Rank-based inference for the single-index model

Long Feng, Changliang Zou, Zhaojun Wang
School of Mathematical Sciences, Nankai University, Tianjin, China

ARTICLE INFO

Article history:
Received 26 July 2011
Received in revised form 25 November 2011
Accepted 25 November 2011
Available online 3 December 2011

MSC:
62G05
62G08
62G20

Keywords:
Asymptotic efficiency
Local polynomial regression
Outer product of gradients
Wilcoxon rank loss function

ABSTRACT

This paper is concerned with estimating the coefficients in single-index models. We develop a robust estimator, which combines the ideas of rank-based regression inference and outer product of gradients. Both asymptotic and numerical results show that the proposed procedure has better performance than the least-squares-based method when the errors deviate from normal.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Since introduced, the single-index model (SIM) has been widely applied in many scientific areas, such as economics, finance, politics, epidemiology, medical science, ecology and so on. It searches for a single linear combination of \( p \)-covariate \( X \) that can capture most information about the relation between response variable \( Y \) and covariate \( X \), thereby avoiding the “curse of dimensionality”. Specifically, the SIM can be written as

\[
Y = g(\beta_0^T X) + \varepsilon,
\]

where it is assumed that \( E(\varepsilon|X) = 0 \), \( g \) is an unknown link function and \( \beta_0 \) is an unknown unit vector with \( \|\beta_0\| = 1 \) for identification purposes (\( \| \cdot \| \) denotes the Euclidean norm).

Due to its flexibility and interpretability, in the past decades, estimation of the SIM has experienced rapid developments in both theory and methodology. In the literature, two most popular approaches are the average derivative estimation (ADE) (Powell et al., 1989; Härdle and Stoker, 1989) and the simultaneous minimization method (Härdle et al., 1993). An advantage of the ADE approach is that it allows estimating \( \beta \) directly. However, because the high-dimensional kernel estimation method is used, the estimation still suffers from the “curse of dimensionality”. Hristache et al. (2001) improved the ADE approach by lowering the dimension of the kernel gradually. The method of Härdle et al. (1993) is carried out by minimizing a least-squares criterion based on nonparametric estimation of the link \( g \) with respect to \( \beta \) and bandwidth \( h \). However, the minimization is not easy to implement since it relies on an optimization problem in a high-dimensional space. Xia et al.
(2002) proposed a new procedure, termed as minimum average conditional variance estimation (MAVE), to estimate the link function and the index simultaneously and gave a simple algorithm for the estimation. Asymptotic efficiency comparisons of the above methods have been made in Xia (2006). Other related development includes but not limited to Ichimura (1993), Horowitz and Härdle (1996), Carroll et al. (1997), Yu and Ruppert (2002), Huh and Park (2002), Yin and Cook (2005).

The estimation procedures in the aforementioned papers are mainly built on least-squares (LS) type method. Although the LS method is a successful and standard choice in SIM fitting, it may suffer when the errors follow a heavy-tailed distribution or in the presence of outliers. Thus, some efforts have been devoted to construct robust estimators for the SIM. Han (1987) proposed an estimator based on the rank correlation between the observed values and the values fitted by the model. Asymptotic theory for that estimator is revealed by Sherman (1993) and a generalization is given by Cavang and Sherman (1998). Unfortunately, this method requires the link function \(g\) to be strictly monotonic. In Section 2.1, we develop a robust estimating procedure which integrates the outer product of gradients method (OPG; Xia (2006)) and Wilcoxon rank-based regression inference (Hettmansperger and McKean, 2010). Under some mild conditions, the root-\(n\) consistency and asymptotic normality of the estimator are established in Section 2.2. Its asymptotic relative efficiency with respect to its LS-based counterpart is closely related to that of the signed-rank Wilcoxon test in comparison with the \(t\)-test. Simulation studies in Section 3 show that our procedure has better performance than the OPG estimator when the errors deviate from normal. Even in the most favorable case for the LS-based counterpart, i.e., normal distribution, our procedure does not lose much, which coincides with our theoretical analysis.

2. Methodology

2.1. Estimation via rank-based outer product of gradients

Let \(G(x) = E(Y|X = x)\). It is easy to see from model (1) that \(\nabla G(x) = g'(\beta^T x)\beta_0\). Therefore, the derivative of gradient of the regression function at any point has the same direction as \(\beta_0\). Based on this observation, Powell et al. (1989) and Härdle and Stoker (1989) proposed estimating \(\beta_0\) via \(E(\nabla G(X))\) by high-order kernel smoothers, which is known as the ADE method. Clearly, the estimation still suffers the “curse of dimensionality” because the high-dimensional kernel estimation method is used. To overcome this drawback, Xia (2006) novelly considered the outer product of the gradients, \(E(\nabla G(X)\nabla^T G(X))\). Note that

\[
E(\nabla G(X)\nabla^T G(X)) = E((g'(\beta_0^T X))^2)\beta_0\beta_0^T
\]

has only one nonzero eigenvalue. Therefore, index \(\beta_0\) is the eigenvector corresponding to the largest eigenvalue of \(E(\nabla G(X)\nabla^T G(X))\). Based on this idea, Xia (2006) proposed the so-called OPG estimator.

Suppose that \([X_1, Y_1, \ldots, X_n, Y_n]\) is a random sample from model (1). We locally approximate the link function by a Taylor expansion

\[
g(\beta^T X_i) \approx g(\beta^T X_0) + g'(\beta^T X_0)\beta^T (X_i - X_0),
\]

where \(X_0 = X_i - X_i\). Denote \(a_i = g(\beta^T X_0), b_i = g'(\beta^T X_0)\beta^T\). To estimate the gradients \(h_i\), Xia (2006) considered the local linear fitting in the form of the following minimization problem:

\[
\min_{a_i, b_i} \sum_{i=1}^{n} (Y_i - a_i - b_i^T X_0)^2 w_{il},
\]

where \(w_{il}\) is a weight depending on the distance between \(X_i\) and \(X_l\). This method is clearly a least-squares based method.

The key of our proposal is to construct an efficient estimate of \(b_i\) through a rank-based robust method. Motivated by the rank method in linear regression analysis (Jaekel, 1972; Aubuchon and Härdle, 1989), we propose a local Wilcoxon rank-based objective (loss) function which is defined as

\[
\sum_{i=1}^{n} \left( \frac{R(e_{il})}{n + 1} - \frac{1}{2} \right) e_{il} w_{il},
\]

where \(e_{il} = Y_i - a_i - b_i^T X_0\) and \(R(e_{il})\) denotes the rank of \(e_{il}\) among \(\{e_{i1}, \ldots, e_{il}\}\). Minimizing the loss function above is analogous to the least-squares procedure except that the Euclidean norm is substituted by a Wilcoxon-type rank norm. Moreover, the objective function (2) is nonnegative convex function and provides a robust measure of the dispersion of the residuals (cf., Theorems 2.5.1 in Hettmansperger and McKean (2010)). As demonstrated in the literature, the Wilcoxon-type rank norm has been shown to perform remarkably well in a wide variety of settings, e.g., see Leng (2010) for variable selection and Wang et al. (2009) for model estimation in varying coefficient model.

It is also well known that minimizing (2) is equivalent to minimizing the following loss function:

\[
Q_n(b_i) = \frac{2}{n(n - 1)} \sum_{i \neq j} |e_{il} - e_{jl}| w_{il} w_{jl},
\]
From this function it is apparent that the Wilcoxon type loss function cannot produce an estimator for the intercept since \( a_i \) is canceled out in \( e_i - e_i' \). This happens to be a unique feature of using this type of estimator in the present problem because only the gradient \( b_i \) is of our interest. It is worth noting that the above rank-based estimation essentially belongs to the framework of minimum distance estimation (Koul, 1985). See also Chapter 5 of Koul (2002) for more details. Also, the loss function \( Q_b \) is similar to the Eq. (5.3.5) in Koul (2002) or (2.1) in Sievers (1983) although those works focus on the linear regression model rather than the present SIM. Here we follow the idea of Xia et al. (2002) and Xia (2006) to refine the kernel loss function in the framework of minimum distance estimation (Koul, 1985). See also Chapter 5 of Koul (2002) for more details. Also, the asymptotic analysis.

Then, calculate \( \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \hat{b}_{i,b} \hat{b}_{i,b}^\top \), and the first eigenvector of \( \hat{\Sigma} \), \( \hat{b} \), is an estimator of \( \beta_0 \).

It is worth noting that the weight \( w_{il} = K_h(\beta_0^\top X_{il}) \) involves unknown parameter \( \beta \). Following Xia (2006), this issue can be handled by a repeated procedure. Say, set an initial estimate of \( \beta_0 \) and update \( w_{il} \) with the latest estimator of \( b_i \) until convergence. Hereafter the resulting estimator, \( \hat{\beta}_{\text{ROPG}} \), is termed as rank-based OPG estimator (ROPG).

In comparison with local linear fit, obtaining \( \hat{b}_{i,b} \) is more complicated because (4) has no close-form solution. However, this minimization problem can be solved by fitting a weighted \( L_1 \) regression on \( n(n-1)/2 \) pseudo observations \( \{X_{ij}, Y_i - Y_j\} \) with weights \( w_{ij} \). Such an algorithm is quite efficient and reliable in practice because subroutines or functions for carrying out weighted \( L_1 \) regression are available in most of statistical software packages. In our numerical studies, we use the function rq in the \( R \) package quantreg. The \( R \) code for implementing the proposed scheme are available from the authors upon request.

2.2. Theoretical properties

In this subsection, we establish the asymptotic properties of \( \hat{\beta}_{\text{ROPG}} \). We impose the following regularity conditions for asymptotic analysis.

(C1) (Design) The density function \( f(v) \) of \( \beta^\top X \) and its derivatives up to third order are bounded on \( \mathbb{R} \) for all \( \beta : \| \beta - \beta_0 \| < \delta \) where \( \delta > 0 \) is a constant, \( E|X|^6 < \infty \), and \( E|Y|^3 < \infty \).

(C2) (Link function) The conditional mean \( g_\beta(v) = E(Y|\beta^\top X = v) \), and its derivatives up to third order are bounded for all \( \beta : \| \beta - \beta_0 \| < \delta \) where \( \delta > 0 \).

(C3) (Kernel function) \( K(v) \) is a symmetric density function with finite moments of all orders and bounded derivatives. Its Fourier transform is absolutely integrable.

(C4) (Bandwidth) Bandwidth \( h \to 0, nh^2/\log n \to \infty \).

(C5) (Error distribution) The random error \( \epsilon \) has density function \( h(x) \) which has finite Fisher information and there is a constant \( M \) such that \( E(\epsilon^2) \leq M < \infty \).

Remark 2. Conditions (C1)–(C4) are essentially the same as those in Xia (2006). The assumption on the random errors in (C5) are the standard condition for rank analysis in multiple linear regression (Hettmansperger and McKean, 2010). These conditions are quite mild and can be satisfied in many practical situations.

We assume in the following context that the initial value \( \beta^\ast \in \Theta_n = \{ \beta : \| \beta - \beta_0 \| \leq \sigma_0 \} \}. These assumptions are feasible because such an initial estimate is obtainable using the existing methods, such as Xia (2006). Let \( \mu_\beta(x) = E(X|\beta^\top X = x) \), \( \nu_\beta(x) = x - \mu_\beta(x) \), \( \omega_\beta(x) = E(XX^\top|\beta^\top X = x) \), \( W_0(x) = v_{\beta_0}v_{\beta_0}^\top(x) \) and \( W_\beta(x) = w_{\beta}(x) - \mu_\beta(x) \mu_\beta^\top(x) \). Let \( A^+ \) denote the Monre–Penrose inverse of symmetric matrix. We have the following asymptotic results for the estimators.

Theorem 1. Under conditions (C1)–(C5), we have

\[
\sqrt{n}(\hat{\beta}_{\text{ROPG}} - \beta_0) \overset{d}{\to} N(0, \Sigma_{\text{ROPG}}),
\]

where

\[
\Sigma_{\text{ROPG}} = \frac{1}{4\tau} \mathbb{E} \left[ g(\beta_0^\top X)^2W_{0}(X)^+W_{0}(X)W_{0}^+(X)(2H(\epsilon) - 1)^2 \right] / \mathbb{E}[g'(\beta_0^\top X)^2]^2,
\]

\[
\tau = \int h_\epsilon^2(x)dx, \text{ and } H(\cdot) \text{ denotes the cumulative distribution function of } \epsilon.
\]

This theorem enables us to compare the efficiencies of \( \hat{\beta}_{\text{ROPG}} \) and existing approaches. In particular, it is of most interest to compare \( \hat{\beta}_{\text{ROPG}} \) with its LS-based counterpart proposed by Xia (2006), OPG estimator, \( \hat{\beta}_{\text{OPG}} \). According to Xia (2006), the asymptotic covariance of \( \hat{\beta}_{\text{OPG}} \) is given by

\[
\mathbb{E} \left[ g(\beta_0^\top X)^2W_{0}(X)^+W_{0}(X)W_{0}^+(X)\epsilon^2 \right] / \mathbb{E}[g'(\beta_0^\top X)^2]^2.
\]
To select the bandwidth for OPG and ROPG for simplicity, we summarize simulation results by using the ratio of AD, an ideal benchmark in our comparison. Following Xia (2006), the rule of thumb of Mack and Silverman (1982) is used to show that the OPG is an efficient and safe alternative to ADE and MAVE in infinite-sample cases and thus should be used to select the bandwidth for OPG and ROPG for simplicity. We summarize simulation results by using the ratio of AD, RAD = AD(β̂ OPC)/AD(β̂ ROPC). The means of ADs for ROPG and RADs are summarized in Table 1. The theoretical AREs of β̂ ROPC with respect to β̂ OPC for various error distributions (cf., (5)) are also documented in the last column of Table 1.

A few observations can be made from Table 1. The proposed ROPG method is highly efficient for all the distributions under consideration. Its efficiency compared to OPG is high and its RADs are all greater than 1 except for the normal distribution. Even for normal, its RADs are merely slightly smaller than 1. In addition, the theoretical ARE and simulated RADs of ROPG are always matched to certain degree. We also examine other error variance magnitudes for both models and the conclusion is similar. Therefore, in applications, the ROPG should be a reasonable alternative to existing works for non-normal error distributions by taking its convenience and robustness into account.

### Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>Design</th>
<th>AD</th>
<th>RAD</th>
<th>AD</th>
<th>RAD</th>
<th>ARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>(I)</td>
<td>(A)</td>
<td>0.132</td>
<td>0.966</td>
<td>0.152</td>
<td>0.967</td>
<td>0.955</td>
</tr>
<tr>
<td>Laplace</td>
<td></td>
<td></td>
<td>0.118</td>
<td>1.304</td>
<td>0.137</td>
<td>1.340</td>
<td>1.500</td>
</tr>
<tr>
<td>t(3)</td>
<td></td>
<td></td>
<td>0.151</td>
<td>1.571</td>
<td>0.174</td>
<td>1.615</td>
<td>1.900</td>
</tr>
<tr>
<td>T(0.05, 3)</td>
<td></td>
<td></td>
<td>0.138</td>
<td>1.102</td>
<td>0.161</td>
<td>1.121</td>
<td>1.196</td>
</tr>
<tr>
<td>T(0.05, 10)</td>
<td></td>
<td></td>
<td>0.166</td>
<td>2.524</td>
<td>0.188</td>
<td>2.655</td>
<td>4.768</td>
</tr>
<tr>
<td>Model (II)</td>
<td></td>
<td>Design</td>
<td>(B)</td>
<td></td>
<td>(B)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>(I)</td>
<td></td>
<td>0.067</td>
<td>0.958</td>
<td>0.110</td>
<td>0.962</td>
<td>0.955</td>
</tr>
<tr>
<td>Laplace</td>
<td></td>
<td></td>
<td>0.062</td>
<td>1.284</td>
<td>0.102</td>
<td>1.281</td>
<td>1.500</td>
</tr>
<tr>
<td>t(3)</td>
<td></td>
<td></td>
<td>0.079</td>
<td>1.617</td>
<td>0.137</td>
<td>1.644</td>
<td>1.900</td>
</tr>
<tr>
<td>T(0.05, 3)</td>
<td></td>
<td></td>
<td>0.072</td>
<td>1.139</td>
<td>0.122</td>
<td>1.150</td>
<td>1.196</td>
</tr>
<tr>
<td>T(0.05, 10)</td>
<td></td>
<td></td>
<td>0.089</td>
<td>2.966</td>
<td>0.158</td>
<td>3.234</td>
<td>4.768</td>
</tr>
</tbody>
</table>

Assuming that ε is independent of X, it immediately follows from Theorem 1 that the asymptotic relative efficiency (ARE) of β̂ ROPC with respect to β̂ OPC is

\[
\text{ARE}(\hat{\beta}_{\text{ROP}}, \hat{\beta}_{\text{OPC}}) = 12\sigma^2, \tag{5}
\]

where \(\sigma^2 = \mathbb{E}(\varepsilon^2)\).

This asymptotic relative efficiency is the same as that of the signed-rank Wilcoxon test with respect to the t-test. It is well known in the literature of rank analysis that the ARE is as high as 0.955 for normal error distribution, and can be significantly higher than one for many heavier-tailed distributions. For instance, this quantity is 1.5 for the double exponential distribution, and 1.9 for the t distribution with three degrees of freedom. For symmetric error distributions with finite Fisher information, this asymptotic relative efficiency is known to have a lower bound equal to 0.864 (Hettmansperger and McKean, 2010).

### 3. Numerical studies

We study the finite-sample performance of ROPG in this section via a simulation study. We use the Gaussian kernel function and 300 replications for each considered example. For a clear comparison, we adopt all the settings used in Xia (2006). To be more specific, one example (model (I)) is the regression

\[
Y = (\beta_0^T X)^2 \exp(\beta_0^T X) + \sigma \varepsilon, \tag{6}
\]

where \(X = (x_1, \ldots, x_p)^T, \beta_0 = (1, 2, 0, \ldots, 0)^T / \sqrt{5}, x_1, \ldots, x_p, \varepsilon\) are independent. We fix \(p = 10, n = 200, \sigma = 0.2\) in this model; Another example is a Tobit model (model (II)), say,

\[
Y = (\theta^T X + \varepsilon) I\{\theta^T X + \varepsilon > 0\}, \tag{7}
\]

where \(p = 2 and \theta = (1, 2)^T\). In this case, \(n = 100\) is considered. For both two models, we consider two sets of design for \(X\): (A) \((x_k + 1)/2 \sim \text{Beta}(\xi, 1)\) for \(k = 1, \ldots, p\) and (B) \((x_2 + 1)/2 \sim \text{Beta}(\xi, 1)\) and \(P(x_k = \pm 0.5) = 0.5, k = 1, 3, 4, \ldots, p\). In design (A), all the components are continuous, whereas in design (B) all the components are discrete except for the second component \(x_2\). \(\xi\) is chosen as one for brevity.

Five error distributions for \(\varepsilon\) are considered: N(0, 1), Laplace, \(t(3)\), Tukey contaminated normal \(T(0.05; 3)\) and \(T(0.05; 10)\). The performance of estimators is assessed via the absolute deviations (AD(\(\hat{\beta}\))), \(\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j|\). Xia (2006) has shown that the OPG is an efficient and safe alternative to ADE and MAVE in finite-sample cases and thus should be an ideal benchmark in our comparison. Following Xia (2006), the rule of thumb of Mack and Silverman (1982) is used to select the bandwidth for OPG and ROPG for simplicity. We summarize simulation results by using the ratio of AD, RAD = AD(\(\hat{\beta}_{\text{OPC}}\))/AD(\(\hat{\beta}_{\text{ROP}}\)). The means of ADs for ROPG and RADs are summarized in Table 1. The theoretical AREs of \(\hat{\beta}_{\text{ROP}}\) with respect to \(\hat{\beta}_{\text{OPC}}\) for various error distributions (c.f., (5)) are also documented in the last column of Table 1.
Acknowledgments

The authors would like to thank the editor, associate editor, and a referee for their helpful comments that have resulted in significant improvements in the article.

Appendix. Proof of Theorem 1

The proof follows along the same lines of the proof of Theorem 4.1 of Xia (2006), although part of details differs much. Here we only present a sketch of proof of theorems. A detailed technical report is available from the authors.

Thought out this section, we will use the following notations for ease of exposition. Suppose $A_n$ is a matrix. By $A_n = O(a_n)$ (or $A_n = o(a_n)$), we mean that all elements in $A_n$ are $O(a_n)$ (or $o(a_n)$) uniformly for $\beta \in \Theta_n$ and $x \in D_n$. Let $\delta_n = (nh/\log n)^{-1/2}$, $\tau_n = h^2 + \delta_n$, and $\delta_\beta = ||\beta - \beta_0||$. Denote

$$c_i = \sqrt{n}(\beta^*_0 X_i - b_i),$$

$$\Delta_i = g(\beta^*_0 X_i) - g(\beta^*_0 X_i) - g'(\beta^*_0 X_i)X_i.$$

With these notations, we rewrite $Q_n(b_i)$ as

$$Q_n(c_i) = \frac{2}{n(n-1)} \sum_{i,j} \{ c_i - c_j + n^{-1/2}c_i^1 X_{ij} + \Delta_i - \Delta_j \}[K_h(\beta^T X_i)K_h(\beta^T X_j)].$$

Moreover, we use $S_i(c_i)$ to denote the gradient function of $Q_n^*(c_i)$. More specifically,

$$S_i(c_i) = \frac{2}{n(n-1)} \sum_{i,j} \text{sgn}(c_i - c_j + n^{-1/2}c_i^1 X_{ij} + \Delta_i - \Delta_j) X_i K_h(\beta^T X_i)K_h(\beta^T X_j),$$

where $\text{sgn}(\cdot)$ is the sign function. In order to prove the theorem, we firstly state a few necessary lemmas.

Lemma 1. Suppose the Conditions (C1)–(C5) hold, then

$$\sqrt{n}(S_i(c) - S_i(0)) \rightarrow2n^{-1/2} \gamma f^2(\beta^T X_i)W_{\beta}(X_i)c_i + o_p(1).$$

Proof. Let $U_i = \sqrt{n}S_i(c_i) - S_i(0)) = \frac{2}{n(n-1)} \sum_{i,j} W_i(d_i, d_j)$, where

$$W_i(d_i, d_j) = [\text{sgn}(c_i - c_j + n^{-1/2}c_i^1 X_{ij} + \Delta_i - \Delta_j)] \times X_i K_h(\beta^T X_i)K_h(\beta^T X_j),$$

and $d_i = (X_i, c_i)$. Clearly, $U_i$ is of the form of U-statistic since $W_i(\cdot, \cdot)$ is symmetric in its arguments. Note that

$$E[\|W_i(d_i, d_j)\|^2] \leq 4h^{-4}E \left[ X_i^T X_j K^2 \left( \frac{\beta^T X_i h}{h} \right) K^2 \left( \frac{\beta^T X_j h}{h} \right) \right] = O(h^{-2}) = o(n).$$

Thus, $U_n = E[W_i(d_i, d_j)] + o_p(1)$ by Lemma A.1 of Wang et al. (2009). Note that

$$E(W_i(d_i, d_j)) = 2E \left[ \int H(\varepsilon + n^{-1/2}c_i^1 X_{ij} + \Delta_i - \Delta_j) - H(\varepsilon + \Delta_i - \Delta_j) \right] h(\varepsilon) d\varepsilon \times X_i K_h(\beta^T X_i)K_h(\beta^T X_j)$$

$$= 2h^2 n^{-1/2} E \left[ \int h(\varepsilon + \Delta_i - \Delta_j) h(\varepsilon) d\varepsilon X_i^T X_j K \left( \frac{\beta^T X_i h}{h} \right) K \left( \frac{\beta^T X_j h}{h} \right) \right] c_i(1 + o_p(1))$$

$$= 2n^{-1/2} \gamma f^2(\beta^T X_i)W_{\beta}(X_i)c_i + o_p(1),$$

where the last equality is followed by a simple calculation from Lemma 6.1 in Xia (2006). This completes the proof. \square

Moreover, we consider the following quadratic function:

$$B_i(c) = \sqrt{nc_i^T S_i(0) + n^{-1/2} \gamma f^2(\beta^T X_i)c_i^T W_{\beta}(X_i)c_i + \sqrt{n}Q_i^*(0)}.$$

Lemma 2. Suppose that conditions (C1)–(C5) hold. Then for any $\xi > 0$ and $c > 0$,

$$P \left\{ \sup_{\|c_i\| \leq c} \left| \sqrt{n}Q_i^*(c_i) - B_i(c) \right| \geq \xi \right\} \rightarrow 0.$$


Proof. By using Lemma 1, we have
\[ \nabla \left[ \sqrt{n}Q^*(c_i) - B_i(c_i) \right] = \sqrt{n}[S_i(c_i) - S_i(0)] - n^{-1/2} \tau f^2(\beta_i^\top X_i)W_\beta(X_i)c_i = o_p(1). \]

Using similar arguments of diagonal sub sequencing and convexity in the proof of Theorem A.3.7 of Hettmansperger and McKean (2010), we can complete the proof of this lemma. Details are omitted here. \( \square \)

Lemma 3. Suppose that conditions (C1)–(C5) hold, then
\[ \hat{b}_{i,\beta} = g'(\beta_{0i}^\top X_i)\beta_i + \frac{n^{1/2}}{2\tau f^2(\beta_i^\top X_i)}W_\beta^+(X_i)S_{i1}(0) + o_p(n^{-1/2}), \]
where
\[ S_{i1}(0) = \frac{2}{n^{3/2}(n-1)} \sum_{i<j} \sgn(e_i - e_j)X_{ij}K_h(\beta_i^\top X_i)K_h(\beta_j^\top X_j). \]

Proof. Denote \( \tilde{c}_i \) be the minimizer of \( B_i(c_i) \) and \( \hat{c}_i \) be the minimizer of \( Q^*(c_i) \). Since the convex function \( \sqrt{n}[Q^*(c_i) - Q^*(0)] \) converges in probability to the convex function \( \sqrt{n}\tilde{c}_i^\top S_i(0) + n^{-1/2} \tau f^2(\beta_i^\top X_i)c_i^\top W_\beta(X_i)c_i \) by Lemma 2, it follows from the convexity lemma (Pollard, 1991) that
\[ \hat{c}_i = \tilde{c}_i + o_p(1) = \frac{n}{2\tau f^2(\beta_i^\top X_i)}W_\beta^+(X_i)S_i(0) + o_p(1). \]

Let us re-express \( S_i(0) \) as
\[ S_i(0) = \frac{2}{n^{3/2}(n-1)} \sum_{i<j} \sgn(e_i - e_j + \Delta_i - \Delta_j)X_{ij}K_h(\beta_i^\top X_i)K_h(\beta_j^\top X_j) + \frac{2}{n^{3/2}(n-1)} \sum_{i<j} \sgn(e_i - e_j)\zeta_{ij} \]
\[ = S_{i2}(0) + S_{i1}(0), \]
where we denote \( \zeta_{ij} = X_{ij}K_h(\beta_i^\top X_i)K_h(\beta_j^\top X_j) \) for simplicity.

Taking the same procedure as the proof of Lemma 1, we can easily obtain that
\[ nS_{i2}(0) = O(h^2) = o_p(1), \]
from which the lemma follows immediately. \( \square \)

Proof of Theorem 1. Suppose we start with \( \beta^{(0)} = \beta \in \Theta_n \). Following by Lemma 6.6 and 6.7 in Xia (2006), we have
\[ M_{\beta_0} = n^{-1/2} \sum_{i=1}^n \frac{g'(\beta_{0i}^\top X_i)}{2\tau f^2(\beta_i^\top X_i)}W_\beta^+(X_i)S_{i1}(0) \]
\[ = \frac{1}{n^2(n-1)} \sum_{i=1}^n \sum_{i<j} \frac{g'(\beta_{0i}^\top X_i)}{2\tau f^2(\beta_i^\top X_i)}W_\beta^+(X_i)[\sgn(e_i - e_j)X_{ij}K_h(\beta_i^\top X_i)K_h(\beta_j^\top X_j)] \]
\[ = \frac{1}{n^2} \sum_{i=1}^n \sum_{i<j} \frac{g'(\beta_{0i}^\top X_i)}{2\tau f^2(\beta_i^\top X_i)}W_\beta^+(X_i)[2H(e_i) - 1](X_i - \mu_{\beta}(X_i))K_h(\beta_j^\top X_j) + o(n^{-1/2}) \]
\[ = \frac{1}{n} \sum_{i=1}^n \frac{1}{2\tau} g'(\beta_{0i}^\top X_i)W_\beta^+(X_i)[2H(e_i) - 1]v_{\beta_0}(X_i) + o(n^{-1/2}). \]

It follows from (9) to (11) in Xia (2006) and Lemma 3 that
\[ \frac{1}{n} \sum_{i=1}^n \hat{b}_{i,\beta} \hat{b}_{i,\beta} = n^{-1} \sum_{i=1}^n \left[ 1 + O(\tau_n + \delta_\beta) \right] g'(\beta_{0i}^\top X_i)^2 \| (\beta_0 + \tilde{M}_{\beta_0}) \|^2 \tilde{\beta}^\top \beta + o_p(n^{-1/2}), \]
where \( \tilde{\beta} = (\beta_0 + \tilde{M}_{\beta_0})/\| \beta_0 + \tilde{M}_{\beta_0} \| \) and
\[ \tilde{M}_{\beta_0} = \left[ Eg'(\beta_{0i}^\top X_i)^2 \right]^{-1} \frac{1}{n} \sum_{i=1}^n \frac{1}{2\tau} g'(\beta_{0i}^\top X_i)W_\beta^+(X_i)[2H(e_i) - 1]v_{\beta}(X_i). \]
Since $\beta_0^\top W^+(x) = 0$, so $\beta_0^\top \tilde{M}_{\beta_0} = 0$. In addition,
\[
\|\beta_0 + \tilde{M}_{\beta_0}\| = 1 + \|\tilde{M}_{\beta_0}\|^2 = 1 + o(n^{-1/2}).
\]
Thus, we have
\[
\tilde{\beta} = \beta_0 + \tilde{M}_{\beta_0} + o(n^{-1/2}).
\]

The estimator $\beta^{(k+1)}$ in the next iteration is the eigenvector of
\[
\frac{1}{n} \sum_{i=1}^{n} \tilde{b}_i \tilde{b}_i^\top.
\]
By Lemma 3.1 in Bai et al. (1991), we have
\[
\beta^{(k+1)} - \tilde{\beta} = o(n^{-1/2}), \text{ i.e.,}
\]
\[
\beta^{(k+1)} = \beta_0 + \tilde{M}_{\beta_0} + o(n^{-1/2}).
\]

Therefore, Theorem 1 follows from the preceding equation and the central limit theorem for $\tilde{M}_{\beta_0}$.  

\textbf{References}


