Supplemental Materials for "Separation of Covariates into Nonparametric and Parametric Parts in High-Dimensional Partially Linear Additive Models"

by Heng Lian, Hua Liang and David Ruppert

Nanyang Technological University, George Washington University, and Cornell University

Proofs of Main Results

In this Appendix, we present the conditions, prepare several preliminary results, and give the proofs of the main results.

A.1 Notations and assumptions

We first introduce some notations and additional definitions. In our proofs, C denotes a generic positive constant that might assume different values at different places. $b_0 = (b_{01}^T, \ldots, b_{0p}^T)^T$ denotes a pK-dimensional vector that satisfies $\|f_{0j} - b_{0j}^T B_j\| = O(K^{-d})$ for $1 \le j \le p_1$ and $f_{0j} = b_{0j}^T B_j$ for $j > p_1$. Due to Proposition 1, we will frequently use centered covariate $X_{ij} - \sum_i X_{ij}/n$. For simplicity in notation, we assume such centering has been done and we still use X_{ij} to denote it. Let $Z^{(1)} = (Z_1, \ldots, Z_{p_1})$ be the $n \times p_1 K$ submatrix of Z containing the columns corresponding to truly nonparametric components, and similarly let $Z^{(2)}$ be the submatrix corresponding to parametric components and $Z^{(3)}$ the submatrix corresponding to zero components. In the same spirit, we can define $X^{(1)}, X^{(2)}, X^{(3)}$ as suitable submatrices of the $n \times p$ covariate matrix X. Similar notations are also applied to other vectors such as $b = (b^{(1)T}, b^{(2)T}, b^{(3)T})^T$.

Let $\mathcal{A}$ denote the subspace of functions on R^{p_1} that take an additive form

$$\mathcal{A} := \{h(x^{(1)}) : h(x^{(1)}) = h_1(x_1) + \cdots + h_{p_1}(x_{p_1}), Eh_j(X_j)^2 < \infty \text{ and } Eh_j(X_j) = 0\}$$

and for any random variable W with $E(W^2) < \infty$, let $E_{\mathcal{A}}(W)$ denote the projection of W onto $\mathcal{A}$ in the sense that

$$E\left[\{W - E_{\mathcal{A}}(W)\}\{W - E_{\mathcal{A}}(W)\}^T\right] = \inf_{h \in \mathcal{A}} E\left[\{W - h(X^{(1)})\}\{W - h(X^{(1)})\}^T\right].$$

Definition of $E_{\mathcal{A}}(W)$ trivially extends to the case W is a random vector by componentwise projection.

In the theoretical studies of our estimator, we will use the decomposition

$$X^{(2)} = \theta(X^{(1)}) + U = \theta(X^{(1)}) - h(X^{(1)}) + h(X^{(1)}) + U, \tag{A.1}$$

with $\theta(X^{(1)}) = E(X^{(2)}|X^{(1)})$, $h(X^{(1)}) = E_{\mathcal{A}}(X^{(2)})$ and $U = X^{(2)} - E(X^{(2)}|X^{(1)})$. In the decomposition above, each component of $h(X^{(1)}) = (h_{(1)}(X^{(1)}), \ldots, h_{(p_2)}(X^{(1)}))^T$ can be written in the form $h_{(s)}(x) = \sum_{j=1}^{p_1} h_{(s)j}(x_j)$ for some $h_{(s)j} \in S_j^0$.

Note that since the conditional expectation $E(X^{(2)}|X^{(1)})$ can be interpreted as projection onto the space $\{h(X^{(1)}) : Eh^2 < \infty\}$ of which $\mathcal{A}$ is a subspace, we see that we also have $h(X^{(1)}) = E_{\mathcal{A}}(\theta(X^{(1)}))$. Let $\Xi = E\{X^{(2)} - h(X^{(1)})\}^{\otimes 2}$.

In some of the proofs below we will make use of the concept of subdifferential and thus we first mention the following frequently used fact. For any matrix A and vector b (as long as the dimensions are compatible), the subdifferential of $||Ab||$ is

$$\partial||Ab|| = \begin{cases} \{A^T Ab/||Ab||\} & \text{if } Ab \neq 0 \\ \{A^T Aa : ||Aa|| \leq 1\} & \text{if } Ab = 0. \end{cases}$$

Note that the subdifferential is a set and its elements are called subgradients. When $Ab = 0$ the subgradient is not unique. In the following we will use the same notation $\partial||Ab||$ to denote either subdifferential or subgradient, when the specific element selected has no significance in our proofs.

The following regularity conditions are used.

(c1) The covariate vector X has a continuous density supported on $[0,1]^p$. Furthermore, the marginal densities for $X_j, 1 \leq j \leq p$ are all bounded from below and above by two fixed positive constants respectively.

(c2) The noises ϵ_i are independent of covariates, have mean zero, variance σ^2, and have sub-Gaussian tails.

(c3) The number of nonzero components s is fixed. $Ef_j(X_j) = 0, 1 \leq j \leq s$. $f_j(x)$ is linear in x for $p_1 + 1 \leq j \leq s$, and $f_j \equiv 0$ for $j > s$.

(c4) For $g = f_j, 1 \leq j \leq p_1$ or $g = h_{(s)j}, 1 \leq s \leq p_2, 1 \leq j \leq p_1$, g satisfies a Lipschitz condition of order $d > 1/2$: $|g^{(\lfloor d \rfloor)}(t) - g^{(\lfloor d \rfloor)}(s)| \leq C|s-t|^{d-\lfloor d \rfloor}$,

where $\lfloor d \rfloor$ is the biggest integer strictly smaller than d and $g^{(\lfloor d \rfloor)}$ is the $\lfloor d \rfloor$-th derivative of g. The order of the B-spline used satisfies $q \geq d + 2$.

(c5) For some fixed positive constants c, C, $c \leq \lambda_{\min}(A_j) \leq \lambda_{\max}(A_j) \leq C$, where $\lambda_{\min}(.)$ and $\lambda_{\max}(.)$ denote the smallest and largest eigenvalue respectively.

(c6) For any $b \in R^K$, $\|b\|_{D_j} = 0$ if and only if $b^T B_j(x)$ is a linear function. D_j is nonnegative definite, of rank $K - 1$, and all its $K - 1$ nonzero eigenvalues are bounded and bounded away from zero. With abuse of notation, we use $\lambda_{\min}(D_j)$ to denote its minimal *positive* eigenvalue.

(c7) $\sqrt{n/K}\{\sqrt{\log(pK)}+\sqrt{K+n/K^{2d}}+\sqrt{nK}(\lambda_1||w_1^0||+\lambda_2||w_2^0||)\} = o_p(n\lambda_2 w_{2j})$ for $p_1+1 \leq j \leq s$ and $\sqrt{n/K}\{\sqrt{\log(pK)}+\sqrt{K+n/K^{2d}}+\sqrt{nK}(\lambda_1||w_1^0||+\lambda_2||w_2^0||)\} = o_p(n\lambda_1 w_{1j})$ for $s+1 \leq j \leq p$.

(c8) The eigenvalues of Ξ are bounded away from zero and infinity.

(c9) $\min_{1\leq j\leq s} \|f_{0j}\|$ and $\min_{1\leq j\leq p_1} \inf_{a,b\in R} \|f_{0j}(x) - ax - b\|^2$ are bounded away from zero.

Most of the assumptions are standard in the literature. We assumed s is fixed in (c3) as in Huang et al. (2010). Some discussions on relaxing this are contained in the Supplementary Material. Assumption (c4) is used to control approximation error. Assumption (c7) looks quite complicated. These expressions roughly require that the weights $w_{1j}, j > s$ associated with zero components and the weights $w_{2j}, p_1 < j \leq s$ associated with parametric components should be sufficiently large. When the weights are defined by the initial lasso estimator, the expressions in (c7) can be made clearer as discussed in Section 2.3. Assumption (c8) is necessary for identifiability of the linear components, as argued in Li (2000) for partially linear additive models. Assumption (c9) roughly speaking imposes condition to distinguish different types of components. Assumptions (c5) and (c6) may seem more restrictive than they really are. What is really meant is that the maximum and minimum positive eigenvalues of A_j and D_j are of the same asymptotic order. For example, if the (k, k') entry of A_j is $\int B_k B_{k'}$, its minimum and maximum eigenvalues are of order $1/K$ (Huang et al. (2010)). A simple multiplication by K makes A_j satisfy the assumption. In practice, this kind of rescaling is of course unnecessary and assuming eigenvalues to be bounded and bounded away

from zero, instead of that they are of the same order, is only used for convenience in proof.

A.2 Proof of Proposition 1

If $\|b_j\|_{D_j} = 0$ then by the stated assumption we have $b_j^T B_j(x) \equiv a_j x + c_j$ for some $a_j, c_j \in R$. Since $b_j^T B_j(x)$ is centered (that is, $b_j^T B_j(x) \in S_j^0$), we have $c_j = -a_j \bar{X}_j$, and thus $\|b_j\|^2/K \sim \|b_j^T B_j(x)\|^2 = \|a_j(x - \bar{X}_j)\|^2 \sim a_j^2$. The linearity of the mapping is obvious. The uniqueness is also easy to show since $B_{jk}, k = 1, \ldots, K$ are assumed to be linearly independent. The other direction is obvious.

Equivalence between (ii) and (iii) is trivial. We also note that such ξ_j is unique and $\|\xi_j\| \sim \sqrt{K}$ by part (ii). □

A.3 Some preliminary results on regularized oracle estimator

The strategy of proof for our main results is by way of considering the following "regularized oracle estimator". If we had the additional knowledge regarding which components are zero or linear, we could take into account this information and instead minimize the following constrained problem:

$$\begin{aligned} \min_b \quad & \frac{1}{2}||Y - Zb||^2 + n\lambda_1 \sum_{j=1}^{p} w_{1j}||b_j||_{A_j} + n\lambda_2 \sum_{j=1}^{p} w_{2j}||b_j||_{D_j}, \\ s.t. \quad & \|b_j\|_{A_j} = 0, j = s+1, \ldots, p \quad \text{and} \quad \|b_j\|_{D_j} = 0, j = p_1 + 1, \ldots, s. \end{aligned} \tag{A.2}$$

We will show in Section A.4 that the solutions to (1.4) and (A.2) are the same with probability approaching one under appropriate conditions, and thus use $\hat{b}$ to denote the minimizer for both. As an immediate corollary, the zero and linear components are correctly identified with probability approaching one, and the convergence rates and asymptotic normality results stated below for (A.2) also apply to estimators obtained from (1.4).

The following lemma is the key to our theoretical investigations, which characterize the solution to the regularized oracle estimator.

Lemma A.1 *A sufficient and necessary condition for* $b = (b^{(1)T}, b^{(2)T}, b^{(3)T})^T$ *to be*

the solution of (A.2) is that

$$0 \in -Z_j^T(Y - Zb) + n\lambda_1 w_{1j}\partial\|b_j\|_{A_j} + n\lambda_2 w_{2j}\partial\|b_j\|_{D_j},\ j = 1, \ldots, p_1, \quad \text{(A.3)}$$

$$-Z_j^T(Y - Zb) + n\lambda_1 w_{1j}\partial\|b_j\|_{A_j} \in span(D_j),\ j = p_1 + 1, \ldots, s, \text{ and} \quad \text{(A.4)}$$

$$b^{(3)} = 0, \|b_j\|_{D_j} = 0,\ j = p_1 + 1, \ldots, s,$$

where $span(D_j)$ *is the linear subspace of* R^K *spanned by columns of* D_j.

Proof. Denote $Z^{(1,2)} = (Z^{(1)}, Z^{(2)})$ and similarly $b^{(1,2)} = (b^{(1)T}, b^{(2)T})^T$. Due to the constraints, (A.2) is obviously equivalent to $b^{(3)} = 0$ together with the following minimization problem for $b^{(1,2)}$:

$$\min_b \quad Q(b) = \frac{1}{2}||Y - Z^{(1,2)}b^{(1,2)}||^2 + n\lambda_1 \sum_{j=1}^{s} w_{1j}||b_j||_{A_j} + n\lambda_2 \sum_{j=1}^{p_1} w_{2j}||b_j||_{D_j}, \quad \text{(A.5)}$$

$$s.t. \quad \|b_j\|_{D_j} = 0, j = p_1 + 1, \ldots, s.$$

Denote $F : R^K \to R \cup \{\infty\}$ with $F(x) = 0$ if $x = 0$ and ∞ otherwise. Note F is a convex function (see for example section III.4 in Ekeland and Turnbull (1983) for the definition of convex function that can take value ∞). Using F, the constrained problem (A.5) can be written as an unconstrained convex problem

$$\min_{b^{(1,2)}} \frac{1}{2}\|Y - Z^{(1,2)}b^{(1,2)}\|^2 + \lambda_1 \sum_{j=1}^{s} w_{1j}\|b_j\|_{A_j} + \lambda_2 \sum_{j=1}^{p_1} w_{2j}\|b_j\|_{D_j} + \sum_{j=p_1+1}^{s} F(D_j b_j).$$

Using the KKT condition (Proposition III.3.1 in Ekeland and Turnbull (1983)), a sufficient and necessary condition for $b^{(1,2)}$ to be its solution is that

$$0 \in -Z_j^T(Y - Z^{(1,2)}b^{(1,2)}) + n\lambda_1 w_{1j}\partial\|b_j\|_{A_j} + n\lambda_2 w_{2j}\partial\|b_j\|_{D_j},\ j = 1, \ldots, p_1 \quad \text{(A.6)}$$

$$0 \in -Z_j^T(Y - Z^{(1,2)}b^{(1,2)}) + n\lambda_1 w_{1j}\partial\|b_j\|_{A_j} + \partial F(D_j b_j),\ j = p_1 + 1, \ldots, s \quad \text{(A.7)}$$

Furthermore, we have, by Proposition III.2.12 in Ekeland and Turnbull (1983),

$$\partial F(D_j b_j) = \begin{cases} span(D_j) & \text{if } D_j b_j = 0, \\ \emptyset & \text{otherwise .} \end{cases}$$

and thus (A.7) is same as (A.4) together with $D_j b_j = 0,\ j = p_1 + 1, \ldots, s.$ □

We present the following two results regarding the convergence of the regularized

oracle estimator. These are obviously the same as those stated in Theorems 2 and 3 respectively for the doubly penalized estimator.

Lemma A.2 *(Convergence rates) Under conditions (c1)–(c6) and that* $K \log K/n \to 0, K \to \infty$*, the estimator obtained from (A.2) satisfies*

$$\sum_{j=1}^{s} ||\hat{f}_j - f_{0j}||^2 = O_p\left(\frac{K}{n} + \frac{1}{K^{2d}} + (\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2)K\right).$$

For the parametric components, under the additional assumption (c8) and that $\sqrt{n}/K^{2d} \to 0$*, we have the faster rate*

$$\sum_{j=p_1+1}^{s} (\hat{\beta}_j - \beta_{0j})^2 = O_p\left(\frac{1}{n} + (\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2)K\right).$$

Proof. The first part is relatively easy to establish.

Using (A.5) and the definition of $\hat{b}$, we have

$$\begin{aligned}
0 &\geq Q(\hat{b}) - Q(b_0) \\
&\geq ||Y - Z^{(1,2)}\hat{b}^{(1,2)}||^2/2 - ||Y - Z^{(1,2)}b_0^{(1,2)}||^2/2 \\
&\quad -n\lambda_1 \sum_{j=1}^{s} w_{1j}||\hat{b}_j - b_{0j}||_{A_j} - n\lambda_2 \sum_{j=1}^{p_1} w_{2j}||\hat{b}_j - b_j^0||_{D_j} \\
&= (Y - Z^{(1,2)}b_0^{(1,2)})^T Z^{(1,2)}(b_0^{(1,2)} - \hat{b}^{(1,2)}) + ||Z^{(1,2)}(b_0^{(1,2)} - \hat{b}^{(1,2)})||^2/2 \\
&\quad -n\lambda_1 \sum_{j=1}^{s} w_{1j}||\hat{b}_j - b_{0j}||_{A_j} - n\lambda_2 \sum_{j=1}^{p_1} w_{2j}||\hat{b}_j - b_{0j}||_{D_j}. \qquad \text{(A.8)}
\end{aligned}$$

Let $\eta = P_{Z^{(1,2)}}(Y - Z^{(1,2)}b_0^{(1,2)})$, where $P_{Z^{(1,2)}} = Z^{(1,2)}(Z^{(1,2)T}Z^{(1,2)})^{-1}Z^{(1,2)T}$ is the matrix of projection onto the columns of $Z^{(1,2)}$. Denote $r_i = \sum_{j=1}^{p_1} f_{0j}(X_{ij})$ and $r = (r_1, \ldots, r_n)^T$. We have $Y - Zb_0 = \epsilon + (r - Z^{(1)}b_0^{(1)})$ and $||\eta||^2 \leq 2||P_{Z^{(1,2)}}\epsilon||^2 + 2||r - Z^{(1)}b_0^{(1)}||^2$. By the approximation property of splines, $||r - Z^{(1)}b_0^{(1)}||^2 = O_p(n/K^{2d})$. Also, $E||P_{Z^{(1,2)}}\epsilon||^2 = E(\epsilon^T P_{Z^{(1,2)}}\epsilon) = \sigma^2 tr(P_{Z^{(1,2)}}) = O_p(K)$ and an application of Markov inequality gives

$$\|\eta\|^2 = O_p(K + n/K^{2d}). \qquad \text{(A.9)}$$

Using the Cauchy-Schwartz inequality, equation (A.8) can be continued as

$$0 \geq -|O_p(K+n/K^{2d})| + \frac{1}{4}||Z^{(1,2)}(b_0^{(1,2)} - \hat{b}^{(1,2)})||^2 \\ -n\lambda_1 \sum_{j=1}^{s} w_{1j}||\hat{b}_j - b_j^0||_{A_j} - n\lambda_2 \sum_{j=1}^{p_1} w_{2j}||\hat{b}_j - b_j^0||_{D_j}. \quad \text{(A.10)}$$

Using now properties of the matrix Z as in Huang et al. (2010), we get $||Z^{(1,2)}(b_0^{(1,2)} - \hat{b}^{(1,2)})||^2 \sim (n/K)||b_0^{(1,2)} - \hat{b}^{(1,2)}||^2$. Furthermore, since $\|b_0^{(1,2)} - \hat{b}^{(1,2)}\|_{A_j} \leq C\|b_0^{(1,2)} - \hat{b}^{(1,2)}\|$, it follows from Cauchy-Schwartz inequality that, for a sufficiently large $C > 0$,

$$n\sum_{j=1}^{s} \lambda_1 w_{1j}||b_0^{(1,2)} - \hat{b}^{(1,2)}||_{D_j} \leq \frac{CKn}{4}\sum_{j=1}^{s}(\lambda_1 w_{1j})^2 + (n/CK)||b_0^{(1,2)} - \hat{b}^{(1,2)}||^2$$
$$n\sum_{j=1}^{p_1} \lambda_2 w_{2j}||b_0^{(1,2)} - \hat{b}^{(1,2)}||_{D_j} \leq \frac{CKn}{4}\sum_{j=1}^{s}(\lambda_1 w_{1j})^2 + (n/CK)||b_0^{(1,2)} - \hat{b}^{(1,2)}||^2,$$

which along with (A.10) implies $||b_0^{(1,2)} - \hat{b}^{(1,2)}||^2 = O_p(K^2/n + 1/K^{2d-1} + (\lambda_1^2 \sum_{j=1}^{s} w_{1j}^2 + \lambda_2^2 \sum_{j=1}^{p_1} w_{2j}^2)K^2)$. Using the definition of b_0, we get the rates given in (1.5).

Now consider the faster convergence rates for the parametric components, which we show by profiling out $b^{(1)}$ in (A.5). For any given $b^{(2)}$ that satisfies $\|b^{(2)}\|_{D_j} = 0$, let $\hat{b}^{(1)}(b^{(2)})$ be the minimizer of (A.5) when $b^{(2)}$ is fixed. By the KKT condition, we know that $\hat{b}^{(1)}(b^{(2)})$ satisfies

$$0 \in -Z_j^T(Y - Z^{(1)}b^{(1)} - Z^{(2)}b^{(2)}) + n\lambda_1 w_{1j}\partial||b_j||_{A_j} + n\lambda_2 w_{2j}\partial||b_j||_{D_j}, j = 1, \ldots, p_1.$$

By Proposition 1, there exists a unique p_2-dimensional vector β such that $Z^{(2)}b^{(2)} = X^{(2)}\beta$, and thus we can write $\hat{b}^{(1)}(\beta)$ instead of $\hat{b}^{(1)}(b^{(2)})$. By this change of notation using β, in the rest of the proof we write $\hat{b}$, b, $\hat{b}(\beta)$ in place of $\hat{b}^{(1)}$ $b^{(1)}$, $\hat{b}^{(1)}(\beta)$ for simplicity.

From the above displayed expression we get

$$\hat{b}(\beta) = (Z^{(1)T}Z^{(1)})^{-1}Z^{(1)T}(Y - X^{(2)}\beta) + (Z^{(1)T}Z^{(1)})^{-1}v(\beta), \quad \text{(A.11)}$$

where $v(\beta)$ is a p_1K-dimensional vector with its components given by $n\lambda_1 w_{1j}\partial||\hat{b}_j(\beta)||_{A_j} + n\lambda_2 w_{2j}\partial||\hat{b}_j(\beta)||_{D_j}$, $1 \leq j \leq p_1$.

Let β_0 be the true slope parameter for the linear components and under the corre-

spondence given in Proposition 1 we have some $b_0^{(2)}$ that satisfies $Z^{(2)}b_0^{(2)} = X^{(2)}\beta_0$. Consider any $\hat{b}^{(2)} \in R^{p_2K}$ given by $b_0^{(2)}+\gamma u_b$ with $\gamma = C\sqrt{K}(\sqrt{1/n}+\sqrt{K}(\lambda_1\|w_1^0\|+\lambda_2\|w_2^0\|)$ and $\|u_b\| = 1, \|u_b\|_{D_j} = 0$. Again by Proposition 1, we can write this equivalently in terms of p_2-dimensional vectors $\hat{\beta} = \beta_0 + \gamma_1 u$ under the correspondence, where $\gamma_1 = C(\sqrt{1/n} + \sqrt{K}(\lambda_1||w_1^0|| + \lambda_2||w_2^0||))$ for some $C > 0$, and $||u|| = 1$. We will show that $\inf_{||u||=1} Q(\hat{b}(\hat{\beta}), \hat{\beta}) - Q(\hat{b}(\beta_0), \beta_0) > 0$ with probability approaching 1 for C large enough and (1.6) will follow.

Using the closed-form expression for $\hat{b}(\beta)$, we get

$$
\begin{aligned}
& Q(\hat{b}(\hat{\beta}), \hat{\beta}) - Q(\hat{b}(\beta_0), \beta_0) \\
= \; & -(\tilde{Y} - \tilde{X}^{(2)}\beta_0)(\gamma_1\tilde{X}^{(2)}u + Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\hat{\beta})) \\
& +(1/2)||\gamma_1\tilde{X}^{(2)}u + Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\hat{\beta})||^2 \\
& +(\tilde{Y} - \tilde{X}^{(2)}\beta_0)^T Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta_0) - (1/2)||Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta_0)||^2 \\
& +n\lambda_1 \sum_{j=1}^{p_1} w_{1j}||\hat{b}_j(\hat{\beta})||_{A_j} + n\lambda_2 \sum_{j=1}^{p_1} w_{2j}||\hat{b}_j(\hat{\beta})||_{D_j} + n\lambda_1 \sum_{j=p_1+1}^{s} w_{1j}|||\hat{b}_j\|_{A_j} \\
& -n\lambda_1 \sum_{j=1}^{p_1} w_{1j}||\hat{b}_j(\beta_0)||_{A_j} - n\lambda_2 \sum_{j=1}^{p_1} w_{2j}||\hat{b}_j(\beta_0)||_{D_j} - n\lambda_1 \sum_{j=p_1+1}^{s} w_{1j}|||b_{0j}\|_{A_j},
\end{aligned}
\tag{A.12}
$$

where for any random matrix W with n rows, we set $\tilde{W} = Q_{Z^{(1)}}W = W - P_{Z^{(1)}}W$ to be the projection of columns of W onto the orthogonal complement of the column space of $Z^{(1)}$, where $P_{Z^{(1)}} = Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}Z^{(1)T}$.

Using that $Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}Z^{(1)T}v(\beta)$ is inside the column space of $Z^{(1)}$, while all variables with $\tilde{}$ are orthogonal to it, the first four terms in (A.12) are simplified to

$$
\begin{aligned}
-(\tilde{Y} - \tilde{X}^{(2)}\beta_0)^T(\gamma_1\tilde{X}^{(2)}u) + (1/2)||\gamma_1\tilde{X}^{(2)}u||^2 + (1/2)||Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\hat{\beta})||^2 \\
-(1/2)||Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta_0)||^2.
\end{aligned}
$$

We will derive the orders in three steps for these four terms: (i) $||(\tilde{Y}-\tilde{X}^{(2)}\beta_0)^T(\tilde{X}^{(2)}u)|| = O_p(\sqrt{n})$, (ii) $||Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta)|| = O_p(\sqrt{nK}(\lambda_1||w_1^0|| + \lambda_2||w_2^0||))$, (iii) the terms in the last two terms in (A.12) involving the penalty terms are of order $O_p(n\sqrt{K}\lambda_1||w_1^0||\gamma_1 + nK(\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2))$.

Step 1. Proof of $||(\tilde{Y} - \tilde{X}^{(2)}\beta_0)^T(\tilde{X}^{(2)}u)|| = O_p(\sqrt{n})$. We first write down the de-

composition

$$X^{(2)} = \Theta - H + H + U.$$

The above uppercase letters represent $n \times p_2$ matrices, and correspond to the decomposition in (A.1) evaluated at n observations. After projection, we have $\tilde{X}^{(2)} = \tilde{\Theta} - \tilde{H} + \tilde{H} + \tilde{U}$. Together with the decomposition $\tilde{Y} - \tilde{X}^{(2)}\beta_0 = \tilde{\epsilon} + (r - Z^{(1)}b_0)$ (same as in the proof of Lemma A.2, $r = (r_1, \ldots, r_n)^T$ with $r_i = \sum_{j=1}^{p_1} f_{0j}(X_{ij})$, b_0 contains the spline coefficients that achieve optimal approximation of $f_{0j}, 1 \le j \le p_1$), the bound for $||(\tilde{Y} - \tilde{X}^{(2)}\beta_0)^T\tilde{X}^{(2)}||$ is obtained from the following expressions.

$$\begin{aligned}
||\epsilon^T Q_{Z^{(1)}} X^{(2)}|| &= O_p(\sqrt{n}), && \text{(A.13a)}\\
||(r - Z^{(1)}b_0)^T Q_{Z^{(1)}}(\Theta - H)|| &= O_p(\sqrt{\frac{n}{K^{2d}}}), && \text{(A.13b)}\\
||(r - Z^{(1)}b_0)^T Q_{Z^{(1)}} U|| &= O_p(\sqrt{\frac{n}{K^{2d}}}), && \text{(A.13c)}\\
||(r - Z^{(1)}b_0)^T Q_{Z^{(1)}} H|| &= O_p(\frac{n}{K^{2d}}) = O_p(\sqrt{n}), && \text{(A.13d)}
\end{aligned}$$

where (A.13a) is obvious from condition (c8), (A.13b) is based on that entries of $\Theta - H$ have mean zero and are orthogonal to $\mathcal{A}$ while entries of $(r - Z^{(1)}b_0)^T$ and $Z^{(1)}$ are inside $\mathcal{A}$ and thus we can calculate the bound by considering its variance, (A.13c) is obtained similarly, and finally $(A.13d)$ is obtained from $\|r - Z^{(1)}b_0\| = O_p(\sqrt{n/K^{2d}})$ and $||Q_{Z^{(1)}}H|| = O_p(\sqrt{n/K^{2d}})$ by condition (c4).

Step 2. Proof of $||Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta)|| = O_p(\sqrt{nK}(\lambda_1||w_1^0|| + \lambda_2||w_2^0||))$. Using the fact that $\partial||\hat{b}_j||_{A_j}$ and $\partial||\hat{b}_j||_{D_j}$ have l_2 norm bounded by $\lambda_{\max}(A_j^{1/2})$ and $\lambda_{\max}(D_j^{1/2})$ respectively, it easily follows from the definition of $v(\beta)$ that $||v(\beta)||^2 = O_p(n^2(\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2))$.

Step 3. Proof for that the last two lines in (A.12) involving the penalty terms is of order $O_p(n\sqrt{K}\lambda_1||w_1^0||\gamma_1 + nK(\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2))$. We have

$$\begin{aligned}
n\lambda_1 \sum_{j=1}^{p_1} w_{1j}||\hat{b}_j(\hat{\beta}) - \hat{b}_j(\beta_0)||_{A_j} &\le Cn\lambda_1||w_1^0|| \cdot ||\hat{b}(\hat{\beta}) - \hat{b}(\beta_0)||\\
&\le Cn\lambda_1||w_1^0|| \cdot (||(Z^{(1)T}Z^{(1)})^{-1}Z^{(1)T}(\hat{\beta} - \beta_0)|| + ||(Z^{(1)T}Z^{(1)})^{-1}(v(\hat{\beta}) - v(\beta_0))||)\\
&= Cn\lambda_1||w_1^0||(\gamma_1\sqrt{K/n} + K(\lambda_1||w_1^0|| + \lambda_2||w_2^0||)\\
&= O_p(\sqrt{nK}\lambda_1||w_1^0||\gamma_1 + nK(\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2)),
\end{aligned}$$

where in the 1st line above we used Cauchy-Schwartz inequality, in the 2nd line we used (A.11), in the 3rd line we used (ii) above. We can bound $n\lambda_2 \sum_{j=1}^{p_1} w_{2j}||\hat{b}_j(\hat{\beta}) - \hat{b}_j(\beta_0)||_{D_j}$ in a similar way. Finally, we have $n\lambda_1 \sum_{j=p_1+1}^{s} w_{1j}\|\hat{b}_j - b_{0j}\|_{A_j} \leq Cn\lambda_1\sqrt{K}\|w_1^0\|\gamma_1$ using Cauchy-Schwartz inequality.

Since the eigenvalues of $\tilde{X}^{(2)T}\tilde{X}^{(2)}/n$ are bounded away from zero as shown in part I in the proof of Theorem 1 in Li (2000), $Q(\hat{b}(\hat{\beta}), \hat{\beta}) - Q(\hat{b}(\beta_0), \beta_0)$ is bounded below by $C_1 n\gamma_1^2 - C_2 a_n\gamma_1 - C_3 b_n$, for some positive constants C_1, C_2, C_3 and $a_n = \sqrt{n} + n\sqrt{K}\lambda_1\|w_1^0\|, b_n = nK(\lambda_1^2\|w_1^0\|^2 + \lambda_2^2\|w_2^0\|^2)$. Thus if $\gamma_1 = C\max\{a_n/n, \sqrt{b_n/n}\}$ for $C > 0$ sufficiently large, the above displayed expression will be positive. The expression $\max\{a_n/n, \sqrt{b_n/n}\}$ is exactly the same as $\sqrt{1/n} + \sqrt{K}(\lambda_1||w_1^0|| + \lambda_2||w_2^0||)$ as in the statement of the Theorem. □

Lemma A.3 *(Asymptotic normality) Under the same assumption as in Lemma A.2, and that* $\sqrt{nK}(\lambda_1\|w_1^0\| + \lambda_2\|w_2^0\|) \to 0$, $\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, \sigma^2\Xi^{-1})$ *in distribution*.

Proof. Lemma A.1 states that the KKT conditions for (A.2) are

$$0 \in -Z_j^T(Y - Z^{(1)}b^{(1)} - Z^{(2)}b^{(2)}) + n\lambda_1 w_{1j}\partial||b_j||_{A_j} + n\lambda_2 w_{2j}\partial||b_j||_{D_j}, j \leq p_1, \quad \text{(A.14)}$$

$$-Z_j^T(Y - Z^{(1)}b^{(1)} - Z^{(2)}b^{(2)}) + n\lambda_1 w_{1j}\partial\|b_j\|_{A_j} \in span(D_j), j = p_1 + 1, \ldots, s. \quad \text{(A.15)}$$

By pre-multiplying the second equation above by ξ_j^T which is defined in Proposition 1, (A.15) becomes $-X_j^T(Y - Z^{(1)}b^{(1)} - Z^{(2)}b^{(2)}) + n\lambda_1 w_{1j}\xi_j^T\partial\|b_j\|_{A_j} = 0, j = p_1 + 1, \ldots, s$.

Since we have the constraint $\|b_j\|_{D_j} = 0, j = p_1 + 1, \ldots, s$, similar as in the proof of Lemma A.2 we can use a change of parameter (and a similar simplification of notation as in the proof of Lemma A.2) to write the above as

$$-X_j^T(Y - Z^{(1)}b - X^{(2)}\beta) + \chi_j = 0, \quad \text{(A.16)}$$

where $\chi_j = n\lambda_1 w_{1j}\xi_j^T\partial\|b_j\|_{A_j}$ and we note $|\chi_j| = O_p(n\lambda_1 w_{1j}\sqrt{K})$ due to that $\|\xi_j\| = O(\sqrt{K})$.

Since $Y = r + X\beta_0 + \epsilon$ where $r = (r_1, \ldots, r_n)^T$ with $r_i = \sum_{j=1}^{p_1} f_{0j}(X_{ij})$, and denote by b_0 the vector containing the spline coefficients that achieve optimal approximation of $f_{0j}(x), 1 \leq j \leq p_1$, and set $a = r - Z^{(1)}b_0$, (A.16) is rewritten as $-X_j^T(\epsilon + a - Z^{(1)}(b - b^0) - X^{(2)}(\beta - \beta_0)) + \chi_j = 0, j = p_1 + 1, \ldots, s$. From (A.14), we get

$Z^{(1)}(b-b_0) = Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}Z^{(1)T}(\epsilon+a-X^{(2)}(\beta-\beta_0))+Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta)$ ($v(\beta)$ defined right after equation (A.11)) and plug into the above displayed equation we get

$$-X_j^T(\epsilon + a - Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}[Z^{(1)T}(\epsilon + a - X^{(2)}(\beta - \beta_0)) + v(\beta)] - X^{(2)}(\beta - \beta_0)) + \chi_j = 0, j = p_1 + 1, \ldots, s,$$

that is,

$$-X_j^T(\widetilde{\epsilon + a} - \tilde{X}^{(2)}(\beta - \beta_0) - Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta)) + \chi_j = 0, j = p_1 + 1, \ldots, s,$$

from which we get

$$\begin{aligned} & \sqrt{n}(\hat{\beta} - \beta_0) \\ = \; & \sqrt{n}(\tilde{X}^{(2)T}\tilde{X}^{(2)})^{-1}\tilde{X}^{(2)T}(\epsilon + a) + \sqrt{n}(\tilde{X}^{(2)T}\tilde{X}^{(2)})^{-1}X^{(2)T}Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta) \\ & + \sqrt{n}(\tilde{X}^{(2)T}\tilde{X}^{(2)})^{-1}\Lambda, \end{aligned} \tag{A.17}$$

where Λ is a p_2-dimensional vector with components $\chi_j, j = p_1+1, \ldots, s$. By part (1) in the proof of Theorem 1 in Li (2000), we can replace $(\tilde{X}^{(2)T}\tilde{X}^{(2)}/n)^{-1}$ by Ξ^{-1} which only results in a multiplicative factor $1 + o_p(1)$ and thus does not disturb the asymptotic distribution.

Using $||\tilde{X}^{(2)T}a|| = O_p(\sqrt{n/K^{2d}} + n/K^{2d})$ (combining (A.13b)-(A.13d)) and $||X^{(2)T}Z^{(1)}(Z^{(1)T}Z^{(1)})^{-1}v(\beta)|| = O_p(n\sqrt{K}(\lambda_1||w_1^0|| + \lambda_2||w_2^0||))$ ((ii) in the proof of Lemma A.2) and $||\Lambda|| = O_p(n\lambda_1\sqrt{K}||w_1^0||)$, all terms in (A.17) are $o_p(1)$ except $\sqrt{n}(\tilde{X}^{(2)T}\tilde{X}^{(2)})^{-1}\tilde{X}^{(2)T}\epsilon$, which can be shown to converge to $N(0, \sigma^2\Xi^{-1})$ by Lindeberg-Feller central limit theorem using standard arguments. □.

A.4 Proofs of the main results

Proofs of Theorems 1-3. As explained before, we only need to show that the solution $\hat{b} = (\hat{b}^{(1)}, \hat{b}^{(2)}, \hat{b}^{(3)})$ to (A.2) is also the solution to (1.4). Since $\hat{b}$ solves the optimization

problem (A.5), by Lemma A.1 we have

$$-Z_j^T(Y - Z^{(1)}\hat{b}^{(1)} - Z^{(2)}\hat{b}^{(2)}) + n\lambda_1 w_{1j}\partial||\hat{b}_j||_{A_j} + n\lambda_2 w_{2j}\partial||\hat{b}_j||_{D_j} = 0, j = 1, \ldots, p_1, \tag{A.18}$$

$$-Z_j^T(Y - Z^{(1)}\hat{b}^{(1)} - Z^{(2)}\hat{b}^{(2)}) + n\lambda_1 w_{1j}\partial||\hat{b}_j||_{A_j} \in span(D_j), j = p_1 + 1, \ldots, s. \tag{A.19}$$

We remind the readers that these equations mean "there exists some subgradient that makes the left hand side satisfy the condition", in case the subgradient is not unique.

In order to show that the pK-dimensional vector $\hat{b}$ is also the solution to (1.4), we only need to verify the corresponding KKT conditions

$$-Z_j^T(Y - Z^{(1)}\hat{b}^{(1)} - Z^{(2)}\hat{b}^{(2)} - Z^{(3)}\hat{b}^{(3)}) + n\lambda_1 w_{1j}\partial||\hat{b}_j||_{A_j} + n\lambda_2 w_{2j}\partial||\hat{b}_j||_{D_j} = 0, j = 1, \ldots, p. \tag{A.20}$$

First, for $1 \leq j \leq p_1$, (A.20) is obviously the same as (A.18) and there is nothing to show.

Next, for $p_1 + 1 \leq j \leq s$, we first show that

$$||Z_j^T(Y - Z\hat{b})|| + n\lambda_1 w_{1j} = o_p(n\lambda_2 w_{2j}). \tag{A.21}$$

In fact, we have $||Z_j^T(Y - Z\hat{b})|| \leq ||Z_j^T\epsilon|| + ||Z_j^T Z^{(1,2)}(\hat{b}^{(1,2)} - b_0^{(1,2)})|| + ||Z_j^T(r - Z^{(1)}b_0^{(1)})||$, where r is as defined as in the proof of Lemma A.2. Using exactly the same arguments as in showing (A.9) and Theorem 1 of Huang et al. (2010), we have $\max_j ||Z_j^T\epsilon|| = O_p(\sqrt{(n/K)\log(pK)})$. Besides, using Lemma A.2 we obtain $||Z_j^T Z^{(1,2)}(\hat{b}^{(1,2)} - b_0^{(1,2)})|| + ||Z_j^T(r - Z^{(1)}b_0^{(1)})|| = O_p\left(\sqrt{(n/K)(K + n/K^{2d} + nK(\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2))}\right)$. Thus, $||Z_j^T(Y - Z\hat{b})|| = o_p(n\lambda_2 w_{2j})$ by assumption (c7). Finally, assumption (c7) also trivially implies that $\lambda_1 w_{1j} \leq \lambda_1||w_1^0|| = o_p(\lambda_2 w_{2j}), p_1 + 1 \leq j \leq s$, and (A.21) is proved. Since $\hat{b}$ satisfies (A.19), we can write

$$-Z_j^T(Y - Z^{(1)}\hat{b}^{(1)} - Z^{(2)}\hat{b}^{(2)} - Z^{(3)}\hat{b}^{(3)}) + n\lambda_1 w_{1j}\partial||\hat{b}_j||_{A_j} + D_j^{1/2}a = 0 \tag{A.22}$$

for some $a \in R^K$, where $D_j^{1/2}$ is the matrix square root of D_j (note that $span(D_j)$ is the same as $span(D_j^{1/2})$) and furthermore by (A.21) we have $\|D_j^{1/2}a\|/(n\lambda_2 w_{2j}) = o_p(1)$. Let P_D be the projection matrix onto the column span of D_j, obviously we have $D_j^{1/2}P_D a = D_j^{1/2}a$. Then $\|P_D a\| \leq \|D_j^{1/2}P_D a\|/\lambda_{\min}(D_j) = \|D_j^{1/2}a\|/\lambda_{\min}(D_j) =$

$o_p(n\lambda_2 w_{2j})$. Setting $\tilde{a} = D_j^{1/2} P_D a/(n\lambda_2 w_{2j})$, (A.22) can be rewritten as

$$-Z_j^T(Y - Z^{(1)}\hat{b}^{(1)} - Z^{(2)}\hat{b}^{(2)} - Z^{(3)}\hat{b}^{(3)}) + n\lambda_1 w_{1j}\partial||\hat{b}_j||_{A_j} + n\lambda_2 w_{2j}\tilde{a} = 0,$$

which can be seen to verify (A.20) for $p_1+1 \le j \le s$ since it is easy to see $\tilde{a} \in \partial\|\hat{b}_j\|_{D_j}$.

For $s+1 \le j \le p$, using similar arguments, we only need to verify $\|Z_j^T(Y - Z\hat{b})\| = o_p(n\lambda_1 w_{1j})$, which can be shown in the same way as before. □

Proof of Theorem 4. For any given pair of regularization parameters $\lambda = (\lambda_1, \lambda_2)$, we denote by $\hat{b}_\lambda$ the minimizer of (1.4), and by $\hat{b}$ the minimizer when the optimal sequence of regularization parameters is chosen such that $\hat{b}$ results in a consistent model selection. We consider the underfitting and overfitting case separately below.

Case 1. Underfitting. We only consider the case where some nonparametric components are estimated as nonzero parametric component in $\hat{b}_\lambda$ (other cases, such as estimating a nonzero linear component as zero, are similar). Similar to the calculations performed in the proof of Lemma A.2, we have

$$\frac{1}{2n}||Y - Z\hat{b}_\lambda||^2 - \frac{1}{2n}||Y - Z\hat{b}||^2 \ge -\frac{1}{n}||P_Z(Y - Z\hat{b})||^2 + \frac{1}{4n}||Z(\hat{b} - \hat{b}_\lambda)||^2.$$

Since there is some $1 \le j \le p_1$ for which $\hat{b}_j$ represents a nonparametric component with convergence rate given by Lemma A.2, while $\hat{b}_{\lambda j}$ satisfies $\|\hat{b}_{\lambda j}\|_{D_j} = 0$, it is easy to show that $||Z(\hat{b} - \hat{b}_\lambda)||^2/n \ge C||Z_j(\hat{b}_j - \hat{b}_{\lambda j})||^2/n$ is bounded away from zero by condition (c9). Besides, $||P_Z(Y - Z\hat{b})||/n = o_p(1)$ (using the same arguments as in proving (A.9) as well as the convergence rates stated in Lemma A.2) and the penalty terms in BIC are all of order $o_p(1)$. Also note that it is easy to show $\frac{1}{2n}||Y - Z\hat{b}||^2$ is bounded away from zero, which implies that $\log||Y - Z\hat{b}_\lambda||^2 - \log||Y - Z\hat{b}||^2 \ge O_p(1)$. Thus the eBIC when λ is used is bigger than the eBIC when the optimal regularization sequence is used.

Case 2. Overfitting. We only consider the case where some zero or linear components are estimated as nonparametric in $\hat{b}_\lambda$. Let $\hat{b}^*$ be the minimizer of the least square $||Y - Zb||^2$ under the additional constraint that the model identified by $\hat{b}_\lambda$ is used when

minimizing the least square. We have that

$$\begin{aligned}\frac{1}{2n}||Y - Z\hat{b}_\lambda||^2 - \frac{1}{2n}||Y - Z\hat{b}||^2 &\geq \frac{1}{2n}||Y - Z\hat{b}^*||^2 - \frac{1}{2n}||Y - Z\hat{b}||^2 \\ &= \frac{1}{n}(Y - Z\hat{b})^T Z(\hat{b} - \hat{b}^*) + \frac{1}{2n}||Z(\hat{b} - \hat{b}^*)||^2 \\ &\geq \frac{1}{n}(Y - Z\hat{b})^T Z(\hat{b} - \hat{b}^*). \end{aligned} \tag{A.23}$$

By the definition of $\hat{b}^*$ and the fact that we only search over models with size bounded by some constant of order $O(1)$, the convergence rate of $\hat{b}^*$ can be obtained using similar arguments as Lemma A.2 but without the terms involving λ_1 and λ_2 appearing. Arguments similar to those used in showing (A.21) in the proof of Theorem 1 can be used to show that the (A.23) is bounded below by a negative term whose absolute value is of order

$$\frac{1}{n}\sqrt{(n\log(pK) + \frac{n^2}{K^{2d+1}}) \cdot (\frac{K^2}{n} + \frac{1}{K^{2d-1}})} = o((\log(n/K) + \log p)/(n/K)).$$

Thus eBIC cannot select such λ, similar as in Case 1. □

A discussion for the case $s \to \infty$

The reason we need to assume s is fixed is that we need to use the property that the eigenvalues of $Z^{(1,2)}$ are of order n/K in the proof. For fixed s, this property has been shown in Lemma 3 of Huang et al. (2010). It is not clear how this lemma can be extended to the case $s \to \infty$. Ravikumar et al. (2009) assumed directly the order of eigenvalues of $Z^{(1,2)}$ which is the reason why they can let $s \to \infty$. As long as we assume eigenvalues of $Z^{(1,2)}$ are of order n/K, the proof can be straightforwardly modified in a few places to allow $s \to \infty$. We now briefly mention the changes required in the proof for the theoretical results in Section 2.2.

To incorporate diverging s, assumption (c7) is slightly changed to the following

(c7') $\sqrt{n/K}\{\sqrt{\log(pK)} + \sqrt{Ks + ns/K^{2d}} + \sqrt{nK}(\lambda_1||w_1^0|| + \lambda_2||w_2^0||)\} = o_p(n\lambda_2 w_{2j})$ for $p_1 + 1 \leq j \leq s$ and $\sqrt{n/K}\{\sqrt{\log(pK)} + \sqrt{Ks + ns/K^{2d}} + \sqrt{nK}(\lambda_1||w_1^0|| + \lambda_2||w_2^0||)\} = o_p(n\lambda_1 w_{1j})$ for $s + 1 \leq j \leq p$.

In Lemma A.2, the rates will now be

$$\sum_{j=1}^{s} ||\hat{f}_j - f_{0j}||^2 = O_p\left(\frac{Ks}{n} + \frac{s}{K^{2d}} + (\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2)K\right).$$

$$\sum_{j=p_1+1}^{s} (\hat{\beta}_j - \beta_{0j})^2 = O_p\left(\frac{p_2}{n} + (\lambda_1^2||w_1^0||^2 + \lambda_2^2||w_2^0||^2)K\right).$$

The appearance of s in the first equation above comes from (A.9), which now will be

$$\|\eta\|^2 = O_p(Ks + ns/K^{2d}).$$

For the rates of the parameter part (now we require $\sqrt{np_2}/K^{2d} \to 0$), an additional $\sqrt{q_2}$ factor will appear in the bounds (A.13a)-(A.13d). By changing the definitions of γ and γ_1 (in the paragraph following (A.11)) to be $\gamma = C\sqrt{K}(\sqrt{p_2/n} + \sqrt{K}(\lambda_1\|w_1^0\| + \lambda_2\|w_2^0\|)$ and $\gamma = C(\sqrt{p_2/n} + \sqrt{K}(\lambda_1\|w_1^0\| + \lambda_2\|w_2^0\|)$, the rates can be shown following the same lines.

For asymptotic normality, due to the diverging dimension of β, the asymptotic normality is more appropriately stated as

$$\sqrt{n}a^T\Xi^{1/2}(\hat{\beta} - \beta_0) \to N(0, \sigma^2) \text{ in distribution,}$$

for any deterministic p_2-vector a with $\|a\| = 1$, using basically the same proof as before.

Supplementary tables

Table 1.4: Model identification results with $n = 50$

		#N	#NT	#L	#LT
n=50	BIC	$29.96_{2.6570}$	$4.98_{0.1414}$	0_{0}	0_{0}
p=50	EBIC	$3.54_{2.0723}$	$2.38_{1.1409}$	0_{0}	0_{0}
σ=0.2	BIC/BIC	$14.94_{3.0865}$	2_{0}	$1.9_{1.8763}$	$0.22_{0.5455}$
	EBIC/EBIC	$1.22_{0.6788}$	$1.14_{0.3505}$	$0.98_{0.9792}$	$0.86_{0.8084}$
	BIC/EBIC	$2.42_{0.8593}$	$1.92_{0.2740}$	$3.28_{1.7266}$	$2.24_{0.8609}$
n=50	BIC	$31.38_{3.1161}$	$4.96_{0.1979}$	0_{0}	0_{0}
p=50	EBIC	$3.44_{1.3273}$	$2.46_{0.9304}$	0_{0}	0_{0}
σ=0.5	BIC/BIC	$17.76_{2.2818}$	2_{0}	$1.68_{1.3915}$	$0.28_{0.6074}$
	EBIC/EBIC	$1.28_{0.5360}$	$1.18_{0.3881}$	$1.12_{0.9176}$	$0.96_{0.7548}$
	BIC/EBIC	$1.7_{0.7354}$	$1.56_{0.5014}$	$3.26_{2.4396}$	$2.08_{1.1400}$
n=50	BIC	$37.86_{7.6212}$	$4.86_{0.5349}$	0_{0}	0_{0}
p=100	EBIC	$3.54_{1.6189}$	$2.28_{1.0110}$	0_{0}	0_{0}
σ=0.2	BIC/BIC	$15.6_{5.1070}$	$1.92_{0.2740}$	$2.32_{1.7195}$	$0.7_{0.8864}$
	EBIC/EBIC	$1.24_{0.4764}$	$1.22_{0.4185}$	$0.98_{0.9581}$	$0.84_{0.8657}$
	BIC/EBIC	$2.48_{1.5151}$	$1.8_{0.4041}$	$2.56_{1.4450}$	$2.2_{0.9689}$
n=50	BIC	$38.6_{10.0306}$	$4.58_{0.9916}$	0_{0}	0_{0}
p=100	EBIC	$4.26_{2.0584}$	$2.26_{1.0461}$	0_{0}	0_{0}
σ=0.5	BIC/BIC	$16.2_{5.8310}$	$1.8_{0.4041}$	$2.32_{1.7893}$	$0.46_{0.6131}$
	EBIC/EBIC	$1.2_{0.4518}$	$1.16_{0.3703}$	$0.84_{1.1132}$	$0.68_{0.7939}$
	BIC/EBIC	$1.44_{0.6440}$	$1.34_{0.4785}$	$1.46_{1.4316}$	$1.22_{1.0934}$
n=50	BIC	$25.84_{21.9920}$	$3.12_{1.6117}$	0_{0}	0_{0}
p=200	EBIC	$4.78_{2.3586}$	$2.1_{1.1294}$	0_{0}	0_{0}
σ=0.2	BIC/BIC	$9.38_{7.9972}$	$1.54_{0.5035}$	$1.26_{1.6880}$	$0.34_{0.6581}$
	EBIC/EBIC	$1.12_{0.3854}$	$1.1_{0.3030}$	$0.84_{1.1493}$	$0.56_{0.7045}$
	BIC/EBIC	$1.26_{0.5272}$	$1.22_{0.4185}$	$1.6_{2.2406}$	$1.02_{1.1865}$
n=50	BIC	$7.08_{4.8481}$	$3.58_{1.7853}$	0_{0}	0_{0}
p=200	EBIC	$5.54_{2.3142}$	$2.04_{0.7814}$	0_{0}	0_{0}
σ=0.5	BIC/BIC	$10.2_{8.3103}$	$1.56_{0.5014}$	$1.22_{1.2824}$	$0.3_{0.5803}$
	EBIC/EBIC	1_{0}	1_{0}	$0.56_{0.8369}$	$0.42_{0.6091}$
	BIC/EBIC	$2.26_{1.2906}$	$1.7_{0.4629}$	$1.84_{1.8335}$	$1.4_{1.3093}$

Table 1.5: Root mean squared errors for the first six components with $n = 50$

		Oracle	Our Estimator	Sparse Additive
n=50	f_1	$0.3485_{0.05593}$	$0.3721_{0.08058}$	$0.4075_{0.09441}$
p=50	f_2	$0.1175_{0.03685}$	$0.2214_{0.16259}$	$0.2598_{0.18056}$
σ=0.2	f_3	$0.0603_{0.04675}$	$0.0966_{0.13139}$	$0.2374_{0.16736}$
	f_4	$0.0470_{0.03255}$	$0.1420_{0.14434}$	$0.2447_{0.16145}$
	f_5	$0.0575_{0.04406}$	$0.1304_{0.14230}$	$0.2482_{0.16582}$
	f_6	$0.0000_{0.00000}$	$0.0014_{0.00961}$	$0.0181_{0.05207}$
n=50	f_1	$0.3645_{0.03977}$	$0.4744_{0.12462}$	$0.4818_{0.09948}$
p=50	f_2	$0.1930_{0.07206}$	$0.4756_{0.24023}$	$0.3817_{0.18565}$
σ=0.5	f_3	$0.0687_{0.05014}$	$0.2551_{0.29540}$	$0.3758_{0.20696}$
	f_4	$0.0882_{0.06131}$	$0.2990_{0.22895}$	$0.3531_{0.14074}$
	f_5	$0.0742_{0.06118}$	$0.2582_{0.21352}$	$0.3504_{0.12749}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0164_{0.05671}$
n=50	f_1	$0.3548_{0.03924}$	$0.3963_{0.09781}$	$0.4141_{0.07939}$
p=100	f_2	$0.1156_{0.04314}$	$0.2826_{0.23158}$	$0.3022_{0.23694}$
σ=0.2	f_3	$0.0453_{0.02931}$	$0.1463_{0.20384}$	$0.3110_{0.22518}$
	f_4	$0.0464_{0.03426}$	$0.1750_{0.18963}$	$0.2900_{0.17832}$
	f_5	$0.0493_{0.04085}$	$0.2102_{0.19378}$	$0.3308_{0.19054}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
n=50	f_1	$0.3848_{0.08284}$	$0.5571_{0.18669}$	$0.5359_{0.15456}$
p=100	f_2	$0.1823_{0.06987}$	$0.5580_{0.23714}$	$0.4629_{0.23394}$
σ=0.5	f_3	$0.0839_{0.06235}$	$0.4698_{0.34764}$	$0.5294_{0.27693}$
	f_4	$0.0848_{0.06894}$	$0.4382_{0.20924}$	$0.4508_{0.15754}$
	f_5	$0.0820_{0.05971}$	$0.4311_{0.21770}$	$0.4528_{0.17077}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0064_{0.02569}$
n=50	f_1	$0.3521_{0.03791}$	$0.5091_{0.14830}$	$0.5144_{0.14475}$
p=200	f_2	$0.1126_{0.04105}$	$0.6174_{0.21791}$	$0.6227_{0.20919}$
σ=0.2	f_3	$0.0602_{0.05229}$	$0.5246_{0.36954}$	$0.6655_{0.27089}$
	f_4	$0.0564_{0.04377}$	$0.4392_{0.21413}$	$0.5282_{0.12445}$
	f_5	$0.0517_{0.03874}$	$0.4503_{0.19922}$	$0.5215_{0.13808}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
n=50	f_1	$0.3380_{0.02206}$	$0.3761_{0.06544}$	$0.3791_{0.06867}$
p=200	f_2	$0.1139_{0.04679}$	$0.3152_{0.27003}$	$0.3038_{0.26639}$
σ=0.5	f_3	$0.0577_{0.03835}$	$0.3423_{0.37147}$	$0.3808_{0.34359}$
	f_4	$0.0523_{0.03508}$	$0.2606_{0.24863}$	$0.3012_{0.22242}$
	f_5	$0.0483_{0.03864}$	$0.2784_{0.25221}$	$0.3179_{0.21913}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$

Table 1.6: Model identification results with $n = 200$

		#N	#NT	#L	#LT
n=200	BIC	$5.48_{0.6773}$	5_0	0_0	0_0
p=50	EBIC	$5.76_{0.9596}$	5_0	0_0	0_0
σ=0.2	BIC/BIC	$2.26_{0.5272}$	2_0	$2.76_{0.5175}$	$2.74_{0.5272}$
	EBIC/EBIC	2_0	2_0	$3.04_{0.1979}$	3_0
	BIC/EBIC	2_0	2_0	$3.04_{0.1979}$	3_0
n=200	BIC	$10.62_{2.3724}$	5_0	0_0	0_0
p=50	EBIC	$7.26_{2.2114}$	5_0	0_0	0_0
σ=0.5	BIC/BIC	$2.18_{0.6289}$	2_0	$3.1_{0.8631}$	$2.84_{0.5095}$
	EBIC/EBIC	$2.08_{0.2740}$	2_0	$2.92_{0.2740}$	$2.92_{0.2740}$
	BIC/EBIC	$2.02_{0.1414}$	2_0	$3.04_{0.2828}$	$2.98_{0.1414}$
n=200	BIC	$6.2_{1.2289}$	5_0	0_0	0_0
p=100	EBIC	$5.96_{1.1599}$	5_0	0_0	0_0
σ=0.2	BIC/BIC	$2.34_{0.6581}$	2_0	$2.68_{0.6833}$	$2.66_{0.6581}$
	EBIC/EBIC	2_0	2_0	$3.02_{0.1414}$	3_0
	BIC/EBIC	$2.02_{0.1414}$	2_0	$3_{0.2020}$	$2.98_{0.1414}$
n=200	BIC	$11.36_{4.6325}$	5_0	0_0	0_0
p=100	EBIC	$5.42_{1.2469}$	5_0	0_0	0_0
σ=0.5	BIC/BIC	$2.16_{0.3703}$	2_0	$3.18_{0.9409}$	$2.84_{0.3703}$
	EBIC/EBIC	$2.06_{0.2399}$	2_0	$2.94_{0.2399}$	$2.94_{0.2399}$
	BIC/EBIC	2_0	2_0	$3.06_{0.2399}$	3_0
n=200	BIC	$7.64_{1.8818}$	5_0	0_0	0_0
p=200	EBIC	$7.38_{1.5894}$	5_0	0_0	0_0
σ=0.2	BIC/BIC	$2.22_{0.6788}$	2_0	$2.86_{0.7562}$	$2.78_{0.6788}$
	EBIC/EBIC	$2.02_{0.1414}$	2_0	$3.02_{0.2466}$	$2.98_{0.1414}$
	BIC/EBIC	2_0	2_0	$3.04_{0.1979}$	3_0
n=200	BIC	$7.3_{4.6258}$	5_0	0_0	0_0
p=200	EBIC	$5.06_{0.6824}$	$4.92_{0.5656}$	0_0	0_0
σ=0.5	BIC/BIC	$2.28_{0.6402}$	2_0	$2.72_{0.6402}$	$2.72_{0.6402}$
	EBIC/EBIC	$1.98_{0.1414}$	$1.98_{0.1414}$	$2.94_{0.4243}$	$2.94_{0.4243}$
	BIC/EBIC	$2.04_{0.1979}$	2_0	$2.98_{0.2466}$	$2.96_{0.1979}$

Table 1.7: Root mean squared errors for the first six components with $n = 200$

		Oracle	Our Estimator	Sparse Additive
n=200	f_1	$0.3180_{0.00631}$	$0.3193_{0.00718}$	$0.3181_{0.00632}$
p=50	f_2	$0.0592_{0.01535}$	$0.0682_{0.01832}$	$0.0610_{0.01585}$
σ=0.2	f_3	$0.0238_{0.01989}$	$0.0235_{0.01939}$	$0.0480_{0.02082}$
	f_4	$0.0190_{0.01782}$	$0.0186_{0.01730}$	$0.0487_{0.01755}$
	f_5	$0.0237_{0.01850}$	$0.0221_{0.01834}$	$0.0544_{0.01753}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
n=200	f_1	$0.3233_{0.01070}$	$0.3261_{0.01165}$	$0.3238_{0.01087}$
p=50	f_2	$0.0858_{0.02775}$	$0.1121_{0.04732}$	$0.0893_{0.02958}$
σ=0.5	f_3	$0.0387_{0.02630}$	$0.0396_{0.02703}$	$0.0823_{0.02861}$
	f_4	$0.0304_{0.02401}$	$0.0326_{0.02399}$	$0.0794_{0.03167}$
	f_5	$0.0362_{0.02498}$	$0.0396_{0.03214}$	$0.0751_{0.03022}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
n=200	f_1	$0.3169_{0.00640}$	$0.3184_{0.00736}$	$0.3171_{0.00622}$
p=100	f_2	$0.0568_{0.01452}$	$0.0709_{0.02163}$	$0.0584_{0.01491}$
σ=0.2	f_3	$0.0199_{0.01731}$	$0.0187_{0.01751}$	$0.0481_{0.02285}$
	f_4	$0.0245_{0.01816}$	$0.0242_{0.01714}$	$0.0522_{0.01769}$
	f_5	$0.0228_{0.01855}$	$0.0239_{0.01849}$	$0.0510_{0.02038}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
n=200	f_1	$0.3231_{0.00907}$	$0.3249_{0.01008}$	$0.3228_{0.00864}$
p=100	f_2	$0.0823_{0.03117}$	$0.1007_{0.03671}$	$0.0837_{0.03341}$
σ=0.5	f_3	$0.0331_{0.02828}$	$0.0327_{0.02958}$	$0.0804_{0.02941}$
	f_4	$0.0306_{0.02241}$	$0.0334_{0.02503}$	$0.0809_{0.03025}$
	f_5	$0.0289_{0.02232}$	$0.0313_{0.02738}$	$0.0794_{0.03067}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
n=200	f_1	$0.3164_{0.01045}$	$0.3180_{0.00974}$	$0.3168_{0.01047}$
p=200	f_2	$0.0596_{0.01189}$	$0.0786_{0.02325}$	$0.0604_{0.01303}$
σ=0.2	f_3	$0.0218_{0.02014}$	$0.0227_{0.02022}$	$0.0545_{0.02163}$
	f_4	$0.0253_{0.02066}$	$0.0256_{0.01809}$	$0.0544_{0.01828}$
	f_5	$0.0175_{0.01577}$	$0.0177_{0.01447}$	$0.0484_{0.01697}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$
n=200	f_1	$0.3246_{0.01130}$	$0.3273_{0.01200}$	$0.3255_{0.01192}$
p=200	f_2	$0.0821_{0.02770}$	$0.1225_{0.05857}$	$0.0845_{0.02767}$
σ=0.5	f_3	$0.0344_{0.02430}$	$0.0366_{0.02661}$	$0.0848_{0.03164}$
	f_4	$0.0273_{0.02206}$	$0.0288_{0.02602}$	$0.0756_{0.03741}$
	f_5	$0.0417_{0.03403}$	$0.0426_{0.03486}$	$0.0883_{0.03286}$
	f_6	$0.0000_{0.00000}$	$0.0000_{0.00000}$	$0.0000_{0.00000}$

Table 1.8: Prediction errors for the three estimators on independent simulated test data.

n	p	σ	Oracle	Our Estimator	Sparse Additive
50	50	0.2	0.809	0.958	1.257
50	50	0.5	0.950	1.754	2.092
50	100	0.2	0.819	1.181	1.461
50	100	0.5	0.876	2.155	2.201
50	200	0.2	0.710	2.252	2.464
50	200	0.5	0.775	1.463	1.497
100	50	0.2	0.416	0.429	0.443
100	50	0.5	0.620	0.633	0.709
100	100	0.2	0.366	0.425	0.438
100	100	0.5	0.646	0.776	0.839
100	200	0.2	0.468	0.698	0.714
100	200	0.5	0.581	1.051	1.114
200	50	0.2	0.303	0.287	0.291
200	50	0.5	0.481	0.485	0.494
200	100	0.2	0.276	0.290	0.296
200	100	0.5	0.515	0.521	0.533
200	200	0.2	0.288	0.290	0.296
200	200	0.5	0.520	0.541	0.544

Table 1.9: Standard errors of the estimators on the linear coefficients.

			$\hat{\beta}_3$		$\hat{\beta}_4$		$\hat{\beta}_5$	
n	p	σ	SD	SE	SD	SE	SD	SE
50	50	0.2	0.557	0.192(0.054)	0.302	0.182(0.050)	0.464	0.172(0.050)
50	50	0.5	0.604	0.327(0.064)	0.559	0.245(0.077)	0.648	0.234(0.075)
50	100	0.2	0.481	0.227(0.084)	0.383	0.182(0.052)	0.482	0.179(0.048)
50	100	0.5	0.808	0.375(0.143)	0.588	0.244(0.137)	0.782	0.260(0.133)
50	200	0.2	0.737	0.302(0.121)	0.390	0.192(0.073)	0.525	0.237(0.129)
50	200	0.5	0.750	0.398(0.095)	0.932	0.255(0.114)	0.926	0.374(0.176)
100	50	0.2	0.144	0.134(0.017)	0.183	0.126(0.016)	0.174	0.129(0.015)
100	50	0.5	0.290	0.204(0.025)	0.344	0.191(0.032)	0.228	0.198(0.026)
100	100	0.2	0.179	0.134(0.013)	0.151	0.131(0.013)	0.196	0.132(0.010)
100	100	0.5	0.225	0.205(0.028)	0.285	0.194(0.023)	0.288	0.192(0.026)
100	200	0.2	0.222	0.144(0.027)	0.161	0.131(0.016)	0.271	0.135(0.024)
100	200	0.5	0.286	0.248(0.069)	0.350	0.195(0.031)	0.366	0.207(0.045)
200	50	0.2	0.106	0.093(0.007)	0.091	0.094(0.006)	0.086	0.093(0.007)
200	50	0.5	0.148	0.145(0.010)	0.184	0.145(0.011)	0.172	0.146(0.010)
200	100	0.2	0.097	0.095(0.006)	0.090	0.094(0.006)	0.078	0.094(0.006)
200	100	0.5	0.153	0.147(0.012)	0.185	0.143(0.011)	0.168	0.145(0.011)
200	200	0.2	0.094	0.095(0.008)	0.086	0.094(0.007)	0.099	0.094(0.008)
200	200	0.5	0.156	0.149(0.013)	0.127	0.147(0.014)	0.133	0.149(0.015)