Statistica Sinica **24** (2014), 1215-1238
doi:http://dx.doi.org/10.5705/ss.2012.040

LOCAL POLYNOMIAL AND PENALIZED TRIGONOMETRIC SERIES REGRESSION

Li-Shan Huang and Kung-Sik Chan

National Tsing Hua University and University of Iowa

Abstract: We investigate the connections between local polynomial regression, mixed models, and penalized trigonometric series regression. Expressing local polynomial regression in a projection framework, we derive equivalent kernels for both interior and boundary points. For interior points, it is shown that the asymptotic bias decreases as the order of polynomial increases. Then we show that, under some conditions, the local polynomial projection approach admits an equivalent mixed model formulation where the fixed effects part includes the polynomial functions. The random effects part in the representation is shown to be the trigonometric series, asymptotically. The connections are extended to partial linear models and additive models. These results suggest a new smoothing approach using a combination of unpenalized polynomials and penalized trigonometric functions. We illustrate the potential usefulness of the new approach through several examples.

Key words and phrases: Additive model, ANOVA decomposition, equivalent kernel, mixed model, partial linear model, projection, varying coefficient model.

1. Introduction

The method of local polynomial regression (LPR) (Fan and Gijbels (1996)) has enjoyed wide popularity due to its attractive theoretical properties. For estimating an unknown regression function, it was suggested by Fan and Gijbels (1996) that choosing an odd degree polynomial, e.g. $p = 1$, is better than the next lower even order, e.g. $p = 0$, since the asymptotic bias of $p = 1$ is simpler than that of $p = 0$ without cost in the asymptotic variance. In this paper, we extend the investigation in Huang and Chen (2008) to form a projection framework for LPR and establish two key results. For interior points, the asymptotic bias of local p-th polynomial projection estimates is of order $h^{2(p+1)}$ for $0 \le p \le 3$, where h is the bandwidth. This result is more intuitive and offers a different view than Fan and Gijbels (1996). Further, local polynomial projection has interesting connections to mixed models (Fitzmaurice, Laird, and Ware (2004)), that suggest a new smoothing approach of taking a combination of unpenalized polynomials and penalized trigonometric functions for curve fitting. They also imply that using mixed model software as smoothing tools is not only for penalized splines

(Ruppert, Wand, and Carroll (2003)), but also for more general penalized series methods.

In Section 2, we review local linear projection, first considered in Huang and Chen (2008). Section 3 presents an extension to local polynomial projection. Theorem 1 derives equivalent kernels for local quadratic and cubic polynomial projection estimates and establishes the asymptotic bias results. Theorem 2 investigates equivalent kernels for a general class of two-step projection-type kernel estimators and includes results for estimation at the boundary region. Theorem 3 in Section 4 shows that the local polynomial projection approach admits an equivalent mixed effects model formulation in which the fixed effects part includes the polynomial functions. The trigonometric series is shown in Theorem 4 to form a penalized basis, asymptotically, for the random effects. In addition, when $0 \leq p \leq 3$, each trigonometric function receives different penalty weight: the higher the frequency, the larger the penalty weight. The mixed model connection is further extended to partial linear models and additive models, with the latter linked to backfitting estimation.

These results renew using the trigonometric functions as a regression tool (Graybill (1976); Eubank (1988); Dette and Melas (2003)) and suggest a different smoothing approach by taking a combination of unpenalized polynomials and penalized trigonometric functions for curve fitting. In the literature, Graybill (1976) and Eubank and Speckman (1990) considered polynomial-trigonometric regression with both parts unpenalized, while the trigonometric spline approach (Wang (2011)) does not include polynomials except for an intercept. The orthogonality of the trigonometric functions allows one to avoid problems of collinearity so the proposed approach is a practical alternative to the use of such smoothing methods as penalized splines. In Section 5, we illustrate the potential usefulness of the new approach for several models, univariate nonparametric regression, partial linear models, and additive models, and heuristically explore extensions to varying coefficient models and semiparametric mixed models. A simulated example demonstrates using AICc (Hurvich, Simonoff, and Tsai (1998)) to choose the penalty and the number of basis functions, and the results suggest that the choice of penalty may be more important than the choice of the number of basis functions. With trigonometric functions widely applied in many fields, it is expected that smoothing with penalized trigonometric series will bring up many interesting problems and applications. A summary of our results and some concluding remarks are given in Section 6.

2. Local Linear Projection

Assume data pairs (X_i, Y_i), $i = 1, \ldots, n$, are independently drawn as

$$Y = m(X) + \varepsilon, \tag{2.1}$$

where both X and Y are 1-dimensional, X and ε are independent, and ε has zero mean and unknown variance σ^2. For estimating the conditional mean $m(x) = E(Y|X=x)$ in (2.1), fitting a p-th order LPR (Fan and Gijbels (1996)) involves solving the weighted least squares problem

$$\min_{\beta} n^{-1} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \Big)^2 K_h (X_i - x), \tag{2.2}$$

where $\beta = (\beta_0, \ldots, \beta_p)^T$, $K_h(\cdot) = K(\cdot/h)/h$, and the dependence of β on x and h is suppressed. The function $K(\cdot)$ is a nonnegative weight function, typically a symmetric probability density function, and h is the smoothing parameter determining the neighborhood size for local fitting. Let $\hat{\beta}_j$, $j = 0, \ldots, p$, denote the solution to (2.2). It is clear that $\hat{\beta}_0$ estimates the regression function $m(x)$ of interest and $\hat{\beta}_j$, $j = 1, \ldots, p$, may be used to estimate derivatives $\beta_j(x) \equiv m^{(j)}(x)/j!$. The conventional LPR focuses on the intercept $\hat{\beta}_0(x)$ for estimating $m(x)$, while the information from $\hat{\beta}_j(x)$, $j \geq 1$, is generally ignored unless there is a need for estimating derivatives.

2.1. Local polynomial ANOVA

In an attempt to develop an analysis of variance (ANOVA) framework for LPR, Huang and Chen (2008) proposed quantities based on integrated sums of squares:

$$\begin{aligned} SSE_p(h) &= n^{-1} \int \sum_{i=1}^{n} \Big(Y_i - \sum_{j=0}^{p} \hat{\beta}_j(x)(X_i - x)^j \Big)^2 K_h \Big(X_i - x \Big) dx, \\ SSR_p(h) &= n^{-1} \int \sum_{i=1}^{n} \Big(\sum_{j=0}^{p} \hat{\beta}_j(x)(X_i - x)^j - \bar{Y} \Big)^2 K_h \Big(X_i - x \Big) dx, \end{aligned} \tag{2.3}$$

and show that the ANOVA decomposition

$$SSE_p(h) + SSR_p(h) = SST \equiv n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{2.4}$$

holds under some conditions. In a matrix form, (2.3) is written as

$$SSE_p(h) = n^{-1} \mathbf{y}^T (I - H_p^*) \mathbf{y}, \qquad SSR_p(h) = n^{-1} \mathbf{y}^T (H_p^* - J) \mathbf{y}, \tag{2.5}$$

and $SST = n^{-1} \mathbf{y}^T (I - J) \mathbf{y}$, where $\mathbf{y} = (Y_1, \ldots, Y_n)^T$, J is an $n \times n$ matrix with entries $1/n$, and H_p^* is a symmetric matrix depending only on X_i's, bandwidth h, and the kernel function $K(\cdot)$. More explicitly,

$$H_p^* = \int W X_p (X_p^T W X_p)^{-1} X_p^T W dx, \tag{2.6}$$

where W is an n-dimensional diagonal matrix with $K_h(X_i - x)$ as its diagonal elements, and X_p is the $n \times (p+1)$ design matrix in (2.2) with the $(j+1)$-th column $((X_1-x)^j, \ldots, (X_n-x)^j)^T$, $j = 0, \ldots, p$. The dependence of W and X_p on x is suppressed and the integration in (2.6) is performed element by element in the resulting matrix product.

The ANOVA investigation by Huang and Chen (2008) gives rise to a new smoother (projection) matrix H_p^* as well as a projection estimate $H_p^*\mathbf{y}$, that differ from the conventional smoother matrix (Hastie and Tibshirani (1990)) S_p and $S_p\mathbf{y}$ with $S_p\mathbf{y} = (\hat{\beta}_0(X_1), \ldots, \hat{\beta}_0(X_n))^T$. We call $H_p^*\mathbf{y}$ a local polynomial projection estimate. Some properties of H_p^* are given in Huang and Chen (2008). (i) Based on a symmetric $K(\cdot)$, H_p^* is symmetric (see (2.6)), while S_p is generally non-symmetric. (ii) $H_p^*\mathbf{x}^j = \mathbf{x}^j$, where $\mathbf{x}^j$ denotes $(X_1^j, \ldots, X_n^j)^T$, $j = 0, \ldots, p$. Hence elements of $H_p^*\mathbf{y}$ are weighted averages of $\mathbf{y}$. We note that S_p also satisfies $S_p\mathbf{x}^j = \mathbf{x}^j$, $j = 0, \ldots, p$. (iii) The projection estimate $H_p^*\mathbf{y}$ has clearer geometric interpretations in projecting $\mathbf{y}$ into the neighborhood space spanned by local polynomials (column space of X_p) and then combining all local projections by a weighted integral as

$$H_p^*\mathbf{y} = \begin{pmatrix} \int \Big(\sum_{j=0}^{p} \hat{\beta}_j(x)(X_1 - x)^j \Big) K_h(X_1 - x)dx \\ \vdots \\ \int \Big(\sum_{j=0}^{p} \hat{\beta}_j(x)(X_n - x)^j \Big) K_h(X_n - x)dx \end{pmatrix}. \tag{2.7}$$

2.2. Smaller bias, larger boundary

In the case of local linear regression ($p = 1$), Huang and Chen (2008) showed that, as $h \to 0$ and $nh \to \infty$ as $n \to \infty$, "interior" elements in $H_1^*\mathbf{y}$ have an asymptotic bias of order h^4, smaller than the order h^2 of $\hat{\beta}_0(X_i)$, and an asymptotic variance of the same order $O(n^{-1}h^{-1})$ as that of $\hat{\beta}_0(X_i)$, though with a larger constant factor. Under Conditions (A1) and (A2) in the Appendix, the "interior" region is $[2h, 1-2h]$, not the conventional $[h, 1-h]$, as will be explained. To better understand (2.7) in the local linear case, Huang and Davidson (2010) derived its "equivalent kernel" (see Fan and Gijbels (1996) for definition). Let $H_1^*\mathbf{y} = (Y_1^*, \ldots, Y_n^*)^T$ and write Y_i^* as a weighted estimator, $Y_i^* = \sum_{k=1}^n W_h(X_k - X_i)Y_k$ with $W_h(\cdot) = W(\cdot/h)/h$. Then $Y_i^* = [(1 + o_P(1))/(nf(X_i))] \sum_{k=1}^n W_h^*(X_k - X_i)Y_k$ and the equivalent kernel $W^*(\cdot)$ when $p = 1$ is

$$\left(K_0^*(\cdot) - \mu_2^{-1} K_1^*(\cdot) \right), \tag{2.8}$$

where $K_0^*(\cdot)$ is the convolution of the kernel function $K(\cdot)$ and itself, $K_1^*(\cdot)$ the convolution of $uK(u)$ and itself, and μ_j denotes the j-th moment of $K(\cdot)$. The equivalent kernel (2.8) is of order $(0,4)$ (see Gasser, Müller, and Mammitzsch (1985) for definition), which yields an asymptotic bias of order h^4 in Huang and Chen (2008).

In case $p=1$, He and Huang (2009) explained why the boundary range for $H_1^*\mathbf{y}$ is $[0,2h)\bigcup(1-2h,1]$, since local linear estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ suffer boundary effects when $x \in [0,h)\bigcup(1-h,1]$. When $X_i \in [h,2h)$, part of $\hat{\beta}_j(x)$, $j=0,1$, falls in $[0,h)$ (see (2.7)) and hence the boundary region is $[0,2h)$ (similarly for $X_i \in (1-2h,1-h]$.) He and Huang (2009) showed that the boundary bias for $X_i \in [h,2h)$ is of order h^2 but with a smaller constant multiplier than that of the local linear estimator. For $X_i \in [0,h)$, the boundary bias is to the constant order and some corrections are suggested, e.g. using a one-sided kernel.

3. Equivalent Kernels for Local Polynomial Projection

We extend the asymptotic results of local linear projection in Section 2 to local polynomial projection in this section by deriving equivalent kernels for $H_p^*\mathbf{y}$ ((2.7)). First, we discuss some properties of H_p^* due to its symmetry.

a. H_p^* is a shrinking matrix (see Lemma 1 in the Appendix), similar to the smoother matrix for smoothing splines. Since S_p for LPR is generally non-symmetric, its eigenvalues may be complex and therefore we conjecture that S_p for LPR may not be a shrinking smoother. Buja, Hastie, and Tibshirani (1989) showed that the smoother for local linear fits with nearest neighbors is not a shrinking matrix. Being symmetric and shrinking is a desirable property for the application of backfitting algorithms in estimation of additive models (Hastie and Tibshirani (1990)).
b. As indicated in Linton and Jacho-Chávez (2010), an estimator with a symmetric smoother matrix, such as $H_p^*\mathbf{y}$, is admissible with respect to trace mean squared error (Cohen (1966)), whereas the local polynomial class $\hat{\beta}_0(x)$ with asymmetric S_p is inadmissible under the same criteria.
c. When the X_i's are equally spaced and ordered monotonically, under Condition (A2) in the Appendix, the H_p^* matrix is asymptotically a banded Toeplitz matrix, i.e., a banded centrosymmetric matrix. This property may be convenient for developing methodology using H_p^* for discrete time series data.

3.1. It is not an odd world

Expanding the results in Section 2, we show that a projection estimate by local quadratic regression ($p=2$) has an asymptotic bias of order h^6, while that for a local cubic regression ($p=3$) has a bias of order h^8.

Theorem 1. *Under Conditions* (A) *in the Appendix, the following results hold for interior points:*

(a) *The equivalent kernel for elements in* $H_0^*\mathbf{y}$ *is asymptotically* $K_0^*(\cdot)$. *For* $p = 2, 3$, *the equivalent kernels for* $H_p^*\mathbf{y}$ *are asymptotically*

$$\frac{1}{\mu_4 - \mu_2^2}\left\{\mu_4 K_0^*(\cdot) - 2\mu_2 K_{0,2}^*(\cdot) + K_2^*(\cdot)\right\} - \frac{1}{\mu_2}K_1^*(\cdot), \tag{3.1}$$

$$\begin{aligned}&\frac{1}{\mu_4 - \mu_2^2}\left\{\mu_4 K_0^*(\cdot) - 2\mu_2 K_{0,2}^*(\cdot) + K_2^*(\cdot)\right\}\\ &- \frac{1}{\mu_2\mu_6 - \mu_4^2}\left\{\mu_6 K_1^*(\cdot) - 2\mu_4 K_{1,3}^*(\cdot) + \mu_2 K_3^*(\cdot)\right\},\end{aligned} \tag{3.2}$$

respectively, where $K_{0,2}^*(\cdot)$ *denotes the convolution of* $K(u)$ *and* $u^2K(u)$, $K_{1,3}^*(\cdot)$ *denotes the convolution of* $uK(u)$ *and* $u^3K(u)$, *and* $K_3^*(\cdot)$ *denotes the convolution of* $u^3K(u)$ *and itself.*

(b) *Local polynomial projection estimates* $H_p^*\mathbf{y}$ *have an asymptotic bias of order* $h^{2(p+1)}$ *for* $p = 0, 1, 2, 3$.

(c) *The asymptotic variance of* $H_p^*\mathbf{y}$, $p = 0, 1, 2, 3$, *is of order* $O(n^{-1}h^{-1})$ *with the constant multiplier* $\int W_p^{*^2}(u)du$, *where* $W_p^*(\cdot)$ *is the corresponding equivalent kernel.*

See the Appendix for an outline of the proof of Theorem 1. The results in Theorem 1 and (2.8) offer a different view from conventional LPR that an odd p is better than an even p ("it is an odd world", as described in Fan and Gijbels (1996)). For interior points of $H_p^*\mathbf{y}$, the higher the order of local polynomials fits, the smaller the asymptotic bias, which is consistent with common knowledge in classical global polynomial models. These results also offer a way to construct higher order $(0, 4)$, $(0, 6)$, and $(0, 8)$ kernels by using a symmetric kernel $K(u)$ of order (0,2) and its associated functions $uK(u)$, $u^2K(u)$, and $u^3K(u)$.

When K is the Gaussian kernel, it is easy to check that the resulting equivalent kernels (2.8), (3.1) and (3.2) are the Gaussian-based kernels in Wand and Schucany (1990). When K is the Epanechnikov kernel, the resulting equivalent kernels are polynomial kernels of order 7, 9, and 11, respectively, and have support $[-2, 2]$ due to convolution. After rescaling them to $[-1, 1]$, the resulting functions do not correspond to optimal kernels as given in Table 5.7 of Müller (1988), nor do those in Berlinet (1993).

3.2. Two-step smoothing

The local polynomial projection estimator (2.7) may be interpreted as a two-step estimator: local p-th degree polynomial with $K(\cdot)$ and h is fitted to obtain $\hat{\beta}_j$'s; all fitted polynomials are combined by a locally weighted integral with weight $K(\cdot)$ and bandwidth h. It is clear that $K(\cdot)$ and h in the second step do not need to be the same as those in the first step, yielding a general class of projection estimators at a target point t as

$$\int \Big(\sum_{j=0}^{p} \hat{\beta}_j(x)(t-x)^j \Big) L_g(t-x)dx, \tag{3.3}$$

where L and g denote the kernel function and bandwidth, respectively, in the second step. He and Huang (2009) investigated (3.3) when $p=1$ and $g=h$. They showed that the bias for $t \in [h, 2h)$ is of order h^3 if $L = K_+$ ($K_+(u) = 2K(u), u \in [0,1]$ is a one-sided kernel), and for $t \in [0,h)$, bias is of order h^2 if $L = K_+$. In Theorem 2 below, we discuss the asymptotic properties of (3.3) for a general p in two scenarios: (a) When $g = h$, expressions of the equivalent kernels are given for both interior and boundary points, and the order of the asymptotic bias can be derived accordingly based on moments of L and K. (b) When $g \neq h$, the first order term of the asymptotic bias is given for interior points, which, in contrast with Theorem 1, does not offer more advantages than the case with $g = h$.

Theorem 2. *Assume that L is supported on a compact interval. Under Conditions* (A) *in the Appendix, the following results hold for* (3.3).

(a) $g = h$.

(i) *For interior points $t \in [2h, 1-2h]$, the equivalent kernel is*

$$\sum_{j=0}^{p} u^j L(\cdot) * K_j^e(\cdot), \tag{3.4}$$

where $*$ denotes convolution and $K_j^e(\cdot)$ is the equivalent kernel for $\hat{\beta}_j$, $j = 0, \ldots, p$.

(ii) *For points $t = (1+c)h \in [h, 2h)$ with $0 < c < 1$, the equivalent kernel is*

$$\sum_{j=0}^{p} \left\{ \int_{-1}^{c} u^j L(u) K_j^e(u-v)du + \int_{c}^{1} u^j L(u) K_{j,(1+c-u)}^e(u-v)du \right\}, \tag{3.5}$$

where $K_{j,d}^e(\cdot)$ is the equivalent kernel for $\hat{\beta}_j(dh)$, $j = 0, \ldots, p$ with $0 < d < 1$. The range of v is the convolution range of K and L.

(iii) *For points $t = ch \in [0, h]$ with $0 < c < 1$, the equivalent kernel is*

$$\sum_{j=0}^{p} \left\{ \int_{-1}^{c-1} u^j L(u) K_j^e(u-v)du + \int_{0}^{1} u^j L(u) K_{j,(c-u)}^e(u-v)du \right\}. \tag{3.6}$$

The range of v is the convolution range of K and L.

(b) *When $g \neq h$ with $g \to 0$ and $ng \to \infty$ as $n \to \infty$, and L is symmetric, the first order term of the asymptotic bias of (3.3) for interior points is*

(i) *for $p = 1$,*

$$\beta_2(t)(h^2\mu_2 - g^2\mu_{2,L}), \tag{3.7}$$

where $\mu_{j,L}$ denotes the j-th moment of L;

(ii) *for $p = 2$,*

$$h^4\left\{\beta_3(t)\frac{f'(t)}{f(t)} + \beta_4(t)\right\}c_1 + h^2g^2\left\{-\beta_3(t)\frac{f'(t)}{f(t)}\frac{c_1\mu_{2,L}}{\mu_2}\right.$$
$$\left.+\beta_4(t)\left(4\frac{\mu_4\mu_{2,L}}{\mu_2} + c_2\mu_{2,L}\right)\right\} - g^4 5\beta_4(t)\mu_{4,L}, \tag{3.8}$$

where $c_1 = (\mu_4^2 - \mu_2\mu_6)/(\mu_4 - \mu_2^2)$ and $c_2 = (\mu_6 - \mu_2\mu_4)/(\mu_4 - \mu_2^2)$;

(iii) *for $p = 3$,*

$$\beta_4(t)\{h^4c_1 + h^2g^2c_2\mu_{2,L} - g^4\mu_{4,L}\}. \tag{3.9}$$

See Fan and Gijbels (1996) for a partial listing of expressions for $K_j^e(u)$ and $K_{j,d}^e(u)$ in Theorem 2(a). Theorem 2(a)(i) extends Theorem 1 by providing a general form of equivalent kernels. The results in Theorem 2(a)(ii) and (iii) can be used to obtain similar results for boundary points in $[1-2h, 1]$ by symmetry. Theorem 2(b) shows that (3.3) does not generally enjoy bias reduction effects when $g \neq h$ for interior points. When $p = 1$, choosing $g = h(\mu_2/\mu_{2,L})^{1/2}$ reduces the bias to $o(h^2 + g^2)$. In (3.8) and (3.9), the bias order depends on the orders of g and h. For example, when $g = o(h)$, taking $c_1 = 0$ reduces the bias to $o(h^4)$. Thus the general form (3.3) does not seem to be more beneficial than the case of $g = h$, and hence the case of $g \neq h$ for boundary points is not considered further.

4. Local Polynomial Projection and Penalized Trigonometric Series

4.1. Local polynomial projection as mixed models

For a linear mixed model (see, e.g., Fitzmaurice, Laird, and Ware (2004)),

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \begin{pmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{pmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \sigma_\varepsilon^2\mathbf{I}\end{bmatrix}\right),$$

where $\mathbf{b}$ contains fixed effects coefficients for predictors in $\mathbf{X}$, $\mathbf{u}$ the random effects coefficients with corresponding matrix $\mathbf{Z}$, and $\mathbf{G}$ may be taken as $\sigma_u^2\mathbf{I}$ for simplicity. The fitting criterion of a mixed model can be written as

$$\min_{\mathbf{b},\mathbf{u}} \frac{1}{\sigma_\varepsilon^2}||\mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{Z}\mathbf{u}||^2 + \mathbf{u}^T D\mathbf{u}, \tag{4.1}$$

where $D = \mathbf{G}^{-1}$. Minimization in (4.1) involves least squares with a penalty term on the random effects. Penalized splines, e.g. linear splines, correspond to (4.1) with $\mathbf{X}$ consisting of lines $(1, x)$, $\mathbf{Z}$ piecewise linear functions with knots κ_k, $(x - \kappa_k)_+$, where x_+ is x if x is positive and 0 otherwise, and $D = (\lambda^2/\sigma_\varepsilon^2)\mathbf{I}$, where λ^2 is the smoothing parameter. A larger value of λ^2 leads to a smoother fit, and as $\lambda^2 \to \infty$, it forces the coefficients corresponding to all $(x-\kappa_k)_+$ to be 0 and results in a least squares line. Ruppert, Wand, and Carroll (2003) discussed REML estimation of σ_u^2 and σ_ε^2 to obtain an estimate of λ^2.

The connections of $H_p^*\mathbf{y}$ to mixed model estimators is based on the eigen-decomposition of H_p^* and a fact mentioned in Huang and Davidson (2010) that $H_p^*\mathbf{y}$ is the solution to the penalized least squares problem,

$$\min_{\mathbf{f}} (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \mathbf{f}^T(H_p^{*^{-1}} - I)\mathbf{f}, \tag{4.2}$$

where $H_p^{*^{-1}}$ denotes the inverse of H_p^*, and $\mathbf{f}$ denotes some estimate of $(m(X_1), \ldots, m(X_n))^T$ under the constraint that $\mathbf{f}$ lies in the space spanned by H_p^*.

Theorem 3. *Under Conditions* (A1)$-$(A4)*, the penalized least squares form* (4.2) *for fitting local polynomial projection can be expressed in a form of a mixed model criterion* (4.1)*: the corresponding* $\mathbf{X}$ *consists of eigenvectors of* H_p^* *with eigenvalue 1, the corresponding* $\mathbf{Z}$ *consists of eigenvectors of* H_p^* *with eigenvalue* λ_k *such that* $0 < \lambda_k < 1$*, and the corresponding* D *matrix is a diagonal matrix with entries* $(1/\lambda_k - 1)$.

Theorem 3 is based on the eigen-decomposition of H_p^* since H_p^* is symmetric; this approach is similar to Demmler and Reinsch (1975) on the eigen-decomposition for linear smoothing splines. In addition, the Demmler-Reinsch basis functions are closely related to sines and cosines; see Eubank (1999, Sec. 5.2). The proof of Theorem 3 is quite straightforward and is given in the Appendix. Since $H_p^*\mathbf{x}^j = \mathbf{x}^j$, $j = 0, \ldots, p$, the unit eigenvalue is at least of multiplicity $(p+1)$ and the space spanned by $\mathbf{X}$ includes p-th order polynomials (we conjecture that it is possible that the rank of $\mathbf{X} > (p + 1)$). We note that smoothing splines and penalized splines involve polynomials too. The penalty term in Theorem 3 involves non-uniform weights $(1/\lambda_k - 1)$, while penalized splines often adopt uniform weights for convenience. Note that Theorem 3 holds in finite-sample cases. We next explore "asymptotically" the form of the eigenvectors of H_p^* corresponding to $0 < \lambda_k < 1$ and the order of penalty weights in D.

4.2. Penalizing trigonometric series

Theorem 4. *Conditioned on* $\mathbf{x}$*, under Conditions* (A)*, the following hold.*

(a) *The eigenvectors of* H_p^* *are asymptotically the trigonometric polynomials* $\cos(2k\pi\mathbf{x})$ *and* $\sin(2k\pi\mathbf{x})$ *in the sense that*

$$H_p^* \cos(2k\pi\mathbf{x}) = \cos(2k\pi\mathbf{x})(1 + O(k^{2(p+1)}h^{2(p+1)})) \tag{4.3}$$

(*similarly for* $\sin(2k\pi\mathbf{x})$), *where* $k = 1, 2, \ldots,$ *such that* k *satisfies* $kh \to 0$.

(b) *When* $p = 0, 1, 2$, *the constant multiplier for the term* $O(k^{2(p+1)}h^{2(p+1)})$ *in (4.3) is negative, which implies that the corresponding penalty weights (entries of the D matrix in (4.1)) are larger for higher-frequency trigonometric series. The last statement for* $p = 3$ *holds when the kernel function satisfies* $(\mu_8 - \mu_4^2)(\mu_4 - \mu_2^2) > (\mu_6 - \mu_4\mu_2)^2$, *which is true for the Gaussian and Epanechnikov kernels.*

(c) *For those* k *such that* $kh \to$ *constant or* $kh \to \infty$, *the elements in* $H_p^* \cos(2k\pi\mathbf{x})$ *and* $H_p^* \sin(2k\pi\mathbf{x})$ *are asymptotically* 0.

See the Appendix for the proof of Theorem 4. The trigonometric polynomials $\cos(2k\pi\mathbf{x})$ and $\sin(2k\pi\mathbf{x})$ form an orthogonal basis for functions defined on $[0, 1]$ (Condition (A1)). For regression functions defined on an interval, say $[a, a + \tau]$, the corresponding orthogonal basis is $\cos(2k\pi\mathbf{x}/\tau)$ and $\sin(2k\pi\mathbf{x}/\tau)$.

From Theorem 4(a), the number of trigonometric eigenvectors has an upper bound of $O(h^{-1})$. Hence, the number of basis functions tends to infinity as $n \to \infty$. Since the optimal rate of h is $O(n^{-1/5})$ when $p = 1$ (Fan and Gijbels (1996)), we conjecture that the corresponding optimal rate of K is $O(n^{1/(5+\delta)})$ with $0 < \delta < 1$. It was shown in Eubank (1988, Sec. 3.4.2) that the optimal rate of K is $O(n^{1/(5+\delta)})$ for the optimal convergence rate $n^{-4/5}$ of unpenalized Fourier series estimators. From (4.3) and Theorem 4(c), the eigenvalues of H_p^* converge to 1 for those k with $kh \to 0$, or 0 for those k with $kh \to$ *constant* or $kh \to \infty$. Thus the penalty weights for trigonometric eigenvectors with $kh \to 0$ converge to 0 asymptotically. In other words, as $n \to \infty$, the local projection estimates tend to polynomial-trigonometric regression (Graybill (1976); Eubank and Speckman (1990)) with no penalty and the number of trigonometric terms smaller than $O(h^{-1})$. Comparing with penalized splines, Claeskens, Krivobokova, and Opsomer (2009), Kauermann, Krivobokova, and Fahrmeir (2009), and Li and Ruppert (2008) provided asymptotic results when the number of knots as well as the penalty weight go to infinity as $n \to \infty$. Results in Theorem 4(a)(b) imply an adaptive penalty for the trigonometric eigenvectors, which appears to be analogous to Demmler and Reinsch (1975) for the smoothing splines on the oscillation property of eigenvectors and the decreasing trend of the eigenvalues with increasing order.

Theorem 4 raises a practical question: is regularization by selecting the number of trigonometric series K or the penalty term a better way for polynomial-trigonometric regression, though asymptotics leads to selection of K? In the context of Fourier series regression, Droge (1998), suggested that regularizing by K or penalty depends on the magnitude of coefficients ($\mathbf{u}$ in (18)). For large coefficients, selection is superior to penalty, and the opposite holds for small and moderate coefficients. This seems to imply that including both terms is a more adaptive approach for finite-sample cases.

4.3. Extension

Theorems 3 and 4 are extended to the partial linear models

$$Y = \mathbf{t}^T\alpha + m(X) + \varepsilon, \tag{4.4}$$

where $\mathbf{t}$ is a q-dimensional covariate vector, and X is a one-dimensional predictor. Assume that data $(\mathbf{t}_i, X_i, Y_i), i = 1, \ldots, n$, are drawn independently from (4.4), and let $\mathcal{T}_{n\times q}$ be the centered data matrix for the $\mathbf{t}_i$'s and $\mathbf{m} = (m(X_1), \ldots, m(X_n))^\top$. Huang and Davidson (2010) showed that estimators $\hat{\alpha} = (\mathcal{T}^\top(I - H_p^*)\mathcal{T})^{-1}\mathcal{T}^\top(I - H_p^*)\mathbf{y}$ and $\hat{\mathbf{m}} = (\hat{m}(X_1), \ldots, \hat{m}(X_n))^\top = H_p^*(\mathbf{y} - \mathcal{T}\hat{\alpha})$, are the solutions to the penalized least squares equation

$$\min_{\alpha,\mathbf{f}} (\mathbf{y} - \mathcal{T}\alpha - \mathbf{f})^\top(\mathbf{y} - \mathcal{T}\alpha - \mathbf{f}) + \mathbf{f}^\top({H_p^*}^{-1} - I)\mathbf{f}. \tag{4.5}$$

Hence the equivalence of (4.5) to mixed models holds for partial linear models as well.

Similarly, we extend our results to the bivariate additive models

$$Y = \alpha + m_1(X_1) + m_2(X_2) + \varepsilon, \tag{4.6}$$

with constraints $E(m_1(X_1)) = 0$ and $E(m_2(X_2)) = 0$. Smoother matrices H_{p1}^* and H_{p2}^* can be constructed for X_1 and X_2 with bandwidths, h_1 and h_2, respectively. To satisfy the constraints, centered smoothers $\tilde{H}_{p1}^* = (H_{p1}^* - J)$ and $\tilde{H}_{p2}^* = (H_{p2}^* - J)$ are used. Following arguments in Hastie and Tibshirani (1990), it is clear that the solutions to the penalized least squares equation

$$\min_{\alpha,\mathbf{f}_1\mathbf{f}_2} (\mathbf{y} - \alpha - \mathbf{f}_1 - \mathbf{f}_2)^\top(\mathbf{y} - \alpha - \mathbf{f}_1 - \mathbf{f}_2) + \mathbf{f}_1^\top({\tilde{H}_{p1}^*}{}^{-1} - I)\mathbf{f}_1 + \mathbf{f}_2^\top({\tilde{H}_{p2}^*}{}^{-1} - I)\mathbf{f}_2 \tag{4.7}$$

correspond to the solutions by backfitting with smoother matrices $\tilde{H}_{p1}^*$ and $\tilde{H}_{p2}^*$. The eigenvectors of centered smoothers are the same as the un-centered version except that the intercept vector is not included. As $\tilde{H}_{p1}^*$ and $\tilde{H}_{p2}^*$ are one-dimensional smoothers, (4.7) can be written in the form of mixed models and the results in Theorems 3 and 4 continue to apply to $\tilde{H}_{p1}^*$ and $\tilde{H}_{p2}^*$. The bivariate case is easily extended to general multiple-component additive models. Hence for additive models, Theorems 3 and 4 not only provide a penalized framework but also connect to backfitting estimators using one-dimensional centered smoothers. This last interpretation appears to be new and has not been discussed in penalized splines for additive models. Numerical performance of backfitting estimators with $\tilde{H}_p^*$ will be pursued in a future paper.

Theorems 3 and 4 may be applicable to time series data. Consider the trend-plus-noise model: $Y_{t_i} = m(t_i) + \epsilon_{t_i}, i = 1, \ldots, n$ where the t_i's are equally spaced, $m(\cdot)$ is a smooth function, and $\{\epsilon_{t_i}, i = 1, \ldots, n\}$ is a correlated process of zero mean and covariance matrix $\Sigma(\theta)$ parametrized by θ. Since H_p^* is asymptotically Toeplitz, its (j, k)-th element is asymptotically some deterministic function of $|j-k|$, say $c_{|j-k|}$. (Consequently, the sequence $\{c_j, j = 0, 1, 2, \ldots\}$ is implicitly defined.) Let $\gamma(x) = c_0 + 2\sum_{k=1}^{n} c_k \cos(kx)$, $-\pi \le x \le \pi$. It follows from Grenander and Rosenblatt (1957, pp.103-105) that, under some regularity conditions, the distribution of the eigenvalues of H_p^* is approximately equal to that of $\gamma(U)$ where U has a uniform distribution over the interval $(-\pi, \pi)$. Theorems 3 and 4 then allow us to treat the above trend-plus-noise model as a mixed effects model with two random effects, one given by trigonometric basis functions with known asymptotic variance and the other implied by the covariance structure of the error process. See Proietti (2007) for a review of signal extraction with time series data via the mixed effects framework. This approach will be further investigated elsewhere.

Theorem 4 shows that the Fourier basis is "asymptotically" the penalizing basis for local polynomial projection. The Fourier basis has a long history in applied mathematics, and in the smoothing literature, there are estimators and hypothesis tests based on Fourier series, e.g., Eubank (1988), but only in the fixed effects part.

Figure 1 illustrates Theorem 4. We simulated 1,000 samples of $n = 100$ data points from the Uniform$(0, 1)$. Using the Epanechnikov kernel and $h = 0.2$, the H_1^* matrix based on local linear regression was obtained for each sample and the eigenvalues and eigenvectors of H_1^* were examined. Among the 1,000 H_1^*'s, all had a unit eigenvalue of multiplicity 2, 941 samples had 7, and 59 samples had 6 eigenvalues in the range $[0.05, 1)$. Ordering the eigenvalues and excluding the unit eigenvalue, the mean (standard deviation) of the third to ninth eigenvalues were 0.9913 (0.0012), 0.9400 (0.0063), 0.8067 (0.0194), 0.5978 (0.0300), 0.3665 (0.0312), 0.1797 (0.0216), and 0.0665 (0.0101). The averages of the first nine eigenvectors (including those two corresponding to the unit eigenvalue) are plotted in Figure 1, where the average was calculated element by element for each eigenvector. The first two eigenvectors look like lines, and the rest look like cosine and sine functions.

Though Theorem 3 holds in finite-sample cases and Theorem 4 holds asymptotically, these results may be combined to develop a new smoothing approach

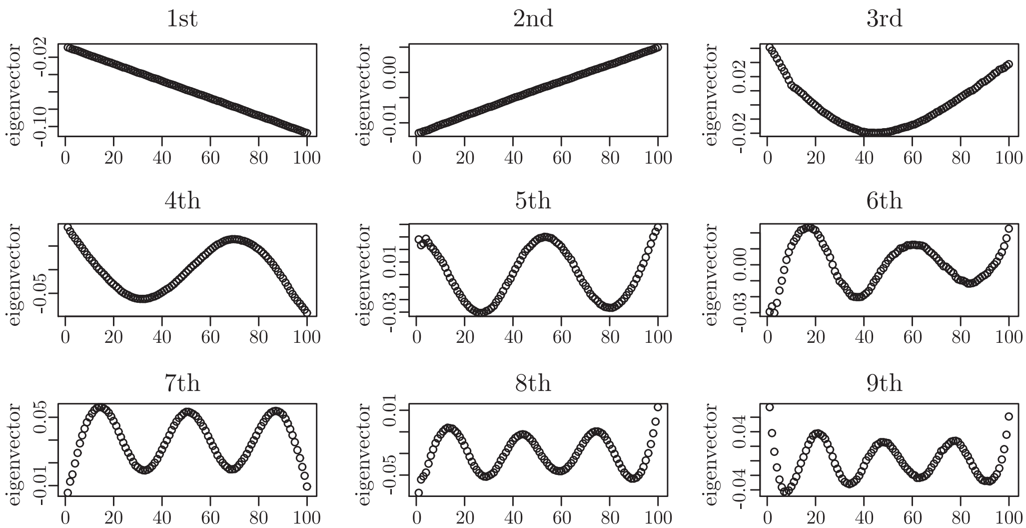

Figure 1. Plot of the averages of the first 9 eigenvectors of H_1^*'s using the Epanechnikov kernel and $h = 0.2$ based on 1,000 samples of $n = 100$ data points from Uniform$(0, 1)$.

based on combinations of unpenalized polynomials and penalized trigonometric basis. This approach is analogous to pseudosplines (Hastie (1996)) in which the original high-rank smoother matrix from smoothing splines is replaced by a low-rank approximation. In other words, the high-rank H_p^* matrix may be replaced by a low-rank polynomials and trigonometric basis approximation, as will be illustrated in Section 5.

5. Examples Using Penalized Trigonometric Series

In this section, some numerical examples illustrate smoothing with a penalized trigonometric basis and unpenalized lines $(1, x)$, abbreviated as penalized trigonometric series. Examples 1 through 3 demonstrate penalized trigonometric series on data sets for univariate smoothing, partial linear models, and generalized additive models, respectively. Example 4 concerns a small simulation study to confirm the results in Theorem 4 numerically and for comparison with penalized splines as implemented by Wood (2006). Examples 5 and 6 explore using penalized trigonometric series for more sophisticated data applications in varying coefficient models (Hastie and Tibshirani (1993)) and semiparametric mixed models, respectively, while their theoretical connections will be studied in future research.

For data examples, we used available mixed model software for penalized trigonometric series as described in Ruppert, Wand, and Carroll (2003) for penalized linear splines except that spline functions were replaced by up to K cosine and K sine functions in the $\mathbf{Z}$ matrix, and the penalty was constrained so that $\cos(2k\pi\mathbf{x})$ and $\sin(2k\pi\mathbf{x})$ shared the same penalty parameter at the same k, but with different penalty at different k. For multiple smoothed functions such as additive models with $\mathbf{x}_1$ and $\mathbf{x}_2$, the penalty was allowed to be different between Fourier functions of $\mathbf{x}_1$ and $\mathbf{x}_2$. The REML method from fitting mixed models was used to obtain the penalty term. For exploration purpose only, the AICc (equation (2.5) of Hurvich, Simonoff, and Tsai (1998)) criterion was adopted for selecting K, $K = 1, \ldots, Kmax$, where $Kmax$ was set arbitrarily as 15 except in Example 3, $Kmax = 9$ due to its small sample size. For model parsimony, we chose the smallest K such that the corresponding AICc differed from the minimum of AICc by less than 1% of the minimum value.

In Examples 1$-$4, we also took a penalty form αk^2 at frequency k for computational convenience since it does not require nonlinear optimizations in REML. The quadratic penalty form αk^2 appears in the representation of linear smoothing splines by the Demmler-Reinsch basis, which is a damped (weighted) cosine series estimator (Eubank (1988, p.234)). Again, the AICc criterion was used for selecting α and K. First, α was determined by where the minimum of AICc occurred, and then the smallest K such that the corresponding AICc differed from the minimum of AICc by less than 1% of the minimum value was chosen. A detailed investigation on the choice of K and penalty will be pursued in the future.

1. *Univariate Smoothing.* For model (2.1), the intercept column and the X variable are included in the fixed effects covariate matrix $\mathbf{X}$. The first example is based on the LIDAR (light detection and ranging) data from Ruppert, Wand, and Carroll (2003). The independent variable, *range*, is the distance traveled before the light is reflected back to its source. The dependent variable, *logratio*, is the logarithm of ratio of received light from two laser sources. The data is of size 221. The trigonometric basis is re-scaled as $\cos(2k\pi \times range/330)$ and $\sin(2k\pi \times range/330)$ since the *range* variable takes on integer values from 390 to 720.

For each $K = 1, \ldots 15$, the REML criterion in mixed model software was used to automatically select the penalty parameters, and the AICc was calculated with the degrees of freedom (d.f.) 2 (for the linear term) plus the sum of $1/(1+penalty)$ based on Theorem 3. We observe that the AICc decreased when $K \leq 5$ and then slightly increased in the 4th decimal place. In this example, when $K \geq 5$, REML selected large penalty parameters for high-frequency functions and the d.f. stayed around 2.885 for $K = 5$ to 14. $K = 3$ was selected with d.f. 2.864 and the resulting estimate is shown in Figure 2(a) (solid line), while the penalty parameter at k in log 10 scale is plotted in Figure 2(b). We see that the penalty is monotonically increasing as k increases.

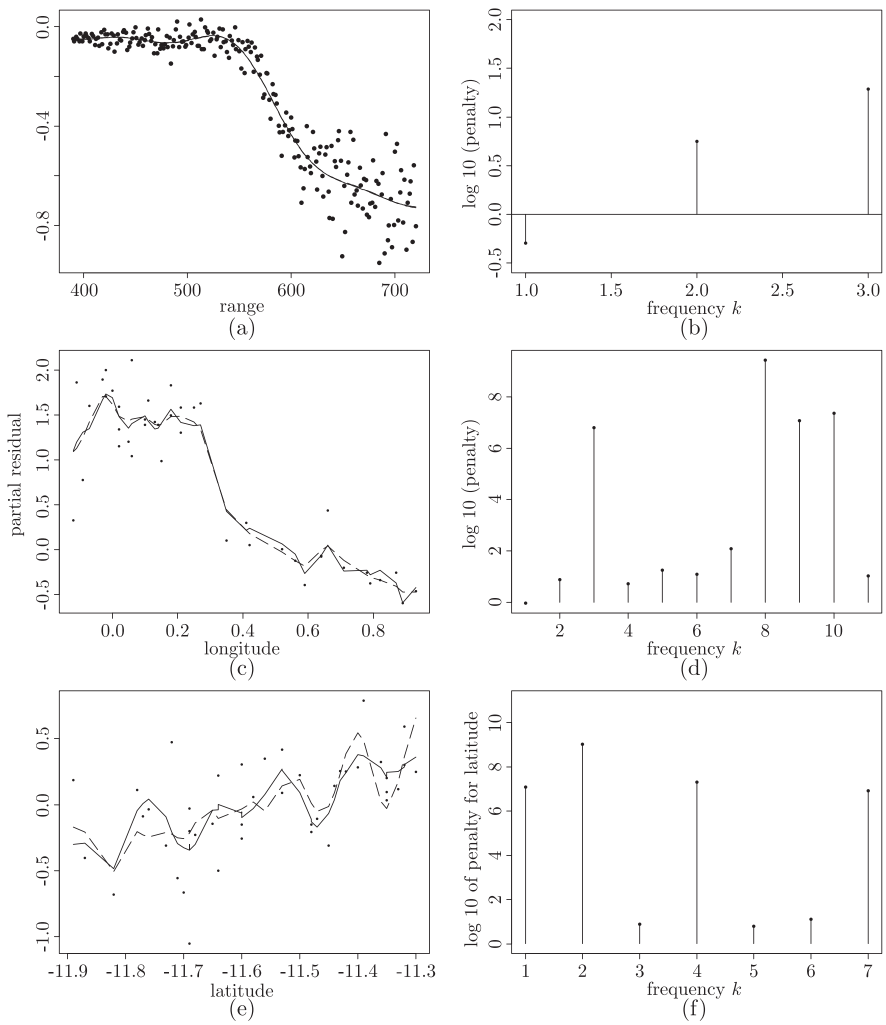

Figure 2. (a) and (b) LIDAR data: (a) Smoothing fit using $K = 3$ penalized trigonometric functions with automatic REML selection of penalty (solid line), fit by penalty$=1.389 \times k^2$ with $K = 3$ (dashed line). (b) The values of the penalty in logarithm scale by REML when $K = 3$. (c)$-$(f) Great Barrier Reef data: (c) fit by a partial linear model with nonlinear functions of longitude estimated by using $K = 11$ penalized trigonometric functions with penalty by REML (solid line) and by specifying penalty as $0.720 \times k^2$ with $K = 11$ (dashed line). (d) the values of the penalty by REML in logarithm scale. (e) fit by an additive model with nonlinear functions of latitude estimated by using $K = 7$ penalized trigonometric functions with penalty by REML (solid line) and $K = 9$ with both penalty $0.518 \times k^2$ (dashed line). The penalty by REML in logarithm scale is shown in (f) for latitude.

For the second penalty form αk^2, the AICc was calculated based on a grid of α that consisted of 50 logarithmically equally-spaced points in the interval $[0.01, 10^5]$ with $K = 1, \ldots, 15$. The smallest AICc occurred when $\alpha = 1.389$ and with this α, $K = 3$ was selected with d.f. 3.290. The resulting fit is given in Figure 2(a) (dashed line). The two curves are visually indistinguishable.

2. *Partial Linear Models.* In fitting (4.4), the only difference from fitting (2.1) using mixed model software is that in the fixed effects matrix $\mathbf{X}$, in addition to the intercept column and the X variable, the parametric covariates $\mathbf{t}$ are also included. We use the Great Barrier Reef data from Bowman and Azzalini (1997) to illustrate fitting (4.4) with penalized trigonometric series. The data is derived from a survey of the fauna on the sea bed lying between the coast of northern Queensland and the Great Barrier Reef. The response variable is a score, which combines information across species, and the explanatory variables are latitude and longitude. Bowman and Azzalini (1997) suggested a partial linear model for the data, with a linear effect of latitude and a nonparametric term of longitude. The data is of size 42 if restricted to year 1992 and closed zones (where commercial fishing is not allowed).

For smoothing on longitude with automatic selection of the penalty by REML, the AICc was calculated at $K = 1, \ldots 15$, and $K = 11$ was selected with d.f. 4.022. Figures 2(c) shows the marginal partial residual plot of the nonlinear estimate in longitude plus the estimated intercept (solid line), and the amount of penalty by REML is given in Figure 2(d). It is evident that the curve connecting fitted values is a bit jagged and the penalty terms at $k = 3, 8, 9$, and 10 are quite large. We also tried specifying the penalty as αk^2 and the minimum AICc occurred at $\alpha = 0.720$. With this α, $K = 11$ was selected with d.f. 6.032 (dashed line in Figure 2(c)). At some places, the dashed curve is a bit smoother than the solid curve. Though both fits chose $K = 11$, the d.f. differs due to the different penalty forms.

3. *Additive Models.* The Great Barrier Reef data was used again to illustrate fitting a bivariate additive model (4.6). Assuming the same K for $m_1(X_1)$ and $m_2(X_2)$, the AICc with REML selected $K = 7$ with a total d.f. 4.240. The estimate for longitude was similar to the curves by partial linear models and hence is omitted. The curve for latitude is shown in Figure 2(e) (solid line) with its corresponding penalty in Figure 2(f). There is some oscillating trend in latitude and the curve is not smooth. We then tried penalty αk^2 and AICc chose $\alpha = 0.518$ and $K = 9$ with a total d.f. 8.927 (dashed line in Figure 2(e)).

In the literature, REML tends to undersmooth for smoothing splines (Wahba (1985)) and oversmooth for penalized splines (Ruppert, Wand, and Carroll (2003)). From Examples 1-3, REML for penalty with AICc for K seems to undersmooth in Examples 2 and 3, while providing a satisfactory fit in Example 1. We conjecture that the behavior in Examples 2 and 3 could be due to the small sample size of $n = 42$. Some further work is needed to investigate the performance of REML and AICc in both theory and practice.

4. *A Simulated Example.* Data of size $n = 100$ were simulated from

$$Y = X + 2\exp(-16X^2) + \varepsilon,$$

with $X \sim$ Uniform $[-2, 2]$, $\varepsilon \sim N(0, .4^2)$, and X and ε independent. This example is taken from Fan and Gijbels (1996) and the number of replications is 500. Each set of data was smoothed by local linear regression using the Epanechnikov kernel and a fixed bandwidth 0.35, and then by penalized trigonometric series with penalty αk^2 based on a choice of $K = 1, \ldots, 15$ and a grid of α as in Example 1. To confirm Theorem 4 numerically, α was chosen so that with this α, one of the d.f.'s at $K = 1, \ldots, 15$, was the closest to that of local linear regression, and then the smallest K was selected that had a difference < 0.1 in d.f. from local linear regression. The d.f.'s were quite close between the two approaches, and the sum of the squared differences of the two fitted values at the 100 data points was calculated. Based on 500 simulations, the mean (sd) on the d.f.'s was 12.552 (0.173) and 12.572 (0.184) for local linear and penalized trigonometric series respectively, and the sum of squared differences had a mean of 0.414 and sd 0.205. On average, the difference $\sqrt{0.414/100} = 0.0643$ was quite small relative to the mean of Y, which had a range of $[-2, 2.007]$. The results provide numerical evidence in support of Theorem 4 in finite-sample cases.

Based on the same 500 sets of simulated data, we investigated how the choice of α and K affect the d.f. and AICc. Figure 3 plots α vs. the average d.f. at different K. As expected, when α is small, the d.f. generally increases as K increases and when α is large, the d.f. stays about the same no matter what K is. Figure 4 shows α vs. the average AICc at different K. It appears that, for a fixed K, there is an α that minimizes AICc, and the minimum among $K = 1, \ldots, 15$, occurs at $\alpha = 0.139$ and $K = 7$. When α is large, AICc stays about the same no matter what K is. Figure 5 plots K vs. the average AICc at 4 values of α. We observe that there is a value of K that minimizes AICc when α is not too large. From Figures 3−5, we conclude that for the penalty form αk^2, the choice of K is not sensitive while the choice of α is more important.

Using the same data sets, we compared the performance of penalized trigonometric series to penalized splines, with the latter based on Wood (2006). The default setting of penalized splines includes GCV smoothness selection (Wahba (1990)). For penalized trigonometric series, the penalty αk^2 was adopted and the choice of α and K was based on the same approach as in Examples 1−3.

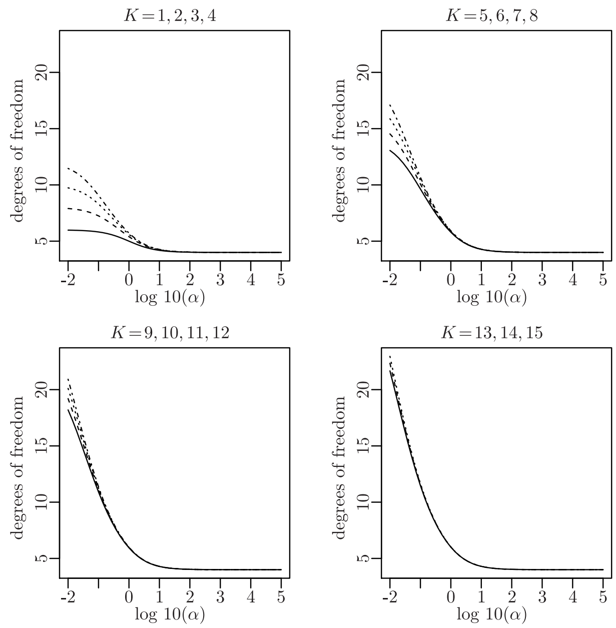

Figure 3. Plot of α vs. the average d.f. in Example 4: for K=1, 5, 9, and 13 (solid line); 2, 6, 10, and 14 (dashed line); 3, 7, 11, and 15 (dotted line); 4, 8, and 12 (dash-dotted line).

Table 1. Example 4, comparison of penalized trigonometric series to penalized splines.

	penalized trigonometric series			penalized splines		
	Q1	mean (sd)	Q3	Q1	mean (sd)	Q3
d.f.	8.368	9.310 (1.299)	10.070	9.234	9.353 (0.339)	9.571
SSE	1.790	2.408 (0.845)	2.891	3.425	4.097 (0.919)	4.612
AICc	-1.716	-1.623 (0.143)	-1.526	-1.547	-1.451 (0.151)	-1.350
GCV	0.157	0.174 (0.0256)	0.190	0.185	0.207 (0.0313)	0.226

The mean (sd) of 500 selected K's was 6.328 (1.571) with a minimum of 4 and maximum 13, while that of 500 selected α's was 0.163 (0.067) with a minimum of 0.0373 and maximum 0.517. The results are summarized in Table 1. The sum of squared errors (SSE) was calculated as $\sum_{i=1}^{100}(\hat{m}(X_i) - m(X_i))^2$, and Q1 and Q3 are the 25- and 75-percentiles among 500 values. The mean d.f. of the two approaches were close, while the sd of penalized trigonometric series was larger. Table 1 shows that the two approaches are comparable in terms of the criteria SSE, AICc, and GCV, and the penalized trigonometric series approach slightly outperforms the penalized spline approach, in terms of having smaller mean values of the preceding three criteria.

Next we explore using penalized trigonometric series for more sophisticated

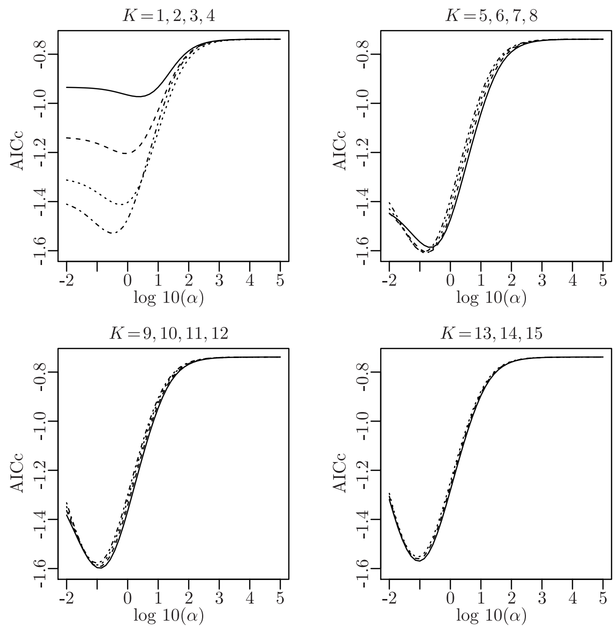

Figure 4. Plot of α vs. the average AICc in Example 4: for K=1, 5, 9, and 13 (solid line); 2, 6, 10, and 14 (dashed line); 3, 7, 11, and 15 (dotted line); 4, 8, and 12 (dash-dotted line).

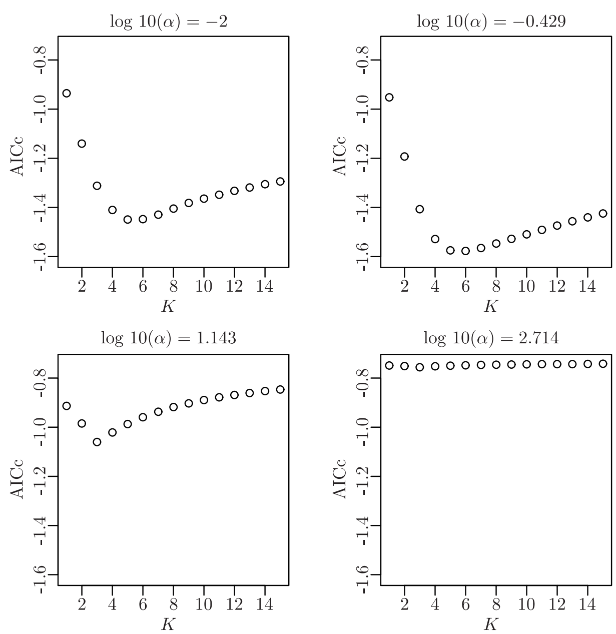

Figure 5. Plot of K vs. the average AICc in Example 4 at 4 values of α.

applications in Examples 5 and 6, while their theoretical connections will be explored in future research.

5. *Varying Coefficient Models.* The varying coefficient model is a regression model which is additive in the regressors, but the relationship between each regressor and the outcome is allowed to vary as a smooth function of an additional, effect modifying variable Z (Hastie and Tibshirani (1993)):

$$Y = \alpha(Z) + m(Z) \times X + \varepsilon. \tag{5.1}$$

Following Ruppert, Wand, and Carroll (2003), we use the ethanol data set for illustration. The data were collected to analyze oxides of nitrogen in automobile exhaust. An experiment was done on a one-cylinder engine fueled by ethanol. Two engine factors were studied: the equivalence ratio, a measure of the richness of the air and fuel mixture, and the compression ratio to which the engine is set. Here we examine the relationship between the oxides of nitrogen and the compression ratio with the equivalence ratio as a modifying variable. AICc chooses $K = 10$ trigonometric functions with REML and the curves with a total d.f. 5.582 are shown in Figure 6(a)(b). The slope of the compression ratio varies at different levels of the equivalence ratio, which hints at interactions between the equivalence ratio and the compression ratio. The respective penalty weights are given in Figure 6(c)(d).

6. Semiparametric Mixed Models. The FEV data from Fitzmaurice, Laird, and Ware (2004) is used to illustrate fitting semiparametric mixed models. The data is a subset of data from the Six Cities Study of Air Pollution and Health (Dockery et al. (1983)) with 13,379 children. The study was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents, and the factors that influence lung function growth including age and height. The subset consists of 300 girls, with a minimum of one and a maximum of twelve observations over time. The outcome FEV1 was log-transformed. A semiparametric mixed model is

$$\log(FEV1_{ij}) = \beta_0 + U_i + \beta_1 height_{ij} + m(age_{ij}) + \varepsilon_{ij}, \quad 1 \le j \le n_i, 1 \le i \le 300, \tag{5.2}$$

where U_i's are i.i.d. random intercepts. The AICc chooses $K = 1$ trigonometric function with REML and the curve estimate for age is shown in Figures 6(e). The total d.f. is 4.478 and the penalty by REML is 20.323. The mean response of age shows a slightly nonlinear shape.

6. Concluding Remarks

This paper presents asymptotic properties of local polynomial projection estimates and connections between local polynomial projection, mixed models, and

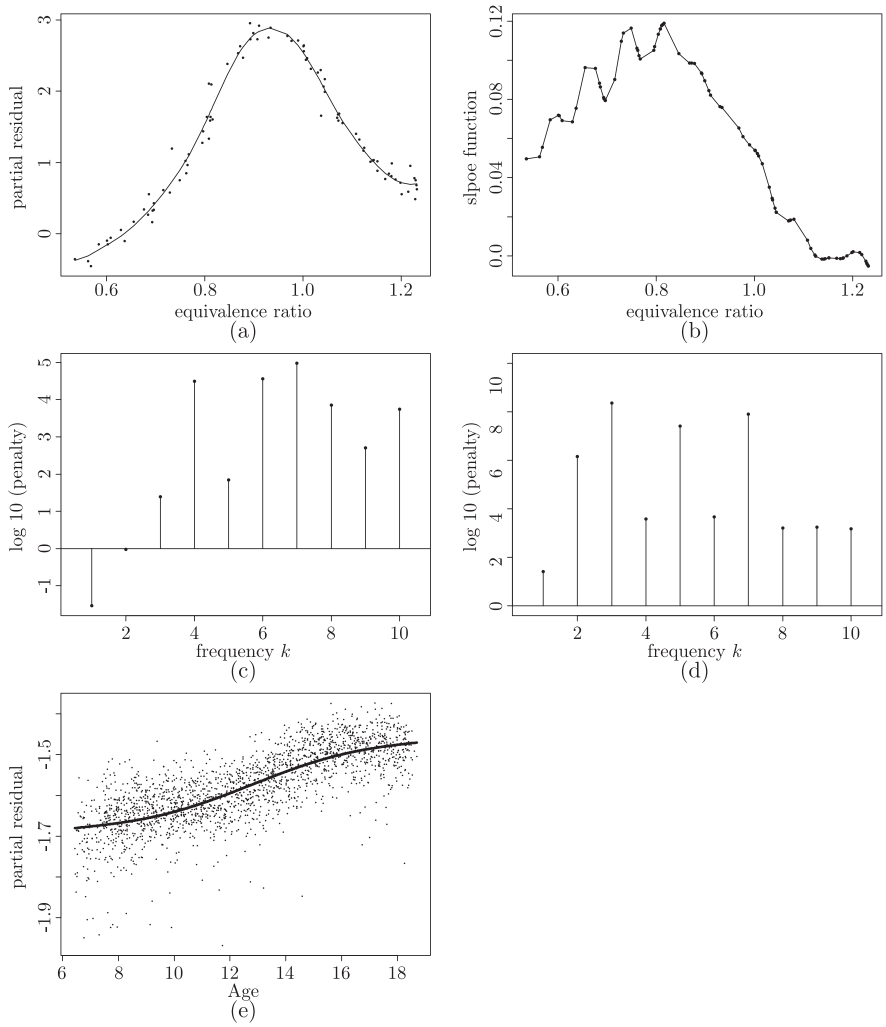

Figure 6. (a)-(d): A varying coefficient model for the ethanol data. Estimates are obtained by using $K = 10$ penalized trigonometric functions with penalty by REML. (a) The nonlinear function of the equivalence ratio; (b) slope function for the compression ratio which varies with the equivalence ratio; (c) and (d) are the values of the penalty applied respectively to estimates in (a) and (b) in logarithm scale. (e) A semiparametric mixed model for the FEV data: the nonlinear function of age (solid line) estimated by using $K = 1$ penalized trigonometric functions with penalty by REML.

smoothing by combinations of unpenalized polynomials and penalized trigonometric series. Results in Theorem 1 indicate that it is no longer an odd world when the whole fitted local polynomials are utilized to form local projection estimates. With the Fourier series widely applied in many fields, it is expected that smoothing with a penalized Fourier basis will bring up many interesting problems and applications. We are exploring if the connections hold in the settings of varying coefficient models, mixed models, and generalized semiparametric models, and if the projection framework of LPR continues to help establish the equivalence. Results in Section 3 of this paper are based on a global bandwidth and how to generalize the results for the case of varying h is a problem for future research. We have explored using the AICc criterion to select K and/or α in this paper, and further investigation is needed. Strictly speaking, local polynomial projection is a special case of general penalized series estimators; Theorems 3 and 4 show the equivalence with specified $\mathbf{X}$, $\mathbf{Z}$, and penalty, not for any choices of basis $\mathbf{Z}$ and penalty. Properties under a general mixed effects framework with penalized series remain to be investigated. The extension to spatial smoothing is the focus of ongoing research.

Acknowledgement

We are grateful to an associate editor and two referees for their helpful comments that substantially improved the presentation of the paper. Huang's research was partially supported by an NTHU faculty research grant 99N82508E1 and by National Science Council (NSC 100-2118-M-007 -001). KSC would like to thank the US National Science Foundation (NSF 0934617) and the US National Institute of Health (NIH R01 HL112986-01A1) for partial support.

References

Berlinet, A. (1993). Hierarchies of higher order kernels. *Probab. Theory Related Fields* **94**, 489-504.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis.* Oxford University Press, Oxford.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453-555.

Claeskens, G., Krivobokova, T. and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* **96**, 529-544.

Cohen, A. (1966). All admissible linear estimates of the mean vector. *Ann. Math. Statist.* **37**, 458-463.

Demmler, A. and Reinsch, C. (1975). Oscillation matrices and spline smoothing. *Numer. Math.* **24**, 375-382.

Dette, H. and Melas, V. B. (2003). Optimal designs for estimating individual coefficients in Fourier regression models. *Ann. Statist.* **31**, 1669-1692.

Dockery, D. W., Berkey, C. S., Ware, J. H., Speizer, F. E. and Ferris, B. G. (1983). Distribution of FVC and FEV1 in children 6 to 11 years old. *Am. Rev. Respir. Dis.* **128**, 405-412.

Droge, B. (1998). Minimax regret analysis of orthogonal series regression estimation: Selection versus shrinkage. *Biometrika* **85**, 631-643.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* Marcel Dekker, New York.

Eubank, R. L. and Speckman, P. (1990). Curve fitting by polynomial-trigonometric regression. *Biometrika* **77**, 1-9.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis.* Wiley, New York.

Gasser, T., Müller, H.-G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47**, 238-252.

Graybill, F. (1976). *Theory and Application of the Linear Model.* Duxbury, North Scituate.

Grenander, U. and Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series.* Wiley, New York.

Hastie, T. J. (1996). Pseudosplines. *J. Roy. Statist. Soc. Ser. B* **58**, 379-396.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

He, H. and Huang, L.-S. (2009). Double-smoothing for bias reduction in local linear regression. *J. Statist. Plann. Inference* **139**, 1056-1072.

Huang, L.-S. and Chen, J. (2008). Analysis of Variance, Coefficient of Determination, and F-test for Local Polynomial Regression. *Ann. Statist.* **36**, 2085-2109.

Huang, L.-S. and Davidson, P. W. (2010). Analysis of variance and F-tests for partial linear models with applications to environmental health data. *J. Amer. Statist. Assoc.* **105**, 991-1004.

Hurvich, C. M, Simonoff, J. S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. Ser. B* **60**, 271-293.

Kauermann, G., Krivobokova, T. and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *J. Roy. Statist. Soc. Ser. B* **71**, 487-503.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika* **95**, 415-436.

Linton, O. B. and Jacho-Chávez, D. T. (2010). On internally corrected and symmetrized kernel estimators for nonparametric regression. *Test* **19**, 166-186.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Springer Verlag, New York.

Proietti, T. (2007). Signal extraction and filtering by linear semiparametric methods. *Comput. Statist. Data Anal.* **52**, 935-958.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378-1402.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Wand, M. P. and Schucany, W. R. (1990). Gaussian-based kernels. *Canad. J. Statist.* **18**, 197-204.

Wang, Y. (2011). *Smoothing Splines.* CRC Press, Boca Raton.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* CRC Press, Boca Raton.

Institute of Statistics, National Tsing Hua University, No.101 Kuang-Fu Road, Section 2, Hsinchu City, TAIWAN 30013.

E-mail: lhuang@stat.nthu.edu.tw

Department of Statistics and Actuarial Science, University of Iowa, 241 Schaeffer Hall, Iowa City, IA 52242-1409, USA.

E-mail: kchan@stat.uiowa.edu

(Received February 2012; accepted July 2013)