Supplement for "Forward Additive Regression for Ultrahigh Dimensional Nonparametric Additive Models"

Wei Zhong[1], Sunpeng Duan[1] and Liping Zhu[2]

Xiamen University[1] *and Renmin University of China*[2]

We require the following lemmas to prove Theorem 1.

Lemma 1. *For any candidate model $\mathcal{M}$ which is bounded, where $|\mathcal{M}| < C_1 < \infty$, under Assumption (A2), with probability converging to one,*

$$c_1 m_n^{-1} \leq \min_{|\mathcal{M}|<C_1} \lambda_{\min}\left(\frac{\mathbf{U}_{\mathcal{M}}^T \mathbf{U}_{\mathcal{M}}}{n}\right) \leq \max_{|\mathcal{M}|<C_1} \lambda_{\max}\left(\frac{\mathbf{U}_{\mathcal{M}}^T \mathbf{U}_{\mathcal{M}}}{n}\right) \leq c_2 m_n^{-1}$$

where c_1 and c_2 are two positive constants.

Lemma 1 comes from Huang, Horowitz and Wei (2010) which was based on Zhou, Shen and Wolfe (1998). It restricts the eigenvalue of B-spline matrix.

Lemma 2. *Let $X_1, \cdots, X_n$ be the triangular array of i.i.d. zero-mean random variables. Suppose that $M_n = (EX_1^2)^{1/2}/(E|X_1|^3)^{1/3} > 0$ and that for some $b_n \to \infty$ slowly, $n^{1/6} M_n / b_n \geq 1$. Then uniformly on $0 \leq x \leq$*

$n^{1/6}M_n/b_n - 1$, *we have*

$$\left| \frac{P(|S_n/V_n| \geq x)}{2[1-\Phi(x)]} - 1 \right| \leq \frac{A}{b_n^3} \to 0,$$

where $S_n = \sum_{i=1}^n X_i$, $V_n^2 = \sum_{i=1}^n X_i^2$, $\Phi(\cdot)$ *is the cumulative distribution function of the standard normal distribution and* A *is a positive constant.*

Lemma 2 follows Lemma 5 in Belloni et al. (2012) and Theorem 7.4 in de la Pena, Lai and Shao (2009). This lemma was also used in Fan and Zhong (2016).

Lemma 3. *For two candidate models* $\mathcal{M}_1$,$\mathcal{M}_2$, *with* $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$, *we have*

$$\begin{aligned} nc_1 m_n^{-1} &\leq \inf_{\|\mathbf{u}\|=1} \mathbf{u}^T \mathbf{U}_{(\mathcal{M}_1)}^T \mathbf{Q}_{(\mathcal{M}_2)} \mathbf{U}_{(\mathcal{M}_1)} \mathbf{u} \\ &\leq \sup_{\|\mathbf{u}\|=1} \mathbf{u}^T \mathbf{U}_{(\mathcal{M}_1)}^T \mathbf{Q}_{(\mathcal{M}_2)} \mathbf{U}_{(\mathcal{M}_1)} \mathbf{u} \leq nc_2 m_n^{-1}, \end{aligned}$$

where $\mathbf{Q}_{(\mathcal{M})} = \mathbf{I}_n - \mathbf{U}_{(\mathcal{M})}\{\mathbf{U}_{(\mathcal{M})}^T \mathbf{U}_{(\mathcal{M})}\}^{-1}\mathbf{U}_{(\mathcal{M})}^T$ *and* c_1 *and* c_2 *are two positive constants defined in Lemma 1.*

Proof. First, we prove $\sup_{\|\mathbf{u}\|=1} \mathbf{u}^{\mathrm{T}} \mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}} \mathbf{Q}_{(\mathcal{M}_2)} \mathbf{U}_{(\mathcal{M}_1)} \mathbf{u} \leq nc_2 m_n^{-1}$.

By spectral decomposition, we have $\mathbf{Q}_{(\mathcal{M}_2)} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}^{\mathrm{T}}$, where $\mathbf{C}$ is a $n \times n$ matrix with jth column being $\mathbf{c}_j$ corresponding to the jth eigenvector such

that $\mathbf{c}_j^{\mathrm{T}}\mathbf{c}_j = 1$ and $\mathbf{c}_j^{\mathrm{T}}\mathbf{c}_k = 0$ for $j \neq k$, and $\mathbf{\Lambda}$ is a diagonal matrix with the diagonal elements being the eigenvalues λ_j of $\mathbf{Q}_{(\mathcal{M}_2)}$.

Then

$$\begin{aligned}
& \mathbf{u}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{M}_2)}\mathbf{U}_{(\mathcal{M}_1)}\mathbf{u} \\
= \; & \mathbf{u}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{C}\mathbf{\Lambda}\mathbf{C}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}\mathbf{u} \\
\leq \; & \mathbf{u}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{C}\mathbf{C}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}\mathbf{u} \times \lambda_{\max}[\mathbf{Q}_{(\mathcal{M}_2)}] \\
= \; & \mathbf{u}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{I}_n\mathbf{U}_{(\mathcal{M}_1)}\mathbf{u} \times \lambda_{\max}[\mathbf{Q}_{(\mathcal{M}_2)}] \\
= \; & \mathbf{u}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}\mathbf{u} \times \lambda_{\max}[\mathbf{Q}_{(\mathcal{M}_2)}] \\
\leq \; & nc_2m_n^{-1} \qquad \text{(A.1)}
\end{aligned}$$

where the last inequality from Lemma 1 and note that the eigenvalue of idempotent matrix $\mathbf{Q}_{(\mathcal{M}_2)}$ is 0 or 1.

Next, we will prove $\inf_{\|\mathbf{u}\|=1} \mathbf{u}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{M}_2)}\mathbf{U}_{(\mathcal{M}_1)}\mathbf{u} \geq nc_1m_n^{-1}$. We can

rewrite $\mathbf{Q}_{(\mathcal{M}_2)}\mathbf{U}_{(\mathcal{M}_1)}$ in matrix notation, that is,

$$\begin{aligned}
\mathbf{Q}_{(\mathcal{M}_2)}\mathbf{U}_{(\mathcal{M}_1)} &= \mathbf{U}_{(\mathcal{M}_1)} - \mathbf{U}_{(\mathcal{M}_2)}(\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)})^{-1}\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)} \\
&= (\mathbf{U}_{(\mathcal{M}_1)}, \mathbf{U}_{(\mathcal{M}_2)}) \begin{pmatrix} \mathbf{I} \\ -(\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)})^{-1}\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)} \end{pmatrix} \\
&= \mathbf{U}_{(\mathcal{M}_1 \cup \mathcal{M}_2)} \begin{pmatrix} \mathbf{I} \\ -(\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)})^{-1}\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)} \end{pmatrix}. \quad \text{(A.2)}
\end{aligned}$$

Hence, we have

$$\begin{aligned}
& \mathbf{u}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{M}_2)}\mathbf{U}_{(\mathcal{M}_1)}\mathbf{u} \\
= \; & \mathbf{u}^{\mathrm{T}}[\mathbf{I}, -\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)}(\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)})^{-1}](\mathbf{U}_{\mathcal{M}_1 \cup \mathcal{M}_2}^{\mathrm{T}}\mathbf{U}_{\mathcal{M}_1 \cup \mathcal{M}_2}) \\
& \times \begin{pmatrix} \mathbf{I} \\ -(\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)})^{-1}\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_1)} \end{pmatrix} \mathbf{u} \\
\geq \; & nc_1 m_n^{-1} \times \|[\mathbf{I}, -\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)}(\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)})^{-1}]^{\mathrm{T}}\mathbf{u}\|^2 \\
\geq \; & nc_1 m_n^{-1}, \quad \text{(A.3)}
\end{aligned}$$

where in the first inequality we use Lemma 1 and spectral decomposition, and the second in inequality comes from the fact $\|[\mathbf{I}, -\mathbf{U}_{(\mathcal{M}_1)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)}(\mathbf{U}_{(\mathcal{M}_2)}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M}_2)})^{-1}]^{\mathrm{T}}\mathbf{u}\| \geq \|\mathbf{u}\| = 1$. This completes the proof of Lemma 3. □

Proof of Theorem 1: In the Forward Additive Regression algorithm, we expect to detect all the p_0 relevant predictors in an acceptable number of steps. If we can identify at least one new relevant predictor in every at most K_0 steps, then all relevant predictors will be identified within at most $p_0 K_0$ steps. To prove this conclusion, we assume that no relevant predictor has been detected in the first l steps, given the model $S^{(l_0)}$ has already been selected. We then evaluate how likely at least one relevant predictor will be detected in the next step. To this end, we study what will happen if the $(l+1)$th selected predictor is still irrelevant given the existence of $S^{(l_0)}$ in the model.

For an ease of the presentation, we define $\mathbf{H}_{(\mathcal{M})} = \mathbf{U}_{(\mathcal{M})}\{\mathbf{U}_{(\mathcal{M})}^{\mathrm{T}}\mathbf{U}_{(\mathcal{M})}\}^{-1}\mathbf{U}_{(\mathcal{M})}^{\mathrm{T}}$ for any model $\mathcal{M}$. Then, we have

$$
\begin{aligned}
&\mathbf{H}_{(\mathcal{S}^{(l_0+l+1)})} = \mathbf{U}_{(\mathcal{S}^{(l_0+l+1)})}\{\mathbf{U}_{(\mathcal{S}^{(l_0+l+1)})}^{\mathrm{T}}\mathbf{U}_{(\mathcal{S}^{(l_0+l+1)})}\}^{-1}\mathbf{U}_{(\mathcal{S}^{(l_0+l+1)})}^{\mathrm{T}} \\
= &\left(\mathbf{U}_{(\mathcal{S}^{(l_0+l)})}, \mathbf{U}_{a_{l_0+l+1}}\right)\left[\begin{pmatrix}\mathbf{U}_{(\mathcal{S}^{(l_0+l)})}^{\mathrm{T}} \\ \mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\end{pmatrix}\left(\mathbf{U}_{(\mathcal{S}^{(l)})}, \mathbf{U}_{a_{l_0+l+1}}\right)\right]^{-1}\begin{pmatrix}\mathbf{U}_{(\mathcal{S}^{(l_0+l)})}^{\mathrm{T}} \\ \mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\end{pmatrix} \\
= &\left(\mathbf{U}_{(\mathcal{S}^{(l_0+l)})}, \mathbf{U}_{a_{l_0+l+1}}\right)\begin{pmatrix}\mathbf{U}_{(\mathcal{S}^{(l_0+l)})}^{\mathrm{T}}\mathbf{U}_{(\mathcal{S}^{(l_0+l)})} & \mathbf{U}_{(\mathcal{S}^{(l_0+l)})}^{\mathrm{T}}\mathbf{U}_{a_{l_0+l+1}} \\ \mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\mathbf{U}_{(\mathcal{S}^{(l_0+l)})} & \mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\mathbf{U}_{a_{l_0+l+1}}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{U}_{(\mathcal{S}^{(l_0+l)})}^{\mathrm{T}} \\ \mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\end{pmatrix}, \quad \text{(A.4)}
\end{aligned}
$$

where $\mathcal{S}^{(l_0+l)}$ denotes the union of $\mathcal{S}^{(l_0)}$ and the first l irrelevant predictors

selected after $\mathcal{S}^{(l_0)}$, and a_{l_0+l+1} denotes the index for the selected predictor in the $(l+1)$th step after $\mathcal{S}^{(l_0)}$.

Using the rule of the matrix inversion in block form, we show that

$$\mathbf{H}_{(\mathcal{S}^{(l_0+l+1)})} = \mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{a_{l_0+l+1}}(\mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{a_{l_0+l+1}})^{-1}\mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}+\mathbf{H}_{(\mathcal{S}^{(l_0+l)})}$$

Then, we consider the difference of the residual sum of squares between the models with and without the a_{l_0+l+1}th predictor. Denote $\mathrm{RSS}(\mathcal{S}^{(l_0+l)})$ and $\mathrm{RSS}(\mathcal{S}^{(l_0+l+1)})$ by the residual sums of squares based on the model $\mathcal{S}^{(l_0+l)}$ and $\mathcal{S}^{(l_0+l+1)}$, respectively. We define

$$\begin{aligned}
\Omega(l) &= \mathrm{RSS}(\mathcal{S}^{(l_0+l)}) - \mathrm{RSS}(\mathcal{S}^{(l_0+l+1)}) \\
&= \mathbf{Y}^{\mathrm{T}}\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(l_0+l)})}\}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(l_0+l+1)})}\}\mathbf{Y} \\
&= \mathbf{Y}^{\mathrm{T}}\{\mathbf{H}_{(\mathcal{S}^{(l_0+l+1)})} - \mathbf{H}_{(\mathcal{S}^{(l_0+l)})}\}\mathbf{Y} \\
&= \mathbf{Y}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{a_{l_0+l+1}}(\mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{a_{l_0+l+1}})^{-1}\mathbf{U}_{a_{l_0+l+1}}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{Y} \\
&\geq \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)}} \mathbf{Y}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_j(\mathbf{U}_j^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_j)^{-1}\mathbf{U}_j^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{Y} \\
&\geq \frac{1}{nc_2m_n^{-1}} \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)}} \mathbf{Y}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_j\mathbf{U}_j^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{Y} \qquad \text{(A.5)}
\end{aligned}$$

where the first inequality is because of the assumption that the a_{l_0+l+1}th predictor is also not relevant and a_{l_0+l+1} is the predictor added in the $(l_0+$

$l+1$)th step which corresponds to the minimum RSS and the last inequality is implied by Lemma 3 and the fact that $(\mathcal{T}/\mathcal{S}^{(l_0)})\bigcap\mathcal{S}^{(l_0+l)}=\emptyset$.

We denote $\mathbf{\Psi}_k(X_j)=(\psi_k(X_{1j}),\cdots,\psi_k(X_{nj}))^{\mathrm{T}}$ where $k=1,\ldots,m_n$, then we have $\mathbf{U}_j=(\mathbf{\Psi}_k(X_1),\ldots,\mathbf{\Psi}_{m_n}(X_j))$. Thus,

$$\begin{aligned}
\Omega(l) &\geq \frac{1}{nc_2m_n^{-1}}\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)}}\sum_{k=1}^{m_n}\mathbf{Y}^{\mathrm{T}}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{\Psi}_k(X_j)\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{Y}\\
&\geq \frac{m_n}{nc_2m_n^{-1}}\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n}|\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{Y}|^2, \qquad \text{(A.6)}\\
&\geq \frac{m_n^2}{nc_2}\Bigg(\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n}|\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T})}\boldsymbol{\gamma}_{(\mathcal{T})}|\\
&\qquad -\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n}|\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\boldsymbol{\xi}|\Bigg)^2 \qquad \text{(A.7)}
\end{aligned}$$

where the last inequality because $\mathbf{Y}=\mathbf{U}_{(\mathcal{T})}\boldsymbol{\gamma}_{(\mathcal{T})}+\boldsymbol{\xi}$ which is the matrix form of (2.3), $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)^{\mathrm{T}}\in\mathbb{R}^n$.

Next, we deal with the first term, that's $\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n}|\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T})}\boldsymbol{\gamma}_{(\mathcal{T})}|$. Note that $\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T})}=\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}$ because $\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}\cap\mathcal{S}^{(l_0)})}=$

0. Then, we consider that

$$
\begin{aligned}
&\|\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T})}\boldsymbol{\gamma}_{(\mathcal{T})}\|^2 = \boldsymbol{\gamma}^{\mathrm{T}}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\mathbf{U}^{\mathrm{T}}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})} \\
= &\sum_{j\in\mathcal{T}/\mathcal{S}^{(l_0)}}\sum_{k=1}^{m_n}\gamma_{jk}\boldsymbol{\Psi}^{\mathrm{T}}_k(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})} \\
\leq &\left(\sum_{j\in\mathcal{T}/\mathcal{S}^{(l_0)}}\sum_{k=1}^{m_n}\gamma^2_{jk}\right)^{1/2}\left\{\sum_{j\in\mathcal{T}/\mathcal{S}^{(l_0)}}\sum_{k=1}^{m_n}(\boldsymbol{\Psi}^{\mathrm{T}}_k(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})})^2\right\}^{1/2} \\
\leq &\|\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\|\sqrt{(p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)m_n}\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n}|\boldsymbol{\Psi}^{\mathrm{T}}_k(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}|
\end{aligned}
$$

Because $\mathcal{S}^{(l_0+l)}\cap(\mathcal{T}/\mathcal{S}^{(l_0)})=\emptyset$, it is implied by Lemma 3 that $\|\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T})}\boldsymbol{\gamma}_{(\mathcal{T})}\|^2 = \|\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\|^2 \geq nc_1m_n^{-1}\|\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\|^2$. Hence, we have

$$
\begin{aligned}
&\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n}|\boldsymbol{\Psi}^{\mathrm{T}}_k(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}| \\
\geq &\frac{\|\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T})}\boldsymbol{\gamma}_{(\mathcal{T})}\|^2}{\|\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\|\sqrt{(p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)m_n}} \geq \frac{nc_1m_n^{-1}\|\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\|}{\sqrt{(p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)m_n}}
\end{aligned}
$$

Based on the proof of Theorem 1 in Huang, Horowitz and Wei (2010), there are positive constants c_3 such that $\|\boldsymbol{\gamma}_j\|^2 \geq c_3c_f^2m_n$, where $j\in\mathcal{T}$ and c_f controls the minimum signal of the true components. Thus, $\|\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\| = \sqrt{\sum_{j\in\mathcal{T}/\mathcal{S}^{(l_0)}}\sum_{k=1}^{m_n}\gamma^2_{jk}} = \sqrt{\sum_{j\in\mathcal{T}/\mathcal{S}^{(l_0)}}\|\boldsymbol{\gamma}_j\|^2} \geq \sqrt{(p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)c_3c_f^2m_n}$.

Therefore, the first term in the parenthesis of (A.7)

$$\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n} |\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T})}\boldsymbol{\gamma}_{(\mathcal{T})}| \tag{A.8}$$

$$= \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n} |\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{U}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}\boldsymbol{\gamma}_{(\mathcal{T}/\mathcal{S}^{(l_0)})}|$$

$$\ge \frac{nc_1m_n^{-1}\sqrt{(p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)c_3c_f^2m_n}}{\sqrt{(p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)m_n}} = c_1\sqrt{c_3}c_fnm_n^{-1} \tag{A.9}$$

Next, we can handle the second part in the parenthesis of (A.7).

$$\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n} |\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\boldsymbol{\xi}|$$

$$\le \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n} |\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\boldsymbol{\delta}| + \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n} |\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\boldsymbol{\varepsilon}|$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1,\cdots,\varepsilon_n)^{\mathrm{T}}$, $\boldsymbol{\delta} = (\delta_1,\cdots,\delta_n)^{\mathrm{T}}$ with $\delta_i = \sum_{j=1}^p(f_j(X_{ij}) - f_{nj}(X_{ij}))$.

For ease of the presentation, we define $\mathbf{\Psi}_k^*(X_j) = \mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\mathbf{\Psi}_k(X_j) \in \mathbb{R}^{n\times 1}$, where $\mathbf{\Psi}_k^*(X_j) = (\psi_k^*(X_{1j}),\cdots,\psi_k^*(X_{nj}))^{\mathrm{T}}$. Note that $\|\mathbf{\Psi}_k^*(X_j)\| \le$

$\|\mathbf{\Psi}_k(X_j)\|$ and the centered B-splines $|\psi_k(X_{ij})| \leq 2$, then

$$\begin{aligned} & \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n} |\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\boldsymbol{\delta}| = \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n} |\mathbf{\Psi}_k^{*T}(X_j)\boldsymbol{\delta}| \\ = & \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n} |\sum_{i=1}^{n} \psi_k^*(X_{ij})\delta_i| \leq n \max_{1\leq i\leq n} \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n} |\psi_k^*(X_{ij})| \cdot |\delta_i| \\ \leq & \ 2nO_p(m_n^{-d}) = O_p(nm_n^{-d}), \end{aligned} \tag{A.10}$$

where the last inequality follows from Lemma 1 of Huang, Horowitz and Wei (2010). That is, suppose that $f \in \mathcal{F}$ and $\mathrm{E}f(X_j) = 0$, then under Assumption (A2), there exists an $f_n \in \mathcal{S}_n$ satisfying$\|f_n - f\|_2 = O_p(m_n^{-d} + m_n^{1/2}n^{1/2})$. In particular, if we choose $m_n = O(n^{1/(2d+1)})$, then $\|f_n - f\| = O(m_n^{-d})$.

On the other hand, we have

$$\begin{aligned} & \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n} |\mathbf{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\boldsymbol{\varepsilon}| \\ \leq & \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n} \left| \frac{\sum_{i=1}^n \psi_k^*(X_{ij})\varepsilon_i}{\sqrt{\sum_{i=1}^n \psi_k^{*2}(X_{ij})\varepsilon_i^2}} \right| \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\leq k\leq m_n} \sqrt{\sum_{i=1}^{n} \psi_k^{*2}(X_{ij})\varepsilon_i^2} \end{aligned}$$

Note that

$$
\begin{aligned}
& P\left(\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n}\left|\frac{\sum_{i=1}^n\psi_k^*(X_{ij})\varepsilon_i}{\sqrt{\sum_{i=1}^n\psi_k^{*2}(X_{ij})\varepsilon_i^2}}\right|>\sqrt{2m_n/a}\right)\\
\le\ & (p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)m_n\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n}P\left(\left|\frac{\sum_{i=1}^n\psi_k^*(X_{ij})\varepsilon_i}{\sqrt{\sum_{i=1}^n\psi_k^{*2}(X_{ij})\varepsilon_i^2}}\right|>\sqrt{2m_n/a}\right)\\
\le\ & (p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)m_n2[1-\Phi(\sqrt{2m_n/a})](1+o(1))\\
\le\ & (p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)m_n\frac{2\exp[-(\sqrt{2m_n/a})^2/2]}{\sqrt{2\pi}\sqrt{2m_n/a}}(1+o(1))\\
=\ & \frac{(p_0-|\mathcal{T}\cap\mathcal{S}^{(l_0)}|)\sqrt{m_na}}{\sqrt{\pi}\exp(m_n/a)}(1+o(1))\to 0, \qquad \text{(A.11)}
\end{aligned}
$$

as $n\to\infty$, uniformly for all $0<a\le 1$, where the second inequality follows the above Lemma 2 on moderate deviation inequality for self-normalized sums and the last inequality follows the fact that $P(Z>z)\le \exp(z^2/2)/(z\sqrt{2\pi})$ for a standard normal random variable Z.

Since $E(\varepsilon_i^2)$ is bounded,

$$
\begin{aligned}
& \max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n}\sqrt{\sum_{i=1}^n\psi_k^{*2}(X_{ij})\varepsilon_i^2}=\sqrt{n}\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n}\sqrt{\frac{1}{n}\sum_{i=1}^n\psi_k^{*2}(X_{ij})\varepsilon_i^2}\\
\le\ & \sqrt{n}\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n}2\sqrt{\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2}=\sqrt{n}O_p(1)=O_p(n^{1/2}) \qquad \text{(A.12)}
\end{aligned}
$$

Then, both (A.11) and (A.12) imply that

$$\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n} |\boldsymbol{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\boldsymbol{\varepsilon}| \le O_p(n^{1/2}m_n^{1/2}). \tag{A.13}$$

Thus, together with (A.10) and (A.13), we have

$$\max_{j\in\mathcal{T}/\mathcal{S}^{(l_0)},1\le k\le m_n} |\boldsymbol{\Psi}_k^{\mathrm{T}}(X_j)\mathbf{Q}_{(\mathcal{S}^{(l_0+l)})}\boldsymbol{\xi}| \le O_p(nm_n^{-d} + n^{1/2}m_n^{1/2}). \tag{A.14}$$

Based on (A.9) and (A.14) as well as the assumption $d > 1$, the second part in the parenthesis of (A.7) is dominated by the first term. Thus, with the probability tending to one, we have

$$\Omega(l) \ge \frac{m_n^2}{nc_2}(c_1\sqrt{c_3}c_f nm_n^{-1})^2 = c_1^2c_3c_f^2 n/c_2, \tag{A.15}$$

for every l. If we run the total K_0 steps with the existence of $S^{(l_0)}$ in the model, then we have

$$n^{-1}\|\mathbf{Y}\|^2 \ge n^{-1}\sum_{l=1}^{K_0}\Omega(l) \ge K_0c_1^2c_3c_f^2/c_2, \tag{A.16}$$

which is contradicted with the assumption that $K_0 > c_2\mathrm{var}(Y)/c_1^2c_3c_f^2$. Therefore, we can detect at least one relevant predictor within every K_0

steps. So all relevant predictors can be detected within p_0K_0 steps with probability tending to one.

Proof of Theorem 2 The Proof of this theorem is parallel to Theorem 2 of Wang (2009). Define $l_{min} = \min_{1\le l\le [\frac{n}{m_n}]}\{l : \mathcal{T} \subset \mathcal{S}^{(l)}\}$. And by Theorem 1, we know that $l_{min} \le p_0K_0$. Thus, we only prove that $P(\hat{m} < l_{min}) \to 1$ as $n \to \infty$. To this end, it suffices to show that

$$P\left(\min_{1\le l<l_{min}}\{\mathrm{BIC}(\mathcal{S}^{(l)}) - \mathrm{BIC}(\mathcal{S}^{(l+1)})\} > 0\right) \to 1. \tag{B.1}$$

And we note that

$$\begin{aligned}
& \mathrm{BIC}(\mathcal{S}^{(l)}) - \mathrm{BIC}(\mathcal{S}^{(l+1)}) \\
= \;& \log\left(\frac{\hat{\sigma}^2_{(\mathcal{S}^{(l)})}}{\hat{\sigma}^2_{(\mathcal{S}^{(l+1)})}}\right) - n^{-1}m_n(\log n + 2\log p m_n) \\
\ge \;& \log\left(1 + \frac{\hat{\sigma}^2_{(\mathcal{S}^{(l)})} - \hat{\sigma}^2_{(\mathcal{S}^{(l+1)})}}{\hat{\sigma}^2_{(\mathcal{S}^{(l+1)})}}\right) - 3n^{-1}m_n\log p - 2n^{-1}m_n\log m_n \\
\ge \;& \log\left(1 + \frac{n^{-1}\Omega(l)}{n^{-1}\|\mathbf{Y}\|^2}\right) - n^{-1}m_nO(n^{c_p}) - 2n^{-1}m_n\log m_n \\
\ge \;& \log\left(1 + \frac{c_1^2c_3c_f^2/c_2}{n^{-1}\|\mathbf{Y}\|^2}\right) - n^{-1}m_nO(n^{c_p}) - 2n^{-1}m_n\log m_n
\end{aligned}$$

where we use the fact $\hat{\sigma}^2_{(\mathcal{S}^{(l+1)})} \le n^{-1}\|\mathbf{Y}\|^2$ and the assumption that $\log p =$

$O(n^{c_p})$ with $0 < c_p < 2d/(2d+1)$.

Using the elementary inequality $\log(1+x) \geq \min\{\log 2, x/2\}$, we have

$$\begin{aligned} & \mathrm{BIC}(\mathcal{S}^{(l)}) - \mathrm{BIC}(\mathcal{S}^{(l+1)}) \\ \geq & \min\left\{\log 2, \frac{c_1^2 c_3 c_f^2/c_2}{2n^{-1}\|\mathbf{Y}\|^2}\right\} - n^{-1}m_n O(n^{c_p}) - 2n^{-1}m_n \log m_n \\ \to & \min\left\{\log 2, \frac{c_1^2 c_3 c_f^2/c_2}{2\mathrm{var}(Y)}\right\} > 0 \end{aligned} \tag{B.2}$$

as $n \to \infty$. This completes the proof. □

References

Belloni, A., Chen, D., Chernozhukov V. and Hansen, C. (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, **80**, 2369–2429.

de la Pena, V., Lai, T., and Shao, Q. (2009) *Self-normalized processes. Probability and its Applications (New York). Springer-Verlag, Berlin. Limit theory and statistical applications.*

Fan, Q. and Zhong, W. (2016), "Nonparametric Additive Instrumental Variable Estimator: a Group Shrinkage Estimation Perspective," *Journal of Business and Economic Statistics*, forthcoming.

Huang, J., Horowitz, J. and Wei, F. (2010), "Variable Selection in Nonparametric Additive

Models," *The Annals of Statistics,* **38**, 2282–2313.

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, **104**, 1512–1524.

Zhou, S., Shen, X. and Wolfe D. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, **26**, 1760–1782.

Wei Zhong, Wang Yanan Institute for Studies in Economics, Department of Statistics, School of Economics, Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen 361005, China.

E-mail: wzhong@xmu.edu.cn

Sunpeng Duan, Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen 361005, China.

E-mail: fredduan.dsp@gmail.com

Liping Zhu, Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

E-mail: zhu.liping@ruc.edu.cn