Biometrics 68, 437–445
June 2012
DOI: 10.1111/j.1541-0420.2011.01672.x

Identification of Partially Linear Structure in Additive Models with an Application to Gene Expression Prediction from Sequences

Heng Lian,* Xin Chen, and Jian-Yi Yang

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore
**email:* henglian@ntu.edu.sg

Summary. The additive model is a semiparametric class of models that has become extremely popular because it is more flexible than the linear model and can be fitted to high-dimensional data when fully nonparametric models become infeasible. We consider the problem of simultaneous variable selection and parametric component identification using spline approximation aided by two smoothly clipped absolute deviation (SCAD) penalties. The advantage of our approach is that one can automatically choose between additive models, partially linear additive models and linear models, in a single estimation step. Simulation studies are used to illustrate our method, and we also present its applications to motif regression.

Key words: Bayesian information criterion; Oracle property; Partially linear additive models; Structure identification.

1. Introduction

The additive model is commonly used for analysis of data that provides a desirable trade-off between stringent linear models and fully nonparametric models. In practice, however, we might have reasons to believe that only some of the components are truly nonparametric, besides that many components contain predictors unrelated to responses. For the latter, several works have considered using penalization approaches to select significant components (Ravikumar et al., 2008; Meier, Van de Geer, and Buhlmann, 2009; Xue, 2009; Huang, Horowitz, and Wei, 2010). However, there are much fewer studies that focus on the problem of how to identify the parametric components automatically.

The partially linear additive models (PLAM) (Liang et al., 2008; Liu, Wang, and Liang, 2011; Ma and Yang, 2011) can be seen as a special case of additive models where both parametric and nonparametric components coexist. In Opsomer and Ruppert (1999), the authors argued for the advantage of PLAM, including that there is less worry of overfitting, that they are more easily interpretable, and that the estimator is more efficient for the parametric components. However, the advantages of the partially linear models can only be realized if the model is correctly specified, which is obviously not easy to do in practice. Thus it would be a significant step forward if one can perform variable selection, parametric component identification, and estimation all at the same time.

Our study is motivated by a recent success story in predicting gene expression from sequence information only (Beer and Tavazoie, 2004). We are interested in a related but different problem called "motif regression" in Meier et al. (2009). With several "motif scores" calculated on a sequence, a regression model is set up to predict the gene expression. However, it is not clear whether the covariates have linear or nonlinear effects on the responses, and maybe some of them have linear effects while others have nonlinear effects.

In this article, we set out to study the possibility of achieving multiple goals simultaneously as discussed. We start by specifying a general additive model. Using a double penalization approach, where one penalty is used for variable selection and the other for parametric component identification, we demonstrate that our approach can automatically produce a PLAM that is correct with probability approaching one. Besides, the estimator possesses the oracle property in the sense that it is as efficient as the estimator when the true model is known prior to statistical analysis.

We will apply the group smoothly clipped absolute deviation (SCAD) penalization procedure to additive models. After the introduction of lasso in Tibshirani (1996), the penalization approach for variable selection has become increasingly popular (Zou and Hastie, 2005; Zhao and Yu, 2006; Zhang and Huang, 2008; Bickel, Ritov, and Tsybakov, 2009; Wu and Liu, 2009; Yuan, Joseph, and Zou, 2009; Peng et al., 2010). Furthermore, both SCAD penalty and adaptive lasso (Fan and Li, 2001; Zou, 2006; Zhang and Lu, 2007; Wang, Li, and Huang, 2008; Wang and Xia, 2009; Xie and Huang, 2009; Huang, Horowitz, and Wei, 2010) were proposed to address the consistency issues of the lasso. Unlike previous works on the same model that only perform variable selection and estimation of the nonzero components, we adopt a double penalization framework that can also identify the linear components. In the next section, we propose the double group SCAD procedure, and present its theoretical properties. In particular, we show that the procedure can identify the true model with probability approaching one, the usual one-dimensional nonparametric convergence rate is achieved on the component functions and furthermore the slope parameters for the linear parametric components actually converge faster and are asymptotically normal. These results show that our estimator has the oracle property. We also propose to use Bayesian information criterion (BIC; also adapted to our context that

includes two penalty terms) to choose the regularization parameters. In Section 3, some simulations are carried out to assess the performance of the proposed method and we also apply the method to some real datasets as illustrations. The technical proofs for the main theoretical results are provided in the Appendix.

2. Methodology

2.1 *Spline Estimator*

At the start of the analysis, we do not know which component functions are linear or actually zero and thus the following general additive model is used initially:

$$Y_i = \mu + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i, i = 1, \ldots, n, \tag{1}$$

where $(Y_i, \mathbf{X}_i)$ are independent and identically distributed with the same distribution as $(Y, \mathbf{X})$, $\mathbf{X} = (X_1, \ldots, X_p)^T$ is the p-dimensional predictor and ϵ_i is the noise with mean zero and variance σ^2.

We use polynomial splines to approximate the components (de Boor, 2001). Without loss of generality, we assume the distribution of X_j is supported on [0, 1] and also impose the condition $Ef_j(X_j) = 0$ for identifiability. Let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K'} < 1 = \tau_{K'+1}$ partition [0, 1] into subintervals $[\tau_k, \tau_{k+1})$, $k = 0, \ldots, K'$ with K' internal knots. We only restrict our attention to equally spaced knots although data-driven choice can be considered such as putting knots at certain sample quantiles of the observed covariate values. A polynomial spline of order q is a function whose restriction to each subinterval is a polynomial of degree $q-1$ and is globally $q-2$ times continuously differentiable on [0, 1]. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\{B_1(x), \ldots, B_{\tilde{K}}(x)\}$ with $\tilde{K} = K' + q$. Because of the centering constraint $Ef_j(X_j) = 0$, we instead focus on the subspace of spline functions $S_j^0 = \{s : s = \sum_{k=1}^{\tilde{K}} b_{jk} B_k(x), \sum_{i=1}^{n} s(X_{ij}) = 0\}$ with basis $\{B_{jk}(x) = B_k(x) - \sum_{i=1}^{n} B_k(X_{ij})/n, k = 1, \ldots, K = \tilde{K} - 1\}$ (the subspace is $K = \tilde{K} - 1$ dimensional due to the empirical version of the constraint). Using spline expansions, we can approximate the components by $f_j(x) \approx g_j(x) = \sum_{k=1}^{K} b_{jk} B_{jk}(x)$. Note that it is possible to specify different K for each component but we assume they are the same for simplicity. In this article, we will use $\|.\|$ to denote both the L_2 norm of the function and the l_2 (Euclidean) norm of a vector. Hopefully this will not cause any confusion in contexts.

Our main goal is to identify both the insignificant components (i.e., $f_j \equiv 0$) and the linear components (i.e., f_j is a linear function). The former can be achieved by shrinking $\|g_j\| = (\int g_j^2(x)dx)^{1/2}$ to zero. For the latter, we want to shrink the second derivative $\|g_j''\|$ to zero instead. This suggests the following minimization problem

$$(\widehat{\mu}, \widehat{\mathbf{b}}) = \arg\min_{\mu, b} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mu - \sum_j \sum_k b_{jk} B_{jk}(X_{ij}) \right)^2 + \sum_{j=1}^{p} p_{\lambda_1}(\|g_j\|) + \sum_{j=1}^{p} p_{\lambda_2}(\|g_j''\|), \tag{2}$$

where p_{λ_1} and p_{λ_2} are two penalties used to identify the zero and the linear coefficients, respectively, with two regularization parameters λ_1 and λ_2, and $g_j = \mathbf{b}_j^T \mathbf{B}_j$ with $\mathbf{b}_j = (b_{j1}, \ldots, b_{jk})^T$, $\mathbf{B}_j = (B_{j1}, \ldots, B_{jK})^T$. The estimated component functions are $\widehat{f}_j = \widehat{\mathbf{b}}_j^T \mathbf{B}_j$. Note that $\|g_j\|$ and $\|g_j''\|$ can be equivalently written as $\sqrt{\mathbf{b}_j^T \mathbf{D}_j \mathbf{b}_j}$ and $\sqrt{\mathbf{b}_j^T \mathbf{E}_j \mathbf{b}_j}$, respectively, with the (k, k') entry of $\mathbf{D}_j$ being $\int_0^1 B_{jk}(x) B_{jk'}(x)dx$ and the (k, k') entry of $\mathbf{E}_j$ being $\int_0^1 B_{jk}''(x) B_{jk'}''(x)dx$. However, in the following we will take $\mathbf{E}_j = K^{-4} \int_0^1 B_{jk}''(x) B_{jk'}''(x)dx$ after appropriate normalization. Penalizing the second derivative is commonly used in smoothing spline estimation (Wahba, 1990) as well as functional linear regression (Ramsay and Silverman, 2005). However, the purpose there is to encourage smoothness of the estimated nonparametric function and no model selection as we aim for here can be achieved. Accordingly, in smoothing spline literature the *square* of $\|g_j''\|$ is used as penalty, which is quite different from using SCAD penalty as done here. Finally, we mention that in Ni, Zhang, and Zhang (2009), for partially linear models, the authors also used a double penalization strategy with one penalty for sparsity in the linear part and the other for smoothness of the nonparametric component. This is also different from our aim here, which is to identify the nonparametric and parametric components starting from a model where all components are nonparametric initially.

It is easy to see that the estimator for the intercept term μ is given by $\overline{Y} = \sum_{i=1}^{n} Y_i/n$. By defining

$$\mathbf{Z}_j = \begin{pmatrix} B_{j1}(X_{1j}) & B_{j2}(X_{1j}) & \cdots & B_{jK}(X_{1j}) \\ \vdots & \vdots & \vdots & \vdots \\ B_{j1}(X_{nj}) & B_{j2}(X_{nj}) & \cdots & B_{jK}(X_{nj}) \end{pmatrix}_{n \times K},$$

$\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$, and $\mathbf{Y} = (Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})^T$, the above can be written in matrix from as

$$\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}} \frac{1}{n} \|\mathbf{Y} - \mathbf{Z}\mathbf{b}\|^2 + \sum_{j=1}^{p} p_{\lambda_1}(\|g_j\|) + \sum_{j=1}^{p} p_{\lambda_2}(\|g_j''\|). \tag{3}$$

For later use we denote the objective function on the right-hand side above as $Q(\mathbf{b})$. There are more than one way to specify the penalty functions and here we only focus on the SCAD penalty function (Fan and Li, 2001), defined by its first derivative

$$p_\lambda'(x) = \lambda \left\{ I(x \leqslant \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda) \right\},$$

with $a > 2$ and $p_\lambda(0) = 0$. We will use $a = 3.7$ as suggested in Fan and Li (2001). Other choices of penalty, such as adaptive lasso (Zou, 2006) or minimax concave penalty (Zhang, 2010), are expected to produce similar results in both theory and practice.

2.2 *Local Quadratic Approximation*

Following Fan and Li (2001) and Wang et al. (2008), we use an iterative local quadratic approximation algorithm to find the minimum of (3). Using a simple Taylor expansion, given an initial estimate $\mathbf{b}_j^{(0)}$ (equivalently given $g_j^{(0)}$), if $||g_j^{(0)}|| > 0$

and $||g_j^{(0)''}|| > 0$, we approximate the regularization terms by

$$p_{\lambda_1}(||g_j||) \approx p_{\lambda_1}\big(\big|\big|g_j^{(0)}\big|\big|\big) + \frac{1}{2}\frac{p'_{\lambda_1}\big(\big|\big|g_j^{(0)}\big|\big|\big)}{||g_j^{(0)}||}\big\{||g_j||^2 - \big|\big|g_j^{(0)}\big|\big|^2\big\},$$

and

$$p_{\lambda_2}(||g_j''||) \approx p_{\lambda_2}\big(\big|\big|g_j^{(0)''}\big|\big|\big) + \frac{1}{2}\frac{p'_{\lambda_2}\big(\big|\big|g_j^{(0)''}\big|\big|\big)}{\big|\big|g_j^{(0)''}\big|\big|}\big\{\big|\big|g_j''\big|\big|^2 - \big|\big|g_j^{(0)''}\big|\big|^2\big\}.$$

After removing some irrelevant terms, the criterion becomes

$$Q(\mathbf{b}) \approx \frac{1}{n}||\mathbf{Y} - \mathbf{Z}\mathbf{b}||^2 + \frac{1}{2}\mathbf{b}^T(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)\mathbf{b}, \tag{4}$$

for two $pK \times pK$ matrices $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ defined by

$$\mathbf{\Omega}_1 = \text{diag}\left(\frac{p'_{\lambda_1}\big(\big|\big|g_1^{(0)}\big|\big|\big)}{\big|\big|g_1^{(0)}\big|\big|}\mathbf{D}_1, \ldots, \frac{p'_{\lambda_1}\big(\big|\big|g_p^{(0)}\big|\big|\big)}{\big|\big|g_p^{(0)}\big|\big|}\mathbf{D}_p\right),$$

and

$$\mathbf{\Omega}_2 = \text{diag}\left(\frac{p'_{\lambda_2}\big(\big|\big|g_1^{(0)''}\big|\big|\big)}{\big|\big|g_1^{(0)''}\big|\big|}\mathbf{E}_1, \ldots, \frac{p'_{\lambda_2}\big(\big|\big|g_p^{(0)''}\big|\big|\big)}{\big|\big|g_p^{(0)''}\big|\big|}\mathbf{E}_p\right).$$

Note that (4) is a quadratic function and thus there exists a closed-form solution.

The algorithm repeatedly solves the minimization criterion (4) and updates $\mathbf{b}^{(m)}$ to $\mathbf{b}^{(m+1)}$, $m = 0, 1, \ldots,$ until convergence. That is, in the mth iteration, we solve (4), where $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ are as defined above but with $g_j^{(0)}$ replaced by the current estimate $g_j^{(m)} = \mathbf{b}_j^{(m)T}\mathbf{B}_j$. The solution obtained from (4) is the new estimate $\mathbf{b}_j^{(m+1)}$. If some $||g_j^{(m)}||$ or $||g_j^{(m)''}||$ falls below a threshold $\epsilon > 0$ (we set $\epsilon = 10^{-6}$ in our implementation), we take it to be zero.

The standard errors for the estimated nonzero linear parameters can be obtained directly by the sandwich formula, as suggested by Fan and Li (2001). The covariance of $\widehat{\mathbf{b}}$ is estimated by

$$(\mathbf{Z}^T\mathbf{Z} + n(\mathbf{\Omega}_1 + \mathbf{\Omega}_2))^{-1}\mathbf{Z}^T\widehat{Cov}(\mathbf{Y})\mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + n(\mathbf{\Omega}_1 + \mathbf{\Omega}_2))^{-1}, \tag{5}$$

where $\mathbf{\Omega}_1$, $\mathbf{\Omega}_2$ are as defined previously but with initial estimates $g_j^{(0)}$ there replaced by the estimates of the component functions after convergence is reached, and $\widehat{Cov}(\mathbf{Y})$ is the estimate of variance based on empirical residuals from the fitted model. The above sandwich formula will be shown in our simulation study to have good accuracy for moderate sample sizes.

Finally, we note that if $||\widehat{f}_j''|| = 0$, then $\widehat{f}_j$ is a linear function and implicitly we actually get an estimate $\widehat{\beta}_j$ for the slope parameter.

2.3 *Tuning Parameter Selection*

In practice, to achieve good numerical performances, we need to choose several parameters appropriately. We fix the spline order to be $q = 4$, that is we use cubic splines in all our numerical examples. For the number of basis K, we first fit the additive model without any penalization and use 10-fold crossvalidation to select K. With K determined, we propose to use a BIC-type criterion to select the regularization parameters λ_1 and λ_2 simultaneously. In our context, a natural BIC-type criterion is defined by

$$BIC_\lambda = \log\left\{\frac{1}{n}||\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}_\lambda||^2\right\} + d_1\frac{\log n}{n} + d_2\frac{\log(n/K)}{n/K},$$

where $\widehat{\mathbf{b}}_\lambda$ is the regularized estimate when $\lambda = (\lambda_1, \lambda_2)$ is used as the smoothing parameters, d_1 is the number of components estimated as parametric and d_2 is the number of components estimated as nonparametric. Thus the nonparametric components and the parametric components are penalized differently. The factor n/K for the nonparametric components is because in nonparametric problems it can be regarded as the effective sample size. Related to this, for kernel regression in varying-coefficient models, Wang and Xia (2009) used nh as the effective sample size in their definition of the BIC criterion, where h is the bandwidth in the kernel. The BIC-type criterion defined above will be shown to work well in our numerical examples.

2.4 *Asymptotic Results*

For convenience, we assume that f_j is truly nonparametric for $1 \leqslant j \leqslant p_1$, is linear for $p_1 + 1 \leqslant j \leqslant s = p_1 + p_2$ and zero for $s + 1 \leqslant j \leqslant p$. The true components are denoted by f_{0j}, $1 \leqslant j \leqslant p$ and the true slope parameters for the parametric components are denoted by $\boldsymbol{\beta}_0 = (\beta_{0,p_1+1}, \ldots, \beta_{0s})^T$.

Let $\mathbf{X}^{(1)} = (X_1, \ldots, X_{p_1})^T$ and $\mathbf{X}^{(2)} = (X_{p_1+1}, \ldots, X_s)^T$. Let $\mathcal{A}$ denote the subspace of functions on R^{p_1} that take an additive form

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

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

$$E\{(W - E_{\mathcal{A}}(W))(W - E_{\mathcal{A}}(W))\} = \inf_{h\in\mathcal{A}} E\big\{\big(W - h\big(\mathbf{X}^{(1)}\big)\big)\big(W - h\big(\mathbf{X}^{(1)}\big)\big)\big\}.$$

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

Let $\mathbf{h}(\mathbf{X}^{(1)}) = E_{\mathcal{A}}(\mathbf{X}^{(2)})$. Because $\mathbf{h}(\mathbf{X}^{(1)})$ is the projection of $\mathbf{X}^{(2)}$ onto the space of functions with an additive form, each component of $\mathbf{h}(\mathbf{X}^{(1)}) = (h_{(1)}(\mathbf{X}^{(1)}), \ldots, h_{(p_2)}(\mathbf{X}^{(1)}))^T$ can be written in the form $h_{(s)}(\mathbf{x}) = \sum_{j=1}^{p_1} h_{(s)j}(x_j)$ for some $h_{(s)j} \in S_j^0$. Denote $\mathbf{\Sigma} = E\{(\mathbf{X}^{(2)} - \mathbf{h}(\mathbf{X}^{(1)}))(\mathbf{X}^{(2)} - \mathbf{h}(\mathbf{X}^{(1)}))^T\}$. It was shown in Li (2000) that the positive definiteness of $\mathbf{\Sigma}$ is necessary for the identifiability of the model.

The following standard regularity assumptions are used.

(A1) The covariate vector $\mathbf{X}$ has a continuous density supported on $[0, 1]^p$. Furthermore, the marginal densities for X_j, $1 \leqslant j \leqslant p$ are all bounded from below and above by two fixed positive constants, respectively.

(A2) The noises ϵ_i are independent of covariates, have mean zero and variance σ^2.

(A3) $Ef_j(X_j) = 0$, $1 \leqslant j \leqslant s$. $f_j(x)$ is linear in x for $p_1 + 1 \leqslant j \leqslant s$, and $f_j \equiv 0$ for $j > s$.

(A4) For $g = f_j$, $1 \leqslant j \leqslant p_1$ or $g = h_{(s)j}$, $1 \leqslant s \leqslant p_2$, $1 \leqslant j \leqslant p_1$, g satisfies a Lipschitz condition of order $d >$

1/2: $|g^{(\lfloor d \rfloor)}(t) - g^{(\lfloor d \rfloor)}(s)| \leqslant C|s-t|^{d-\lfloor d \rfloor}$, 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 $q \geqslant d+2$.

(A5) The matrix $\mathbf{\Sigma}$ is positive definite.

All of the assumptions above are standard in the literature and in particular similar to those in Li (2000) and Huang, Horowitz, and Wei (2010). Assumptions (A3) and (A5) are necessary for the identifiability of the model. Assumption (A4) guarantees that the functions can be well approximated by polynomial splines.

THEOREM 1: *Assume (A1)–(A4), and that* $K \to \infty$, $K/n \to 0$, $\lambda_1, \lambda_2 \to 0$, *we have the rate of convergence*

$$||f_{0j} - \widehat{f}_j||^2 = O(K/n + 1/K^{2d}), 1 \leqslant j \leqslant p,$$

where $\widehat{f}_j = \widehat{\mathbf{b}}_j^T \mathbf{B}_j$ *is the estimated component function.*

The next theorem shows that when λ_1, λ_2 are appropriately specified, we can identify the true PLAM with high probability.

THEOREM 2: *In addition to the assumptions in Theorem 1, we assume* $((K/n)^{1/2} + K^{-d})^{-1}\min\{\lambda_1, \lambda_2\} \to \infty$. *Then with probability approaching 1,*

(a) $\widehat{f}_j \equiv 0, s+1 \leqslant j \leqslant p$,
(b) $\widehat{f}_j$ *is a linear function for* $p_1 + 1 \leqslant j \leqslant s$.

Finally, we show that for the linear components, the estimator for the slope parameter is asymptotically normal (this estimator $\widehat{\beta}_j$ is implicitly defined by $\widehat{f}_j$ when $\widehat{f}_j$ represents a linear function). We note that the asymptotic variance is the same as the estimate obtained when the true model is known beforehand, thus our estimator has the so-called oracle property.

THEOREM 3 (*Asymptotic Normality*): *Under the same set of assumptions as in Theorem 2, together with (A5) and that* $\sqrt{n}/K^{2d} \to 0$, *we have* $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(0, \sigma^2\Sigma^{-1})$ *in distribution.*

REMARK 1: From the convergence rate in Theorem 1, where the first term represents the stochastic error and the second term represents the approximation error, it is easily seen that $K \sim n^{1/(2d+1)}$ is the optimal choice of K. Using this theoretically optimal value of K, we see the assumption $((K/n)^{1/2} + K^{-d})\min\{\lambda_1, \lambda_2\} \to \infty$ in Theorem 2 becomes $n^{d/(2d+1)}\lambda_1 \to \infty$, $n^{d/(2d+1)}\lambda_2 \to \infty$. Furthermore, $\sqrt{n}/K^{2d} \to 0$ in Theorem 3 is satisfied with $K \sim n^{1/(2d+1)}$ as soon as $d > 1/2$.

3. Numerical Examples

3.1 *Simulation Studies*

We generate data from the model

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i,$$

with $f_1(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$, $f_2(x) = 4x(1-x)$, $f_3(x) = 2x$, $f_4(x) = x$, $f_5(x) = -x$, and $p = 10$ so that five components are actually zero.

Thus, the number of nonlinear components is $p_1 = 2$ and the number of linear components is $p_2 = 3$. To generate covariates, we first let X_{ij} be marginally standard normal with correlations given by $Cov(X_{ij_1}, X_{ij_2}) = (1/2)^{|j_1 - j_2|}$, and then apply the cumulative distribution function of the standard normal distribution to transform X_{ij} to be marginally uniform on [0, 1]. The noises are generated from mean zero normal distribution with standard deviation σ. We consider sample sizes n = 80, 150 and noise levels σ = 0.1, 0.3, resulting in four scenarios. The two noise levels give signal to noise ratio of about 9 : 1 and 3 : 1, respectively, with signal to noise ratio defined by standard deviation of the true regression function over that of the noise. For comparison, we compute the oracle estimator where we fit the true PLAM with five component functions, we also compute the sparse additive model where only the first penalty is used (and thus parametric components cannot be identified). For the oracle estimator we also use 10-fold crossvalidation to select K. During the review process, a referee raised the interesting question of whether an estimator that sequentially detects zero components and linear components performs as well as our simultaneous approach. Thus we also include this comparison with the sequential estimation procedure, where in the first stage only the first penalty in (2) is added, which removes the zero components, and in a second stage the second penalty in (2) is used to identify the partially linear structure.

For all four scenarios, 300 datasets are generated and the results are summarized in Tables 1 and 2. In Table 1, we show the model selection results for our doubly penalized estimator together with those for the sparse additive model with one single SCAD penalty and the sequential approach. In the table, under "PLAM" for each scenario, the first line shows the results for the doubly penalized estimator, while the second line shows the result of the sequential estimator. The standard errors reported are the sample standard deviations computed from the 300 replications. This demonstrates the good performance of BIC for tuning parameter selection. In terms of identifying the significant variables, the three methods perform similarly. However, the sparse additive model cannot detect the parametric components. We also see that the sequential method is inferior to doubly penalized estimator, with more linear components incorrectly incorporated into the nonparametric part.

In Table 2, we present the root mean squared errors for the first six component functions (note the sixth component is actually zero), which is defined by

$$\text{rmse} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\widehat{f}_j(t_i) - f_j(t_i))^2},$$

on a fine grid $(t_1, \ldots, t_T)$ consisting of 500 points equally spaced on [0, 1]. On the nonparametric components (f_1 and f_2), the errors for doubly penalized estimator, the sequential estimator and the sparse additive estimator are all similar. On the other hand, for the truly linear components, our doubly penalized estimators have obvious advantages, with rmse

Table 1

Model selection results for different estimators, using BIC for model selection. NV: average number of variables selected; NVT: average number of variables selected that are truly significant; NN: average number of nonlinear components selected; NNT: average number of nonlinear components selected that are truly nonlinear; NL: average number of linear components selected; NLT: average number of linear components selected that are truly linear. The numbers after "±" are the corresponding standard errors. For PLAM, results for both doubly penalized estimator (first line under each scenario) and sequential estimator (second line) are reported

	Sparse additive		PLAM			
	NV	NVT	NN	NNT	NL	NLT
$n = 80\,\sigma = 0.1$	5.28 ± 0.76	5 ± 0	2.65 ± 0.80	2 ± 0	2.65 ± 1.08	2.60 ± 0.73
			3.45 ± 1.70	2 ± 0	1.84 ± 1.30	1.81 ± 1.27
$n = 80\,\sigma = 0.3$	6.34 ± 1.63	5 ± 0	3.24 ± 2.14	2 ± 0	2.90 ± 1.56	2.85 ± 1.19
			5.18 ± 2.41	2 ± 0	1.15 ± 1.50	1.02 ± 1.21
$n = 150\,\sigma = 0.1$	4.95 ± 0.27	4.93 ± 0.22	2.19 ± 0.49	2 ± 0	2.88 ± 0.55	2.82 ± 0.47
			2.34 ± 0.75	2 ± 0	2.61 ± 0.74	2.60 ± 0.74
$n = 150\,\sigma = 0.3$	5.18 ± 0.47	5 ± 0	2.56 ± 0.84	2 ± 0	2.74 ± 1.04	2.60 ± 0.73
			2.98 ± 1.11	2 ± 0	2.19 ± 1.04	2.12 ± 0.95

about 40 ∼ 50% smaller than that for the sparse additive model, and also smaller than the sequential estimator.

We now test the accuracy of the standard error formula for the linear part. Table 3 presents the results for the three linear coefficients. The sample standard deviation of the estimated coefficients in the 300 simulations can be regarded as the benchmark (indicated by SD in the table). The average of the 300 estimated SDs based on (5) is shown under the column indicated by "SE." When calculating SDs and SEs, the cases where the components are incorrectly identified as nonparametric are discarded. Table 3 suggests that the sandwich formula performs reasonably well, with an underestimation bias.

3.2 *Motif Regression*

Gene expressions are regulated by transcription factors (TF) binding to the sequence elements at the upstream of genes. Toward decoding this regulatory mechanism, Beer and Tavazoie (2004) raised an interesting and very challenging question about whether gene expression can be predicted from the gene's upstream sequence. They made the first attempt to answer this question by first clustering genes into a small

Table 2

Root mean squared errors for $f_1, \ldots, f_6$. The numbers after ± are the corresponding standard errors estimated from 300 Monte Carlo replications

		Oracle	Simultaneous	Sequential	Sparse additive
$n = 80$	f_1	7.90 ± 0.63	8.09 ± 0.74	8.07 ± 0.69	7.98 ± 0.63
$\sigma = 0.1$	f_2	3.12 ± 1.04	3.41 ± 1.01	3.39 ± 0.99	3.41 ± 1.14
	f_3	1.49 ± 1.34	1.48 ± 1.15	2.10 ± 1.64	3.04 ± 1.32
	f_4	1.38 ± 0.99	1.56 ± 1.09	1.93 ± 1.48	3.07 ± 1.32
	f_5	1.42 ± 1.23	1.43 ± 1.11	1.84 ± 1.43	3.00 ± 1.17
	f_6	0 ± 0	0.13 ± 0.59	0.26 ± 1.16	0.17 ± 0.78
$n = 80$	f_1	9.92 ± 1.91	11.1 ± 2.74	10.9 ± 2.32	10.7 ± 2.15
$\sigma = 0.3$	f_2	6.98 ± 2.40	7.43 ± 3.04	7.57 ± 2.99	8.22 ± 3.07
	f_3	2.75 ± 2.33	4.82 ± 3.27	6.21 ± 3.81	7.91 ± 2.97
	f_4	3.04 ± 2.33	5.21 ± 3.60	6.77 ± 4.05	8.30 ± 3.31
	f_5	2.93 ± 2.09	5.22 ± 3.17	6.67 ± 4.12	8.38 ± 3.23
	f_6	0 ± 0	1.24 ± 2.90	2.65 ± 4.84	2.12 ± 4.07
$n = 150$	f_1	7.56 ± 0.33	7.64 ± 0.45	7.69 ± 0.42	7.58 ± 0.35
$\sigma = 0.1$	f_2	2.49 ± 0.56	2.74 ± 0.66	2.64 ± 0.62	2.52 ± 0.57
	f_3	0.88 ± 0.74	0.89 ± 0.64	1.00 ± 0.76	1.80 ± 0.83
	f_4	0.92 ± 0.73	0.93 ± 0.75	0.99 ± 0.82	1.81 ± 0.88
	f_5	0.90 ± 0.71	0.90 ± 0.70	0.96 ± 0.75	1.83 ± 0.77
	f_6	0 ± 0	0.03 ± 0.22	0 ± 0	0 ± 0
$n = 150$	f_1	8.91 ± 1.25	9.41 ± 1.56	9.51 ± 1.50	9.43 ± 1.26
$\sigma = 0.3$	f_2	4.92 ± 1.72	5.07 ± 1.63	4.97 ± 1.52	5.36 ± 1.94
	f_3	2.20 ± 1.70	2.54 ± 1.73	2.81 ± 2.07	4.80 ± 1.88
	f_4	2.24 ± 1.65	2.61 ± 2.03	2.79 ± 2.20	5.03 ± 2.06
	f_5	2.12 ± 1.54	2.68 ± 1.95	2.85 ± 2.04	5.21 ± 2.11
	f_6	0 ± 0	0.22 ± 0.99	0.30 ± 1.45	0.14 ± 0.83

Table 3
Standard deviations of our estimator for the coefficients of the three linear components. SE denotes the estimated standard deviation based on sandwich formula, while SD denotes the standard deviation estimated from the 300 Monte Carlo replicates

	f_3		f_4		f_5	
	SE	SD	SE	SD	SE	SD
$n = 80$, $\sigma = 0.1$	0.047	0.050	0.049	0.058	0.045	0.046
$n = 80$, $\sigma = 0.3$	0.149	0.171	0.148	0.173	0.126	0.146
$n = 150$, $\sigma = 0.1$	0.034	0.033	0.034	0.035	0.035	0.042
$n = 150$, $\sigma = 0.3$	0.094	0.101	0.090	0.092	0.092	0.098

number of clusters with similar expression patterns under various experimental conditions, and then using the presence or absence of discovered motifs in the sequence to predict the cluster membership of other genes. Because of the clustering, their prediction problem is discrete in nature and thus they adopted a Bayesian network for training and prediction.

Here we instead focus on the related motif regression problem, where we are directly interested in using motifs in the sequence to predict gene expression levels in microarray experiments, instead of looking at a finite number of typical expression patterns. More specifically, a motif score is calculated based on sequence information which represents the certainty that a motif appears in the sequence, because a "motif" is mathematically a position weight matrix representing a probabilistic model. Using scores of multiple motifs to predict expression level is thus a regression problem. Because this problem is not well investigated in the literature, it is reasonable to try a general additive model rather than a linear model to make the statistical analysis less constrained, as was done in Meier et al. (2009). However, such general additive models with more than just a couple of motif scores still cause the healthy skepticism of overfitting. In particular, one would wonder whether prediction can be improved using a partially linear structure.

We use the ChIP-chip data from Lee et al. (2002), which was also used in Hong et al. (2005). However, unlike here the main goal in those studies was to find the TF motifs. Recall that a ChIP-chip experiment uses chromatin immunoprecipitation (ChIP), followed by the detection of enriched fragments using DNA microarray hybridization, to determine the genomic-binding location of TF. Forty datasets, each containing genes targeted by one TF in *Saccharomyces cerevisiae*, have been obtained using ChIP-chip P-value 0.001 as the cutoff in the study of Hong et al. (2005). The sizes of these datasets range from 25 to 176 genes. For each gene, its promoter sequence is taken up to 800 bps upstream, but not overlapping with the previous gene. This results in between 25 and 176 positive sequences for different TFs. On the other hand, negative sequences were selected as those with ChIP-chip ratio $\leqslant 1$ and ChIP-chip P-value $\geqslant 0.05$, producing several thousand negative sequences for each TF.

For each TF, using the algorithm W-AlignACE (Chen et al., 2008), we obtain a candidate list of 10 motifs that are potentially predictive of expression levels. For each (sequence, motif) pair, we also obtain a score x_{ij} that represents how well the sequence matches the motif as in Conlon et al. (2003). For our study, we use only 24 datasets (24 TFs) for which W-AlignACE can find the true motif as in Chen et al. (2008), and we randomly selected the same number of negative sequences as the number of positive sequences. Half of the sequences (positive and negative combined) are used for training and predictions of binding intensities are made on the rest. The prediction errors for these 24 TFs are shown in Table 4, where we also listed the number of positive sequences in each dataset. Three estimators are compared in terms of prediction errors. These include our estimator, the sparse additive estimator where only one penalty is used, and a linear model with SCAD penalty (Fan and Li, 2001). It is seen that the nonparametric methods perform better than the linear method and our estimator is the best, achieving the smallest prediction error among the three estimators for 14 out of 24 datasets.

Table 4
Prediction errors for the 24 TFs, comparing three estimators

TF	#seq	Our estimator	Sparse additive	Sparse linear
ABF1	176	**4.25**	4.46	5.73
ACE2	46	**0.77**	0.86	1.61
BAS1	31	1.01	**0.90**	1.84
CAD1	27	0.62	**0.57**	0.68
CBF1	28	1.53	**1.40**	1.91
DIG1	35	1.22	**1.16**	2.88
FHL1	124	**3.86**	3.87	4.38
FKH2	72	**2.53**	2.70	4.98
GAL4	25	**3.03**	3.94	6.10
GCN4	56	**3.96**	4.15	5.35
HAP4	42	0.86	**0.80**	0.99
HSF1	33	**2.70**	2.95	4.34
MBP1	74	0.76	0.97	**0.62**
MCM1	58	2.13	2.15	**1.76**
NDD1	66	**2.09**	2.21	2.59
RAP1	127	**1.97**	2.14	3.28
REB1	89	0.92	**0.89**	1.06
STE12	54	**2.72**	2.91	3.42
SUM1	41	0.82	1.05	**0.81**
SWI4	90	**3.03**	3.62	4.07
SWI5	72	**0.89**	1.15	1.24
SWI6	65	**0.28**	0.37	0.33
YAP1	35	**1.49**	1.66	2.02
YAP5	55	4.74	**4.35**	4.68

Note: The smallest prediction errors are presented in boldface.

4. Conclusion

In previous studies on PLAM, the specification of the model is mainly based on prior domain knowledge, which is not always easy to obtain. One reasonable strategy is to put discrete covariates in the linear part. However, this does not apply to our expression prediction study because all predictors, the motif scores, are continuous and separation of parametric and nonparametric part cannot be guided by domain knowledge. We successfully used a novel penalization term that can help us identify the parametric part automatically, and in a principled way. When variable selection is also desired as in our case, the usual group penalty is added resulting in a doubly

penalized model. We also show that this double penalization strategy performs better than sequentially detecting zero and parametric components in two steps. By correctly identifying the partially linear structure, the resulting model is shown to be more accurate in predicting gene expressions based on sequence information.

After the first round of review, it was brought to our attention by one of the referees that at least two papers have addressed similar problems. In Zhang, Cheng, and Liu (2011), the authors used smoothing splines to decompose a nonparametric component into a linear part and another part orthogonal to it. However, they did not satisfactorily demonstrate the model identification consistency of their approach due to the difficulty in dealing with smoothing splines (they only proved consistency for the special case where the nonparametric components are periodic functions and conjectured that it is generally true). They also did not show asymptotic normality of the linear part. In terms of computation, using smoothing splines the number of basis coefficients is proportional to the sample size, while for polynomial splines the number of basis coefficients is proportional to K, which is typically chosen to be less than 10 in the literature. Thus use of polynomial splines might have advantages for large datasets, although a detailed comparison is out of the scope of the current article. Finally, they used the Boston housing data to illustrate the method while our application is related to biostatistics. In Huang, Wei, and Ma (2010), the authors used truncated power basis functions to approximate the nonlinear components; however, no variable selection is performed. They used the minimax concave penalty instead of the SCAD penalty. We expect the theoretical and empirical properties of using these two penalties would be similar. The use of B-splines in our current study makes the theoretical investigations more tractable and also our penalization of the second derivatives seems a natural approach to shrinkage toward linear functions. In particular, other series expansion approaches could be easily implemented by adopting the strategy of penalizing second derivatives.

Acknowledgements

We thank Professor Russell Millar, the associate editor, and two referees for their insightful comments and suggestions that have greatly improved both presentation and technical contents of this article. The research of HL is supported by a Singapore MOE Tier 2 grant. The research of XC is supported by Singapore MOE AcRF Tier 1 grant RG78/08.

References

Beer, M. and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell* **117,** 185–198.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* **37,** 1705–1732.

Chen, X., Guo, L., Fan, Z., and Jiang, T. (2008). W-alignace: An improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/chip-chip data. *Bioinformatics* **24,** 1121–1128.

Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 3339–3344.

de Boor, C. (1993). B(asic)spline basics. In *Fundamental Developments of Computer-Aided Geometric Modelling*, L. Piegl (ed.), 27–49. San Diego, California: Academic Press.

de Boor, C. (2001). *A Practical Guide to Splines*, rev. edition. New York: Springer-Verlag.

Fan, J. Q. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Hong, P., Liu, X. S., Zhou, Q., Lu, X., Liu, J. S., and Wong, W. H. (2005). A boosting approach for motif modeling using chip-chip data. *Bioinformatics* **21,** 2636–2643.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics* **38,** 2282–2313.

Huang, J., Wei, F., and Ma, S. (2010). *Semiparametric regression pursuit*. Technical Report, University of Iowa, Iowa City, Iowa.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., and Simon, I. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298,** 799–804.

Li, Q. (2000). Efficient estimation of additive partially linear models. *International Economic Review* **41,** 1073–1092.

Liang, H., Thurston, S., Ruppert, D., Apanasovich, T., and Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika* **95,** 667–678.

Liu, X., Wang, L., and Liang, H. (2011). Variable selection and estimation for semiparametric additive partial linear models. *Statistica Sinica* **21,** 1225–1248.

Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *Journal of Statistical Planning and Inference* **141,** 204–219.

Meier, L., Van de Geer, S., and Buhlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics* **37,** 3779–3821.

Ni, X., Zhang, H. H., and Zhang, D. (2009). Automatic model selection for partially linear models. *Journal of Multivariate Analysis* **100,** 2100–2111.

Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics* **8,** 715–732.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics* **4,** 53–77.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd edition, *Springer Series in Statistics*. New York: Springer.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2008). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems 20*, J. Platt and D. Koller, Y. Singer, and S. Roweis (eds.), 1201–1208. Cambridge, Massachusetts: MIT Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B—Methodological* **58,** 267–288.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Wang, H. S. and Xia, Y. C. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* **104,** 747–757.

Wang, L. F., Li, H. Z., and Huang, J. H. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association* **103,** 1556–1569.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica* **19,** 801–817.

Xie, H. L. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *Annals of Statistics* **37,** 673–696.

Xue, L. (2009). Consistent variable selection in additive models. *Statistica Sinica* **19,** 1281–1296.

Yuan, M., Joseph, V. R., and Zou, H. (2009). Structured variable selection and estimation. *Annals of Statistics* **3,** 1738–1757.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38,** 894–942.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* **36,** 1567–1594.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94,** 691–703.

Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear of nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association* to appear.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7,** 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320.

Received February 2011. *Revised July* 2011. *Accepted July* 2011.

Appendix

In the proofs, C denotes a generic constant that might assume different values at different places. We first note that, based on well-known properties of B-splines, $\mathbf{D}_j$ has eigenvalues of order $1/K$, $\mathbf{E}_j$ is of rank $K-1$ and all its positive eigenvalues are of order $1/K$ (de Boor, 1993, 2001; Huang, Horowitz, and Wei, 2010).

Proof of Theorem 1. Let $\mathbf{b}_0 = (b_{01}^T, \ldots, b_{0p}^T)^T$ be a pK dimensional vector that satisfies $\|f_{0j} - \mathbf{b}_{0j}^T \mathbf{B}_j\| = O(K^{-d}), 1 \leqslant j \leqslant p_1$ and $f_{0j} = \mathbf{b}_{0j}^T \mathbf{B}_j, j > p_1$. By the definition of $\widehat{\mathbf{b}}$, we have

$$\begin{aligned}
0 &\geqslant nQ(\widehat{\mathbf{b}}) - nQ(\mathbf{b}_0) \\
&= ||\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}||^2 - ||\mathbf{Y} - \mathbf{Z}\mathbf{b}_0||^2 \\
&\quad + n\sum_j p_{\lambda_1}\big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{D}_j \widehat{\mathbf{b}}_j}\big) - n\sum_j p_{\lambda_1}\big(\sqrt{\mathbf{b}_{0j}^T \mathbf{D}_j \mathbf{b}_{0j}}\big) \\
&\quad + n\sum_j p_{\lambda_2}\big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}\big) - n\sum_j p_{\lambda_2}\big(\sqrt{\mathbf{b}_{0j}^T \mathbf{E}_j \mathbf{b}_{0j}}\big) \\
&\geqslant ||\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}||^2 - ||\mathbf{Y} - \mathbf{Z}\mathbf{b}_0||^2 \\
&\quad - Cn(\lambda_1 + \lambda_2)\sum_j ||\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}||/\sqrt{K} \\
&= 2(\mathbf{Y} - \mathbf{Z}\mathbf{b}_0)^T \mathbf{Z}(\mathbf{b}_0 - \widehat{\mathbf{b}}) + ||\mathbf{Z}(\mathbf{b}_0 - \widehat{\mathbf{b}})||^2 \\
&\quad - Cn(\lambda_1 + \lambda_2)\sum_j ||\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}||/\sqrt{K}, \qquad \text{(A1)}
\end{aligned}$$

where in the second inequality above we used $|p_\lambda(|s|) - p_\lambda(|t|)| \leqslant \lambda|s-t|$.

Let $\boldsymbol{\eta} = \mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{Z}\mathbf{b}_0)$, where $\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$, be the projection of $\mathbf{Y} - \mathbf{Z}\mathbf{b}_0$ onto the columns of $\mathbf{Z}$, we will show that

$$||\boldsymbol{\eta}||^2 = O_p(K + n/K^{2d}). \qquad \text{(A2)}$$

In fact, if we denote $r_i = \sum_{j=1}^{p_1} f_{0j}(X_{ij})$ and $\mathbf{r} = (r_1, \ldots, r_n)^T$, we have $\mathbf{Y} - \mathbf{Z}\mathbf{b}_0 = \boldsymbol{\epsilon} + (\mathbf{r} - \mathbf{Z}\mathbf{b}_0)$ and $||\boldsymbol{\eta}||^2 \leqslant 2||\mathbf{P}_\mathbf{Z}\boldsymbol{\epsilon}||^2 + 2||\mathbf{r} - \mathbf{Z}\mathbf{b}_0||^2$. By the approximation property of splines, $||\mathbf{r} - \mathbf{Z}\mathbf{b}_0||^2 = O_p(n/K^{2d})$. Also, $E||\mathbf{P}_\mathbf{Z}\boldsymbol{\epsilon}||^2 = E(\boldsymbol{\epsilon}^T\mathbf{P}_\mathbf{Z}\boldsymbol{\epsilon}) = \sigma^2 tr(\mathbf{P}_\mathbf{Z}) = O(K)$ and (A2) is proved by an application of Markov inequality.

Based on this and using the Cauchy–Schwartz inequality, (A1) can be continued as

$$\begin{aligned}
0 \geqslant -|O_p(K + n/K^{2d})| - \frac{1}{2}||\mathbf{Z}(\mathbf{b}_0 - \widehat{\mathbf{b}})||^2 + ||\mathbf{Z}(\mathbf{b}_0 - \widehat{\mathbf{b}})||^2 \\
- Cn(\lambda_1 + \lambda_2)\sum_j ||\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}||/\sqrt{K}. \qquad \text{(A3)}
\end{aligned}$$

Using now Lemma A.1 in Wang et al. (2008), which implies that $||\mathbf{Z}(\mathbf{b}_0 - \widehat{\mathbf{b}})||^2 \sim n/K||\mathbf{b}_0 - \widehat{\mathbf{b}}||^2$, together with the Cauchy–Schwartz inequality $n(\lambda_1 + \lambda_2)\sum_j ||\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}||/\sqrt{K} \leqslant (CKn/4)(\lambda_1 + \lambda_2)^2/K + (n/CK)||\mathbf{b}_0 - \widehat{\mathbf{b}}||^2$ with a sufficiently large $C > 0$, (A3) gives $||\widehat{\mathbf{b}} - \mathbf{b}_0||^2 = O_p(K^2/n + 1/K^{2d-1} + (\lambda_1 + \lambda_2)^2 K)$.

The above convergence rate can be further improved to $||\widehat{\mathbf{b}} - \mathbf{b}_0||^2 = O_p(K^2/n + 1/K^{2d-1})$ as follows. First note that because the model is fixed as $n \to \infty$, we can find a constant $C > 0$ such that $\mathbf{b}_{0j}^T \mathbf{D}_j \mathbf{b}_{0j} > C$ when $j \leqslant s$ and $\mathbf{b}_{0j}^T \mathbf{E}_j \mathbf{b}_{0j} > C$ when $j \leqslant p_1$. Because $||\widehat{\mathbf{b}} - \mathbf{b}_0||^2 = o_p(K)$ by the convergence rates proved above and that $\lambda_1, \lambda_2 = o(1)$, we have $P(p_{\lambda_1}(\sqrt{\mathbf{b}_{0j}^T \mathbf{D}_j \mathbf{b}_{0j}}) = p_{\lambda_1}(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{D}_j \widehat{\mathbf{b}}_j})) \to 1$ if $j \leqslant s$. Similarly

$$P\big(p_{\lambda_2}\big(\sqrt{\mathbf{b}_{0j}^T \mathbf{E}_j \mathbf{b}_{0j}}\big) = p_{\lambda_2}\big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}\big)\big) \to 1 \text{ if } j \leqslant p_1.$$

These facts imply that

$$n\sum_{j=1}^{p} p_{\lambda_1}\big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{D}_j \widehat{\mathbf{b}}_j}\big) - n\sum_{j=1}^{p} p_{\lambda_1}\big(\sqrt{\mathbf{b}_{0j}^T \mathbf{D}_j \mathbf{b}_{0j}}\big) \geqslant 0,$$

and

$$n\sum_{j=1}^{p} p_{\lambda_2}\big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}\big) - n\sum_{j=1}^{p} p_{\lambda_2}\big(\sqrt{\mathbf{b}_{0j}^T \mathbf{E}_j \mathbf{b}_{0j}}\big) \geqslant 0,$$

with probability tending to 1. Removing the regularizing terms in (A1) and using the same reasoning as before, the rates are improved to $||\widehat{\mathbf{b}} - \mathbf{b}_0||^2 = O_p(K^2/n + 1/K^{2d-1})$.

The rates of convergence for $||\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}||^2$ immediately imply the rates for $||\widehat{f}_j - f_{0j}||$ by the property (de Boor, 2001)

$$\begin{aligned}
C_1 K\big\|\widehat{\mathbf{b}}_j^T \mathbf{B}_j - \mathbf{b}_{0j}^T \mathbf{B}_j\big\|^2 \leqslant ||\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}||^2 \\
\leqslant C_2 K\big\|\widehat{\mathbf{b}}_j^T \mathbf{B}_j - \mathbf{b}_{0j}^T \mathbf{B}_j\big\|^2,
\end{aligned}$$

for some constants C_1, $C_2 > 0$.

Proof of Theorem 2. We only show part (b) as an illustration and part (a) is similar. Suppose for some $p_1 < j \leqslant s$, $\widehat{\mathbf{b}}_j^T \mathbf{B}_j$

does not represent a linear function. Define $\widehat{\mathbf{b}}^*$ to be same as $\widehat{\mathbf{b}}$ except that $\widehat{\mathbf{b}}_j$ is replaced by its projection onto the subspace $\{\mathbf{b}_j : \mathbf{b}_j^T \mathbf{B}_j$ represents a linear function$\}$. By definition of $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{b}}^*$ we have

$$\begin{aligned}
0 &\geqslant nQ(\widehat{\mathbf{b}}) - nQ(\widehat{\mathbf{b}}^*) \\
&= ||\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}||^2 - ||\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}^*||^2 \\
&\quad + np_{\lambda_1}\Big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{D}_j \widehat{\mathbf{b}}_j}\Big) - np_{\lambda_1}\Big(\sqrt{\widehat{\mathbf{b}}_j^{*T} \mathbf{D}_j \widehat{\mathbf{b}}_j^*}\Big) \\
&\quad + np_{\lambda_2}\Big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}\Big) - np_{\lambda_2}\Big(\sqrt{\widehat{\mathbf{b}}_j^{*T} \mathbf{E}_j \widehat{\mathbf{b}}_j^*}\Big).
\end{aligned}$$

As in the proof of Theorem 1, because $\|\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}\|^2 = o_p(K)$, we have $\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{D}_j \widehat{\mathbf{b}}_j} \geqslant a\lambda$ with probability approaching 1, which implies $P(p_{\lambda_1}(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{D}_j \widehat{\mathbf{b}}_j}) \geqslant p_{\lambda_1}(\sqrt{\widehat{\mathbf{b}}_j^{*T} \mathbf{D}_j \widehat{\mathbf{b}}_j^*})) \to 1$, and thus (because $\widehat{\mathbf{b}}_j^{*T} \mathbf{E}_j \widehat{\mathbf{b}}_j^* = 0$)

$$\begin{aligned}
0 &\geqslant ||\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}||^2 - ||\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}^*||^2 + np_{\lambda_2}\Big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}\Big) \\
&= -2(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}^*)\mathbf{Z}(\widehat{\mathbf{b}} - \widehat{\mathbf{b}}^*) + ||\mathbf{Z}(\widehat{\mathbf{b}} - \widehat{\mathbf{b}}^*)||^2 \\
&\quad + np_{\lambda_2}\Big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}\Big) \\
&\geqslant -2||\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}^*)|| \cdot ||\mathbf{Z}(\widehat{\mathbf{b}} - \widehat{\mathbf{b}}^*)|| \\
&\quad + np_{\lambda_2}\Big(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}\Big). \qquad \text{(A4)}
\end{aligned}$$

We have the bound

$$\begin{aligned}
&||\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{b}}^*)||^2 \\
&\quad \leqslant 2||\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\mathbf{b}_0)||^2 + 2||\mathbf{Z}(\widehat{\mathbf{b}}^* - \mathbf{b}_0)||^2 \\
&\quad = O_p(K + n/K^{2d}) + (n/K)O_p(K^2/n + 1/K^{2d-1}) \\
&\quad = O_p(K + n/K^{2d}), \qquad \text{(A5)}
\end{aligned}$$

using (A2) and that $||\widehat{\mathbf{b}}^* - \mathbf{b}_0|| \leqslant ||\widehat{\mathbf{b}} - \mathbf{b}_0||$ (this is because $\widehat{\mathbf{b}}_j^*$ is the projection of $\widehat{\mathbf{b}}_j$ to the subspace $\{\mathbf{b}_j : \mathbf{b}_j^T \mathbf{B}_j$ represents a linear function$\}$, and $\mathbf{b}_{0j}$ is inside this subspace, thus we have $||\widehat{\mathbf{b}}_j^* - \mathbf{b}_{0j}|| \leqslant ||\widehat{\mathbf{b}}_j - \mathbf{b}_{0j}||$).

On the other hand, because $\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j} = \sqrt{(\widehat{\mathbf{b}}_j - \mathbf{b}_{0j})^T \mathbf{E}_j (\widehat{\mathbf{b}}_j - \mathbf{b}_{0j})} = O_p((K/n)^{1/2} + K^{-d}) = o(\lambda_2)$ by Theorem 1, we have

$$p_{\lambda_2}(\sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}) = \lambda_2 \sqrt{\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}, \text{with probability tending to 1,} \qquad \text{(A6)}$$

by the definition of the SCAD penalty function. Noting that $||\widehat{\mathbf{b}} - \widehat{\mathbf{b}}^*|| = ||\widehat{\mathbf{b}}_j - \widehat{\mathbf{b}}_j^*|| \sim \sqrt{K\widehat{\mathbf{b}}_j^T \mathbf{E}_j \widehat{\mathbf{b}}_j}$ and combining (A4)–(A6), we have a contradiction if $\widehat{\mathbf{b}}_j^T E_j \widehat{\mathbf{b}}_j > 0$.

Proof of Theorem 3. We note that because of Theorem 2, we only need to consider a correctly specified PLAM without regularization terms. This reduces the problem to the one studied in Li (2000) and the results there directly apply, showing the asymptotic normality of the slope parameter.