulated samples as described in Section 5. In Example 1, the p-values are 0.623 and 0.437 for erros having Gaussian and Laplace distributions, respectively. While in Example 2, the p-values are 0.847 and 0.501 for erros having Gaussian and Laplace distributions, respectively. Note that at $\mathbf{x} = (0.5, 0.5)$, $f(\mathbf{x}) = x_1^2 + x_2^2 = 0.5$ (in Example 1) and $f(\mathbf{x}) = \exp(x_1 + x_2) = e > 2$ (in Example 2), and hence, the null hypotheses for both cases are expected to be accepted, which is reflected in the obtained p-values.

Further, in order to investigate the performance of the finite sample power of the test based on T_n, we consider the following examples.

Example 3: Let $f(\mathbf{x}) = x_1^2 + x_2^2$, and $\mathbf{x}$ is observed at $\left(\frac{i}{n_1+1}, \frac{j}{n_2+1}\right)$, $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. Suppose that we want to test $H_0 : f(\mathbf{x})|_{\mathbf{x}=(0.5,0.5)} = 0.5$ against $H_1 : f(\mathbf{x})|_{\mathbf{x}=(0.5,0.5)} > 0.5$ (say, $= a > 0.5$), and here, errors follow standard normal distribution.

To implement the test, we compute the critical value using the asymptotic distribution described in Theorem 2.8, i.e., in other words, at $\alpha\%$ level of significance, the critical value will be $(1 - \alpha)$-th quantile of the distribution of $T(\mathbb{W})'(0, \ldots, 0)$. As the the distribution of $T(\mathbb{W})'(0, \ldots, 0)$ is not easily tractable, we compute the quantile of the aforementioned distribution approximately using the procedure described in Section 4.1. For different values of n_1, n_2 and a, the finite sample power of the test is studied, and the result is summarized in Table 2.

Finite Sample Power					
a	50	60	90	150	500
$n_1 = 10$ and $n_2 = 10$	0.050	0.089	0.112	0.223	0.341
$n_1 = 25$ and $n_2 = 25$	0.052	0.123	0.236	0.356	0.622
$n_1 = 50$ and $n_2 = 50$	0.053	0.436	0.567	0.678	0.943
$n_1 = 100$ and $n_2 = 100$	0.051	0.632	0.902	0.945	0.989

Table 2: The results for **Example 3**: The finite sample power of the test based on T_n at 5% level of significance for different values of n_1, n_2 and a.

Example 4: Let $f(\mathbf{x}) = \exp(x_1 + x_2)$, and $\mathbf{x}$ is observed at $\left(\frac{i}{n_1+1}, \frac{j}{n_2+1}\right)$, $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. Suppose that we want to test $H_0 : f(\mathbf{x})|_{\mathbf{x}=(0.5,0.5)} = e$ against $H_1 : f(\mathbf{x})|_{\mathbf{x}=(0.5,0.5)} > e$ (say, $= a^* > e$), and here, errors follow standard normal distribution. For different values of n_1, n_2 and a^*, the finite sample power of the test based on T_n is studied, and the result is summarized in Table 3.

Finite Sample Power					
a^*	e	$2e$	$5e$	$10e$	$100e$
$n_1 = 10$ and $n_2 = 10$	0.050	0.123	0.0.245	0.546	0.659
$n_1 = 25$ and $n_2 = 25$	0.057	0.244	0.386	0.732	0.844
$n_1 = 50$ and $n_2 = 50$	0.055	0.505	0.667	0.789	0.932
$n_1 = 100$ and $n_2 = 100$	0.050	0.713	0.0.888	0.965	0.999

Table 3: The results for **Example 4**: The finite sample power of the test based on T_n at 5% level of significance for different values of n_1, n_2 and a^*.

The figures in Tables 2 and 3 indicate that for a fixed n_1 and n_2 (i.e., sample sizes), the finite sample power increases as the values of a and a^* increase. In other words, for the fixed sample sizes, the test based on T_n will become more powerful as the deviation of the alternative hypothesized value from the null hypothesized value is getting increased. Besides, for a fixed values of a and a^*, for both Examples 3 and 4, the finite sample power of the test increases as the sample sizes (i.e, n_1 and n_2) increase, i.e., this study also indicates that the test based on T_n is consistent.

5 Real Data Analysis

5.1 Auto MPG Data Set

This data set lists the miles-per-gallon (MPG) of 392 automobiles manufactured between 1970 and 1982 with seven covariates. For detailed description of the variables, we refere the readers to `https://archive.ics.uci.edu/ml/datasets/auto+mpg`, and it was earlier analyzed in Luss et al. (2012). In this numerical study, we consider only the continuous covariates, namely, displacement (denote it as x_1), horsepower (denote it as x_2), weight (denote it as x_3) and acceleration (denote it as x_4), and it is seen that MPG (denote it as y) has monotone association with these four variables. Suppose that for this data, we want to study the performance of the proposed estimator when $x_1 = 455$, $x_2 = 225$, $x_3 = 3086$ and $x_4 = 10$, and for this value of (x_1, x_2, x_3, x_4), it is given that $y = 14$. In this study, we generate B many permuted resamples, and for each resample, our proposed estimate $\hat{f}_n(x_1, x_2, x_3, x_4)$ is computed, and let $\hat{f}_{n,i}(x_1, x_2, x_3, x_4)$ be the estimate for the i-th permuted resample, where $i = 1, \ldots, B$. The empirical mean square error (EMSE) of $\hat{f}_{n,i}(x_1, x_2, x_3, x_4)$ is defined as $\frac{1}{B}\sum\limits_{i=1}^{B}\{\hat{f}_{n,i}(x_1, x_2, x_3, x_4) - y\}^2 = \frac{1}{B}\sum\limits_{i=1}^{B}\{\hat{f}_{n,i}(x_1, x_2, x_3, x_4) - 14\}^2$, and we here investigate the behavior of EMSE of $\hat{f}_n(x_1, x_2, x_3, x_4)$ for different values of B. The permutation of the sample is done in the following way. Let $(i_1, \ldots, i_n)$ be one permutation of $(1, \ldots, n)$, and the permuted sample will be $(((x_{1,1}, x_{2,1}, x_{3,1}, x_{4,1}), y_{i_1}), \ldots, ((x_{1,n}, x_{2,n}, x_{3,n}, x_{4,n}), y_{i_n}))$, where $n = 392$, and $x_{k,l}$ is the l-th observation of x_k; here $k = 1, 2, 3, 4$ and $l = 1, \ldots, n$. The result is summarized in Table 4.

It clearly indicates from the figures in Table 4 that the EMSE of the proposed estimator decreases as the number of resamples obtained by permutation decreases. In other words, the least square estimate accurately estimate the actual value of MPG when the number of replications (i.e., B) is sufficiently large. Moreover, the EMSE values of $\hat{f}_n(\mathbf{x})$ of Bodyfat data are larger than that of $\hat{f}_n(\mathbf{x})$ of Auto MPG Data as the Bodyfat data has a few outliers/influential observations.

Besides, for this data, we also test that whether MPG is larger than ten or not when $x_1 = 455$, $x_2 = 225$, $x_3 = 3086$ and $x_4 = 10$ (x_1, x_2, x_3 and x_4 are same as before). We compute the p-value of the test based on T_n (described in Section 3) using the permuted resamples as mentioned in the first paragraph in this data analysis. We obtain the p-value $= 0.661$, i.e., favours the null hypothesis, and it is expected since at $x_1 = 455$, $x_2 = 225$, $x_3 = 3086$ and $x_4 = 10$, the MPG is 14. In this study, we have used our own R code to compute $\hat{f}_n(\mathbf{x})$, which is available to the authors.

B	100	200	300	400	500	1000
Auto MPG Data : EMSE of $\hat{f}_n(\mathbf{x})$	3.227	2.898	2.766	1.868	1.422	1.001
Bodyfat Data : EMSE of $\hat{f}_n(\mathbf{x})$	5.554	5.001	4.998	4.887	4.775	4.776

Table 4: The results for **Auto MPG Data** and **Bodyfat Data**: The EMSE of $\hat{f}_n(\mathbf{x})$ different values of B.

5.2 Bodyfat Data Set

This data set consists of percentage of body fat (obtained from equation from Siri et al. (1956) ; denote it as y), Age (years), Weight (lbs), Height (inches), Neck circumference (cm), Chest circumference (cm), Abdomen 2 circumference (cm), Hip circumference (cm), Thigh circumference (cm), Knee circumference (cm), Ankle circumference (cm), Biceps (extended) circumference (cm), Forearm circumference (cm) and Wrist circumference (cm) of 252 men aged from 22 to 81, and it is available in `http://lib.stat.cmu.edu/datasets/bodyfat`. This data set was earlier analyzed by Dette and Scheder (2006) in the context of multiple isotonic regression; although their proposed methodology was different from the estimator considered here. In that study, they considered two covariates, namely, Weight and Height since bodyfat should be monotonically increasing function of weight and decreasing function of height. Following their idea and along with the fact that the unknown regression function is monotonic with respect to each component in the same direction, Weight (denote it as x_1) and the negative of Height (denote it as x_2) are considered as two covariates in this study.

As we discussed in the earlier data, we here also investigate the behavior of EMSE of $\hat{f}_n(x_1, x_2)$ when $x_1 = 71$ and $x_2 = -209.25$ for different values of B, and for this value of (x_1, x_2), it is given that $y = 1.0468$. The results are summarized in Table 4. Here also, we have observed the same phenomena as the earlier data set that the EMSE decreases as the number of replications increases. Overall, we would like to conclude that if the data has monotone association, it will be expected that $\hat{f}_n(x_1, \ldots, x_n)$ will perform well.

For this data also, we compute the p-value of the test based on T_n to check whether bodyfat (i.e., y) is larger then one or not when $x_1 = 71$ and $x_2 = -209.25$. To compute the p-value, as we did for Auto MPG Data Set, we carry out well-known permutation test and obtain the p-value $= 0.557$ (i.e., favours the said hypothesis) as expected since for $x_1 = 71$ and $x_2 = -209.25$, $y = 1.0468 > 1$.

6 Concluding Remarks

In this article, we propose a least square estimator of the multiple isotonic regression function and study it's asymptotic properties along with applications in the testing of hypothesis and the formulation of the pointwise confidence interval. In this context, we would like to point out that even for the smooth (i.e., differentiable) multiple isotonic regression function, our estimator is not smooth enough; strictly speaking, it is not a differentiable function. To overcome this issue, one may consider the kernalized version of our proposed estimation, which can be defined as follows.

$$\hat{f}_{n,sm}(u_1^*, \dots, u_d^*) = \frac{1}{nh_n} \sum_{i=1}^{n} k\left(\frac{\mathbf{u}^* - \mathbf{x}_i}{h_n}\right) \hat{f}_n(u_1^*, \dots, u_d^*),$$

where k is a sufficiently smooth kernel function with bandwidth $= h_n$, $\mathbf{u}^* = (u_1^*, \dots, u_d^*)$, and $\mathbf{x}_i$ is the i-th covariate. One can hope that under some conditions, the smoothness of $\hat{f}_{n,sm}(u_1^*, \dots, u_d^*)$ will be the same as that of the kernel k, and the isotonicity propoerty of the estimator depends on the choice of the kernel k.

The issue of robustness is another topic of research since $\hat{f}_n(.)$ is an average based estimator, and hence, it is expected that $\hat{f}_n(.)$ will be influenced by the presence of the outliers/influential observations. Technically speaking, the breakdown point or the influence function may give us an idea about how much the estimator will be robust against the outliers. Moreover, one may also consider the median or the trimmed mean type of estimator so that the estimator possess good robustness property. For instance, in the case of univariate isotonic regression function, Dhar (2016) showed that trimmed mean isotonic regression estimator may achieve 25% asymptotic breakdown point.

In order to implement the test described in Section 3, one needs to compute a certain quantile of $T(\mathbb{W})'(0, \dots, 0)$, which is not easily tractable. As we indicated earlier, note here that $T(\mathbb{W})'(0, \dots, 0)$ can be thought as a multivariate extension of well-known Chernoff's distribution; however, unlike the univariate Chernoff's distribution, the accurate computation of the quantiles of multivariate version of Chernoff's distribution is not available in the literature. For univariate case, the readers may see Groeneboom and Wellner (2001). Instead of the simulation based procedure of computing quantiles of $T(\mathbb{W})'(0, \dots, 0)$ described in Section 4.2, it may be of future interest of research about how to compute the quantiles exactly of the distribution of $T(\mathbb{W})'(0, \dots, 0)$ like the univariate Chernoff's distribution.

Moreover, to estimate the partical derivatives of the unknown multiple isotonic regression function, we discussed the cross validation technique to obtain the optimum bandwidth in Section 4.2. Here, in order to obtain the optimum bandwidht involved in the kernel function, one may consider the asymptotic results of having the optimum order of the bandwidth in terms of the sample size.

Acknowledgement

The work of the authors was supported by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Project C1) of the German Research Foundation (DFG). The authors would like to thank Mr. Subhajit Chattopadhyay for carrying out the numerical computations for the paper.

References

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*, 73(4):1175–1204.

Anevski, D. and Hössjer, O. (2006). A general asymptotic scheme for inference under order restrictions. *The Annals of Statistics*, 34(4):1874–1930.

Bagchi, P., Banerjee, M., and Stoev, S. A. (2016). Inference for monotone functions under short-and long-range dependence: Confidence intervals and new universal limits. *Journal of the American Statistical Association*, 111(516):1634–1647.

Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *The Annals of Statistics*, 29(6):1699–1731.

Banerjee, M. and Wellner, J. A. (2005). Confidence intervals for current status data. *Scandinavian Journal of Statistics*, 32(3):405–424.

Barlow, R. (1972). Statistical inference under order restrictions; the theory and application of isotonic regression. Technical report.

Brunk, H. (1958). On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, pages 437–454.

Brunk, H. D. (1970). Estimation of isotonic regression. In Puri, M. L., editor, *Nonparametric Techniques in Statistical Inference*, pages 177–197. Cambridge University Press, London.

Chatterjee, S., Guntuboyina, A., Sen, B., et al. (2018). On matrix estimation under monotonicity constraints. *Bernoulli*, 24(2):1072–1100.

Dette, H. and Scheder, R. (2006). Strictly monotone and smooth nonparametric regression for two or more variables. *Canadian Journal of Statistics*, 34(4):535–561.

Dhar, S. S. (2016). Trimmed mean isotonic regression. *Scand. J. Stat*, 43:202–212.

Groeneboom, P. (1985). Estimating a monotone density. In Lucien, M. L. and Olshen, R. A., editors, *Proceeding of the Berkeley Conference in Honor of Jezry Neyman and Jack Kiefer*, volume II.

Groeneboom, P. (1989). Brownian motion with a parabolic drift and airy functions. *Probability theory and related fields*, 81(1):79–109.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer.

Groeneboom, P. and Wellner, J. A. (2001). Computing chernoff's distribution. *Journal of Computational and Graphical Statistics*, 10(2):388–400.

Han, Q., Wang, T., Chatterjee, S., and Samworth, R. J. (2017). Isotonic regression in general dimensions. *arXiv preprint arXiv:1708.09468*.

Hoffmann, L. (2009). *Multivariate isotonic regression and its algorithms*. PhD thesis, Wichita State University.

Luss, R., Rosset, S., Shahar, M., et al. (2012). Efficient regularized isotonic regression with application to gene–gene interaction search. *The Annals of Applied Statistics*, 6(1):253–283.

Makowski, G. G. (1977). Consistency of an estimator of doubly nondecreasing regression functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 39(4):263–268.

Moolchan, E. T., Hudson, D. L., Schroeder, J. R., and Sehnert, S. S. (2004). Heart rate and blood pressure responses to tobacco smoking among african-american adolescents. *Journal of the National Medical Association*, 96(6):767.

Paranjape, S. and Park, C. (1973). Laws of iterated logarithm of multiparameter wiener processes. *Journal of Multivariate Analysis*, 3(1):132–136.

Resnick, S. I. (1999). *A Probability Path*. Birkhäuser, Boston.

Robertson, T. and Wright, F. (1973). Multiple isotonic median regression. *The Annals of Statistics*, pages 422–432.

Robertson, T., Wright, F., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Siri, W. E. et al. (1956). The gross composition of the body. *Adv Biol Med Phys*, 4(239-279):513.

Wang, Y. and Huang, J. (2002). Limiting distribution for monotone median regression. *Journal of Statistical Planning and Inference*, 107:281–287.

A Appendix: Technical Details

A.1 Properties of d-convex function and d-GCM

Lemma A.1. *If $G \in C_I$, then $\partial_r G$ is non-decreasing in each co-ordinate.*

Proof. This follows from the convexity of G, noting that convex functions are almost everywhere differentiable and at the points of non-differentiability $\partial_\ell G(x_1, \ldots, x_d) < \partial_r G(x_1, \ldots, x_d)$. □

Lemma A.2. *Let $G_n \in \mathcal{C}_I$ for all $n \in \mathbb{N}$ and $\lim\limits_{n\to\infty} G_n(x_1, \ldots, x_d) = G(x_1, \ldots, x_d)$ uniformly on a ball $B((x_1, \ldots, x_d), \epsilon)$, then*

$$\partial_\ell G(x_1, \ldots, x_d) \leq \liminf_{n\to\infty} \partial_\ell G_n(x_1, \ldots, x_d) \leq \limsup_{n\to\infty} \partial_r G_n(x_1, \ldots, x_d) \leq \partial_r G(x_1, \ldots, x_d).$$

Proof. Since $G_n \in \mathcal{C}_I$, for $0 < h < \epsilon$, we have.

$$\begin{aligned}\frac{1}{h^d}\int_{x_1-h}^{x_1}\cdots\int_{x_d-h}^{x_d} dG_n(u_1, \ldots, u_d) &= \int_{x_1-h}^{x_1}\cdots\int_{x_d-h}^{x_d} \partial_\ell G_n(u_1, \ldots, u_d)du_1 \ldots du_d \\ &\leq \partial_\ell G_n(x_1, \ldots, x_d) \leq \ \partial_r G_n(x_1, \ldots, x_d) \\ &\leq \int_{x_1}^{x_1+h}\cdots\int_{x_d}^{x_d+h} \partial_r G_n(u_1, \ldots, u_d)du_1 \ldots du_d \\ &= \frac{1}{h^d}\int_{x_1}^{x_1+h}\cdots\int_{x_d}^{x_d+h} dG_n(u_1, \ldots, u_d).\end{aligned}$$

Letting $n \to \infty$ and $h \downarrow 0$ gives the desired result. □

Lemma A.3. *Let G be a d-convex function on $I \in \mathbb{R}^d$ and $I_d := \{x_d \in \mathbb{R} : (x_1, \ldots, x_d) \in I\}$. Then for any fixed $x_{d,0} \in I_d$, the map $(x_1, \ldots, x_{d-1}) \mapsto G(x_1, \ldots, x_{d-1}, x_{d,0})$ is a $(d-1)$-convex function.*

Proof. Note that for $h > 0$, we have $G_{h,\ell}(x_1, \ldots, x_{d-1}) := \int_{x_{d,0}-h}^{x_{d,0}} \partial_\ell G(x_1, \ldots, x_d)dx_d$ is a non-decreasing in all $(d-1)$ coordinates. The result then follows by letting $h \to 0$. □

Lemma A.4. *For any real-valued function S on $I \subset \mathbb{R}^d$, $T_I(S) \in \mathcal{C}_I$.*

Proof. Let $l_1 < l_2$ be two real numbers. Without loss of generality, assume that $T_I(S)(x_1, \ldots, x_d)$ is differentiable for $x_1 \in (l_1, l_2)$. We can find sequences $\{G_n^1\}, \{G_n^2\} \subset \mathcal{C}_I$ such that $G_n^1(l_1, x_2, \ldots, x_d) \uparrow T_I(S)(l_1, x_2, \ldots, x_d)$ and $G_n^2(l_2, x_2, \ldots, x_d) \uparrow T_I(S)(l_2, x_2, \ldots, x_d)$ as $n \to \infty$. Define $G_n(x_1, \ldots, x_d) = \max\{G_n^1(x_1, \ldots, x_d), G_n^2(x_1, \ldots, x_d)\}$. Note that $G_n(l_i, x_2, \ldots, x_d) \to T_I(S)(l_i, x_2, \ldots, x_d)$ for $i = 1, 2$ as $n \to \infty$ and $G_n \in \mathcal{C}_I$. Then the result follows from Lemma A.2. □

Lemma A.5. *Suppose $S : I \mapsto \mathbb{R}$ is a continuous function on an interval $I \subseteq \mathbb{R}^d$. If $T_I(S)$ and S do not have any touch point in an interval $J \subseteq I$, then $T_I(S)$ is linear at each coordinate on J.*

Proof. Let $J = [x_{1,l}, x_{1,u}] \times \cdots \times [x_{d,l}, x_{d,u}]$, and suppose that the map $x_1 \mapsto T_I(x_1, x_{2,0}, \ldots, x_{d,0})$ is not linear on $[x_{1,l}, x_{1,u}]$, for some $x_{k,0} \in [x_{k,l}, x_{k,u}]$ when $k = 2, \ldots, d$. We will construct a d-convex function G which is a minorant of S on I and is greater than $T_I(S)$.

To this end, let $L(x_1,\dots,x_d)$ be the line joining $T_I(S)(x_{1,l},x_2,\dots,x_d)$ and $T_I(S)(x_{1,u},x_2,\dots,x_d)$, i.e.,

$$L(x_1,\dots,x_d) := \frac{T_I(S)(x_{1,l},x_2,\dots,x_d)(x_{1,u}-x_1)+T_I(S)(x_{1,u},x_2,\dots,x_d)(x_1-x_{1,l})}{x_{1,u}-x_{1,l}}.$$

Note that $L(x_1,\dots,x_d) \geq T_I(S)(x_1,\dots,x_d)$ for all $(x_1,\dots,x_d) \in J$ and $L(x_1,x_{2,0},\dots,x_{d,0}) > T_I(S)(x_1,x_{2,0},\dots,x_{d,0})$ for at least one value of x_1 in $(x_{1,l},x_{1,u})$. Introduce the notation $T_I(S)(x_{1,k},.) =: T^k(.)$ for $k=l,u$. We consider the following two cases separately.

Case 1: $L(x_1,\dots,x_d) \leq S(x_1,\dots,x_d)$ for all $(x_1,\dots,x_d) \in J$. In this case, define

$$G(x_1,\dots,x_d) := T_I(S)(x_1,\dots,x_d)\mathbf{1}(x_1 \notin (x_{1,l},x_{1,u})) + L(x_1,\dots,x_d)\mathbf{1}(x_1 \in (x_{1,l},x_{1,u})).$$

Note that $G \in \mathcal{C}_I$, as for $x_1 \notin (x_{1,l},x_{1,u})$, $G \equiv T_I(S)$ and for $x_1 \in (x_{1,l},x_{1,u})$,

$$\begin{aligned}\partial_\ell G(x_1,\dots,x_d) =& \frac{1}{x_{1,u}-x_{1,l}}\left(\partial_\ell T^u(x_2,\dots,x_d) - \partial_\ell T^l(x_2,\dots,x_d)\right)\\ =& \frac{1}{x_u-x_l}\int_{x_l}^{x_u} \partial_\ell T_I(S)(x_1,\dots,x_d)dx_1.\end{aligned}$$

By construction of $T_I(S)$, the last quantity lies in $(\partial_\ell T_I(S)(x_{1,l},x_2,\dots,x_d), \partial_\ell T_I(S)(x_{1,u},x_2,\dots,x_d))$ and non-decreasing in x_k for $k=2,\dots,d$. Therefore, G is a d-convex minorant of S and $G(x_1,\dots,x_d) > T_I(S)(x_1,\dots,x_d)$ for at least one point by our assertion.

Case 2: $L(x_1,\dots,x_d) > S(x_1,\dots,x_d)$ for at least one $(x_1,\dots,x_d) \in J$. We define D as the distance between L and S, i.e., $D(x_1,\dots,x_d) = L(x_1,\dots,x_d) - S(x_1,\dots,x_d)$. Note that $D_m = D(x_{1,m},\dots,x_{d,m}) := \sup_J D(x_1,\dots,x_d) > 0$ by our assertion. We further define $L_2(x_1,\dots,x_d) = L(x_1,\dots,x_d) - D_m$. Note that for $x_1 \in [x_{1,l},x_{1,u}]$, we have

$$L_2(x_1,\dots,x_d) \leq L(x_1,\dots,x_d) - D(x_1,\dots,x_d) = S(x_1,\dots,x_d).$$

Moreover, $L_2(x_{1,u},x_2,\dots,x_d) < L(x_{1,u},x_2,\dots,x_d) = T_I(S)(x_{1,u},x_2,\dots,x_d)$, and for all $x_2,\dots,x_d$, and L_2 is linear in x_1 with slope $(T_I(S)(x_{1,u},x_2,\dots,x_d) - T_I(S)(x_{1,l},x_2,\dots,x_d))$ which is bounded above by right slope of $x_1 \mapsto T_I(S)(x_1,x_2,\dots,x_d)$ at $x_{1,u}$, due to convexity. Therefore, for $x_1 > x_{1,u}$, we have $L_2(x_1,\dots,x_d) < T_I(S)(x_1,\dots,x_d) \leq S(x_1,\dots,x_d)$. Similarly, we can argue that $L_2(x_1,\dots,x_d) < S(x_1,\dots,x_d)$ for $x_1 < x_{1,l}$. Therefore, L_2 is a minorant of S on I.

Indeed L_2 is d-convex follows from the fact that $L \in \mathcal{C}_I$ and by construction, $L_2(x_{1,m},\dots,x_{d,m}) = S(x_{1,m},\dots,x_{d,m}) > T_I(S)(x_{1,m},\dots,x_{d,m})$, as $T_I(S)$ and S do not have any touch point in J by the assertion of the lemma. Now we define

$$G(x,y) = \max(L_2(x,y), T_I(S)(x,y)).$$

The function G is d-convex by Lemma A.4, and it is a minorant of S, as both L_2 and $T_I(S)$ lie below S. Further $G \geq T_I(S)$ on I with $G(x_{1,m}, \dots, x_{d,m}) = L_2(x_{1,m}, \dots, x_{d,m}) > T_I(S)(x_{1,m}, \dots, x_{d,m})$.

Therefore, the linearity of $T_I(S)$ in x_1 must hold under the assumptions of the Lemma. The linearity of other coordinates can be shown similarly. □

Lemma A.6. *For any compact set $K \subset \mathbb{R}^d$, the map $T_K : C(\mathbb{R}^d) \to C(K)$ is continuous.*

Proof. Suppose for two continuous real valued functions f_1, f_2 on $\mathbb{R}^d$, we have

$$\sup_I |f_1(s_1, \dots, s_d) - f_2(s_1, \dots, s_d)| < \epsilon \tag{2.11}$$

for all compact set $I \subset \mathbb{R}^2$. Using the fact that $T_K(f + a) = T_K(f) + a$ for any real constant a and $T_K(f) \leq T_K(g)$ provided $f \leq g$ we write,

$$\begin{aligned} T_K(f_2) - \sup_K |f_1(s_1, \dots, s_d) - f_2(s_1, \dots, s_d)| =& T_K(f_2 - \sup_K |f_1(s_1, \dots, s_d) - f_2(s_1, \dots, s_d)|) \\ \leq& T_K(f_2 + f_1 - f_2) = T_K(f_1) \\ \leq& T_K(f_2 + \sup_K |f_1(s_1, \dots, s_d) - f_2(s_1, \dots, s_d)|) \\ =& T_K(f_2) + \sup_K |f_1(s_1, \dots, s_d) - f_2(s_1, \dots, s_d)| \end{aligned}$$

Under (2.11) we have for any compact set $I \subset K$

$$\sup_I |T_K(f_1)(s_1, \dots, s_d) - T_K(f_2)(s_1, \dots, s_d)| < \sup_K |f_1(s_1, \dots, s_d) - f_2(s_1, \dots, s_d)| < \epsilon.$$

Hence the result. □

Lemma A.7. *Let S be a continuous real-valued function in $\mathbb{R}^d$. The functions $x_k \mapsto T_{[-c,c]^d}(S)'(x_1, \dots, x_d) - T(S)'(x_1, \dots, x_d)$ are non-decreasing on $[-c, c]$ for each coordinate $k = 1, \dots, d$.*

Proof. Let $\{J_l\}$ be a sequence of open rectangles on $(-c, c)^d$ such that their union covers $(-c, c)^d$. Without loss of generality, either J_l does not have any point of touch for $T_{[-c,c]^d}(S)$ and $T(S)$ or have a simply connected set of touch points Ω_{J_l}. (Ω_{J_l} is of the form $I_1 \times \cdots \times I_d$, where I_k's are either a singleton set or an interval in $\mathbb{R}$). If Ω_{J_l} is empty, by Lemma A.5, $T(S)$ is co-ordinate wise linear in the interval, so $T(S)'$ is constant and consequently, $T_{[-c,c]^d}(S)' - T(S)'$ is co-ordinate wise non-decreasing. If Ω_{J_l} is non-empty, as $T(S)$ is a convex minorant on $[-c, c]^d$, we have $T_{[-c,c]^d}(S) \geq T(S)$. Therefore, $T_{[-c,c]^d}(S)' - T(S)'$ is co-ordinate wise non-decreasing. □

A.2 Proof of Theorem 2.2

We start by noting that Theorem 2.2 holds for $d = 1$ in view of the assertion in Theorem 1.2.1 from Robertson et al. (1988). We use method of induction to prove the result for general d. Before starting the main proof, we state one additional result characterizing the d-GCM of the partial sum process S_n as described in Section 2.1.

To this end, for any fixed $x_{d,0} \in [0,1]$, introduce the function $S_n^{x_{d,0}} : [0,1]^{d-1} \mapsto \mathbb{R}$ as the restriction of S_n at $x_d = x_{d,0}$, i.e., $S_n^{x_{d,0}}(x_1, \ldots, x_{d-1}) = S_n(x_1, \ldots, x_{d-1}, x_{d,0})$.

Lemma A.8. *The map* $(x_1, \ldots, x_{d-1}) \mapsto T_{[0,1]^d}(S_n)(x_1, \ldots, x_{d-1}, 1)$ *is the* $(d-1)$*-GCM of* S_n^1*.*

Proof. Suppose that the assertion is not true. For notational convenience, we denote the map $(x_1, \ldots, x_{d-1}) \mapsto T_{[0,1]^d}(S_n)(x_1, \ldots, x_{d-1}, 1)$ by G. Let G^1 be the GCM of S_n^1. Note that by Lemma A.3, G is a $(d-1)$-convex minorant of S_n^1 and therefore, we have $G^1(x_1, \ldots, x_{d-1}) \geq G(x_1, \ldots, x_{d-1})$ and $G^1(x_1, \ldots, x_{d-1}) > G(x_1, \ldots, x_{d-1})$ for some point in $[0,1]^{d-1}$. Define $G_1(x_1, \ldots, x_d) = G(x_1, \ldots, x_d)$ for $(x_1, \ldots, x_d) \in [0,1]^{d-1} \times [0, x_{d,n_d-1}]$. For every x_d, in the interval $(x_{d,n_d-1}, 1]$, join $T_{[0,1]^d}(S_n)(x_1, \ldots, x_{d-1}, x_{d,n_d-1})$ and $G^1(x_1, \ldots, x_{d-1})$ linearly. More precisely, for $(x_1, \ldots, x_d) \in [0,1]^{d-1} \times (x_{d,n_d-1}, 1]$,

$$G_1(x_1, \ldots, x_d) = n_d[T_{[0,1]^d}(S_n)(x_1, \ldots, x_{d-1}, x_{d,n_d-1})(1 - x_d) + G^1(x_1, \ldots, x_{d-1})(x_d - x_{d,n_d-1})].$$

Claim 1. G_1 *is in* $\mathcal{C}_{[0,1]^d}$*.*

Proof: Similar to proof of $G \in \mathcal{C}_I$ in Lemma A.5.

Claim 2. $G_1(x_1, \ldots, x_d) \leq S_n(x_1, \ldots, x_d)$ *for all* $(x_1, \ldots, x_d) \in [0,1]^d$*.*

Proof: This is trivially true if $x_d \leq x_{d,n_d-1}$. Moreover, for every $(x_1, \ldots, x_{d-1}) \in [0,1]^{d-1}$, we have

$$G_1(x_1, \ldots, x_{d-1}, x_{d,n_d-1}) = T_{[0,1]^d}(S_n)(x_1, \ldots, x_{d-1}, x_{d,n_d-1}) \leq S_n(x_1, \ldots, x_{d-1}, x_{d,n_d-1}),$$

$$G_1(x_1, \ldots, x_{d-1}, 1) = G^1(x_1, \ldots, x_{d-1}) \leq S_n^1(x_1, \ldots, x_{d-1}) = S_n(x_1, \ldots, x_{d-1}, 1)$$

and both $x_d \mapsto S_n(x_1, \ldots, x_d)$ and $x_d \mapsto G_1(x_1, \ldots, x_d)$ are linear in between $(x_{d,n_d-1}, 1]$. Hence, $G_1(x_1, \ldots, x_d) \leq S_n(x_1, \ldots, x_d)$ for $x_d \in (x_{d,n_d-1}, 1]$.

Claim 3. $G_1(x_1, \ldots, x_d) \geq T_{[0,1]^d}(S_n)(x_1, \ldots, x_d)$ *for all* $(x_1, \ldots, x_d) \in [0,1]^d$*.*

Proof: Note that the map $x_d \mapsto T_{[0,1]^d}(S_n)(x_1, \ldots, x_d)$ is indeed convex, and as a consequence for $x_d \in (x_{d,n_d-1}, 1]$, the function $T_{[0,1]^d}(S_n)(x_1, \ldots, x_d)$ lies below the line joining $T_{[0,1]^d}(S_n)(x_1, \ldots, x_{d-1}, x_{d,n_d-1})$ and $T_{[0,1]^d}(S_n)(x_1, \ldots, x_{d-1}, 1)$, which is further below G_1 because of the fact $G^1(x_1, \ldots, x_{d-1}) \geq T_{[0,1]^d}(S_n)(x_1, \ldots, x_{d-1}, 1)$ For $x_d \leq x_{d,n_d-1}$, the claim is trivially true.

So G_1 constructed this way is a d-convex minorant of S_n such that $G_1 \geq T_{[0,1]^d}(S_n)$, with strict inequality at at least one point because of our assertion. This is a contradiction to the fact that $T_{[0,1]^d}(S_n)$ is GCM of S_n. □

Now back to the proof of Theorem 2.2, let $g(x_1, \ldots, x_d) := \partial_\ell G(x_1, \ldots, x_d)$. We will show that for any co-ordinate wise non-decreasing real valued function f on $[0,1]^d$. We now have

$$\begin{aligned}\sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} (Z_{i_1 \ldots i_d} - f(x_{1,i_1}, \ldots, x_{d,i_d}))^2 \geq & \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} (Z_{i_1 \ldots i_d} - g(x_{1,i_1}, \ldots, x_{d,i_d}))^2 \\ & + \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} (g(x_{1,i_1}, \ldots, x_{d,i_d}) - f(x_{1,i_1}, \ldots, x_{d,i_d}))^2.\end{aligned} \tag{2.12}$$

To show (2.12), it is enough to show that

$$\frac{1}{n} \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} (Z_{i_1 \ldots i_d} - g(x_{1,i_1}, \ldots, x_{d,i_d}))(g(x_{1,i_1}, \ldots, x_{d,i_d}) - f(x_{1,i_1}, \ldots, x_{d,i_d})) \geq 0. \tag{2.13}$$

To this end, we introduce some notations. For all i_k and $k \in \{1, \ldots, d\}$, define

$$S_{n,i_d}(x_{1,i_1}, \ldots, x_{d-1,i_{d-1}}) := \frac{1}{n} \sum_{l=1}^{i_d} Z_{i_1 \ldots i_{d-1} l},$$

$$G_{i_d}(x_{1,i_1}, \ldots, x_{d-1,i_{d-1}}) := \frac{1}{n} \sum_{l=1}^{i_d} g(x_{1,i_1} \ldots x_{d-1,i_{d-1}}, x_{d,l}).$$

With this notation, we have

$$\sum_{l_1=1}^{i_1} \cdots \sum_{l_{d-1}=1}^{i_{d-1}} S_{n,i_d}(x_{1,l_1}, \ldots, x_{d-1,l_{d-1}}) = S_n(x_{1,i_1}, \ldots, x_{d,i_d}),$$

$$\sum_{l_1=1}^{i_1} \cdots \sum_{l_{d-1}=1}^{i_{d-1}} G_{i_d}(x_{1,l_1}, \ldots, x_{d-1,l_{d-1}}) = G(x_{1,i_1}, \ldots, x_{d,i_d}).$$

Using Abel's partial summation formula on the left hand side of (2.13), we get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d}(Z_{i_1\dots i_d}-g(x_{1,i_1},\dots,x_{d,i_d}))(g(x_{1,i_1},\dots,x_{d,i_d})-f(x_{1,i_1},\dots,x_{d,i_d}))\\
=&\sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d}(f(x_{1,i_1},\dots,x_{d,i_d})-f(x_{1,i_1},\dots,x_{d,i_d-1}))(S_{n,i_d-1}(x_{1,i_1},\dots,x_{d-1,i_{d-1}})-G_{i_d-1}(x_{1,i_1},\dots,x_{d-1,i_{d-1}}))\\
&+\sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d}(g(x_{1,i_1},\dots,x_{d,i_d-1})-g(x_{1,i_1},\dots,x_{d,i_d}))(S_{n,i_d-1}(x_{1,i_1},\dots,x_{d-1,i_{d-1}})-G_{i_d-1}(x_{1,i_1},\dots,x_{d-1,i_{d-1}}))\\
&+\sum_{i_1=1}^{n_1}\cdots\sum_{i_{d-1}=1}^{n_{d-1}}(g(x_{1,i_1},\dots,x_{d-1,i_{d-1}},x_{d,n_d})-f(x_{1,i_1},\dots,x_{d-1,i_{d-1}},x_{d,n_d}))(S_{n,n_d}(x_{1,i_1},\dots,x_{d-1,i_{d-1}})\\
&\qquad\qquad\qquad\qquad\qquad\qquad-G_{n_d}(x_{1,i_1},\dots,x_{d-1,i_{d-1}}))\\
=&I+II+III
\end{aligned}
$$

Treating each of these terms separately,

$$
\begin{aligned}
I\ge&\sum_{i_d=1}^{n_d}\min_{i_d}(f(x_{1,i_1},\dots,x_{d,i_d})-f(x_{1,i_1},\dots,x_{d,i_d-1}))\sum_{i_1=1}^{n_1}\cdots\sum_{i_{d-1}=1}^{n_{d-1}}(S_{n,i_d-1}(x_{1,i_1},\dots,x_{d-1,i_{d-1}})-G_{i_d-1}(x_{1,i_1},\dots,x_{d-1,i_{d-1}}))\\
=&\sum_{i_d=1}^{n_d}\min_{i_d}(f(x_{1,i_1},\dots,x_{d,i_d})-f(x_{1,i_1},\dots,x_{d,i_d-1}))(S_n(x_{1,n_1},\dots,x_{d-1,n_{d-1}},x_{d,i_d})-G(x_{1,n_1},\dots,x_{d-1,n_{d-1}},x_{d,i_d})).
\end{aligned}
$$

Every terms inside the last summation is non-negative and hence, $I\ge 0$. Using Abel's formula again and treating the first $(d-1)$ summation as one sum over $i=1,\dots,n_1+\cdots+n_{d-1}$, we write

$$
\begin{aligned}
II=&\sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d}(g(x_{1,i_1},\dots,x_{d,i_d-1})-g(x_{1,i_1},\dots,x_{d,i_d}))(S_{n,i_d-1}(x_1,..,x_{i_{d-1}})-G_{i_d-1}(x_1,..,x_{i_{d-1}}))\\
=&\sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d}(g(x_{1,i_1},...,x_{d,i_d})-g(x_{1,i_1-1},...,x_{d,i_d}))(S_n(x_{1,i_1},...,x_{d-1,i_{d-1}},x_{d,i_d-1})-G(x_{1,i_1},...,x_{d-1,i_{d-1}},x_{d,i_d-1}))\\
&+\sum_{i_1=1}^{n_1}...\sum_{i_d=1}^{n_d}(g(x_{1,i_1-1},...,x_{d,i_d-1})-g(x_{1,i_1},...,x_{d,i_d-1}))(S_n(x_{1,i_1},...,x_{d-1,i_{d-1}},x_{d,i_d-1})-G(x_{1,i_1},...,x_{d-1,i_{d-1}},x_{d,i_d-1}))\\
&+\sum_{i_2=1}^{n_2}...\sum_{i_d=1}^{n_d}(g(x_{1,n_1},x_{2,i_2}...,x_{d,i_d})-g(x_{1,n_1},x_{2,i_2-1},...,x_{d,i_d}))(S_n(x_{1,n_1},x_{2,i_2}...,x_{d,i_d-1})-G(x_{1,n_1},x_{2,i_2},...,x_{d,i_d-1}))\\
&+\sum_{i_2=1}^{n_2}\cdots\sum_{i_d=1}^{n_d}(g(x_{1,n_1},x_{2,i_2-1},\dots,x_{d,i_d-1})-g(x_{1,n_1},x_{2,i_2},\dots,x_{d,i_d-1}))(S_n(x_{1,n_1},\dots,x_{d,i_d-1})-G(x_{1,n_1},\dots,x_{d,i_d-1}))\\
&\quad\vdots\\
&+\sum_{i_d=1}^{n_d}(g(x_{1,n_1},\dots,x_{d,i_d-1})-g(x_{1,n_1},\dots,x_{d,i_d})(S_n(x_{1,n_1},\dots,x_{d-1,n_{d-1}},x_{d,i_d-1})-G(x_{1,n_1},\dots,x_{d-1,n_{d-1}},x_{d,i_d-1}))
\end{aligned}
$$

The odd-numbered terms (first, third etc), except the last term in the above equation is non-negative because g is co-ordinate wise non-decreasing, and G is a minorant of S_n. For the second term, note that is

$G(x_{1,i_1},\dots,x_{d,i_d}) < S_n(x_{1,i_1},\dots,x_{d,i_d})$, then by Lemma A.5, we have $g(x_{1,i_1},\dots,x_{d,i_d}) = g(x_{1,i_1-1},\dots,x_{d,i_d})$. Therefore, this term is zero. All the even numbered terms and the last term can similarly be shown to be zero.

The treatment of the third term is exactly similar to II. Similar arguments as above along with the observation that $G(x_{1,n_1},\dots,x_{d,n_d}) = S_n(x_{1,n_1},\dots,x_{d,n_d})$ by Lemma A.8 yield that $III \geq 0$. This shows (2.13) holds and hence, we have the result.

A.3 Technical Details for Section 2.2

In this section, we present technical details necessary to establish the asymptotic properties of the isotonic regression estimator. The crucial element for this is Proposition 2.6. The arguments are essentially an extension of the proof of Theorem A.1 from Anevski and Hössjer (2006).

Proof of Proposition 2.6: We will prove the second part of the Proposition, and the proof of the first part is similar. For readability, we present the proof for $d = 2$. The general case can be proven similarly with some additional notational complexity.

Let $K = [-1,1]^2$, and $\delta > 0$ be arbitrary. Consider the set,

$$A(n,M,\kappa,\tau) = \left\{ \sup_{s_1,s_2 \in K} |\mathbb{W}_n(s_1,s_2)| \leq M \right\} \bigcup \left\{ \inf_{|s_1|>\tau,|s_2|>\tau} (\mathbb{W}_n(s_1,s_2) - \kappa|s_1 s_2|) > 0 \right\}.$$

As $\mathbb{W}_n$ converges in distribution $\mathbb{W}$ on $C(I)$, and given compact set K and $\delta > 0$, we can find M, such that $\mathbb{P}\left(\sup_{s_1,s_2\in K} |\mathbb{W}(s_1,s_2)| > M\right) < \delta/2$, we have

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s_1,s_2\in K} |\mathbb{W}_n(s_1,s_2)| > M \right) < \delta.$$

Also by properties of Brownian motion as $|s_1|$ and $|s_2|$ approach infinity, we have $\mathbb{B}(s_1,s_2)/|s_1 s_2| \leq \mathbb{B}(s_1,s_2)/|s_1 + s_2| \to 0$ with high probability. Therefore, $\mathbb{W}(s_1,s_2)/|s_1 s_2| \to \infty$ as $|s_1|, |s_2| \to \infty$ for almost every sample path. Hence, for sufficiently large τ (for all $\tau \geq \tau(\delta,\kappa)$), we have

$$\limsup_{n\to\infty} \mathbb{P}\left(\inf_{|s_1|>\tau,|s_2|>\tau} (\mathbb{W}_n(s_1,s_2) - \kappa|s_1 s_2|) \leq 0 \right) < \delta.$$

Therefore, we can find M and τ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(A(n,M,\kappa,\tau)^c\right) < 2\delta. \tag{2.14}$$

Define sets

$$B(n,I,c,\epsilon) := \left\{ \sup_{I} |T_c(\mathbb{W}_n)'(.) - T_{I_n}(\mathbb{W}_n)'(.)| < \epsilon \right\}.$$

By Lemma A.7, for any $I \subset [-\tau, \tau]^2$ with $\tau < c$, we have

$$\bigcap_{i=1}^{4} B(n, I_{\tau,i}, c, \epsilon) \subset B(n, I, c, \epsilon),$$

where $I_{\tau,1} = \{(-\tau, -\tau)\}, I_{\tau,3} = \{(\tau, -\tau)\}, I_{\tau,3} = \{(-\tau, \tau)\}, I_{\tau,4} = \{(\tau, \tau)\}$. Next, we show that for given $\delta > 0$ and large enough c, for $i = 1, \ldots, 4$,

$$\limsup_{n\to\infty} \mathbb{P}\left(B(n, I_{\tau,i}, c, \epsilon)^c \cap A(n, M, \kappa, \tau)\right) < \delta. \tag{2.15}$$

We will show (2.15) for $i = 4$, other cases can be tackled similarly. Without loss of generality, τ can be chosen so large that $\tau > M/\kappa$, and n is large enough so that $[-c, c]^2 \subset I_n$. Then on $A(n, M, \kappa, \tau)$, we have

$$\inf_{|s_1|>\tau, |s_2|>\tau} \mathbb{W}_n(s_1, s_2) \geq M \geq \sup_{(s_1,s_2)\in K} \mathbb{W}_n(s_1, s_2).$$

Let $\Gamma_{n,c,\tau}(.,.)$ be the tangent plane of $T_c(\mathbb{W}_n)(s_1, s_2)$ at $(s_1, s_2) = (\tau, \tau)$ with slope $T_c(\mathbb{W}_n)'(\tau, \tau)$. Then, there can be three possible scenario. If $c > \tau$, we can have

1. $\Gamma_{n,c,\tau}(s_1, s_2) \leq \mathbb{W}_n(s_1, s_2)$ for all $(s_1, s_2) \notin [-c, c]^2$;

2. $\Gamma_{n,c,\tau}(s_1, s_2) \leq \mathbb{W}_n(s_1, s_2)$ for all (s_1, s_2) such that either $s_1 \leq -c$ or $s_2 \leq -c$.
   $\Gamma_{n,c,\tau}(s_1, s_2) > \mathbb{W}_n(s_1, s_2)$ for some (s_1, s_2) with $s_1 \geq c$ and $s_2 \geq c$.

3. $\Gamma_{n,c,\tau}(s_1, s_2) > \mathbb{W}_n(s_1, s_2)$ for some (s_1, s_2) with $s_1 \leq -c$ or $s_2 \leq -c$.
   $\Gamma_{n,c,\tau}(s_1, s_2) \leq \mathbb{W}_n(s_1, s_2)$ for all (s_1, s_2) such that either $s_1 \geq c$ or $s_2 \geq c$.

In case 1, $\Gamma_{n,c,\tau}$ is a convex minorant for $\mathbb{W}_n$. As $T_{I_n}(\mathbb{W}_n)$ is the GCM, we have $T_c(\mathbb{W}_n)(\tau, \tau) = \Gamma_{n,c,\tau}(\tau, \tau) \leq T_{I_n}(\mathbb{W}_n)(\tau, \tau)$. As $T_{I_n}(\mathbb{W}_n)$ is a doubly convex function on $[-c, c]^2$, we have $T_{I_n}(\mathbb{W}_n)(\tau, \tau) \leq T_c(\mathbb{W}_n)(\tau, \tau)$. Therefore, $T_c(\mathbb{W}_n)(\tau, \tau) = \Gamma_{n,c,\tau}(\tau, \tau) = T_{I_n}(\mathbb{W}_n)(\tau, \tau)$. As $\Gamma_{n,c,\tau}$ is a convex minorant, and $T_{I_n}(\mathbb{W}_n)$ is the GCM for $\mathbb{W}_n$ on I_n, this implies $T_c(\mathbb{W}_n)'(\tau, \tau) \leq T_{I_n}(\mathbb{W}_n)'(\tau, \tau)$. As $T_{I_n}(\mathbb{W}_n)$ is a convex minorant and $T_c(\mathbb{W}_n)$ is the GCM of $\mathbb{W}_n$ on $[-c, c]^2$, we have $T_c(\mathbb{W}_n)'(\tau, \tau) \geq T_{I_n}(\mathbb{W}_n)'(\tau, \tau)$. Therefore, in this case, we have $T_c(\mathbb{W}_n)'(\tau, \tau) = T_{I_n}(\mathbb{W}_n)'(\tau, \tau)$.

Under case 2, if $c > 2\tau$, on $A(n, M, \kappa, \tau)$, we have

$$\begin{aligned}
\inf_{s_1<-c, s_2<-c} \frac{T_{I_n}(\mathbb{W}_n)(\tau,\tau) - \mathbb{W}_n(s_1,s_2)}{(\tau - s_1)(\tau - s_2)} &\leq T_{I_n}(\mathbb{W}_n)'(\tau,\tau) \leq T_c(\mathbb{W}_n)'(\tau,\tau) \\
&\leq \inf_{\tau<s_i<c} \frac{\mathbb{W}_n(s_1,s_2) - T_c(\mathbb{W}_n)(\tau,\tau)}{(s_1-\tau)(s_2-\tau)} \\
&\leq \inf_{\tau<s_i<c} \frac{\mathbb{W}_n(s_1,s_2) - T_{I_n}(\mathbb{W}_n)(\tau,\tau)}{(s_1-\tau)(s_2-\tau)} \\
&\leq \left| \frac{\mathbb{W}_n(2\tau,2\tau) - T_{I_n}(\mathbb{W}_n)(\tau,\tau)}{\tau^2} \right|.
\end{aligned}$$

Let $\sup_{|s_1|\leq 2\tau, |s_1|\leq 2\tau} \mathbb{W}_n(s_1,s_2) \leq \widetilde{M}$ with probability at least $1 - \delta/4$. Then, the right hand side is bounded above by $2\widetilde{M}/\tau^2$, and we then have $T_{I_n}(\mathbb{W}_n)'(\tau,\tau) = T_c(\mathbb{W}_n)'(\tau,\tau)$ unless

$$\inf_{s_1<-c, s_2<-c} \frac{T_{I_n}(\mathbb{W}_n)(\tau,\tau) - \mathbb{W}_n(s_1,s_2)}{(\tau - s_1)(\tau - s_2)} \leq \frac{2\widetilde{M}}{\tau^2}. \tag{2.16}$$

Now, assume that (2.16) is true. For sufficiently large $\tilde{\tau} > \tau$, we write

$$\begin{aligned}
&|T_{I_n}(\mathbb{W}_n)'(\tau,\tau) - T_c(\mathbb{W}_n)'(\tau,\tau)| \\
\leq &\inf_{\tilde{\tau}\leq s_1,s_2\leq c} \frac{\mathbb{W}_n(s_1,s_2) - T_c(\mathbb{W}_n)(\tau,\tau)}{(s_1-\tau)(s_2-\tau)} - \inf_{\tilde{\tau}\leq s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2) - T_{I_n}(\mathbb{W}_n)(\tau,\tau)}{(s_1-\tau)(s_2-\tau)} \\
\leq &\frac{2\widetilde{M}}{(\tilde{\tau}-\tau)^2} - \inf_{\tilde{\tau}\leq s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2)}{(s_1-\tau)(s_2-\tau)} + \inf_{\tilde{\tau}\leq s_1,s_2\leq c} \frac{\mathbb{W}_n(s_1,s_2)}{(s_1-\tau)(s_2-\tau)} \\
\leq &\frac{2\widetilde{M}}{(\tilde{\tau}-\tau)^2} - \inf_{\tilde{\tau}\leq s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2)}{s_1 s_2} + \inf_{\tilde{\tau}\leq s_1,s_2\leq c} \frac{\mathbb{W}_n(s_1,s_2)}{(s_1-\tau)(s_2-\tau)} \\
\leq &\frac{2\widetilde{M}}{(\tilde{\tau}-\tau)^2} + \left(\inf_{\tilde{\tau}\leq s_1,s_2\leq c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1 s_2} - \inf_{\tilde{\tau}\leq s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2)}{s_1 s_2} \right) \\
&+ \left(\inf_{\tilde{\tau}\leq s_1,s_2\leq c} \frac{\mathbb{W}_n(s_1,s_2)}{(s_1-\tau)(s_2-\tau)} - \inf_{\tilde{\tau}\leq s_1,s_2\leq c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1 s_2} \right).
\end{aligned}$$

As $\lim_{s_1,s_2\to\infty} \mathbb{B}_n(s_1,s_2)/(as_1^2|s_2| + bs_2^2|s_1|) \xrightarrow{p} 0$ as $n \to \infty$ for any $a, b > 0$, given any $\epsilon' > 0$, for large enough n with very high probability, we have

$$\inf_{\tilde{\tau}\leq s_1,s_2\leq c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1 s_2} \leq \inf_{\tilde{\tau}\leq s_1,s_2\leq c} (1+\epsilon') \frac{|s_1| f_1(u_1,u_2) + |s_2| f_2(u_1,u_2)}{2} = \frac{(1+\epsilon')\tilde{\tau}}{2}(f_1(u_1,u_2) + f_2(u_1,u_2)),$$

and

$$\inf_{c\leq s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2)}{s_1 s_2} \geq \inf_{c\leq s_1,s_2} (1-\epsilon') \frac{|s_1| f_1(u_1,u_2) + |s_2| f_2(u_1,u_2)}{2} = \frac{(1-\epsilon')c}{2}(f_1(u_1,u_2) + f_2(u_1,u_2)).$$

Hence, for $c > \tilde{\tau}$, we in fact have with very high probability

$$\inf_{\tilde{\tau}\le s_1,s_2\le c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} - \inf_{c\le s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} < (f_1(u_1,u_2)+f_2(u_1,u_2))\tilde{\tau}\epsilon'.$$

Therefore, with appropriate choice of ϵ', we can write with probability at least $1-\delta/4$,

$$\inf_{\tilde{\tau}\le s_1,s_2\le c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} - \inf_{c\le s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} < \epsilon/3,$$

and consequently,

$$\inf_{\tilde{\tau}\le s_1,s_2\le c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} - \inf_{\tilde{\tau}\le s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} < \epsilon/3.$$

Also, in this situation

$$\begin{aligned}\inf_{\tilde{\tau}\le s_1,s_2\le c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} &\le \inf_{c\le s_1,s_2} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} + \epsilon/3\\ &\le 2\frac{(c-\tau)^2}{c^2}\inf_{c\le s_1,s_2}\frac{\mathbb{W}_n(s_1,s_2)}{(s_1-\tau)(s_2-\tau)} + \epsilon/3\\ &\le 2\frac{(c-\tau)^2}{c^2}\left(\inf_{c\le s_1,s_2}\frac{\mathbb{W}_n(s_1,s_2)-T(\mathbb{W}_n)(\tau,\tau)}{(s_1-\tau)(s_2-\tau)} + \frac{\widetilde{M}}{(c-\tau)^2}\right) + \epsilon/3\\ &\le 2\frac{(c-\tau)^2}{c^2}\left(\frac{2\widetilde{M}}{\tau^2} + \frac{\widetilde{M}}{(c-\tau)^2}\right) + \epsilon/3 \le \frac{6\widetilde{M}}{\tau^2} + \epsilon/3,\end{aligned}$$

and hence,

$$\begin{aligned}\inf_{\tilde{\tau}\le s_1,s_2\le c} \frac{\mathbb{W}_n(s_1,s_2)}{(s_1-\tau)(s_2-\tau)} - \inf_{\tilde{\tau}\le s_1,s_2\le c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2} &\le \left(\frac{\tilde{\tau}^2}{(\tilde{\tau}-\tau)^2)-1}\right)\inf_{\tilde{\tau}\le s_1,s_2\le c} \frac{\mathbb{W}_n(s_1,s_2)}{s_1s_2}\\ &\le \frac{\tau(2\tilde{\tau}-\tau)}{(\tilde{\tau}-\tau)^2}\left(\frac{6\widetilde{M}}{\tau^2}+\epsilon/3\right).\end{aligned}$$

Therefore, for choice of large enough $\tilde{\tau}$, we have $|T_{I_n}(\mathbb{W}_n)'(\tau,\tau) - T_c(\mathbb{W}_n)'(\tau,\tau)| < \epsilon$ with probability at least $1-\delta/4$ under (2.16). So under case 2, the probability $\mathbb{P}\left(B(n,\{(\tau,\tau)\},c,\epsilon)^c \cap A(n,M,\kappa,\tau)\right) \le \delta/2$, and hence, we have (2.15). Combining (2.14) and (2.15), we have $\mathbb{P}\left(B(n,I,c,\epsilon)^c\right) < 6\delta$. As the choice of $\delta > 0$ is arbitrary, we have the desired result. □

Next we prove some properties of the process $\mathbb{W}$ appearing in the limit distribution.

Proposition A.9. *Let $\mathcal{Q}^C = \{Q_1 \times \cdots \times Q_d : Q_j$ is either $[0,C]$ or $[-C,0]\}$ and for any $Q \subset \mathbb{R}^d$,*

$$M_Q := \inf_{(s_1,\ldots,s_d)\in Q} \mathbb{W}(s_1,\ldots,s_d).$$

Then $\mathbb{P}(M_Q < 0) = 1$ for all $Q \in \mathcal{Q}^C$.

Proof. Let $Q = [0, C]^d$, then

$$\mathbb{P}(M_Q \geq 0) = \mathbb{P}(\mathbb{W}(s_1, \dots, s_d) + |s_1 \dots s_d|(a_1|s_1| + \dots a_d|s_d|) \geq 0, \forall C \geq s_1, \dots, s_d > 0)$$

for some $a_1, \dots, a_d > 0$. Using self-similarity of Brownian motion, we can write

$$\begin{aligned}
&\mathbb{P}(\mathbb{W}(s_1, \dots, s_d) + |s_1 \dots s_d|(a_1|s_1| + \dots a_d|s_d|) \geq 0, \forall\, C \geq s_1, \dots, s_d > 0) \\
=&\mathbb{P}(\kappa^{-d}\mathbb{W}(\kappa s_1, \dots, \kappa s_d) + |s_1 \dots s_d|(a_1|s_1| + \dots a_d|s_d|) \geq 0, \forall\, C \geq s_1, \dots, s_d > 0) \\
=&\mathbb{P}(\mathbb{W}(\tau_1, \dots, \tau_d) \geq -\kappa^{-1}|\tau_1 \dots \tau_d|(a_1|\tau_1| + \dots a_d|\tau_d|), \forall\, \kappa C \geq \tau_1, \dots, \tau_d > 0) \\
=&\mathbb{P}(A_\kappa) \equiv const.
\end{aligned}$$

Note that $A_\kappa \to \{\mathbb{W}(\tau_1, \dots, \tau_d) \geq 0, \forall \tau_1, \dots, \tau_d > 0\}$ as $\kappa \to \infty$. As the probability of A_κ does not depend on c, it is equal to $\mathbb{P}(\mathbb{W}(\tau_1, \dots, \tau_d) \geq 0, \forall \tau_1, \dots, \tau_d > 0)$. The last probability is zero by law of iterated logarithm for multivariate Wiener processes. (Theorem 3 and 4 from Paranjape and Park (1973)). Hence, the result is established. □

Corollary A.10. *With probability 1, $T_{[-c,c]^d}(\mathbb{W})$ and $\mathbb{W}$ do not touch at $(0, \dots, 0)$ for any $c > 0$.*

Proof. Let M_Q be as defined in Proposition A.9 and $M := \min_{Q \in \mathcal{Q}^c} M_Q$. Note that the function $G(s_1, \dots, s_d) \equiv M$ for $(s_1, \dots, s_d) \in \mathbb{R}^d$ is a convex minorant of $\mathbb{W}$ and touches $\mathbb{W}$ at some point say $(u_1, \dots, u_d)$. Therefore the GCM $T_{[-c,c]^d}(\mathbb{W})$ and $\mathbb{W}$ also have a touch point at $(u_1, u_2, \dots, u_d)$. Therefore if $T_{[-c,c]^d}(\mathbb{W})(0, \dots, 0) = 0$, by convexity it has to be positive on some $Q \in \mathcal{Q}^c$. This cannot happen if $M_Q < 0$ for all $Q \in \mathcal{Q}^c$. Hence, we have the desired result. □