Nonparametric Interaction Selection

Yushen Dong and Yichao Wu

University of Illinois at Chicago

Supplementary Material

Supplementary Material contains implementation codes, technical conditions, and proofs of Dong and Wu (2020).

S0 Implementation codes

NIScodes.zip contains all R and supporting C codes for implementation of our nonparametric interaction selection, as well as a demo code for one simulation example to illustrate how the codes are run.

S1 Technical conditions

Condition 1. The kernel function $K(x)$ is bounded and Lipschitz-continuous with a bounded support.

Condition 2. The density function $f_j(x_j)$ of X_j is Lipschitz-continuous and bounded away from 0, and has a bounded support Ω_j for $j = 1 \dots d$.

Condition 3. For an arbitrary quadruple $(X_{ji_1}, X_{ji_2}, X_{ji_3}, X_{ji_4})$, the joint density of any two, three, or all of them is Lipschitz-continuous on its support.

Condition 4. For all main and interaction components, $m_j(\cdot)$ and $m_{jk}(\cdot,\cdot)$, $1 \leq j,k \leq d$, their 1st derivatives (or partial derivatives) exist, and are bounded and continuous.

Condition 5. The random error has a finite fourth moment, $E(|\epsilon|^4) < \infty$.

S2 Normal Equation

In this section, we solve the additive model (1.1) with identifiability condition (1.2) and (1.3) in a theoretical way. Let $\mathcal{H}$ be the space of square-integrable functions of $X_1, X_2, \ldots, X_d$. For each $j = 1, 2, \ldots, d$, $\mathcal{H}_j$ denotes the Hilbert spaces of univariate square-integrable functions $\phi(\cdot)$ satisfying $\phi(x_{j,0}) = 0$ with inner product $\langle \phi_1, \phi_2 \rangle = E(\phi_1(X_j)\phi_2(X_j))$ for any $\phi_1, \phi_2 \in \mathcal{H}_j$. For each $1 \leq j < k \leq d$, $\mathcal{H}_{jk}$ denotes the Hilbert space of bivariate square-integrable function $\psi(\cdot,\cdot)$ satisfying $\psi(x_{j,0}, \cdot) = 0$, $\psi_{jk}(\cdot, x_{k,0}) = 0$, $\psi(x_{j,0}, x_{k,0}) = 0$ with inner product $\langle \psi_1, \psi_2 \rangle = E(\psi_1(X_j, X_k)\psi_2(X_j, X_k))$ for any $\psi_1, \psi_2 \in \mathcal{H}_{jk}$. Obviously, $\mathcal{H}_j$ $(j = 1, 2, \ldots, d)$ and $\mathcal{H}_{jk}$ $(1 \leq j < k \leq d)$ are subspaces of $\mathcal{H}$. Moreover, $\mathcal{H}^{\oplus} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \ldots \oplus \mathcal{H}_d \oplus \mathcal{H}_{1,2} \oplus \mathcal{H}_{1,3} \oplus \ldots \oplus \mathcal{H}_{(d-1)d}$ is a subspace of $\mathcal{H}$ and is closed under some technical assumptions. These Hilbert spaces are also subspaces of $\mathcal{H}_{YX}$, the space of square-integrable functions of Y, X_1, .., X_d.

The optimization problem is to minimize the mean squared error $E(Y-\alpha-m(\boldsymbol{X}))^2$ with constraint $m(\boldsymbol{X})=\sum_{j=1}^d m_j(X_j)+\sum_{1\le j<k\le d} m_{jk}(X_j,X_k)\in\mathcal{H}^{\oplus}$. Denote the conditional expectation operators $E(\cdot|X_j)-E(\cdot|X_j=x_{j,0})$ and $E(\cdot|X_j,X_k)-E(\cdot|X_j=x_{j,0},X_k)-E(\cdot|X_j,X_k=x_{k,0})+E(\cdot|X_j=x_{j,0},X_k=x_{k,0})$ on $\mathcal{H}_{YX}$ by P_j and P_{jk}, and they are orthogonal projections onto $\mathcal{H}_j$ and $\mathcal{H}_{jk}$. We also denote the expectation operator $E(\cdot)$ on $\mathcal{H}_{YX}$ by P_0, a project onto the space of constant functions.

Denote the minimizer of $m(\boldsymbol{X})$ in the aforementioned optimization problem by $\widehat{m}(\boldsymbol{X})$. Then the minimizer of intercept α is $\widehat{\alpha}=E(Y)-E(\widehat{m}(\boldsymbol{X}))=P_0(Y-\sum_{j=1}^d m_j(X_j)-\sum_{1\le j<k\le d} m_{jk}(X_j,X_k))$. The residual $Y-\widehat{\alpha}-\widehat{m}(\boldsymbol{X})$ is orthogonal to $\mathcal{H}^{\oplus}$. Consequently $Y-\widehat{\alpha}-\widehat{m}(\boldsymbol{X})$ is orthogonal to $\mathcal{H}_j$, $j=1,\dots,d$ and $\mathcal{H}_{jk}$, $1\le j<k\le d$. Equivalently, we have $P_j(Y-\widehat{\alpha}-\widehat{m}(\boldsymbol{X}))=0$, $j=1,\dots,d$ and $P_{jk}(Y-\widehat{\alpha}-\widehat{m}(\boldsymbol{X}))=0$, $1\le j<k\le d$. The equations can be rewritten as

$$m_j(X_j)=P_j(Y-\widehat{\alpha}-\sum_{s\neq j}\widehat{m}_s(X_s)-\sum_{1\le s<t\le d}\widehat{m}_{st}(X_s,X_t))$$

for each main effect term and

$$m_{jk}(X_j,X_k)=P_{jk}(Y-\widehat{\alpha}-\sum_{s=1}^d\widehat{m}_s(X_s)-\sum_{s<t\,:\,(s,t)\neq(j,k)}\widehat{m}_{st}(X_s,X_t))$$

for each interaction effect term.

Then we have the following normal equation for α and $(m_1, \ldots, m_d, m_{1,2}, \ldots, m_{(d-1)d})^T$ to minimize MSE:

$$\begin{bmatrix} 1 & P_0 & P_0 & \cdots & P_0 & \cdots & P_0 \\ P_1 & 1 & P_1 & \cdots & P_1 & \cdots & P_1 \\ P_2 & P_2 & 1 & \cdots & P_2 & \cdots & P_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_d & P_d & P_d & \cdots & 1 & \cdots & P_d \\ P_{1,2} & P_{1,2} & P_{1,2} & \cdots & P_{1,2} & \cdots & P_{1,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{(d-1)d} & P_{(d-1)d} & P_{(d-1)d} & \cdots & P_{(d-1)d} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ m_1 \\ m_2 \\ \vdots \\ m_d \\ m_{1,2} \\ \vdots \\ m_{(d-1)d} \end{bmatrix} = \begin{bmatrix} P_0 \\ P_1 \\ P_2 \\ \vdots \\ P_d \\ P_{1,2} \\ \vdots \\ P_{(d-1)d} \end{bmatrix} Y, \qquad \text{(S2.1)}$$

where $P_{j,k}$ is the same as P_{jk} and is used to avoid potential confusion.

S3 Backfitting estimator

Denote $\boldsymbol{m}_j = (m_j(X_{1j}), \ldots, m_j(X_{nj}))^T$, $\boldsymbol{m}_{jk} = (m_{jk}(X_{1j}, X_{1k}), \ldots, m_{jk}(X_{nj}, X_{nk}))^T$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $K_h(x) = h^{-1}K(\frac{x}{h})$. Recall that the smoothing matrices for local constant regression are

$\mathbf{S}_j = (\mathbf{s}_j(X_{1j}), \ldots, \mathbf{s}_j(X_{nj}))^T$ for main effect term;

$\tilde{\mathbf{S}}_{jk} = (\tilde{\mathbf{s}}_{jk}(X_{1j}, X_{1k}), \ldots, \tilde{\mathbf{s}}_{jk}(X_{nj}, X_{nk}))^T$ for interaction effect term;

$\tilde{\mathbf{S}}_{j0k} = (\tilde{\mathbf{s}}_{jk}(x_{j,0}, X_{1k}), \ldots, \tilde{\mathbf{s}}_{jk}(x_{j,0}, X_{nk}))^T$,

$\tilde{\mathbf{S}}_{jk0} = (\tilde{\mathbf{s}}_{jk}(X_{1j}, x_{k,0}), \ldots, \tilde{\mathbf{s}}_{jk}(X_{nj}, x_{k,0}))^T$ for shifting.

Here

$$\mathbf{s}_j(x_j) = \left(K_{h_j}(X_{1j}-x_j), \ldots, K_{h_j}(X_{nj}-x_j)\right)^T / \sum_{i=1}^{n} K_{h_j}(X_{ij}-x_j)$$

and

$$\tilde{\mathbf{s}}_{jk}(x_j, x_k) = \frac{1}{\sum_{i=1}^{n} K_{\tilde{h}_{jk}}(X_{ij}-x_j)K_{\tilde{h}_{jk}}(X_{ik}-x_k)} \begin{pmatrix} K_{\tilde{h}_{jk}}(X_{1j}-x_j)K_{\tilde{h}_{jk}}(X_{1k}-x_k) \\ K_{\tilde{h}_{jk}}(X_{2j}-x_j)K_{\tilde{h}_{jk}}(X_{2k}-x_k) \\ \vdots \\ K_{\tilde{h}_{jk}}(X_{nj}-x_j)K_{\tilde{h}_{jk}}(X_{nk}-x_k) \end{pmatrix}.$$

We now define the smoothing matrices after the shifting

$$\mathbf{S}_j^* = \mathbf{S}_j - \mathbf{1}(\mathbf{s}_j(x_{j,0}))^T \text{ and } \tilde{\mathbf{S}}_{jk}^* = \tilde{\mathbf{S}}_{jk} - \tilde{\mathbf{S}}_{j0k} - \tilde{\mathbf{S}}_{jk0} + \mathbf{1}(\tilde{\mathbf{s}}_{jk}(x_{j,0}, x_{k,0}))^T$$

for each main effect and interaction effect term, respectively. The main term $\boldsymbol{m}_i$ and interaction term $\boldsymbol{m}_{jk}$ can be estimated through the solutions to the sample version of a part of the normal equation (S2.1)

$$\begin{bmatrix} \boldsymbol{I}_n & \mathbf{S}_1^* & \cdots & \mathbf{S}_1^* & \mathbf{S}_1^* & \cdots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \boldsymbol{I}_n & \cdots & \mathbf{S}_2^* & \mathbf{S}_2^* & \cdots & \mathbf{S}_2^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d^* & \mathbf{S}_d^* & \cdots & \boldsymbol{I}_n & \mathbf{S}_d^* & \cdots & \mathbf{S}_d^* \\ \tilde{\mathbf{S}}_{1,2}^* & \tilde{\mathbf{S}}_{1,2}^* & \cdots & \tilde{\mathbf{S}}_{1,2}^* & \boldsymbol{I}_n & \cdots & \tilde{\mathbf{S}}_{1,2}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{S}}_{(d-1)d}^* & \tilde{\mathbf{S}}_{(d-1)d}^* & \cdots & \tilde{\mathbf{S}}_{(d-1)d}^* & \tilde{\mathbf{S}}_{(d-1)d}^* & \cdots & \boldsymbol{I}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \\ \vdots \\ \boldsymbol{m}_d \\ \boldsymbol{m}_{1,2} \\ \vdots \\ \boldsymbol{m}_{(d-1)d} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_d^* \\ \tilde{\mathbf{S}}_{1,2}^* \\ \vdots \\ \tilde{\mathbf{S}}_{(d-1)d}^* \end{bmatrix} \boldsymbol{Y} \qquad \text{(S3.1)}$$

by noting $P_j\alpha = 0$ and $P_{jk}\alpha = 0$. Similarly $\tilde{\mathbf{S}}_{j,k}^*$ is the same as $\tilde{\mathbf{S}}_{jk}^*$.

In practice, the backfitting algorithm is used to solve the normal equation (S3.1) and the backfitting estimators converge to the solution

$$\begin{bmatrix} \widehat{\boldsymbol{m}}_1 \\ \widehat{\boldsymbol{m}}_2 \\ \vdots \\ \widehat{\boldsymbol{m}}_d \\ \widehat{\boldsymbol{m}}_{1,2} \\ \vdots \\ \widehat{\boldsymbol{m}}_{(d-1)d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_n & \mathbf{S}_1^* & \dots & \mathbf{S}_1^* & \mathbf{S}_1^* & \dots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \boldsymbol{I}_n & \dots & \mathbf{S}_2^* & \mathbf{S}_2^* & \dots & \mathbf{S}_2^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d^* & \mathbf{S}_d^* & \dots & \boldsymbol{I}_n & \mathbf{S}_d^* & \dots & \mathbf{S}_d^* \\ \tilde{\mathbf{S}}_{1,2}^* & \tilde{\mathbf{S}}_{1,2}^* & \dots & \tilde{\mathbf{S}}_{1,2}^* & \boldsymbol{I}_n & \dots & \tilde{\mathbf{S}}_{1,2}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{S}}_{(d-1)d}^* & \tilde{\mathbf{S}}_{(d-1)d}^* & \dots & \tilde{\mathbf{S}}_{(d-1)d}^* & \tilde{\mathbf{S}}_{(d-1)d}^* & \dots & \boldsymbol{I}_n \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_d^* \\ \tilde{\mathbf{S}}_{1,2}^* \\ \vdots \\ \tilde{\mathbf{S}}_{(d-1)d}^* \end{bmatrix} \boldsymbol{Y}$$

$$\equiv \boldsymbol{M}^{-1}\boldsymbol{C}\boldsymbol{Y},$$

provided that the inverse of $\boldsymbol{M}$ exist. The intercept α can be estimated by $\overline{\boldsymbol{Y}} - \sum_{j=1}^{d} \overline{\widehat{\boldsymbol{m}}}_i - \sum_{1 \le j < k \le d} \overline{\widehat{\boldsymbol{m}}}_{jk}$, where $\overline{\boldsymbol{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

For the convenience of presentation, in the following we will not distinguish main effect term and interaction effect term if they share a same property. In that case, putting them together we have $\check{d} = d + \frac{d(d-1)}{2}$ terms in total. We use $\check{\mathbf{S}}_j$ to denote the (either univariate or bivariate) local constant smoothing matrix, $\check{\mathbf{S}}_j^*$ to denote the local constant smoothing matrix after shifting, $\check{\boldsymbol{m}}_j$ to denote the component function, and $\check{h}_j$ to denote the bandwidth for each of these $\check{d}$ terms.

Then the backfitting smoothing matrix for term $j \in \{1, 2, \dots, \check{d}\}$ is defined by $\boldsymbol{W}_j = \boldsymbol{E}_j \boldsymbol{M}^{-1} \boldsymbol{C}$. Here $\boldsymbol{E}_j$ is a zero matrix of dimension $n \times n\check{d}$ except its jth block being an $n \times n$ identity matrix when treated as blocks of $n \times n$ submatrices. The backfitting estimator is given by $\widehat{\check{\boldsymbol{m}}}_j = \boldsymbol{W}_j \boldsymbol{Y}$. We also define $\boldsymbol{m} = \sum_{j=1}^{\check{d}} \check{\boldsymbol{m}}_j$, $\boldsymbol{W} = \sum_{j=1}^{\check{d}} \boldsymbol{W}_j$, and the backfitting estimator

$\widehat{\boldsymbol{m}} = \boldsymbol{W}\boldsymbol{Y}$.

To state the conditions for the existence and uniqueness of the backfitting estimators, we define $\boldsymbol{W}^{[-j]}$ as the smoothing matrix for the model without the jth term. Then if $||\check{\mathbf{S}}_j^* \boldsymbol{W}^{[-j]}|| < 1$ for $j = 1, \ldots, \check{d}$ and a matrix norm $||\cdot||$, by Corollary 4.3 of Buja et al. (1989) and Lemma 2.1 of Opsomer (2000), the backfitting estimators exist, are unique and

$$\begin{aligned} \boldsymbol{W}_j &= \boldsymbol{I}_n - (\boldsymbol{I}_n - \mathbf{S}_j^* \boldsymbol{W}^{[-j]})^{-1} (\boldsymbol{I}_n - \mathbf{S}_j^*) \\ &= (\boldsymbol{I}_n - \mathbf{S}_j^* \boldsymbol{W}^{[-j]})^{-1} \mathbf{S}_j^* (\boldsymbol{I}_n - \boldsymbol{W}^{[-j]}). \end{aligned} \tag{S3.2}$$

S4 PROOFS

Lemma 1. Under Conditions 1-3 in Appendix, if $\check{h}_j \to 0$ and $n\check{h}_j^2 \to \infty$ as $n \to \infty$, the following asymptotic approximation holds uniformly over all elements of the matrices

$$\check{\mathbf{S}}_j^* = \check{\mathbf{S}}_j - \mathbf{1}\mathbf{1}^T/n + o_p(\mathbf{1}\mathbf{1}^T/n) \ .$$

Proof. For main effect term, recall that

$$\mathbf{s}_j(x_j) = \left(K_{h_j}(X_{1j} - x_j), \ldots, K_{h_j}(X_{nj} - x_j)\right)^T / \sum_{i=1}^n K_{h_j}(X_{ij} - x_j).$$

Consider $\sum_{i=1}^{n} K_{h_j}(X_{ij}-x_j)$, we have

$$
\begin{aligned}
E(\sum_{i=1}^{n} K_{h_j}(X_{ij}-x_j)) &= \sum_{i=1}^{n}\int \frac{1}{h_j}K(\frac{t-x_j}{h_j})f_j(t)dt \\
&= n\int K(u)f_j(x_j+h_ju)du \\
&= n\int K(u)(f_j(x_j)+O(h_j))du \\
&= n(f_j(x_j)+O(h_j))
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{i=1}^{n} E((K_{h_j}(X_{ij}-x_j))^2) &= \sum_{i=1}^{n}\int \frac{1}{h_j^2}K^2(\frac{t-x_j}{h_j})f_j(t)dt \\
&= \sum_{i=1}^{n}\int \frac{1}{h_j}K^2(u)f_j(x_j+h_ju)du \\
&= \frac{n}{h_j}\int K^2(u)(f_j(x_j)+O(h_j))du \\
&= \frac{n}{h_j}O(1).
\end{aligned}
$$

Consequently we have

$$
\begin{aligned}
\sum_{i=1}^{n} K_{h_j}(X_{ij}-x_j) &= n(f_j(x_j)+O(h_j))+O_p(\sqrt{\frac{n}{h_j}}) \\
&= nf_j(x_j)(1+O(h_j)+O_p(\sqrt{\frac{1}{nh_j}})) \\
&= nf_j(x_j)(1+o_p(1)).
\end{aligned}
$$

As defined before, $\mathbf{S}_j^* = \mathbf{S}_j - \mathbf{1}(\mathbf{s}_j(x_{j,0}))^T$. For vector $(\mathbf{s}_j(x_{j,0}))^T$, its sth entry denoted by $s_{j,s}(x_{j,0})$ is

$$\begin{aligned} s_{j,s}(x_{j,0}) &= (\sum_{t=1}^{n} K_{h_j}(X_{tj} - x_{j,0}))^{-1} K_{h_j}(X_{sj} - x_{j,0}) \\ &= \frac{1}{n}(f_j(x_{j,0})^{-1} K_{h_j}(X_{sj} - x_{j,0}) + o_p(1)). \end{aligned}$$

Similarly, we have

$$\begin{aligned} E(f_j(x_{j,0})^{-1} K_{h_j}(X_{sj} - x_{j,0})) &= \int \frac{1}{h_j} f_j(x_{j,0})^{-1} f_j(t) K(\frac{t - x_{j,0}}{h_j}) dt \\ &= \int f_j(x_{j,0})^{-1} f_j(x_{j,0} + h_j u) K(u) du \\ &= \int f_j(x_{j,0})^{-1} (f_j(x_{j,0}) + O(h_j)) K(u) du \\ &= 1 + O(h_j) \end{aligned}$$

and

$$\begin{aligned} E((f_j(x_{j,0})^{-1} K_{h_j}(X_{sj} - x_{j,0}))^2) &= \int \frac{1}{h_j^2} f_j(x_{j,0})^{-2} f_j(t) K^2(\frac{t - x_{j,0}}{h_j}) dt \\ &= \int \frac{1}{h_j} f_j(x_{j,0})^{-2} f_j(x_{j,0} + h_j u) K^2(u) du \\ &= \frac{1}{h_j} O(1). \end{aligned}$$

Then we have $f_j(x_{j,0})^{-1} K_{h_j}(X_{sj} - x_{j,0}) = 1 + O(h_j) + O_p(\sqrt{\frac{1}{h_j}})$.

Combining the results before, we get

$$\begin{aligned} s_{j,s}(x_{j,0}) &= \frac{1}{n}(f_j(x_{j,0})^{-1} K_{h_j}(X_{sj} - x_{j,0}) + o_p(1)). \\ &= \frac{1}{n}(1 + O(h_j) + O_p(\sqrt{\frac{1}{h_j}}) + o_p(1)) \end{aligned}$$

$$= \frac{1}{n} + o_p(\frac{1}{n}).$$

Consequently we have shown that $\mathbf{S}_j^* = \mathbf{S}_j - \mathbf{1}\mathbf{1}^T/n + o_p(\mathbf{1}\mathbf{1}^T/n)$.

For interaction effect term,

$$\tilde{\mathbf{s}}_{jk}(x_j, x_k) = \frac{1}{\sum_{i=1}^n K_{\tilde{h}_{jk}}(X_{ij}-x_j)K_{\tilde{h}_{jk}}(X_{ik}-x_k)} \begin{pmatrix} K_{\tilde{h}_{jk}}(X_{1j}-x_j)K_{\tilde{h}_{jk}}(X_{1k}-x_k) \\ K_{\tilde{h}_{jk}}(X_{2j}-x_j)K_{\tilde{h}_{jk}}(X_{2k}-x_k) \\ \vdots \\ K_{\tilde{h}_{jk}}(X_{nj}-x_j)K_{\tilde{h}_{jk}}(X_{nk}-x_k) \end{pmatrix}.$$

Consider $\sum_{i=1}^n K_{\tilde{h}_{jk}}(X_{ij}-x_j)K_{\tilde{h}_{jk}}(X_{ik}-x_k)$, we have

$$\begin{aligned} & E(\sum_{i=1}^n K_{\tilde{h}_{jk}}(X_{ij}-x_j)K_{\tilde{h}_{jk}}(X_{ik}-x_k)) \\ = & \sum_{i=1}^n \iint \frac{1}{\tilde{h}_{jk}^2} K(\frac{s-x_j}{\tilde{h}_{jk}}) K(\frac{t-x_k}{\tilde{h}_{jk}}) f_{jk}(s,t) ds dt \\ = & \ n \iint K(u)K(v) f_{jk}(x_j + \tilde{h}_{jk}u, x_k + \tilde{h}_{jk}v) du dv \\ = & \ n \iint K(u)K(v)(f_{jk}(x_j, x_k) + O(\tilde{h}_{jk})) du dv \\ = & \ n(f_{jk}(x_j,x_k) + O(\tilde{h}_{jk})) \end{aligned}$$

and

$$\begin{aligned} & \sum_{i=1}^n E((K_{\tilde{h}_{jk}}(X_{ij}-x_j)K_{\tilde{h}_{jk}}(X_{ik}-x_k))^2) \\ = & \sum_{i=1}^n \iint \frac{1}{\tilde{h}_{jk}^4} K^2(\frac{s-x_j}{\tilde{h}_{jk}}) K^2(\frac{t-x_k}{\tilde{h}_{jk}}) f_{jk}(s,t) ds dt \\ = & \ n \iint \frac{1}{\tilde{h}_{jk}^2} K^2(u) K^2(v) f_{jk}(x_j + \tilde{h}_{jk}u, x_k + \tilde{h}_{jk}v) du dv \end{aligned}$$

$$
\begin{aligned}
&= \frac{n}{\tilde{h}_{jk}^2} \iint K^2(u)K^2(v)(f_{jk}(x_j, x_k) + O(\tilde{h}_{jk}))dudv \\
&= \frac{n}{\tilde{h}_{jk}^2} O(1).
\end{aligned}
$$

Then

$$
\begin{aligned}
\sum_{i=1}^{n} K_{\tilde{h}_{jk}}(X_{ij} - x_j)K_{\tilde{h}_{jk}}(X_{ik} - x_k) &= n(f_{jk}(x_j, x_k) + O(\tilde{h}_{jk})) + O_p(\sqrt{\frac{n}{\tilde{h}_{jk}^2}}) \\
&= n(f_{jk}(x_j, x_k) + O(\tilde{h}_{jk}) + O_p(\sqrt{\frac{1}{n\tilde{h}_{jk}^2}})) \\
&= n(f_{jk}(x_j, x_k) + o_p(1)).
\end{aligned}
$$

For $\tilde{\mathbf{S}}_{jk}^* = \tilde{\mathbf{S}}_{jk} - \tilde{\mathbf{S}}_{j0k} - \tilde{\mathbf{S}}_{jk0} + \mathbf{1}(\tilde{\mathbf{s}}_{jk}(x_{j,0}, x_{k,0}))^T$ as defined above, its entry in the ith row and the sth column, $\tilde{S}_{jk,(i,s)}^*$, is

$$
\tilde{S}_{jk,(i,s)}^* = \tilde{S}_{jk,(i,s)} - \tilde{S}_{j0k,(i,s)} - \tilde{S}_{jk0,(i,s)} + \tilde{s}_{jk,i}(x_{j,0}, x_{k,0}),
$$

where

$$
\begin{aligned}
\tilde{S}_{j0k,(i,s)} &= (\sum_{t=1}^{n} K_{\tilde{h}_{jk}}(X_{tj} - x_{j,0})K_{\tilde{h}_{jk}}(X_{tk} - X_{ik}))^{-1} K_{\tilde{h}_{jk}}(X_{sj} - x_{j,0})K_{\tilde{h}_{jk}}(X_{sk} - X_{ik}) \\
&= \frac{1}{n}(f_{jk}(x_{j,0}, X_{ik})^{-1} K_{\tilde{h}_{jk}}(X_{sj} - x_{j,0})K_{\tilde{h}_{jk}}(X_{sk} - X_{ik}) + o_p(1)), \\
\tilde{S}_{jk0,(i,s)} &= (\sum_{t=1}^{n} K_{\tilde{h}_{jk}}(X_{tj} - X_{ij})K_{\tilde{h}_{jk}}(X_{tk} - x_{k,0}))^{-1} K_{\tilde{h}_{jk}}(X_{sj} - X_{ij})K_{\tilde{h}_{jk}}(X_{sk} - x_{k,0}) \\
&= \frac{1}{n}(f_{jk}(X_{ij}, x_{k,0})^{-1} K_{\tilde{h}_{jk}}(X_{sj} - X_{ij})K_{\tilde{h}_{jk}}(X_{sk} - x_{k,0}) + o_p(1)), \\
\tilde{s}_{jk,i}(x_{j,0}, x_{k,0}) &= (\sum_{t=1}^{n} K_{\tilde{h}_{jk}}(X_{tj} - x_{j,0})K_{\tilde{h}_{jk}}(X_{tk} - x_{k,0}))^{-1} K_{\tilde{h}_{jk}}(X_{sj} - x_{j,0})K_{\tilde{h}_{jk}}(X_{sk} - x_{k,0}) \\
&= \frac{1}{n}(f_{jk}(x_{j,0}, x_{k,0})^{-1} K_{\tilde{h}_{jk}}(X_{sj} - x_{j,0})K_{\tilde{h}_{jk}}(X_{sk} - x_{k,0}) + o_p(1)).
\end{aligned}
$$

By similar argument as in the main effect term case, we obtain

$$\tilde{S}_{j0k,(i,s)} = \frac{1}{n} + o_p(\frac{1}{n}), \quad \tilde{S}_{jk0,(i,s)} = \frac{1}{n} + o_p(\frac{1}{n}), \quad \tilde{s}_{jk,i}(x_{j,0}, x_{k,0}) = \frac{1}{n} + o_p(\frac{1}{n}).$$

Combining the results before, we get

$$S^*_{jk,(i,s)} = S_{jk,(i,s)} - \frac{1}{n} + o_p(\frac{1}{n}).$$

In matrix form, we have $\mathbf{S}^*_{jk} = \mathbf{S}_{jk} - \mathbf{1}\mathbf{1}^T/n + o_p(\mathbf{1}\mathbf{1}^T/n)$ as desired. □

Lemma 2. Under Conditions 1-3 in Appendix, if $\check{h}_j \to 0$ and $n\check{h}_j^4 \to \infty$ as $n \to \infty$, the following asymptotic approximations hold for all main effect and interaction effect terms:

$$\check{\mathbf{S}}^*_j \boldsymbol{W}^{[-j]} = O_p(\mathbf{1}\mathbf{1}^T/n)$$
$$(\boldsymbol{I}_n - \check{\mathbf{S}}^*_j \boldsymbol{W}^{[-j]})^{-1} = \boldsymbol{I}_n + O_p(\mathbf{1}\mathbf{1}^T/n) \ .$$

Proof. We calculate $\check{\mathbf{S}}_j \check{\mathbf{S}}_{j'}$ first. There are three cases: both j and j' are main effect terms; both of them are interaction effect terms; there are one main effect and one interaction effect term. We only show the calculation for case one in detail. The other two cases can be showed by similar arguments.

If both j and j' are main terms, the entry in sth row and tth column,

$[\mathbf{S}_j\mathbf{S}_{j'}]_{st}$, is

$$
\begin{aligned}
[\mathbf{S}_j\mathbf{S}_{j'}]_{st} = &\left\{ \begin{bmatrix} \mathbf{s}_j(X_{1j})^T \\ \vdots \\ \mathbf{s}_j(X_{nj})^T \end{bmatrix} \begin{bmatrix} \mathbf{s}_{j'}(X_{1j'})^T \\ \vdots \\ \mathbf{s}_{j'}(X_{nj'})^T \end{bmatrix} \right\}_{st} \\
=&(\sum_{i=1}^{n}(K_{h_j}(X_{ij}-X_{sj})))^{-1} \begin{bmatrix} K_{h_j}(X_{1j}-X_{sj}) \\ \vdots \\ K_{h_j}(X_{nj}-X_{sj}) \end{bmatrix}^T \begin{bmatrix} (\sum_{i=1}^{n}(K_{h_{j'}}(X_{ij'}-X_{1j'})))^{-1}K_{h_{j'}}(X_{1j'}-X_{tj'}) \\ \vdots \\ (\sum_{i=1}^{n}(K_{h_{j'}}(X_{ij'}-X_{nj'})))^{-1}K_{h_{j'}}(X_{nj'}-X_{tj'}) \end{bmatrix} \\
=&(\sum_{i=1}^{n}(K_{h_j}(X_{ij}-X_{sj})))^{-1} \\
&\sum_{i'=1}^{n}((\sum_{i=1}^{n}(K_{h_{j'}}(X_{ij'}-X_{i'j'})))^{-1}K_{h_j}(X_{i'j}-X_{sj})K_{h_{j'}}(X_{i'j'}-X_{tj'})) \\
=&\frac{1}{n}(f_j(X_{sj})^{-1}+o_p(1))(\sum_{i'=1}^{n}(\frac{1}{n}f_{j'}(X_{i'j'})^{-1}K_{h_j}(X_{i'j}-X_{sj})K_{h_{j'}}(X_{i'j'}-X_{tj'}))+o_p(1)).
\end{aligned}
$$

Then as before, we have

$$
\begin{aligned}
&E(\sum_{i'=1}^{n}(\frac{1}{n}f_{j'}(X_{i'j'})^{-1}K_{h_j}(X_{i'j}-X_{sj})K_{h_{j'}}(X_{i'j'}-X_{tj'}))) \\
= \ & \frac{1}{n}\sum_{i'=1}^{n}\iint K_{h_j}(z_1-X_{sj})K_{h_{j'}}(z_2-X_{tj'})f_{j'}(z_2)^{-1}f_{jj'}(z_1,z_2)dz_1dz_2 \\
= \ & \iint K(u)K(v)f_{j'}(X_{tj'}+h_{j'}v)^{-1}f_{jj'}(X_{sj}+h_ju,X_{tj'}+h_{j'}v)dudv \\
= \ & \iint K(u)K(v)(f_{j'}(X_{tj'})^{-1}f_{jj'}(X_{sj},X_{tj'})+O(h_j+h_{j'}))dudv \\
= \ & f_{j'}(X_{tj'})^{-1}f_{jj'}(X_{sj},X_{tj'})+O(h_j+h_{j'})
\end{aligned}
$$

and

$$
\sum_{i'=1}^{n}E(((\frac{1}{n}f_{j'}(X_{i'j'})^{-1}K_{h_j}(X_{i'j}-X_{sj})K_{h_{j'}}(X_{i'j'}-X_{tj'}))^2)
$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i'=1}^{n}\iint \frac{1}{h_j^2 h_{j'}^2}K^2(\frac{z_1-X_{sj}}{h_j})K^2(\frac{z_2-X_{tj'}}{h_{j'}})f_{j'}(z_2)^{-2}f_{jj'}(z_1,z_2)dz_1dz_2 \\
&= \frac{1}{nh_jh_{j'}}\iint K^2(u)K^2(v)f_{j'}(X_{tj'}+h_{j'}v)^{-1}f_{jj'}(X_{sj}+h_ju,X_{tj'}+h_{j'}v)dudv \\
&= \frac{1}{nh_jh_{j'}}O(1).
\end{aligned}
$$

Consequently we have

$$
\begin{aligned}
&\sum_{i'=1}^{n}(\frac{1}{n}f_{j'}(X_{i'j'})^{-1}K_{h_j}(X_{i'j}-X_{sj})K_{h_{j'}}(X_{i'j'}-X_{tj'})) \\
&= f_{j'}(X_{tj'})^{-1}f_{jj'}(X_{sj},X_{tj'})+O(h_j+h_{j'})+O_p(\sqrt{\frac{1}{nh_jh_{j'}}}) \\
&= f_{j'}(X_{tj'})^{-1}f_{jj'}(X_{sj},X_{tj'})+o_p(1)
\end{aligned}
$$

and

$$
\begin{aligned}
[\mathbf{S}_j\mathbf{S}_{j'}]_{st} &= \frac{1}{n}(f_j(X_{sj})^{-1}+o_p(1))(\sum_{i'=1}^{n}(\frac{1}{n}f_{j'}(X_{i'j'})^{-1}K_{h_j}(X_{i'j}-X_{sj})K_{h_{j'}}(X_{i'j'}-X_{tj'}))+o_p(1)) \\
&= \frac{1}{n}(f_j(X_{sj})^{-1}+o_p(1))(f_{j'}(X_{tj'})^{-1}f_{jj'}(X_{sj},X_{tj'})+o_p(1))+o_p(1)) \\
&= \frac{1}{n}(f_j(X_{sj})^{-1}f_{j'}(X_{tj'})^{-1}f_{j'}(X_{sj},X_{tj'})+o_p(1)).
\end{aligned}
$$

Applying similar arguments to the case with two interactions effect terms and the case with one main effect and one interaction effect term, we get

$$
[\mathbf{S}_j\mathbf{S}_{j'k}]_{st} = \frac{1}{n}(f_j(X_{sj})^{-1}f_{j'k}(X_{tj'},X_{tk})^{-1}f_{jj'k}(X_{sj},X_{tj'},X_{tk})+o_p(1))
$$

and

$$
[\mathbf{S}_{jk}\mathbf{S}_{j'k'}]_{st} = \frac{1}{n}(f_{jk}(X_{sj},X_{sk})^{-1}f_{j'k'}(X_{tj'},X_{tk'})^{-1}f_{jj'kk'}(X_{sj},X_{sk},X_{tj'},X_{tk'})+o_p(1)).
$$

Then by Lemma 1 and a same argument of Lemma 3.2 of Opsomer and Ruppert (1997), we obtain $(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\check{\mathbf{S}}_{j'}^*)^{-1} = \boldsymbol{I}_n + O_p(\mathbf{1}\mathbf{1}^T/n)$ for all main effect terms and interaction effect terms.

By a similar argument of Theorem 3.1 of Opsomer (2000), we obtain

$$\check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]} = O_p(\mathbf{1}\mathbf{1}^T/n)$$

and

$$(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1} = \boldsymbol{I}_n + O_p(\mathbf{1}\mathbf{1}^T/n),$$

which complete the proof. □

Lemma 3. Set $\tilde{\boldsymbol{W}} = (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)\boldsymbol{W}$ and $\boldsymbol{A}_n = (\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)^T(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)$. Under Conditions 1-3 in Appendix, if $\check{h}_j \to 0$ and $n\check{h}_j^4 \to \infty$ as $n \to \infty$, we have $RSS = \tilde{\boldsymbol{Y}}^T\boldsymbol{A}_n\tilde{\boldsymbol{Y}} + O_p(\mathbf{1}^T/\sqrt{n})(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)\tilde{\boldsymbol{Y}} + O_p(1)$, where $RSS = \left\langle \boldsymbol{Y} - \widehat{\boldsymbol{Y}}, \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \right\rangle$ denotes the residual sum of squares for the backfitting estimates, $\tilde{\boldsymbol{Y}}$ is the centered response, and $\boldsymbol{A}_n = \mathbf{S}^T\mathbf{S} - \mathbf{S} - \mathbf{S}^T + \boldsymbol{I}_n + \boldsymbol{R}_n$ with $\mathbf{S} = \sum_{j=1}^{\check{d}} \check{\mathbf{S}}_j$ and $\boldsymbol{R}_n = O_p(\mathbf{1}\mathbf{1}^T/n)$.

Proof. For the backfitting residual vector, we have

$$\begin{aligned}
\boldsymbol{Y} - \widehat{\boldsymbol{Y}} &= \boldsymbol{Y} - \mathbf{1}\widehat{\alpha} - \widehat{\boldsymbol{m}} \\
&= \boldsymbol{Y} - \frac{\mathbf{1}\mathbf{1}^T}{n}\boldsymbol{Y} + \frac{\mathbf{1}\mathbf{1}^T}{n}\boldsymbol{W}\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{Y} \\
&= \tilde{\boldsymbol{Y}} + O_p(\mathbf{1}/\sqrt{n}) - \tilde{\boldsymbol{W}}\boldsymbol{Y}
\end{aligned}$$

$$= \tilde{\boldsymbol{Y}} + O_p(\mathbf{1}/\sqrt{n}) - \tilde{\boldsymbol{W}}(\tilde{\boldsymbol{Y}} + \mathbf{1}E(Y))$$

$$= (\boldsymbol{I}_n - \tilde{\boldsymbol{W}})\tilde{\boldsymbol{Y}} + O_p(\mathbf{1}/\sqrt{n}).$$

Then we have $RSS = \tilde{\boldsymbol{Y}}^T\boldsymbol{A}_n\tilde{\boldsymbol{Y}} + O_p(\mathbf{1}^T/\sqrt{n})(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)\tilde{\boldsymbol{Y}} + O_p(1)$ by definition. Note that we can rewrite $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{m}} + \boldsymbol{\epsilon} = \sum_{j=1}^{\check{d}} \tilde{\boldsymbol{m}}_j + \boldsymbol{\epsilon}$, where $\tilde{\boldsymbol{m}}$ is centered $\check{\boldsymbol{m}}_j$ and $E(\tilde{\boldsymbol{m}}_j) = \mathbf{1}\cdot 0$.

By Lemma 1, Lemma 2 and direct matrix multiplication, we have

$$\tilde{\boldsymbol{W}} = (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)\boldsymbol{W} = (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)\sum_{j=1}^{\check{d}} \boldsymbol{W}_j$$

$$= \sum_{j=1}^{\check{d}}(\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}\check{\mathbf{S}}_j^*(\boldsymbol{I}_n - \boldsymbol{W}^{[-j]})$$

$$= \mathbf{S} + \boldsymbol{U},$$

where $\mathbf{S} = \sum_{j=1}^{\check{d}} \check{\mathbf{S}}_j$ and $\boldsymbol{U} = O_p(\mathbf{1}\mathbf{1}^T/n)$.

Consequently we have

$$\boldsymbol{A}_n = (\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)^T(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)$$

$$= (\mathbf{S} + \boldsymbol{U} - \boldsymbol{I}_n)^T(\mathbf{S} + \boldsymbol{U} - \boldsymbol{I}_n)$$

$$= \mathbf{S}^T\mathbf{S} - \mathbf{S} - \mathbf{S}^T + \boldsymbol{I}_n + \boldsymbol{R}_n,$$

where $\boldsymbol{R}_n = O_p(\mathbf{1}\mathbf{1}^T/n)$ as desired. □

Lemma 4. Set $\boldsymbol{B} = E[\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{m}}|\boldsymbol{X}] = (\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)\tilde{\boldsymbol{m}}$. Under Conditions 1-4 in Appendix, if $\check{h}_j \to 0$ and $n\check{h}_j^4 \to \infty$ as $n \to \infty$, $\boldsymbol{B} = O_p(\sum_{j=1}^{\check{d}} \mathbf{1}\check{h}_j) +$

$O_p(\mathbf{1}/\sqrt{n})$ uniformly over all elements of the vector.

Proof. Applying the same Taylor expansion approximation as in Theorem 2.1 of Ruppert and Wand (1994), we obtain that

$$\begin{aligned}
\check{\mathbf{S}}_j\tilde{\boldsymbol{m}}_j &= \tilde{\boldsymbol{m}}_j + O(\mathbf{1}\check{h}_j),\\
\check{\mathbf{S}}_j^*\tilde{\boldsymbol{m}}_j &= (\check{\mathbf{S}}_j + O_p(\mathbf{1}\mathbf{1}^{\boldsymbol{T}}/N))\tilde{\boldsymbol{m}}_j\\
&= \tilde{\boldsymbol{m}}_j + \bar{\tilde{m}}_j O_p(\mathbf{1}) + O(\mathbf{1}\check{h}_j),\\
(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*)\tilde{\boldsymbol{m}}_j &= \bar{\tilde{m}}_j O_p(\mathbf{1}) + O(\mathbf{1}\check{h}_j) = O_p(\mathbf{1}/\sqrt{N}) + O(\mathbf{1}\check{h}_j).
\end{aligned}$$

Set $\boldsymbol{B} = E[\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{m}}|\boldsymbol{X}] = (\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)\tilde{\boldsymbol{m}}$ and $\boldsymbol{B}^{(j)} = E[\tilde{\boldsymbol{W}}_j\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{m}}_j|\boldsymbol{X}] = \tilde{\boldsymbol{W}}_j\tilde{\boldsymbol{m}} - \tilde{\boldsymbol{m}}_j$.

Combining the above results with the formula (S3.2), we obtain

$$\begin{aligned}
\tilde{\boldsymbol{W}}_j &= (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)\boldsymbol{W}_j = (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I}_n - (\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*))\\
&= \boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n - (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*),\\
(\boldsymbol{I}_n - \tilde{\boldsymbol{W}}_j)\tilde{\boldsymbol{m}}_j &= \bar{\tilde{m}}_j\mathbf{1} + (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*)\tilde{\boldsymbol{m}}_j\\
&= O_p(\mathbf{1}/\sqrt{n}) + (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}(O_p(\mathbf{1}/\sqrt{n}) + O(\mathbf{1}\check{h}_j),\\
(\boldsymbol{I}_n - \tilde{\boldsymbol{W}}_j)\tilde{\boldsymbol{m}}_{(-j)} &= \bar{\tilde{m}}_{(-j)}\mathbf{1} + (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*)\tilde{\boldsymbol{m}}_{(-j)}\\
&= \bar{\tilde{m}}_{(-j)}\mathbf{1} + (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\tilde{\boldsymbol{m}}_{(-j)} + (\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}\check{\mathbf{S}}_j^*(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)\tilde{\boldsymbol{m}}_{(-j)})\\
&= \tilde{\boldsymbol{m}}_{(-j)} + (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}\check{\mathbf{S}}_j^*\boldsymbol{B}_{(-j)},
\end{aligned}$$

where $\tilde{\boldsymbol{m}}_{(-j)}$ and $\boldsymbol{B}_{(-j)}$ are the summation of all component functions and conditional bias for the model without the jth term.

Then by the definition and the previous two formulas of $(\boldsymbol{I}_n - \tilde{\boldsymbol{W}}_j)\tilde{\boldsymbol{m}}_j$ and $(\boldsymbol{I}_n - \tilde{\boldsymbol{W}}_j)\tilde{\boldsymbol{m}}_{(-j)}$, we obtain

$$\begin{aligned}\boldsymbol{B}^{(j)} &= \tilde{\boldsymbol{W}}_j(\tilde{\boldsymbol{m}}_j + \tilde{\boldsymbol{m}}_{(-j)}) - \tilde{\boldsymbol{m}}_j = (\tilde{\boldsymbol{W}}_j - \boldsymbol{I}_n)\tilde{\boldsymbol{m}}_j + \tilde{\boldsymbol{W}}_j\tilde{\boldsymbol{m}}_{(-j)}, \\ &= O_p(\mathbf{1}/\sqrt{n}) + (\boldsymbol{I}_n - \mathbf{1}\mathbf{1}^T/n)(\boldsymbol{I}_n - \check{\mathbf{S}}_j^*\boldsymbol{W}^{[-j]})^{-1}(O_p(\mathbf{1}/\sqrt{n}) + O(\mathbf{1}\check{h}_j) - \mathbf{S}_j^*\boldsymbol{B}_{(-j)}).\end{aligned}$$

Finally, $\boldsymbol{B} = O(\sum_{j=1}^{\tilde{d}} \mathbf{1}\check{h}_j) + O_p(\mathbf{1}/\sqrt{n})$ holds by Lemma 2 and a recursive argument. □

Lemma 5. Under Conditions 1-5 in Appendix, if $\check{h}_j \to 0$ and $n\check{h}_j^4 \to \infty$ as $n \to \infty$, we have

$$RSS/n = \sigma^2 + O_p(\sum_{j=1}^{d} h_j^2) + O_p(\sum_{j=1}^{d-1}\sum_{k=j+1}^{d} \tilde{h}_{jk}^2) + O_p(\sum_{j=1}^{d} \frac{1}{nh_j}) + O_p(\sum_{j=1}^{d-1}\sum_{k=j+1}^{d} \frac{1}{n\tilde{h}_{jk}^2}). \tag{S4.1}$$

Proof. From Lemma 3, we have

$$\begin{aligned} RSS &= \tilde{\boldsymbol{Y}}^T\boldsymbol{A}_n\tilde{\boldsymbol{Y}} + O_p(\mathbf{1}^T/\sqrt{n})(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)\tilde{\boldsymbol{Y}} + O_p(1) \\ &= (\tilde{\boldsymbol{m}} + \boldsymbol{\epsilon})^T\boldsymbol{A}_n(\tilde{\boldsymbol{m}} + \boldsymbol{\epsilon}) + O_p(\mathbf{1}^T/\sqrt{n})(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)(\tilde{\boldsymbol{m}} + \boldsymbol{\epsilon}) + O_p(1) \\ &= \boldsymbol{\epsilon}^T\boldsymbol{A}_n\boldsymbol{\epsilon} + 2\boldsymbol{B}^T(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)\boldsymbol{\epsilon} + \boldsymbol{B}^T\boldsymbol{B} + O_p(1) \\ &= \boldsymbol{\epsilon}^T\boldsymbol{A}_n\boldsymbol{\epsilon} + 2\boldsymbol{B}^T(\mathbf{S} + O_p(\mathbf{1}\mathbf{1}^T/n) - \boldsymbol{I}_n)\boldsymbol{\epsilon} + \boldsymbol{B}^T\boldsymbol{B} + O_p(1). \end{aligned}$$

By calculating the mean and variance, we have $\boldsymbol{B}^T\boldsymbol{B} = O_p(1 + \sum_{j=1}^{\tilde{d}} nh_j^2)$, $\boldsymbol{B}^T\mathbf{S}\boldsymbol{\epsilon} = O_p(1 + \sum_{j=1}^{\tilde{d}} \sqrt{n}h_j)$, $\boldsymbol{B}^T\boldsymbol{\epsilon} = O_p(1 + \sum_{j=1}^{\tilde{d}} \sqrt{n}h_j)$, and hence

$\boldsymbol{B}^T(\tilde{\boldsymbol{W}} - \boldsymbol{I}_n)\boldsymbol{\epsilon} = O_p(1 + \sum_{j=1}^{\tilde{d}} \sqrt{n}h_j)$.

Finally, we consider the term $\boldsymbol{\epsilon}^T\boldsymbol{A}_n\boldsymbol{\epsilon}$,

$$\boldsymbol{\epsilon}^T\boldsymbol{A}_n\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^T\mathbf{S}^T\mathbf{S}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\mathbf{S}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\mathbf{S}^T\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\boldsymbol{I}_n\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\boldsymbol{R}_n\boldsymbol{\epsilon}.$$

We first focus on term $\mathbf{S}$. For main effect term, we have

$$[\mathbf{S}_j]_{st} \approx \frac{1}{nh_j} f_j^{-1}(X_{sj})K(\frac{X_{tj} - X_{sj}}{h_j}),$$

$$\sum_{s=1}^{n} \epsilon_s^2[\mathbf{S}_j]_{ss} \approx \frac{1}{nh_j}\sum_{s=1}^{n} \epsilon_s^2 f_j^{-1}(X_{sj})K(0) \quad \text{with mean } O(\frac{1}{h_j}) \text{ and deviation } O_p(\sqrt{\frac{1}{nh_j^2}}),$$

$$\sum_{s\neq t} \epsilon_s\epsilon_t[\mathbf{S}_j]_{st} \approx \frac{1}{nh_j}\sum_{s\neq t} \epsilon_s\epsilon_t f_j^{-1}(X_{sj})K(\frac{X_{tj} - X_{sj}}{h_j}) \quad \text{with mean 0 and deviation } O_p(\sqrt{\frac{1}{h_i}}).$$

For interaction effect term, we have

$$[\tilde{\mathbf{S}}_{jk}]_{st} \approx \frac{1}{n\tilde{h}_{jk}^2} f_{jk}^{-1}(X_{sj}, X_{sk})K(\frac{X_{tj} - X_{sj}}{\tilde{h}_{jk}})K(\frac{X_{tk} - X_{sk}}{\tilde{h}_{jk}}),$$

$$\sum_{s=1}^{n} \epsilon_s^2[\tilde{\mathbf{S}}_{jk}]_{ss} \approx \frac{1}{n\tilde{h}_{jk}^2}\sum_{s=1}^{n} \epsilon_s^2 f_{jk}^{-1}(X_{sj}, X_{sk})K(0)K(0)$$

$$\text{with mean } O(\frac{1}{\tilde{h}_{jk}^2})\text{and deviation } O_p(\sqrt{\frac{1}{n\tilde{h}_{jk}^4}}),$$

$$\sum_{s\neq t} \epsilon_s\epsilon_t[\tilde{\mathbf{S}}_{jk}]_{st} \approx \frac{1}{n\tilde{h}_{jk}^2}\sum_{s\neq t} \epsilon_s\epsilon_t f_{jk}^{-1}(X_{sj}, X_{sk})K(\frac{X_{tj} - X_{sj}}{\tilde{h}_{jk}})K(\frac{X_{tk} - X_{sk}}{\tilde{h}_{jk}})$$

$$\text{with mean 0 and deviation } O_p(\sqrt{\frac{1}{\tilde{h}_{jk}^2}}).$$

Next we consider term $\mathbf{S}^T\mathbf{S}$. As above, we consider all three possible cases.

For the case of two main effect terms, we have

$$[\mathbf{S}_j^T\mathbf{S}_{j'}]_{st} \approx \frac{1}{n^2h_jh_{j'}}\sum_{i=1}^{n} f_j^{-1}(X_{ij})f_{j'}^{-1}(X_{ij'})K(\frac{X_{sj} - X_{ij}}{h_j})K(\frac{X_{tj'} - X_{ij'}}{h_{j'}}),$$

$$\sum_{s=1}^{n}\epsilon_s^2[\mathbf{S}_j^T\mathbf{S}_{j'}]_{ss} \approx \frac{1}{n^2h_jh_{j'}}\sum_{s=1}^{n}\epsilon_s^2\sum_{i=1}^{n}f_j^{-1}(X_{ij})f_{j'}^{-1}(X_{ij'})K(\frac{X_{sj}-X_{ij}}{h_j})K(\frac{X_{sj'}-X_{ij'}}{h_{j'}})$$

$$\text{with a constant mean and deviation } O_p(\sqrt{\frac{1}{n}+\frac{1}{n^2h_jh_{j'}}}),$$

$$\sum_{s\neq t}\epsilon_s\epsilon_t[\mathbf{S}_j^T\mathbf{S}_{j'}]_{st} \approx \frac{1}{n^2h_jh_{j'}}\sum_{s\neq t}\epsilon_s\epsilon_t\sum_{i=1}^{n}f_j^{-1}(X_{ij})f_{j'}^{-1}(X_{ij'})K(\frac{X_{sj}-X_{ij}}{h_j})K(\frac{X_{tj'}-X_{ij'}}{h_{j'}})$$

$$\text{with mean 0 and deviation } O_p(\sqrt{1+\frac{1}{nh_jh_{j'}}}).$$

For the case with one main effect and one interaction effect term, we have

$$\begin{aligned}[\mathbf{S}_j^T\tilde{\mathbf{S}}_{j'k}]_{st} \approx& \frac{1}{n^2h_j\tilde{h}_{j'k}^2}\sum_{i=1}^{n}f_j^{-1}(X_{ij})f_{j'k}^{-1}(X_{ij'},X_{ik})\\ &K(\frac{X_{sj}-X_{ij}}{h_j})K(\frac{X_{tj'}-X_{ij'}}{\tilde{h}_{j'k}})K(\frac{X_{tk}-X_{ik}}{\tilde{h}_{j'k}}),\\ \sum_{s=1}^{n}\epsilon_s^2[\mathbf{S}_j^T\tilde{\mathbf{S}}_{j'k}]_{ss} \approx& \frac{1}{n^2h_j\tilde{h}_{j'k}^2}\sum_{s=1}^{n}\epsilon_s^2\sum_{i=1}^{n}f_j^{-1}(X_{ij})f_{j'k}^{-1}(X_{ij'},X_{ik})\\ &K(\frac{X_{sj}-X_{ij}}{h_j})K(\frac{X_{sj'}-X_{ij'}}{\tilde{h}_{j'k}})K(\frac{X_{sk}-X_{ik}}{\tilde{h}_{j'k}})\\ &\text{with a constant mean and deviation } O_p(\sqrt{\frac{1}{n}+\frac{1}{n^2h_i\tilde{h}_{jk}^2}}),\\ \sum_{s\neq t}\epsilon_s\epsilon_t[\mathbf{S}_j^T\tilde{\mathbf{S}}_{j'k}]_{st} \approx& \frac{1}{n^2h_j\tilde{h}_{j'k}^2}\sum_{s\neq t}\epsilon_s\epsilon_t\sum_{i=1}^{n}f_j^{-1}(X_{ij})f_{j'k}^{-1}(X_{ij'},X_{ik})\\ &K(\frac{X_{sj}-X_{ij}}{h_j})K(\frac{X_{tj'}-X_{ij'}}{\tilde{h}_{j'k}})K(\frac{X_{tk}-X_{ik}}{\tilde{h}_{j'k}})\\ &\text{with mean 0 and deviation } O_p(\sqrt{1+\frac{1}{nh_i\tilde{h}_{jk}^2}}).\end{aligned}$$

For the case with two interaction effect terms, we have

$$[\tilde{\mathbf{S}}_{jk}^T\tilde{\mathbf{S}}_{j'k'}]_{st} \approx \frac{1}{n^2\tilde{h}_{jk}^2\tilde{h}_{j'k'}^2}\sum_{i=1}^{n}f_{jk}^{-1}(X_{ij},X_{ik})f_{j'k'}^{-1}(X_{ij'},X_{ik'})$$

$$K(\frac{X_{sj}-X_{ij}}{\tilde{h}_{jk}})K(\frac{X_{sk}-X_{ik}}{\tilde{h}_{jk}})K(\frac{X_{tj'}-X_{ij'}}{\tilde{h}_{j'k'}})K(\frac{X_{tk'}-X_{ik'}}{\tilde{h}_{j'k'}}),$$

$$\sum_{s=1}^{n}\epsilon_s^2[\tilde{\mathbf{S}}_{jk}^T\tilde{\mathbf{S}}_{j'k'}]_{ss} \approx \frac{1}{n^2\tilde{h}_{jk}^2\tilde{h}_{j'k'}^2}\sum_{s=1}^{n}\epsilon_s^2\sum_{i=1}^{n}f_{jk}^{-1}(X_{ij},X_{ik})f_{j'k'}^{-1}(X_{ij'},X_{ik'})$$
$$K(\frac{X_{sj}-X_{ij}}{\tilde{h}_{jk}})K(\frac{X_{sk}-X_{ik}}{\tilde{h}_{jk}})K(\frac{X_{sj'}-X_{ij'}}{\tilde{h}_{j'k'}})K(\frac{X_{sk'}-X_{ik'}}{\tilde{h}_{j'k'}})$$
$$\text{with a constant mean and deviation } O_p(\sqrt{\frac{1}{n}+\frac{1}{n^2\tilde{h}_{jk}^2\tilde{h}_{j'k'}^2}}),$$

$$\sum_{s\neq t}\epsilon_s\epsilon_t[\tilde{\mathbf{S}}_{jk}^T\tilde{\mathbf{S}}_{j'k'}]_{st} \approx \frac{1}{n^2\tilde{h}_{jk}^2\tilde{h}_{j'k'}^2}\sum_{s\neq t}\epsilon_s\epsilon_t\sum_{i=1}^{n}f_{jk}^{-1}(X_{ij},X_{ik})f_{j'k'}^{-1}(X_{ij'},X_{ik'})$$
$$K(\frac{X_{sj}-X_{ij}}{\tilde{h}_{jk}})K(\frac{X_{sk}-X_{ik}}{\tilde{h}_{jk}})K(\frac{X_{tj'}-X_{ij'}}{\tilde{h}_{j'k'}})K(\frac{X_{tk'}-X_{ik'}}{\tilde{h}_{j'k'}})$$
$$\text{with mean 0 and deviation } O_p(\sqrt{1+\frac{1}{n\tilde{h}_{jk}^2\tilde{h}_{j'k'}^2}}).$$

Overall, $\boldsymbol{\epsilon}^T\boldsymbol{A}_n\boldsymbol{\epsilon}/n = \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{I}_n\boldsymbol{\epsilon}+\frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{R}_n\boldsymbol{\epsilon}+O_p(\sum_{j=1}^{d}\frac{1}{nh_j})+O_p(\sum_{j=1}^{d-1}\sum_{k=j+1}^{d}\frac{1}{n\tilde{h}_{jk}^2}) = \sigma^2+O_p(\sum_{j=1}^{d}\frac{1}{nh_j})+O_p(\sum_{j=1}^{d-1}\sum_{k=j+1}^{d}\frac{1}{n\tilde{h}_{jk}^2})$ and $\boldsymbol{B}^T\boldsymbol{B}/n = O_p(\sum_{j=1}^{d}h_j^2)+O_p(\sum_{j=1}^{d-1}\sum_{k=j+1}^{d}\tilde{h}_{jk}^2)$. Then $RSS/n = \sigma^2+O_p(\sum_{j=1}^{d}h_j^2)+O_p(\sum_{j=1}^{d-1}\sum_{k=j+1}^{d}\tilde{h}_{jk}^2)+O_p(\sum_{j=1}^{d}\frac{1}{nh_j})+O_p(\sum_{j=1}^{d-1}\sum_{k=j+1}^{d}\frac{1}{n\tilde{h}_{jk}^2})$ holds. □

Proof of Theorem 1. Recall that the sets of important main and interaction effects are denoted by $\mathcal{M} = \{j : m_j(\cdot) \neq 0\}$ and $\mathcal{I} = \{(j,k) : m_{jk}(\cdot,\cdot) \neq 0\}$. Then the sets of unimportant main and interaction effects denote by the complement $\mathcal{M}^c = \{1,\ldots,d\}\backslash\mathcal{M}$ and $\mathcal{I}^c = \{(j,k) : 1 \leq j < k \leq d\}\backslash\mathcal{I}$. We first simplify the formula (S4.1) in Lemma 5 with the following conditions: under Conditions 1-5 in Appendix, if $h_j \to 0$ for $j \in \mathcal{S} = \mathcal{M}\cup\mathcal{I}$, and

$h_j \geq c_0 > 0$ for $j \in \mathcal{N} = \mathcal{M}^c \cup \mathcal{I}^c$ and some $c_0 > 0$, then

$$RSS/n = \sigma^2 + O_p(\sum_{j\in\mathcal{S}} h_j^2) + O_p(\sum_{j\in\mathcal{S}} \frac{1}{nh_j^2}). \tag{S4.2}$$

It can be easily verified by following the detailed proofs of Lemmas 1-5. Here is some heuristic justification. For any $j \in \mathcal{N} = \mathcal{M}^c \cup \mathcal{I}^c$, the corresponding component function (either a main effect or an interaction effect term) is a constant and there is no approximation bias for the backfitting estimator with local constant smoothing. Then the unimportant predictor does not have any contribution to the total approximation bias. At the same time, with the condition $h_j \geq c_0 > 0$ for $j \in \mathcal{N} = \mathcal{M}^c \cup \mathcal{I}^c$, the smoothing bandwidth for unimportant term is bounded away zero. Then the variance of unimportant terms is dominated by the variance of important terms with corresponding smoothing bandwidths shrinking to zero. Consequently we have equation (S4.2).

Now we are ready to show the selection consistency. In the following, we use the notion $\widehat{\lambda}_j = \frac{1}{\widehat{h}_j}$ for $j = 1, \ldots, \check{d}$. We first prove $\widehat{\lambda}_j \to \infty$ for $j \in \mathcal{S}$ by contradiction. Assume $\widehat{\lambda}_{j'}$ is bounded from above for some $j' \in \mathcal{S}$. If $\widehat{\lambda}_{j'}$ is bounded, RSS/n will converge to σ^2 plus a bias. The extra bias comes from the second term in equation (S4.2) as $\widehat{h}_{j'} = \frac{1}{\widehat{\lambda}_{j'}} \not\to 0$ and the corresponding approximation bias does not shrink to zero. It is a suboptimal as equation (S4.2) implies if the bandwidths for all important predictors approach zero, the RSS/n reaches its optimal value σ^2 asymptotically. This proves $\hat{\lambda}_j \to 0$ for $j \in \mathcal{S}$.

Before moving to the unimportant terms part, note that the condition that if $\tau \to \infty$, $\frac{\tau^4}{n} \to 0$ as $n \to \infty$, is used to guarantee that the variance term (third term on the right hand side of formula (S4.2)) is dominated by the bias term (second term on the right hand side of formula (S4.2)). Without loss of generality, we assume $\widehat{\lambda}_j$ converges for $j \in \mathcal{N}$, since otherwise we can consider an convergent subsequence. Next we try to show that $\widehat{\lambda}_j \to 0$ for $j \in \mathcal{N}$ also by contradiction. Assume there are some $\widehat{\lambda}_j \not\to 0$ for $j \in \mathcal{N}$ and set $\acute{\lambda}_j = \widehat{\lambda}_j \tau / (\tau - \sum_{j' \in \mathcal{N}} \widehat{\lambda}_{j'})$ for $j \in \mathcal{S}$ and $\acute{\lambda}_j = 0$ for $j \in \mathcal{N}$. Note that the variance is dominated by the bias and small $\widehat{\lambda}_j$ for $j \in \mathcal{N}$ does not induce bias, then we only need to consider the bias induced by the important terms. Note that $\acute{\lambda}_j$ diverges to infinity faster than $\widehat{\lambda}_j$ for $j \in \mathcal{S}$, hence using the set of $\acute{\lambda}s$ has a smaller bias term. When $\sum_{j \in \mathcal{N}} \widehat{\lambda}_j \to \infty$, the smaller is in the sense of asymptotic order, and when $\sum_{j \in \mathcal{N}} \widehat{\lambda}_j$ is bounded, it is in the sense of the multiplying constant in the asymptotic order. Therefore, $\widehat{\lambda}s$ is a suboptimal. Consequently we have that $\widehat{\lambda}_j \to 0$ for $j \in \mathcal{N}$ and this finishes the proof. □

Bibliography

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *Ann. Statist. 17*(2), 453–510.

Dong, Y. and Y. Wu (2020). Nonparametric interaction selection. manuscript.

Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis 73*(2), 166 – 179.

Opsomer, J. D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist. 25*(1), 186–211.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *Ann. Statist. 22*(3), 1346–1370.