decrease more in south town. And the plot α_{13}(AGE) also indicates that a newer house is more expensive than a older one with similar condition.

Table 3.4 reports the averaged R^2, MAEE, MAPE and the percentage of 'good" predicted price over 10 replicates for SCAD, FULL and parametric model respectively. From Table 3.4, FULL gives a slightly better estimation than SCAD at the expense of a more complex model structure. However for prediction, SCAD outperforms FULL. Therefore, SCAD not only gives a simpler and more interpretable model, but also improves prediction accuracy. Furthermore, the parametric model gives the worst results in both estimation and prediction, indicating that the data contains a nonlinear structure that can not be fully explained by the parametric model. Figure 3.4 plots the randomly selected actual prices against corresponding predicted prices under three models separately, with criteria band enclosing "good" estimates. Again, it suggests both penalize model and full model do much better than the parametric model.

3.7 Proof of Lemmas and Theorems

Based on the oracle estimator $\widehat{\gamma}^{(\mathbf{0})} = (\widehat{\gamma}_{ls}^{(0)}, l = 1, \cdots, d_1, s = 0, \cdots, d_2)$ in Section 3.3, we further define $\widehat{\alpha}_{ls}^{(0)}$ for notation convenience. For $(l, 0) \in S^{(0)}$, $\widehat{\alpha}_{l0}^{(0)} = \widehat{\gamma}_{l0}^{(0)}$. For $s > 0$ and $(l, s) \in S^{(0)}$, $\widehat{\alpha}_{ls}^{(0)} = \widehat{\gamma}_{ls}^{(0)}\mathbf{B}_s$. Here $\mathbf{B}_s$ is the vector of the empirically centered spline basis on x_s defined in Section 3.2. For $(l, s) \notin S^{(0)}$, $\widehat{\alpha}_{ls}^{(0)} = 0$.

For any square matrix $\mathbf{U}$, we denote $\rho_{min}(\mathbf{U})$ and $\rho_{max}(\mathbf{U})$ as the minimal and maximal eigenvalues of $\mathbf{U}$ respectively. For notation simplicity, we use the same c, c_1, c_2 as general notations for positive constants with not necessarily the same value.

3.7.1 Preliminary Lemmas

Lemma 3.7.1. *Under Assumptions (C.3) - (C.4), for each pair of* $(l, s) \in S$*, the eigenvalues of* $\mathbf{W}_{ls}$ *are bounded by two positive constants with probability approaching to* 1.

That is, let $\rho_{min}(\mathbf{W}_{ls})$ *and* $\rho_{max}(\mathbf{W}_{ls})$ *be the minimal and maximal eigenvalues of* $\mathbf{W}_{ls}$ *respectively. Then there exist* $0 < c_1 < c_2$*, such that*

$$P\left(c_1 \leqslant \rho_{min}(\mathbf{W}_{ls}) \leqslant \rho_{max}(\mathbf{W}_{ls}) \leqslant c_2\right) \to 1 \text{ as } n \to \infty. \tag{3.8}$$

Proof. From Lemma 5 of Xue and Qu (2012), we know that (3.8) is true for $(l, s) \in S$ and $s = 0$. Now, for $s = 0$, $\rho\left(\mathbf{W}_{l0}\right) = \rho\left(\frac{\mathbf{t}_l^T \mathbf{t}_l}{n}\right) = \|T_l\|_n^2$. Therefore, according to Assumption (C.4), one has that $\|T_l\|_n^2 \leqslant c^2$. Furthermore, the Central Limit Theorem gives that, $\|T_l\|_n^2 \geqslant E\left(T_l^2\right) - o_P(1) \geqslant \mathrm{Var}(T_l)$. Therefore, by letting $c_1 = \min_{l=1,\cdots,d_1} \mathrm{Var}(T_l)$ and $c_2 = c^2$, Lemma 3.7.1 is proved.

Lemma 3.7.2. *Let* $\mathbf{Z} = (\mathbf{Z}_{ls}, (l, s) \in S)$*,* $\mathbf{W} = \frac{\mathbf{Z}^T \mathbf{Z}}{n}$*, then under Assumptions (C.3) - (C.4), the eigenvalues of* $\mathbf{W}$ *are bounded within two positive constants. That is, there exist* $c_2 > c_1 > 0$ *such that, except on an event whose probability tends to zero, as* $n \to \infty$*,*

$$c_1 \leqslant \rho_{min}(\mathbf{W}) \leqslant \rho_{max}(\mathbf{W}) \leqslant c_2. \tag{3.9}$$

Proof. For any coefficients $\gamma = (\gamma_{ls}^T, (l, s) \in S)^T \in \mathbb{R}^{d_1(1+d_2 J_n)}$, one has

$$\gamma^T \mathbf{W} \gamma = \sum_{l=1}^{d_1} \left(\gamma_{l0} + \sum_{s=1}^{d_2} \sum_{j=1}^{J_n} \gamma_{ls,j} B_{s,j} T_l \right)_n^2 .$$

Lemma A.5 in Xue and Yang (2006b) ensures that there exists a positive constants c_1 such that $\gamma^T \mathbf{W} \gamma \geqslant c_1 \sum_{l=1}^{d_1} \left(\gamma_{l0}^2 + \sum_{s=1}^{d_2} \sum_{j=1}^{J_n} \gamma_{ls,j}^2 \right) = c_1 \|\gamma\|^2$. On the other hand, Cauchy-Schwarz inequality gives that there exists a constant $c > 0$ such that

$$\gamma^T \mathbf{W} \gamma = \sum_{l=1}^{d_1} \left(\gamma_{l0} T_l + \sum_{s=1}^{d_2} \sum_{j=1}^{J_n} \gamma_{ls,j} B_{s,j} T_l \right)_n^2 \leqslant c \sum_{l=1}^{d_1} \left(\gamma_{l0}^T \frac{\mathbf{Z}_{l0}^T \mathbf{Z}_{l0}}{n} \gamma_{l0} + \sum_{s=1}^{d_2} \gamma_{ls}^T \frac{\mathbf{Z}_{ls}^T \mathbf{Z}_{ls}}{n} \gamma_{ls} \right) .$$

Therefore, from Lemma 3.7.1, there exists a positive c_2 such that $\gamma^T \mathbf{W} \gamma \leqslant c_2 \|\gamma\|^2$.

As a consequence of Lemmas 3.7.1 and 3.7.2, one has the following corollary.

Corollary 3.7.1.1. *Let* A *be a subset of the full index pair set* S*. Denote* $\mathbf{Z}_A = (\mathbf{Z}_{ls}, (l, s) \in A)$ *and* $\mathbf{W}_A = \frac{\mathbf{Z}_A^T \mathbf{Z}_A}{n}$*, then under Assumption (C.3) - (C.4), there exist two positive*

constants c_1, c_2, such that

$$P\left(c_1 \leqslant \rho_{min}(\mathbf{W}_A) \leqslant \rho_{max}(\mathbf{W}_A) \leqslant c_2\right) \to 1 \text{ as } n \to +\infty.$$

Lemma 3.7.3. *Under Assumptions (C.3) - (C.4) and (C.6), each additive term of oracle estimators, $\widehat{\alpha}_{ls}^{(0)}$, converges to the corresponding true function α_{ls} in probability. Specifically, we have*

$$\sum_{l=1}^{d1} \widehat{\alpha}_{l0}^{(0)} T_l - \alpha_{l0} T_l \; + \sum_{l=1}^{d1}\sum_{s=1}^{d2} \widehat{\alpha}_{ls}^{(0)}(X_s) T_l - \alpha_{ls}(X_s) T_l \;_2 = O_P \quad N_n^{-(p+1)} + \sqrt{\frac{N_n}{n}} \quad .$$

Proof. Theorem 1 in Xue and Yang (2006) entails that

$$\max_{1\leqslant l\leqslant d_1} \widehat{\alpha}_{l0}^{(0)} - \alpha_{l0} \; + \max_{1\leqslant l\leqslant d_1, 1\leqslant s\leqslant d_2} \widehat{\alpha}_{ls}^{(0)}(X_s) - \alpha_{ls}(X_s) \;_2 = O_P \quad N_n^{-(p+1)} + \sqrt{\frac{N_n}{n}} \quad .$$

Therefore, Lemma follows from Assumption (C.4).

3.7.2 Proof of Theorem 3.3.1

For notation simplicity, denote

$$L_n(\gamma) = \frac{1}{2n} \; \mathbf{Y} - \sum_{l=1}^{d_1}\sum_{s=0}^{d_2} \mathbf{Z}_{ls}(\mathbf{x}_s, \mathbf{t}_l)\gamma_{ls} \;_2^2 + \sum_{l=1}^{d_1}\sum_{s=0}^{d_2} p_{\lambda_n}\left(\|\gamma_{ls}\|_{\mathbf{W}_{ls}}\right).$$

Now, let $\mathbf{Z}_{ls}^* = \mathbf{Z}_{ls}\mathbf{W}_{ls}^{-\frac{1}{2}}$ and $\gamma_{ls}^* = \mathbf{W}_{ls}\gamma_{\mathbf{ls}}$. Consequently one has $\mathbf{W}_{ls}^* = \frac{\mathbf{Z}_{ls}^{*T}\mathbf{Z}_{ls}^*}{n} = \mathbf{I}_{J_n}$ for $s = 0$, and $\mathbf{W}_{l0}^* = 1$. Therefore, (3.3) can be rewritten as

$$\begin{aligned}
&\widehat{\gamma}^* = \underset{\gamma^*}{\text{argmin}} L_n(\gamma^*) \\
= \;& \underset{\gamma^*}{\text{argmin}} \left\{ \frac{1}{2n} \; \mathbf{Y} - \sum_{l=1}^{d_1}\sum_{s=0}^{d_2} \mathbf{Z}_{ls}^*(\mathbf{x}_s, \mathbf{t}_l)\gamma_{ls}^* \;_2^2 + \sum_{l=1}^{d_1}\sum_{s=0}^{d_2} p_{\lambda_n}\left(\|\gamma_{ls}^*\|_{\mathbf{W}_{ls}^*}\right) \right\} \\
= \;& \underset{\gamma^*}{\text{argmin}} \left\{ \frac{1}{2n} \; \mathbf{Y} - \sum_{l=1}^{d_1}\sum_{s=0}^{d_2} \mathbf{Z}_{ls}^*(\mathbf{x}_s, \mathbf{t}_l)\gamma_{ls}^* \;_2^2 + \sum_{l=1}^{d_1}\sum_{s=1}^{d_2} p_{\lambda_n}\left(\|\gamma_{ls}\|_2\right) + \sum_{l=1}^{d_1} p_{\lambda_n}\left(|\gamma_{l0}^*|\right) \right\}.
\end{aligned}$$

Therefore, for any given index pair $(l, s) \in S$, the partial derivative

$$\frac{\partial L_n(\gamma^*)}{\partial \gamma_{ls}^*} = -\frac{1}{n}\mathbf{Z}_{ls}^{*T} \; \mathbf{Y} - \sum_{l'=1}^{d_1}\sum_{s'=0}^{d_2} \mathbf{Z}_{l's'}^*(\mathbf{x}_{s'}, \mathbf{t}_{l'})\gamma_{l's'}^* \; + \partial p_{\lambda_n}\left(\|\gamma_{ls}^*\|_2\right),$$

where $\partial p_{\lambda_n}\left(\|\gamma^*_{ls}\|_2\right)$ is the subgradient of $p_{\lambda_n}\left(\|\gamma^*_{ls}\|_2\right)$.

We denote $\mathbf{C}^*_{ls}(\gamma^*) = -\frac{1}{n}\mathbf{Z}^{*T}_{ls}\left[\mathbf{Y} - \sum_{l'=1}^{d_1}\sum_{s'=0}^{d_2}\mathbf{Z}^*_{l's'}(\mathbf{x}_{s'}, \mathbf{t}_{l'})\gamma^*_{l's'}\right]$. Then the KKT local optimality condition suggests that, any γ^* satisfying the following two conditions must be a local minimum of our penalized objective function,

$$\text{(i)} \quad \mathbf{C}^*_{ls}(\gamma^*) = 0, \|\gamma^*_{ls}\|_2 > a\lambda_n, \text{ for } (l,s) \in S^{(0)};$$

$$\text{(ii)} \quad \|\mathbf{C}^*_{ls}(\gamma^*)\|_2 \leqslant \lambda_n, \|\gamma^*_{ls}\|_2 < \lambda_n, \text{ for } (l,s) \notin S^{(0)}.$$

Equivalently in terms of γ, the vector of untransformed coefficients, the sufficient conditions for a solution to be a local minimum are

$$\text{(i)} \quad \mathbf{C}_{ls}(\gamma) = \mathbf{0}, \|\gamma_{ls}\|_{\mathbf{W}_{ls}} > a\lambda_n, \text{ for } (l,s) \in S^{(0)}; \tag{3.10}$$

$$\text{(ii)} \quad \|\mathbf{C}_{ls}(\gamma)\|_{\mathbf{W}^{-1}_{ls}} \leqslant \lambda_n, \|\gamma_{ls}\|_{\mathbf{W}_{ls}} < \lambda_n, \text{ for } (l,s) \notin S^{(0)}. \tag{3.11}$$

Therefore, to prove $\widehat{\gamma}^{(0)} \in A_n(\lambda_n)$, one only needs to prove (3.10) and (3.11) for $\gamma = \widehat{\gamma}^{(\mathbf{0})}$.

When $\gamma = \widehat{\gamma}^{(\mathbf{0})}$, the first equation in (3.10) and the second inequality in (3.11) naturally hold by the definition of $\widehat{\gamma}^{(0)}$. So one only needs to prove that, except on an event whose probability tends to 0, as $n \to \infty$,

$$\text{(i)} \quad \widehat{\gamma}^{(0)}_{ls}{}_{\mathbf{W}_{ls}} > a\lambda_n, \text{ for } (l,s) \in S^{(0)}; \tag{3.12}$$

$$\text{(ii)} \quad \mathbf{C}_{ls}(\widehat{\gamma}^{(0)}_{ls})_{\mathbf{W}^{-1}_{ls}} \leqslant \lambda_n, \text{ for } (l,s) \notin S^{(0)}. \tag{3.13}$$

We first prove (3.12). Note that $\widehat{\gamma}^{(0)}_{ls}{}^2_{\mathbf{W}_{ls}} = \widehat{\alpha}^{(0)}_{ls}(X_s)T_l{}^2_n$ is the empirical norm of an non-zero additive term of the oracle estimator. Lemma 3.7.3 implies that,

$$\begin{aligned}
\widehat{\gamma}^{(0)}_{ls}{}_{\mathbf{W}_{ls}} &= \widehat{\alpha}^{(0)}_{ls}(X_s)T_l \left.\frac{\widehat{\alpha}^{(0)}_{ls}(X_s)T_l}{\widehat{\alpha}^{(0)}_{ls}(X_s)T_l}\right|_n \\
&\geqslant [\|\alpha_{ls}(X_s)T_l\|_2 - o_P(1)] \left.\frac{\widehat{\alpha}^{(0)}_{ls}(X_s)T_l}{\widehat{\alpha}^{(0)}_{ls}(X_s)T_l}\right|_n.
\end{aligned}$$

Furthermore, Assumptions (C.2), (C.6) and Lemma A.4 in Xue and Yang (2006b) imply that the second multiplicative term converges to 1 in probability. Together with Assumptions (C.1), (C.3), and (iii) of Assumption (C.5), one has that, with probability goes to 1,

$$\begin{aligned}
\widehat{\gamma}_{ls}^{(0)} \quad _{W_{ls}} &\geqslant \|\alpha_{ls}(X_s)T_l\|_2 - o_P(1) \\
&= \sqrt{E\left[\alpha_{ls}^2(X_s)T_l^2\right]} - o_P(1) \\
&= \sqrt{E\left[\alpha_{ls}^2(X_s)E\left\{T_l^2|X_s\right\}\right]} - o_P(1) \\
&\geqslant c\sqrt{E\left[\alpha_{ls}^2(X_s)\right]} - o_P(1) \\
&= c - o_P(1) \geqslant a\lambda_n. \qquad (3.14)
\end{aligned}$$

Therefore (3.12) is proved. Now for (3.13), we define $\mathbf{Z}_{(1)} = (\mathbf{Z}_{ls}, (l,s) \in S^{(0)})$ as the column-wise combination of all $\mathbf{Z}_{ls}$ matrices corresponding to all "nonzero" components, and $\mathbf{Z}_{(2)} = (\mathbf{Z}_{ls}, (l,s) \notin S^{(0)})$ as the column-wise combination of all $\mathbf{Z}_{ls}$ matrices corresponding to all "redundant" components. Recall that $Y = \sum_{l=1}^{d_1} \sum_{s \in S_l} \alpha_{ls}(X_s)T_l + \varepsilon$. One has

$$\begin{aligned}
\mathbf{C}_{ls}(\widehat{\gamma}^{(0)}) &= \frac{1}{n}\mathbf{Z}_{ls}^T \left[\mathbf{I}_n - \mathbf{Z}_{(1)}\left(\mathbf{Z}_{(1)}^T\mathbf{Z}_{(1)}\right)^{-1}\mathbf{Z}_{(1)}^T\right]\mathbf{Y} \\
&= \frac{1}{n}\mathbf{Z}_{ls}^T\mathbf{H}_n\left[\delta + \varepsilon\right],
\end{aligned}$$

where $\mathbf{H}_n = \mathbf{I}_n - \mathbf{Z}_{(1)}\left(\mathbf{Z}_{(1)}^T\mathbf{Z}_{(1)}\right)^{-1}\mathbf{Z}_{(1)}^T$ and $\delta = \sum_{(l,s)\in S^{(0)}} \delta_{ls}$ with $\delta_{ls} = (\delta_{1,ls}, \cdots, \delta_{n,ls})^T$ and $\delta_{i,ls} = \alpha_{ls}(x_{i,s})t_{i,l} - \widehat{\alpha}_{ls}^{(0)}(x_{i,s})t_{i,l}$. Therefore,

$$\begin{aligned}
P\left(\left[\mathbf{C}_{ls}(\widehat{\gamma}^{(0)})\right]_{\mathbf{W}_{ls}^{-1}} > \lambda_n, \exists (l,s) \notin S^{(0)}\right) &\leqslant P\left(\max_{(l,s)\notin S^{(0)}} \frac{1}{n}\left[\mathbf{Z}_{ls}^T\mathbf{H}_n\delta\right]_{\mathbf{W}_{ls}^{-1}} > \frac{\lambda_n}{2}\right) \\
&+ P\left(\max_{(l,s)\notin S^{(0)}} \frac{1}{n}\left[\mathbf{Z}_{ls}^T\mathbf{H}_n\varepsilon\right]_{\mathbf{W}_{ls}^{-1}} > \frac{\lambda_n}{2}\right). \qquad (3.15)
\end{aligned}$$

According to Lemma 3.7.1, one has

$$\max_{(l,s)\notin S^{(0)}} \frac{1}{n}\left[\mathbf{Z}_{ls}^T\mathbf{H}_n\delta\right]_{\mathbf{W}_{ls}^{-1}} \leqslant \max_{(l,s)\notin S^{(0)}} \frac{c}{n}\left[\mathbf{Z}_{ls}^T\mathbf{H}_n\delta\right]_2 \leqslant \frac{c}{\sqrt{n}}\|\mathbf{H}_n\delta\|_2.$$

Note that $\mathbf{H}_n \leqslant \mathbf{I}_n$. That is, $\mathbf{I}_n - \mathbf{H}_n$ is semi-positive definite. The approximation theory in de Boor (2001) (p.149) gives that $\|\delta\|_2 \leqslant c\sqrt{n}N_n^{-(p+1)}$. Therefore, one has

$$\max_{(l,s)\notin S^{(0)}} \frac{1}{n} \ \left\| \mathbf{Z}_{ls}^T \mathbf{H}_n \delta \right\|_{\mathbf{W}_{ls}^{-1}} \leqslant \frac{1}{\sqrt{n}c_1} \|\delta\|_2 \leqslant cN_n^{-(p+1)}.$$

Consequently, (i) of Assumption (C.5) entails that

$$\lim_{n\to\infty} P\left(\max_{(l,s)\notin S^{(0)}} \frac{1}{n} \ \left\| \mathbf{Z}_{ls}^T \mathbf{H}_n \delta \right\|_{\mathbf{W}_{ls}^{-1}} > \frac{\lambda_n}{2} \right) = 0. \tag{3.16}$$

Similarly, one has $\max_{(l,s)\notin S^{(0)}} \frac{1}{n} \ \left\| \mathbf{Z}_{ls}^T \mathbf{H}_n \varepsilon \right\|_{\mathbf{W}_{ls}^{-1}} \leqslant \frac{1}{c_1 n} \max_{(l,s)\notin S^{(0)}} \ \left\| \mathbf{Z}_{ls}^T \mathbf{H}_n \varepsilon \right\|_2$. On the other hand, denote $\mathbf{H}^{(2)T} = \mathbf{Z}_{(2)}^T - \mathbf{Z}_{(2)}^T \mathbf{Z}_{(1)} \left(\mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \right)^{-1} \mathbf{Z}_{(1)}^T$ and let $\mathbf{H}_{ls}^{(2)}$ be the J_n columns of $\mathbf{H}^{(2)}$ corresponding to $\mathbf{Z}_{ls}$ with $\mathbf{H}_{ls}^{(2)} = \mathbf{Z}_{ls}^T \mathbf{H}_n$ for $(l,s) \in S^{(0)}$. Note that $\mathbf{H}^{(2)}\mathbf{H}^{(2)T} = \mathbf{Z}_{(2)}\mathbf{H}_n\mathbf{Z}_{(2)}^T \leqslant \mathbf{Z}_{(2)}\mathbf{Z}_{(2)}^T$. Therefore, one has $\left\| \mathbf{Z}_{ls}^T \mathbf{H}_n \varepsilon \right\|_2 \leqslant \left\| \mathbf{Z}_{ls}^T \varepsilon \right\|_2$. Consequently,

$$\max_{(l,s)\notin S^{(0)}} \frac{1}{n} \ \left\| \mathbf{Z}_{ls}^T \mathbf{H}_n \varepsilon \right\|_{\mathbf{W}_{ls}^{-1}} \leqslant \frac{1}{c_1 n} \max_{(l,s)\notin S^{(0)}} \ \left\| \mathbf{Z}_{ls}^T \varepsilon \right\|_2.$$

According to Lemma 7 in Xue and Qu (2012),

$$E\left(\max_{(l,s)\notin S^{(0)}} \frac{1}{\sqrt{n}} \ \left\| \mathbf{Z}_{ls}^T \varepsilon \right\|_2 \right) = O\left(\sqrt{\log\left(N_n d_1 d_2\right)} \right).$$

Then Markov inequality implies

$$P\left(\max_{(l,s)\notin S^{(0)}} \frac{1}{n} \ \left\| \mathbf{Z}_{ls}^T \mathbf{H}_n \varepsilon \right\|_{\mathbf{W}_{ls}^{-1}} > \frac{\lambda_n}{2} \right) \leqslant C\sqrt{\frac{\log\left(N_n\right)}{n\lambda_n^2}}. \tag{3.17}$$

Then (3.13) follows from (3.16), (3.17) and (ii) of Assumption (C.5).

3.7.3 Proof of Theorem 3.3.2

To prove $\widehat{\gamma}^{(0)}$ converges to the global minimum in (3.3), it suffices to show that

$$P\left(L_n(\gamma) \geqslant L_n(\widehat{\gamma}(0)) \text{ for all } \gamma \in \mathbb{R}^{d_1 d_2 J_n} \right) \to 1 \text{ as } n \to \infty. \tag{3.18}$$

We first define several subset of the index pair set $S = \{(l,s) | l = 1, \cdots, d_1, s = 1, \cdots, d_2\}$. Let $\mathcal{P} = S^{(0)}$ and $\mathcal{N} = S/S^{(0)} = \left\{(l,s) | (l,s) \notin S^{(0)}\right\}$. For any given $\gamma \in \mathbb{R}^{d_1 d_2 J_n}$, let

$$\begin{aligned} \mathcal{P}^+ &= \left\{ (l,s) \in \mathcal{P} = S^{(0)}, \|\gamma_{ls}\|_{\mathbf{W}_{ls}} > a\lambda_n \right\}, \ \mathcal{P}^- = \mathcal{P}/\mathcal{P}^+; \\ \mathcal{N}^+ &= \left\{ (l,s) \in \mathcal{N} = S/S^{(0)}, \|\gamma_{ls}\|_{\mathbf{W}_{ls}} > \lambda_n \right\}, \ \mathcal{N}^- = \mathcal{N}/\mathcal{N}^+. \end{aligned}$$

Zou, H. and Li, R. (2008). *One-step sparse estimates in nonconcave penalized likelihood models*. *Annals of Statistics*, **36**, 1509–1533.