SUPPLEMENT TO "NONLINEAR INTERACTION DETECTION THROUGH MODEL-BASED SUFFICIENT DIMENSION REDUCTION"

Guoliang Fan[1,2], Liping Zhu[2] and Shujie Ma[3]

[1]*Zhejiang Gongshang University,* [2]*Renmin University of China*

[3]*University of California, Riverside*

Abstract

In this supplementary document, we will briefly review some related models and comment on some relevant methods. We will also report some simulation results when $\mathbf{x}$ and $\mathbf{z}$ are correlated. Proofs of Theorems 1-4 are also given here.

1. A REVIEW ON RELATED MODELS AND METHODS

For the purposes of fair comparison, we shall concentrate on the univariate response case in this section unless stated otherwise because many existing models and the existing partial central mean dimension reduction methods are designed for this case. We give a brief review on related models and existing methods. We also emphasize here that we are concerned with the multivariate response case in the present paper.

1.1. Relationship to Existing Models

Model (1.1) with an unspecified d_0 is so flexible that it encompasses many existing semiparametric models. Ma and Song (2015) suggested the following varying index coefficient model

$$E(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = \sum_{k=1}^{q} m_k(\boldsymbol{\beta}_k^{\mathrm{T}} \mathbf{x}) Z_k, \tag{A.1}$$

where $\boldsymbol{\beta}_k$ is a p-vector. Model (A.1) is a special case of model (1.1) if we set $d_0 = q$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$. The single-index coefficient model proposed by Xia and Li (1999) takes the form of

$$E(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = \sum_{k=1}^{q} m_k(\boldsymbol{\beta}_1^{\mathrm{T}} \mathbf{x}) Z_k, \tag{A.2}$$

where a common $\boldsymbol{\beta}_1$ is shared by all m_ks. In the literature the partially linear varying multi-index coefficient model (Liu et al., 2016) is another popular semiparametric model which takes the form of

$$E(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = \sum_{k=1}^{q} \left\{ m_k(\boldsymbol{\alpha}_{k,1}^{\mathrm{T}} \mathbf{x}_1) + (\boldsymbol{\alpha}_{k,2}^{\mathrm{T}} \mathbf{x}_2) \right\} Z_k, \tag{A.3}$$

where $\mathbf{x} = (\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}_k = \mathrm{diag}(\boldsymbol{\alpha}_{k,1}, \boldsymbol{\alpha}_{k,2})$ is a block-diagonal $p \times 2$ matrix. This is an extension of the partially linear single index model (Carroll et al., 1997). Setting $d_0 = q$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$, we can clearly see that the partially linear varying multiple-index coefficient model is again a special case of model (1.1).

Because of such a difference in model specification, these models behave rather differently to characterize interaction effects, which are of central interests in our motivating examples. Some of the key differences are summarized as follows.

1. If all entries in the i-th row of $\boldsymbol{\beta}$ in model (1.1) are identically zero, then X_i does not interact with any components of $\mathbf{z} = (Z_1, \ldots, Z_q)^{\mathrm{T}}$. By contrast, if the i-th entry of $\boldsymbol{\beta}_k$ in model (A.1) or the i-th component of $\boldsymbol{\alpha}_{k,1}$ or $\boldsymbol{\alpha}_{k,2}$ in model (A.3) is zero, then X_i does not interact with Z_k. However, the individual interaction effects between X_i and Z_k may be too weak to be detectable. To detect weak interaction effects, in model (1.1) we use the grouped covariates ($\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$) which strengthen the weak signal level of individual interaction effects. This property makes model (1.1) significantly different from models (A.1) and (A.3). In addition, a single vector $\boldsymbol{\beta}_1$ in model (A.2) may not be sufficient to capture the interaction effects completely. Through choosing an appropriate d_0, model (1.1) aims to retain complete information of the interaction effects.

2. Ma and Song (2015) argued that model (A.1) could be used to assess

arbitrary nonlinear interactions but model (A.2) could not. Ma and Song (2015) illustrated this point through a linear example. This corresponds to a special case of model (1.1) with $d_0 = 1$. However, we allow for a general d_0 in model (1.1) where the linear example is no longer valid. We emphasize here that, model (1.1) can accommodate arbitrary nonlinear interaction effects as well as model (A.1). This property makes model (1.1) significantly different from model (A.2) in that the latter may contain ill-defined interaction effects while the former does not.

3. Liu et al. (2016) and Ma and Song (2015) argued, respectively, that models (A.1) and (A.3) enable to engage different components of $\mathbf{z}$ to modify the slope function of $\mathbf{x}$ through using different $\boldsymbol{\beta}_k$s and $\boldsymbol{\alpha}_{k,j}$s, while model (A.2) with a common $\boldsymbol{\beta}$ and $d_0 = 1$ does not. In model (1.1) we allow for a general d_0 and d_0 must be determined in a data-driven fashion. This also permits different components of $\mathbf{z}$ to modify the slope function of $\mathbf{x}$ differently. For example, if $d_0 = 2$ in model (1.1) and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, then $\mathbf{m}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $\mathbf{m}_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ may vary with $(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})$ and $(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})$, respectively. In this particular example, different components of $\mathbf{z}$ may use different basis of $\text{span}(\boldsymbol{\beta})$ to modify the slope function of $\mathbf{x}$. Therefore, allowing for a general d_0 in model (1.1)

maintains the same flexibility as model (A.1) and has the desirable model fitting and interpretation. This property makes model (1.1) significantly different from model (A.2).

In the Framingham Heart Study in Section 3.3, we let the first component of $\mathbf{z}$ be 1, namely, $\mathbf{z} = (1, Z_2, \ldots, Z_q)^{\mathrm{T}}$. Accordingly, model (1.1) boils down to

$$E(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) \;=\; \mathbf{m}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) + \sum_{k=2}^{q} \mathbf{m}_k(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) Z_k. \tag{A.4}$$

Another related model is

$$E(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) \;=\; \mathbf{m}_1(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + \sum_{k=2}^{q} \mathbf{m}_k(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}) Z_k, \tag{A.5}$$

which is also a special case of model (1.1), if we simply choose $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\mathbf{z} = (1, Z_2, \ldots, Z_p)^{\mathrm{T}}$ in model (1.1). Both our proposed algorithm and the theoretical results can be directly used in these two models.

Because we allow for a general d_0, that all $\mathbf{m}_k$s share a common $\boldsymbol{\beta}$ in model (1.1) is an imperative assumption for the identifiability purposes. It is not a restriction or an assumption. If we used $\boldsymbol{\beta}_k$ in model (1.1) with an unknown column dimension d_k, model (1.1) would no longer be identifiable. Specifically, in the following two cases we can easily choose a different function $\mathbf{m}_k^*$ such that $\mathbf{m}_k(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{x}) = \mathbf{m}_k^*(\boldsymbol{\beta}_k^{*\mathrm{T}}\mathbf{x})$.

1. The first is to increase the column dimension of $\boldsymbol{\beta}_k$ from d_0 to d^* such that $\mathrm{span}(\boldsymbol{\beta}_k) \subseteq \mathrm{span}(\boldsymbol{\beta}_k^*)$. Say, $\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_k, \boldsymbol{\alpha}_k)$ for an arbitrary matrix $\boldsymbol{\alpha}_k$.

2. The second is to multiply $\boldsymbol{\beta}_k$ with a nonsingular $d_0 \times d_0$ matrix $\boldsymbol{\gamma}_k$. Let $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_k\boldsymbol{\gamma}_k$. In this case, the column dimension d_0 remains unchanged.

That is why we assume a common $\boldsymbol{\beta}$ in model (1.1).

Cook (2007) and Cook and Forzani (2008, 2009) assumed that the inverse regression $(\mathbf{x} \mid \mathbf{y})$ admits heterogeneous linear structures and their interest is in estimating the central subspace (Cook, 1998), while we assume that the forward regression of $(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ admits the semiparametric structure (1.1) and our goal is to identify the interaction effects through estimating the partial central mean dimension reduction subspace $\mathrm{span}(\boldsymbol{\beta})$ (Li et al., 2003).

1.2. A Brief Review on Existing Methods

There are some existing works in the literature which aim to estimate the partial central mean dimension reduction subspace $\mathrm{span}(\boldsymbol{\beta})$ such that $E(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = E(\mathbf{y} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \mathbf{z})$ when both $\mathbf{y}$ and $\mathbf{z}$ are univariate random variables and $(\mathbf{x} \mid \mathbf{z})$ satisfies certain distributional assumptions. We review these methods briefly here. Suppose $\mathbf{z}$ is a categorical variable and has C categories, say, $\mathbf{z} = \{1, \cdots, C\}$. Under the linearity assumption that $E(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \mathbf{z})$ is a linear function of $\mathbf{x}$, Li et al. (2003) showed that $\boldsymbol{\beta}(\mathbf{z}) \stackrel{\text{def}}{=} \{\text{cov}(\mathbf{x}, \mathbf{x}^{\mathrm{T}} \mid \mathbf{z})\}^{-1} \{\text{cov}(\mathbf{x}, \mathbf{y} \mid \mathbf{z})\} \subseteq \text{span}(\boldsymbol{\beta})$. Therefore, we can simply recover $\text{span}(\boldsymbol{\beta})$ through the eigen-space of $E\Big\{\boldsymbol{\beta}(\mathbf{z})\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{z})\Big\}$. Following the idea of Zhu et al. (2010), Feng et al. (2013) generalized the work of Li et al. (2003) by allowing $\mathbf{z}$ to be a continuous random variable. To be precise, Feng et al. (2013) recovered $\text{span}(\boldsymbol{\beta})$ through the eigen-space of $E\left\{\widetilde{\boldsymbol{\beta}}(\widetilde{\mathbf{z}})\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}(\widetilde{\mathbf{z}})\right\}$ if the partial central mean dimension reduction subspace is of interest, where $\widetilde{\boldsymbol{\beta}}(\widetilde{\mathbf{z}}) \stackrel{\text{def}}{=} \{\text{cov}(\mathbf{x}, \mathbf{x}^{\mathrm{T}} \mid \mathbf{z} \leq \widetilde{\mathbf{z}})\}^{-1} \{\text{cov}(\mathbf{x}, \mathbf{y} \mid \mathbf{z} \leq \widetilde{\mathbf{z}})\}$ and $\widetilde{\mathbf{z}}$ is an independent copy of $\mathbf{z}$. The same idea can be readily generalized to recover the partial central dimension reduction subspace. The distributional assumptions on $(\mathbf{x} \mid \mathbf{z})$ are relaxed by Ma and Song (2015) and Liu et al. (2016) under different model structures. However, their proposed semiparametric estimates require that $\mathbf{y}$ be univariate. In addition, the linearity condition is violated if some components of $\mathbf{x}$ are categorical. Such requirements are possibly very restrictive. In particular, in the Framingham Heart Study, both $\mathbf{z}$ and $\mathbf{y}$ are continuous and multivariate. Therefore, these existing methods cannot be used directly.

2. SOME ADDITIONAL SIMULATIONS

In this section we conduct some additional simulations where $\mathbf{x}$ and $\mathbf{z}$ are correlated. We revisit model (II) with the following nonlinear link functions:

$$\begin{cases} Y_1 = \sin(4\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})Z_1 + 2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})Z_2 + \varepsilon_1; \\ Y_2 = \cos(2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})Z_2 + \varepsilon_2; \\ Y_3 = 2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})Z_1 + \sin(2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})Z_2 + \varepsilon_3. \end{cases}$$

We draw $\widetilde{\mathbf{x}} \stackrel{\text{def}}{=} (\mathbf{x}^{\mathrm{T}}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}} = (\widetilde{X}_1, \ldots, \widetilde{X}_{p+q})^{\mathrm{T}}$ from multivariate normal distribution with mean zero and covariance matrix $\mathrm{cov}(\widetilde{X}_k, \widetilde{X}_l) = \rho^{|k-l|}$. We set $\rho = 0.2$, 0.5 and 0.8, respectively. We fix $r = 3$, and generate $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3)^{\mathrm{T}}$ from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

We fix $p = 10$, $q = 2$ and $\boldsymbol{\beta} = (1, 0.8, 0.6, 0.4, 0.2, -0.2, -0.4, -0.6, -0.8, 0)^{\mathrm{T}}$. We choose the sample size $n = 200$ and 500 and repeat each simulation 1000 times.

The average of estimation bias ("bias"), the Monte Carlo standard deviation ("std"), the average of estimated standard deviation ("$\widehat{\text{std}}$"), and the empirical coverage probability ("cvp") at the nominal 95% confidence level for all free parameter are summarized in Table S1 - Table S2 for $n = 200$

and $n = 500$, respectively. It can be clearly seen that all estimates have very small biases, and the biases become smaller as the sample size increases. This phenomenon again shows that both the weighted and the unweighted estimates are consistent. In addition, as the correlations between $\mathbf{x}$ and $\mathbf{z}$ increase, both the Monte Carlo standard deviations and the estimated standard deviations increase significantly, indicating that the estimates are more and more unstable. However, the empirical coverage probabilities are still very close to 95%, indicating that the inferential results are still reliable.

3. PROOFS OF THEOREM 1 - THEOREM 4

For notational simplicity, we omit the subscript $_{\mathbf{w}}$ and write $\widehat{\boldsymbol\beta}_{-d} = \widehat{\boldsymbol\beta}_{-d,\mathbf{w}}$, $\widehat{\boldsymbol\beta} = \big(\mathbf{I}_d, \widehat{\boldsymbol\beta}_{-d,\mathbf{w}}^{\mathrm{T}}\big)^{\mathrm{T}}$ and $\widehat{\mathbf{m}}(\widehat{\boldsymbol\beta}^{\mathrm{T}}\mathbf{x}_i) = \widehat{\mathbf{m}}(\widehat{\boldsymbol\beta}^{\mathrm{T}}\mathbf{x}_i, \widehat{\boldsymbol\beta})$. Let $c_n = h^s + \{\log n/(nh^d)\}^{1/2}$.

3.1. Proof of Theorem 1

By conditions (C1) and (C2),

$$
\begin{aligned}
E\big\{\mathbf{z}\mathbf{z}^{\mathrm{T}} \otimes (\boldsymbol\beta^{\mathrm{T}}\mathbf{x} - \mathbf{u})K_h(\boldsymbol\beta^{\mathrm{T}}\mathbf{x} - \mathbf{u})\big\} &= E\big\{\boldsymbol\Omega(\boldsymbol\beta^{\mathrm{T}}\mathbf{x}) \otimes (\boldsymbol\beta^{\mathrm{T}}\mathbf{x} - \mathbf{u})K_h(\boldsymbol\beta^{\mathrm{T}}\mathbf{x} - \mathbf{u})\big\} \\
&= O(h^s). \qquad \text{(C.1)}
\end{aligned}
$$

Applying similar techniques to those used in Mack and Silverman (1982),

Table S1: The simulation results when $\mathbf{x}$ and $\mathbf{z}$ are correlated based on $n = 200$: the average bias of the estimators ("bias"), the Monte Carlo standard deviation ("std"), the average of the estimated standard deviation ("$\widehat{\text{std}}$") based on the theoretical calculation, and the empirical coverage probability ("cvp") at the nominal 95% confidence level. All simulation results reported below are multiplied by 100.

			$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	$\widehat{\beta}_9$	$\widehat{\beta}_{10}$
$\rho_{\mathbf{x},\mathbf{z}}$	$\mathbf{W}$	True value	0.8	0.6	0.4	0.2	-0.2	-0.4	-0.6	-0.8	0
0.2	$\mathbf{I}$	bias	1.30	0.59	0.79	0.45	-0.13	-0.41	-0.75	-0.83	-0.05
		std	8.81	7.15	6.81	6.37	6.68	6.99	7.44	7.93	6.50
		$\widehat{\text{std}}$	9.10	7.64	7.11	6.78	6.78	7.16	7.65	8.37	6.45
		cvp	95.90	96.40	95.60	95.90	94.90	95.50	95.80	95.40	94.90
0.2	$\widehat{\mathbf{\Sigma}}^{-1}$	bias	1.07	0.75	0.35	0.18	-0.13	-0.26	-0.55	-0.85	0.16
		std	5.10	4.42	3.87	3.88	3.92	4.00	4.38	4.71	3.63
		$\widehat{\text{std}}$	5.24	4.39	4.08	3.90	3.90	4.11	4.40	4.81	3.73
		cvp	95.60	94.90	95.90	95.20	94.80	95.00	94.80	95.30	95.60
0.5	$\mathbf{I}$	bias	1.84	1.03	0.25	0.57	-0.49	-0.92	-0.99	-0.80	-0.03
		std	11.63	9.49	8.95	8.67	8.81	9.00	9.78	10.20	7.19
		$\widehat{\text{std}}$	12.64	9.93	9.35	8.99	8.97	9.36	9.96	10.68	7.48
		cvp	96.10	95.70	96.10	95.40	94.70	96.10	95.20	96.10	95.60
0.5	$\widehat{\mathbf{\Sigma}}^{-1}$	bias	1.38	0.62	0.15	0.20	-0.27	-0.46	-0.73	-0.66	-0.08
		std	6.87	5.48	5.20	5.03	5.27	5.23	5.52	5.99	4.45
		$\widehat{\text{std}}$	7.13	5.59	5.27	5.06	5.06	5.28	5.61	6.03	4.25
		cvp	95.30	96.10	94.90	94.80	93.70	95.10	95.10	95.40	94.50
0.8	$\mathbf{I}$	bias	3.79	1.34	1.31	0.76	-0.23	-1.00	-1.70	-2.32	0.23
		std	18.16	16.01	15.43	14.95	15.50	16.05	15.94	16.57	10.70
		$\widehat{\text{std}}$	20.51	16.36	15.66	15.19	15.22	15.68	16.45	17.23	10.81
		cvp	97.00	95.30	95.50	96.10	94.30	94.90	95.90	96.30	95.40
0.8	$\widehat{\mathbf{\Sigma}}^{-1}$	bias	2.62	0.90	1.16	0.35	-0.78	-0.35	-1.25	-1.35	-0.01
		std	11.04	9.44	9.31	9.10	9.23	9.11	9.31	9.69	6.62
		$\widehat{\text{std}}$	11.48	9.12	8.76	8.49	8.51	8.74	9.20	9.63	6.06
		cvp	95.70	94.50	93.50	92.60	92.90	93.50	95.30	95.00	93.20

we obtain

$$\sup_{\mathbf{u}} \left\| h\mathbf{S}_{n1}(\mathbf{u},\boldsymbol{\beta}) - E\Big[\mathbf{z}\mathbf{z}^{\mathrm{T}} \otimes (\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}-\mathbf{u})K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}-\mathbf{u})\Big] \right\| = O_p\Big\{\Big(\frac{\log n}{nh^d}\Big)^{1/2}\Big\}, \tag{C.2}$$

Table S2: The simulation results when $\mathbf{x}$ and $\mathbf{z}$ are correlated based on $n = 500$: the average bias of the estimators ("bias"), the Monte Carlo standard deviation ("std"), the average of the estimated standard deviation ("$\widehat{\text{std}}$") based on the theoretical calculation, and the empirical coverage probability ("cvp") at the nominal 95% confidence level. All simulation results reported below are multiplied by 100.

			$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	$\widehat{\beta}_9$	$\widehat{\beta}_{10}$
$\rho_{\mathbf{x},\mathbf{z}}$	$\mathbf{W}$	True value	0.8	0.6	0.4	0.2	-0.2	-0.4	-0.6	-0.8	0
0.2	$\mathbf{I}$	bias	0.76	0.15	0.39	0.21	-0.27	-0.16	-0.25	-0.41	-0.13
		std	5.39	4.48	4.40	4.13	4.05	4.51	4.79	4.84	3.88
		$\widehat{\text{std}}$	5.73	4.80	4.47	4.24	4.24	4.47	4.81	5.25	4.05
		cvp	96.60	96.20	95.90	95.20	95.80	94.80	94.80	96.20	95.40
0.2	$\widehat{\mathbf{\Sigma}}^{-1}$	bias	0.87	0.38	0.27	0.15	-0.18	-0.28	-0.44	-0.54	-0.02
		std	2.69	2.39	2.26	2.08	2.16	2.38	2.32	2.51	2.14
		$\widehat{\text{std}}$	3.07	2.57	2.39	2.27	2.26	2.39	2.57	2.81	2.17
		cvp	96.40	95.90	95.90	96.70	95.10	95.00	96.60	96.60	95.10
0.5	$\mathbf{I}$	bias	1.15	0.81	0.48	0.29	-0.08	-0.30	-0.94	-1.05	0.13
		std	7.56	6.01	5.40	5.39	5.40	5.69	6.09	6.41	4.62
		$\widehat{\text{std}}$	8.00	6.25	5.86	5.61	5.62	5.86	6.23	6.75	4.66
		cvp	95.30	95.80	97.00	96.50	96.70	95.60	96.10	96.30	95.00
0.5	$\widehat{\mathbf{\Sigma}}^{-1}$	bias	1.17	0.61	0.43	0.21	-0.10	-0.32	-0.68	-0.90	-0.18
		std	3.98	3.31	3.02	3.03	2.94	3.12	3.39	3.47	2.52
		$\widehat{\text{std}}$	4.29	3.35	3.14	3.00	3.00	3.14	3.33	3.61	2.52
		cvp	95.60	95.00	95.40	95.10	94.70	94.20	94.10	95.10	94.50
0.8	$\mathbf{I}$	bias	3.32	0.93	0.84	0.07	0.18	-1.09	-1.05	-1.95	0.06
		std	12.57	9.60	9.38	9.35	9.26	9.35	10.01	10.59	6.51
		$\widehat{\text{std}}$	13.29	10.26	9.79	9.50	9.50	9.78	10.25	10.91	6.70
		cvp	96.60	97.40	95.80	95.30	95.70	96.20	95.10	95.40	94.80
0.8	$\widehat{\mathbf{\Sigma}}^{-1}$	bias	2.78	0.87	0.65	0.28	-0.32	-0.54	-1.08	-1.38	-0.13
		std	6.39	5.35	5.08	5.30	5.13	5.30	5.43	5.59	3.69
		$\widehat{\text{std}}$	7.05	5.44	5.19	5.04	5.04	5.18	5.44	5.78	3.56
		cvp	95.20	95.40	95.10	93.20	94.10	94.20	94.00	95.60	93.20

which together with (C.1) yields that $h\mathbf{S}_{n1}(\mathbf{u},\boldsymbol{\beta}) = O_p(c_n)$. Thus, we have

$$\sum_{j=1}^{q} \widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)Z_{ij} = \big\{\mathbf{S}_{n0}^{-1}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}})\boldsymbol{\xi}_{n0}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}})\big\}^{\mathrm{T}}\mathbf{z}_i\{1+o_p(1)\}$$

$$= \Big[\mathbf{S}_{n0}^{-1}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}})\frac{1}{n}\sum_{j=1}^{n} Z_{ij}\Big\{\sum_{k=1}^{q}\mathbf{m}_k(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)Z_{jk}+\boldsymbol{\varepsilon}_j\Big\}^{\mathrm{T}}K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j-\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\Big]^{\mathrm{T}}\mathbf{z}_i\{1+o_p(1)\}$$

$$\stackrel{\text{def}}{=} \Delta_i\{1+o_p(1)\}.$$

Taylor expansion gives

$$\mathbf{m}_k(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j) = \mathbf{m}_k(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j) + \mathbf{m}_k^{(1)}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j)(\boldsymbol{\beta}_{-d} - \widehat{\boldsymbol{\beta}}_{-d})^{\mathrm{T}}\mathbf{x}_{-d,j} + o_p(\|\widehat{\boldsymbol{\beta}}_{-d} - \boldsymbol{\beta}_{-d}\|).$$

Hence, we have

$$\begin{aligned}
\Delta_i &= n^{-1}\sum_{j=1}^{n}\Big\{\sum_{k=1}^{q}\mathbf{m}_k(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j)Z_{jk}\mathbf{z}_j^{\mathrm{T}}K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\Big\}\mathbf{S}_{n0}^{-1}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}})\mathbf{z}_i\\
&+ n^{-1}\sum_{j=1}^{n}\sum_{k=1}^{q}\mathbf{m}_k^{(1)}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j)(\boldsymbol{\beta}_{-d} - \widehat{\boldsymbol{\beta}}_{-d})^{\mathrm{T}}\mathbf{x}_{-d,j}Z_{j,k}\mathbf{z}_j^{\mathrm{T}}K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\mathbf{S}_{n0}^{-1}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}})\mathbf{z}_i\\
&+ n^{-1}\sum_{j=1}^{n}\boldsymbol{\varepsilon}_j\mathbf{z}_j^{\mathrm{T}}K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\mathbf{S}_{n0}^{-1}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}})\mathbf{z}_i\\
&\overset{\text{def}}{=} \Delta_{i1} + \Delta_{i2} + \Delta_{i3}.
\end{aligned}$$

Similar to the proof of (C.2), we can derive that

$$\begin{aligned}
&\mathbf{S}_{n0}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j,\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j)f(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j) = O_p\Big\{\Big(\frac{\log n}{nh^d}\Big)^{1/2}\Big\},\\
&n^{-1}\sum_{j=1}^{n}\Big\{\sum_{k=1}^{q}\mathbf{m}_k(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j)Z_{jk}\mathbf{z}_j^{\mathrm{T}}K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\Big\} - \mathbf{m}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)f(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\\
&\qquad = O_p\Big\{\Big(\frac{\log n}{nh^d}\Big)^{1/2}\Big\}.
\end{aligned}$$

Thus, we obtain that $\Delta_{i1} = \mathbf{m}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\mathbf{z}_i + O_p\big[\{\log n/(nh^d)\}^{1/2}\big]$. Similarly,

$$\Delta_{i2} \;=\; \sum_{k=1}^{q}\mathbf{m}_k^{(1)}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)(\boldsymbol{\beta}_{-d} - \widehat{\boldsymbol{\beta}}_{-d})^{\mathrm{T}}E(\mathbf{x}_{-d,i}|\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)Z_{ik} + O_p\Big\{\Big(\frac{\log n}{nh^d}\Big)^{1/2}\Big\},$$

$$\Delta_{i3} = O_p\Big\{\Big(\frac{\log n}{nh^d}\Big)^{1/2}\Big\}.$$

Combining the above results, we have

$$\begin{aligned}\mathbf{y}_i - \sum_{j=1}^q \widehat{\mathbf{m}}_j(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)Z_{ij} &= \sum_{j=1}^q \Big\{\mathbf{m}_j(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i) + \mathbf{m}_j^{(1)}(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)(\boldsymbol\beta_{-d} - \widehat{\boldsymbol\beta}_{-d})^{\mathrm T}\mathbf{x}_{-d,i}\Big\}Z_{ij} \\ &\quad - \sum_{j=1}^q \widehat{\mathbf{m}}_j(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)Z_{ij} + \boldsymbol\varepsilon_i \\ &= \sum_{j=1}^q \mathbf{m}_j^{(1)}(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)(\boldsymbol\beta_{-d} - \widehat{\boldsymbol\beta}_{-d})^{\mathrm T}\big\{\mathbf{x}_{-d,i} - E(\mathbf{x}_{-d,i}|\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)\big\}Z_{ij} + \boldsymbol\varepsilon_i + o_p(\|\widehat{\boldsymbol\beta}_{-d} - \boldsymbol\beta_{-d}\|).\end{aligned}$$

Thus it follows that

$$\begin{aligned}&\sum_{i=1}^n \Big\{\mathbf{y}_i - \sum_{j=1}^q \widehat{\mathbf{m}}_j(\mathbf{x}_{d,i} + \boldsymbol\beta_{-d}^{\mathrm T}\mathbf{x}_{-d,i})Z_{ij}\Big\}^{\mathrm T}\mathbf{W}\Big\{\mathbf{y}_i - \sum_{j=1}^q \widehat{\mathbf{m}}_j(\mathbf{x}_{d,i} + \boldsymbol\beta_{-d}^{\mathrm T}\mathbf{x}_{-d,i})Z_{ij}\Big\} \\ &\approx \sum_{i=1}^n \Big[\sum_{j=1}^q \mathbf{m}_j^{(1)}(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)Z_{ij} \otimes \big\{\mathbf{x}_{-d,i} - E(\mathbf{x}_{-d,i}|\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)\big\}^{\mathrm T}\big\{\mathrm{vec}(\boldsymbol\beta_{-d}) - \mathrm{vec}(\widehat{\boldsymbol\beta}_{-d})\big\} + \boldsymbol\varepsilon_i\Big]^{\mathrm T}\mathbf{W} \\ &\quad \times \Big[\sum_{j=1}^q \mathbf{m}_j^{(1)}(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)Z_{ij} \otimes \big\{\mathbf{x}_{-d,i} - E(\mathbf{x}_{-d,i}|\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)\big\}^{\mathrm T}\big\{\mathrm{vec}(\boldsymbol\beta_{-d}) - \mathrm{vec}(\widehat{\boldsymbol\beta}_{-d})\big\} + \boldsymbol\varepsilon_i\Big].\end{aligned}$$

Minimizing the above equation yields that

$$\begin{aligned}\mathrm{vec}(\widehat{\boldsymbol\beta}_{-d}) - \mathrm{vec}(\boldsymbol\beta_{-d}) &= \Big[\sum_{i=1}^n \Big\{\sum_{j=1}^q \mathbf{m}_j^{(1),\mathrm T}(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)Z_{ij} \otimes \big(\mathbf{x}_{-d,i} - E(\mathbf{x}_{-d,i}|\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)\big)\Big\}\mathbf{W} \\ &\quad \cdot \Big\{\sum_{j=1}^q \mathbf{m}_j^{(1)}(\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)Z_{ij} \otimes \big(\mathbf{x}_{-d,i}^{\mathrm T} - E(\mathbf{x}_{-d,i}^{\mathrm T}|\widehat{\boldsymbol\beta}^{\mathrm T}\mathbf{x}_i)\big)\Big\}\Big]^{-1}\end{aligned}$$

$$\times \left[\sum_{i=1}^{n} \left\{ \sum_{j=1}^{q} \mathbf{m}_j^{(1),\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i) Z_{ij} \otimes \big(\mathbf{x}_{-d,i} - E(\mathbf{x}_{-d,i}|\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\big) \right\} \mathbf{W}\boldsymbol{\varepsilon}_i \right]$$

$$+ o_p(\|\widehat{\boldsymbol{\beta}}_{-d} - \boldsymbol{\beta}_{-d}\|) \stackrel{\text{def}}{=} \Psi_{n1}^{-1}\Psi_{n2} + o_p(\|\widehat{\boldsymbol{\beta}}_{-d} - \boldsymbol{\beta}_{-d}\|).$$

By Slutsky's theorem, to prove Theorem 1, we need only to show that

$$n^{-1}\Psi_{n1} \stackrel{p}{\longrightarrow} \mathbf{A}_{\mathbf{w}} \quad \text{and} \quad n^{-1/2}\Psi_{n2} \stackrel{d}{\longrightarrow} \mathcal{N}(0, \mathbf{B}_{\mathbf{w}}).$$

We prove the second part because the proof of fist part is similar. Observe that

$$\begin{aligned}
\Psi_{n2} &= \sum_{i=1}^{n}\sum_{j=1}^{q} \mathbf{m}_j^{(1),\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) Z_{ij} \otimes \widetilde{\mathbf{x}}_{-d,i} \mathbf{W}\boldsymbol{\varepsilon}_i \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{q} \mathbf{m}_j^{(1),\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) Z_{ij} \otimes \left\{ E(\mathbf{x}_{-d,i}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) - E(\mathbf{x}_{-d,i}|\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i) \right\} \mathbf{W}\boldsymbol{\varepsilon}_i \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{q} \left\{ \mathbf{m}_j^{(1)}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i) - \mathbf{m}_j^{(1)}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) \right\}^{\mathrm{T}} Z_{ij} \otimes \widetilde{\mathbf{x}}_{-d,i} \mathbf{W}\boldsymbol{\varepsilon}_i \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{q} \left\{ \mathbf{m}_j^{(1)}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i) - \mathbf{m}_j^{(1)}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) \right\}^{\mathrm{T}} Z_{ij} \otimes \left\{ E(\mathbf{x}_{-d,i}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) - E(\mathbf{x}_{-d,i}|\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i) \right\} \mathbf{W}\boldsymbol{\varepsilon}_i \\
&\stackrel{\text{def}}{=} \sum_{k=1}^{4} \Psi_{n2,k}.
\end{aligned}$$

Obviously, $n^{-1/2}\Psi_{n2,1} \stackrel{d}{\longrightarrow} \mathcal{N}(0, \mathbf{B}_{\mathbf{w}})$, and $\Psi_{n2,k} = o_p(n^{-1/2})$, for $k = 2, 3, 4$. Thus the proof of Theorem 1 is completed. □

3.2. Proof of Theorem 2

Let

$$\begin{aligned} \boldsymbol{\Xi}_1 &= \mathbf{A}_{\mathbf{I}}^{-1}\Big\{\sum_{j=1}^{q}\mathbf{m}^{(1),\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\otimes\widetilde{\mathbf{x}}_{-d}\Big\}\boldsymbol{\Sigma}^{1/2} \text{ and}\\ \boldsymbol{\Xi}_2 &= \mathbf{A}_{\boldsymbol{\Sigma}^{-1}}^{-1}\Big\{\sum_{j=1}^{q}\mathbf{m}^{(1),\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\otimes\widetilde{\mathbf{x}}_{-d}\Big\}\boldsymbol{\Sigma}^{-1/2}. \end{aligned}$$

Simple calculation yields that $0 \le E\big\{(\boldsymbol{\Xi}_1-\boldsymbol{\Xi}_2)(\boldsymbol{\Xi}_1-\boldsymbol{\Xi}_2)^{\mathrm{T}}\big\} = \mathbf{A}_{\mathbf{I}}^{-1}\mathbf{B}_{\mathbf{I}}\mathbf{A}_{\mathbf{I}}^{-1} - \mathbf{A}_{\boldsymbol{\Sigma}^{-1}}^{-1}$, which together with $\mathbf{A}_{\boldsymbol{\Sigma}^{-1}} = \mathbf{B}_{\boldsymbol{\Sigma}^{-1}}$ yields the result of Theorem 2. □

3.3. Proof of Theorem 3

Using similar arguments to that in the proof of Theorem 1, we have

$$\begin{aligned} h^2\mathbf{S}_{n2}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}}) &= \boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)f(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\int u^sK(u)du + o_p(1),\\ h\boldsymbol{\xi}_{n1}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i,\widehat{\boldsymbol{\beta}}) &= n^{-1}\sum_{j=1}^{n}\mathbf{z}_j\otimes\Bigg[\Big(\frac{\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j-\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i}{h}\Big)\Big\{\sum_{k=1}^{q}\Big(\mathbf{m}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j)+\mathbf{m}^{(1)}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{x}_j\\ &+ o_p(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|)\Big)Z_{jk}+\boldsymbol{\varepsilon}_j\Big\}^{\mathrm{T}}\Bigg]K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j-\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\\ &= \mathbf{m}^{(1)}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)f(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\int u^sK(u)du + o_p(1), \end{aligned}$$

which together with $\mathbf{S}_{n1}(\mathbf{u},\boldsymbol{\beta}) = O_p(c_n)$ yields that $\widehat{\mathbf{m}}^{(1)}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i) = \mathbf{m}^{(1)}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) + o_p(1)$. Similarly, we can prove that $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + o_p(1)$, $\widehat{\mathbf{A}}_{\mathbf{w}} = \mathbf{A}_{\mathbf{w}} + o_p(1)$ and $\widehat{\mathbf{B}}_{\mathbf{w}} = \mathbf{B}_{\mathbf{w}} + o_p(1)$. Thus Theorem 3 follows. □

3.4. Proof of Theorem 4

Let

$$\mathcal{L}_1(d) = \sum_{i=1}^{n} \big\{\mathbf{y}_i - \sum_{k=1}^{q} \widehat{\mathbf{m}}_k(\widehat{\boldsymbol{\beta}}_{d,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) Z_{ik}\big\}^{\mathrm{T}} \big\{\mathbf{y}_i - \sum_{k=1}^{q} \widehat{\mathbf{m}}_k(\widehat{\boldsymbol{\beta}}_{d,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) Z_{ik}\big\}.$$

By definition,

$$\begin{aligned}\mathcal{L}_1(d) - \mathcal{L}_1(d_0) &= \sum_{i=1}^{n} \Big\{\mathbf{y}_i - \sum_{j=1}^{q} \widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) Z_{ij}\Big\}^{\mathrm{T}} \Big[\sum_{j=1}^{q} \Big\{\widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d_0,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) - \widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i)\Big\} Z_{ij}\Big] \\ &\quad + \sum_{i=1}^{n} \Big[\sum_{j=1}^{q} \Big\{\widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d_0,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) - \widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i)\Big\} Z_{ij}\Big]^{\mathrm{T}} \Big\{\mathbf{y}_i - \sum_{j=1}^{q} \widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d_0,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) Z_{ij}\Big\} \\ &\stackrel{\text{def}}{=} \Lambda_1 + \Lambda_2.\end{aligned}$$

Model (1.1) implies that $\mathbf{y}_i = \sum_{j=1}^{q} \mathbf{m}_j(\boldsymbol{\beta}_{d_0}^{\mathrm{T}} \mathbf{x}_i) Z_{ij} + \boldsymbol{\varepsilon}_i$. $E(\mathbf{y}_i \mid \boldsymbol{\beta}_d^{\mathrm{T}} \mathbf{x}_i, \mathbf{z}_i) \neq E(\mathbf{y}_i \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{d_0}, \mathbf{z}_i)$ if $d < d_0$ and $E(\mathbf{y}_i \mid \boldsymbol{\beta}_d^{\mathrm{T}} \mathbf{x}_i, \mathbf{z}_i) = E(\mathbf{y}_i \mid \boldsymbol{\beta}_{d_0}^{\mathrm{T}} \mathbf{x}_i, \mathbf{z}_i)$ otherwise. The proof in Theorem 1 implies that $\mathbf{m}(\boldsymbol{\beta}_d^{\mathrm{T}} \mathbf{x}_i) - \widehat{\mathbf{m}}(\boldsymbol{\beta}_d^{\mathrm{T}} \mathbf{x}_i) = O_p(c_n)$, and $\widehat{\mathbf{m}}(\boldsymbol{\beta}_d^{\mathrm{T}} \mathbf{x}_i) - \widehat{\mathbf{m}}(\widehat{\boldsymbol{\beta}}_{d,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) = o_p(n^{-1/2})$. If $d < d_0$,

$$\begin{aligned}\Lambda_2 &= \sum_{i=1}^{n} \Big[\sum_{j=1}^{q} \Big\{\widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d_0,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) - \widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i)\Big\} Z_{ij}\Big]^{\mathrm{T}} \\ &\quad \cdot \Big[\sum_{j=1}^{q} \Big\{\mathbf{m}_j(\widehat{\boldsymbol{\beta}}_{d_0,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i) - \widehat{\mathbf{m}}_j(\widehat{\boldsymbol{\beta}}_{d_0,\mathbf{w}}^{\mathrm{T}} \mathbf{x}_i)\Big\} Z_{ij} + \boldsymbol{\varepsilon}_i\Big] = o_p(n). \\ \Lambda_1 &= \sum_{i=1}^{n} \Big[\sum_{j=1}^{q} \big\{\mathbf{m}_j(\boldsymbol{\beta}_{d_0}^{\mathrm{T}} \mathbf{x}_i) - \mathbf{m}_j(\boldsymbol{\beta}_d^{\mathrm{T}} \mathbf{x}_i)\big\} Z_{ij}\Big]^{\mathrm{T}} \Big[\sum_{j=1}^{q} \big\{\mathbf{m}_j(\boldsymbol{\beta}_{d_0}^{\mathrm{T}} \mathbf{x}_i) - \mathbf{m}_j(\boldsymbol{\beta}_d^{\mathrm{T}} \mathbf{x}_i)\big\}\Big] \\ &\quad + o_p(n).\end{aligned}$$

The first term on the left side of the equation is $O_p(n)$, and is positive. Invoking condition (C3) and $\lambda_n n^{-1/2} \to 0$, we have

$$\mathcal{L}^*(d) - \mathcal{L}^*(d_0) = \{\mathcal{L}_1(d) - \mathcal{L}_1(d_0)\} \Big/ \left\{ \sum_{i=1}^n (\mathbf{y}_i - \overline{\mathbf{y}})^{\mathrm{T}} (\mathbf{y}_i - \overline{\mathbf{y}}) \right\}^{1/2} + p(d - d_0)\lambda_n$$
$$> 0, \text{ in probability, if } d < d_0.$$

Analogously, by condition (C3) and $\lambda_n / \log n \to \infty$, when $d > d_0$,

$$\mathcal{L}^*(d) - \mathcal{L}^*(d_0) = O_p(n^{1/2} h^{2s} + n^{-1/2} h^{-d} \log n) + p(d - d_0)\lambda_n > 0, \text{ in probability.}$$

Hence, $\mathrm{pr}(\widehat{d} = d_0)$ goes to 1 and the proof is completed. □

References

Carroll, R.J., Fan, J., Gijbels, I., and Wand, M.P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92(438)**, 477–489.

Cook, R.D. (1998). *Regression Graphics*. New York: Wiley.

Cook, R.D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science*, **22(1)**, 1–26.

Cook, R.D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, **23(4)**, 485–501.

Cook, R.D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104(485)**, 197–208.

Feng, Z., Wen, X.M., Yu, Z., and Zhu, L. (2013). On partial sufficient dimension reduction with applications to partially linear multi-index models. *Journal of the American Statistical Association*, **108(501)**, 237–246.

Li, B., Cook, R.D., and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Annals of Statistics*, **31(5)**, 1636–1668.

Liu, X., Cui, Y., and Li, R. (2016). Partial linear varying multi-index coefficient model for integrative gene-environment interactions. *Statistica Sinica*, **26**, 1037–1060.

Ma, S. and Song, P.X.K. (2015). Varying index coefficient models. *Journal of the American Statistical Association*, **110(509)**, 341–356.

Mack, Y. and Silverman, B.W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61(3)**, 405–415.

Xia, Y. and Li, W. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association*, **94(448)**, 1275–1285.

Zhu, L., Wang, T., Zhu, L., and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97(2)**, 295–304.