Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

Consistent Model Selection for Marginal Generalized Additive Model for Correlated Data

Lan XUE, Annie QU, and Jianhui ZHOU

We consider the generalized additive model when responses from the same cluster are correlated. Incorporating correlation in the estimation of nonparametric components for the generalized additive model is important because it improves estimation efficiency and increases statistical power for model selection. In our setting, there is no specified likelihood function for the generalized additive model, because the outcomes could be nonnormal and discrete, which makes estimation and model selection very challenging problems. We propose consistent estimation and model selection that incorporate the correlation structure. We establish an asymptotic property with L_2-norm consistency for the nonparametric components, which achieves the optimal rate of convergence. In addition, the proposed model selection strategy is able to select the correct generalized additive model consistently. That is, with probability approaching to 1, the estimators for the zero function components converge to 0 almost surely. We illustrate our method using numerical studies with both continuous and binary responses, along with a real data application of binary periodontal data.

Supplemental materials including technical details are available online.

KEY WORDS: L_2-norm consistency; Model selection; Nonparametric function; Oracle property; Polynomial spline; Quadratic inference function; SCAD.

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

1. INTRODUCTION

Longitudinal or correlated data occur very frequently in biomedical studies. The challenge in analyzing correlated data is that the likelihood function is difficult to specify or formulate for nonnormal responses with large cluster size. To allow richer and more flexible model structures, the generalized additive model (Hastie and Tibshirani 1990) takes the additive form of nonparametric functions. This is appealing for both model interpretation and prediction. However, when some of the predictor variables are redundant, traditional estimation methods for the generalized additive model (Stone 1985; Linton and Härdle 1996; Horowitz and Mammen 2007) are unable to produce an accurate and efficient estimator. Because even a single predictor variable is typically associated with a large number of unknown parameters in the nonparametric functions, inclusion of redundant variables can hinder accuracy and efficiency for both estimation and inference. Therefore, variable selection is a crucial step in generalized additive modeling.

In general, even when the data are not correlated, variable selection with discrete responses is quite challenging, because it involves numerical approximation and can be computationally intensive (Meier, van de Geer, and Bühlmann 2008). Variable selection for the generalized additive model is more challenging, because the dimension for the nonparametric components could be infinite. Huang, Horowitz, and Wei (2010) proposed additive model selection using B-spline bases when the number of variables and additive components is larger than the sample size. Meier, van de Geer, and Bühlmann (2009) developed estimation procedures for high-dimensional generalized additive models by penalizing the sparse terms. Other approaches to variable selection in additive models include those of Huang and Yang (2004) and Xue (2009). However, these approaches either require the specification of the likelihood function or target mainly continuous outcomes using least squares. In general, these methods cannot handle correlated discrete types of outcomes in model selection and estimation, because there is no closed form of the likelihood function and/or it is practically infeasible to perform numerical approximation for infinite-dimensional problems. In this article we are interested in developing a general framework for estimation and model selection for generalized additive models that can handle correlated categorical responses in addition to continuous ones.

Furthermore, the current literature on incorporating correlation for the generalized additive model is rather limited. This is mainly because nonparametric modeling can be very computationally intensive and introduces high-dimensional nuisance parameters associated with nonparametric forms. This makes it difficult to incorporate additional correlation structure into the model. However, ignoring correlation could lead to inefficient estimation and impair statistical power for the selection of correct models. In addition, for smoothing spline and seemingly unrelated kernel approaches (Wang 2003), ignoring the correlation also could lead to biased estimation (Zhu, Fung, and He 2008). Specifically, Wang (2003) reported that selection of the smoothing parameter could fail, because it is rather sensitive for even a small departure from the true correlation structure, which is reflected by overfitting the nonparametric estimator to reduce the overall bias. In contrast to the parametric setting, these problems could be more serious for the nonparametric generalized additive model, because the true model might not be easily verified.

To incorporate correlation, Berhane and Tibshirani (1998) proposed estimating the generalized additive model using a

Lan Xue is Assistant Professor, Department of Statistics, Oregon State University, Corvallis, OR 97331. Annie Qu is Associate Professor, Department of Statistics, University of Illinois at Urbana–Champaign, Champaign, IL 61820 (E-mail: *anniequ@illinois.edu*). Jianhui Zhou is Assistant Professor, Department of Statistics, University of Virginia, Charlottesville, VA 22904. Xue's research was supported by the National Science Foundation (DMS-0906739). Qu's research was supported by the National Science Foundation (DMS-0906660). Zhou's research was supported by the National Science Foundation (DMS-0906665). The authors thank Dr. Julie Stoner of University of Oklahoma for providing the periodontal disease data for the data analysis. They also thank two referees, an associate editor, and the editor for constructive suggestions that substantially improved the article.

© 2010 American Statistical Association
Journal of the American Statistical Association
December 2010, Vol. 105, No. 492, Theory and Methods
DOI: 10.1198/jasa.2010.tm10128

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

smoothing spline generalized estimating equation (GEE) approach (Liang and Zege 1986). This approach does not address the variable selection problem, however. Wang, Li, and Huang (2008) proposed a nonparametric varying coefficient model for longitudinal data, but that model is applicable mainly for continuous outcomes.

The quadratic inference function (QIF) approach (Qu, Lindsay, and Li 2000) can efficiently take the within-cluster correlation into account and is more efficient than the GEE approach when the working correlation is misspecified. In this article we propose consistent estimation and variable selection based on the penalized quadratic inference function for the generalized additive model for correlated data. The proposed method is able to shrink small coefficients of the nonparametric functional components to 0. In addition, it effectively takes the within-cluster correlation into consideration and provides efficient estimators for the nonzero components.

We provide an asymptotic consistency property for nonparametric function estimation, and establish the optimal rate of convergence. Huang (1998) and Xue and Yang (2006) established asymptotic theory for additive models with polynomial splines when the responses are continuous and independent. However, the asymptotic property has not been well studied in the literature for the correlated generalized additive model using polynomial splines. The theoretical technique for the spline approach is very different from the kernel approach (Horowitz 2001) for the generalized additive model. In addition, because the objective function that we minimize is a penalized quadratic distance function, the theoretical proof is very different from the penalized least squares approach (Xue 2009). The techniques involved in asymptotic property development are quite challenging given the curse of the dimensionality of nonparametric functions, the nonlinear relationship between the response and covariates, and the correlated nature of measurements.

The advantage of our approach is that model selection is accomplished through SCAD penalization (Fan and Li 2001), and it achieves the oracle property. That is, the proposed method is able to select the correct generalized additive model consistently, and the estimators of the nonzero components achieve the same rate of L_2 convergence as if the true model is known in advance. In addition, because in our approach the nonparametric function forms are selected groupwise for each functional component, the choice of the basis functions of the polynomial spline is not critical to the final model selection. In addition, our numerical studies show that our approach performs extremely well in estimation efficiency and model selection when the dimension of functional components is high.

The article is organized as follows. Section 2 describes a marginal framework for the generalized additive model for correlated data. Section 3 introduces a penalized quadratic inference function method for simultaneous estimation and variable selection of marginal generalized additive models. Asymptotic theories are developed and issues for implementation are discussed. Section 4 illustrates simulation studies for continuous and binary responses. Section 5 demonstrates model selection and estimation for binary periodontal data. Section 6 provides concluding remarks and discussion. The proofs are given in the Appendix, and more technical details are available in the online supplemental files.

2. A MARGINAL GENERALIZED ADDITIVE MODEL FOR LONGITUDINAL DATA

Suppose that there are n clusters with the ith ($i = 1, \ldots, n$) cluster containing m_i observations. Let Y_{ij} be a response variable, and $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(d)})^T$ be a $d \times 1$ vector of covariates for the jth ($j = 1, \ldots, m_i$) observation in the ith cluster. Write

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{pmatrix}, \qquad \mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{i1}^T \\ \vdots \\ \mathbf{X}_{im_i}^T \end{pmatrix}.$$

Then the basic assumption of the marginal generalized additive model is that observations from different clusters are independent. The first two moments of Y_{ij} are specified as $E(Y_{ij}|\mathbf{X}_{ij}, \mathbf{X}_i) = E(Y_{ij}|\mathbf{X}_{ij}) = \mu_{ij}^0$, and $\mathrm{Var}(Y_{ij}|\mathbf{X}_{ij}, \mathbf{X}_i) = \mathrm{Var}(Y_{ij}|\mathbf{X}_{ij}) = \phi V(\mu_{ij}^0)$, where ϕ is a scale parameter and $V(\cdot)$ is a known variance function, and the marginal mean μ_{ij}^0 relates to the covariates through the known link function $g(\cdot)$,

$$\begin{aligned} \eta_{ij}^0 &= g(\mu_{ij}^0) = \alpha_0 + \sum_{l=1}^{d} \alpha_l\big(x_{ij}^{(l)}\big), \\ \mu_{ij}^0 &= g^{-1}(\eta_{ij}^0). \end{aligned} \tag{1}$$

Here α_0 is an unknown constant, and $\{\alpha_l(\cdot)\}_{l=1}^d$ are unknown smooth functions. Because the estimation of nonparametric functions is usually obtained on a compact support, we assume without loss of generality that the covariates $X_{ij}^{(l)}$ can be scaled into the interval $[0, 1]$, for $1 \le l \le d$. For identification of model (1), without loss of generality, we assume that each $\alpha_l(\cdot)$ is centered with $\int_0^1 \alpha_l(x)\,dx = 0$.

In model (1), the additive nonparametric functions are formulated to model the covariate effects. If each $\alpha_l(\cdot)$ is a linear function, then model (1) reduces to the generalized linear models for clustered data. Berhane and Tibshirani (1998) extended the generalized estimating equation approach and applied the smoothing spline to estimate the model (1). We propose an efficient estimation procedure that approximates the nonparametric functions using polynomial splines. In addition, within-cluster correlation is incorporated to obtain efficient parameter estimation via the quadratic inference function (Qu, Lindsay, and Li 2000). Unlike the GEE approach (Berhane and Tibshirani 1998), our proposed method does not require estimation of the working correlation parameters, and it is more efficient than the GEE approach when the working correlation is misspecified.

Another focus of this article is on performing variable selection for the marginal generalized additive model (1). With advanced technology, massive high-throughput data with large-dimensional covariates are encountered quite frequently. Identifying important variables is a crucial step in analyzing high-dimensional data, because each redundant variable involves an infinite dimension of parameters for nonparametric components. Here a predictor variable X_l is said to be redundant in model (1), if and only if $\alpha_l(X_l) = 0$ almost surely. Otherwise, a predictor variable X_l is said to be relevant. Suppose that only an unknown subset of predictor variables in model (1) is relevant. Our goal is to consistently identify such subsets of relevant variables and estimate their unknown function components in (1) at the same time.

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

3. METHODOLOGY AND THEORY

In our estimation procedure, we approximate the smooth functions $\{\alpha_l(\cdot)\}_{l=1}^d$ in (1) by polynomial splines. For each $1 \le l \le d$, let υ_l be a partition of the interval $[0, 1]$, with N_n interior knots

$$\upsilon_l = \{0 = \upsilon_{l,0} < \upsilon_{l,1} < \cdots < \upsilon_{l,N_n} < \upsilon_{l,N_n+1} = 1\}.$$

Using υ_l as knots, the polynomial splines of order $p+1$ are functions with p-degree (or less) of polynomials on intervals $[\upsilon_{l,i}, \upsilon_{l,i+1}), i = 0, \ldots, N_n - 1$, and $[\upsilon_{l,N_n}, \upsilon_{l,N_n+1}]$, and have $p-1$ continuous derivatives globally. We denote the space of such spline functions by $\varphi_l = \varphi^p([0, 1], \upsilon_l)$. Denote $\varphi_l^0 = \{s \in \varphi_l : \int_0^1 s(x)\,dx = 0\}$, which consists of centered spline functions. The advantage of polynomial splines is that they often provide a good approximation of smooth functions with a limited number of knots.

3.1 An Initial Estimator

Let $\{B_{lj}(\cdot)\}_{j=1}^{J_n}$ be a set of spline bases of φ_l^0 with $J_n = N_n + p$. Then

$$\alpha_l(\cdot) \approx s_l(\cdot) = \sum_{j=1}^{J_n} \beta_{lj} B_{lj}(\cdot),$$

with a set of coefficients $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lJ_n})^T$. As a result,

$$\eta_{ij}^0 \approx \eta_{ij}(\boldsymbol{\beta}) = \beta_0 + \sum_{l=1}^{d} \sum_{j=1}^{J_n} \beta_{lj} B_{lj}\big(x_{ij}^{(l)}\big),$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_d^T)^T$. Or, equivalently, the mean function μ_{ij}^0 in (1) can be approximated by

$$\mu_{ij}^0 \approx \mu_{ij}(\boldsymbol{\beta}) = g^{-1}\left\{\beta_0 + \sum_{l=1}^{d} \sum_{j=1}^{J_n} \beta_{lj} B_{lj}\big(x_{ij}^{(l)}\big)\right\}.$$

In matrix notation, we write $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}), \ldots, \mu_{im_i}(\boldsymbol{\beta}))^T$. To estimate unknown coefficients $\boldsymbol{\beta}$, we apply the quadratic inference function (Qu, Lindsay, and Li 2000), which efficiently incorporates the within-cluster correlation. To simplify notation, we first assume equal cluster size with $m_i = m < \infty$. We discuss implementation for data with unequal cluster size in Section 3.5. Let $\mathbf{R}$ be a common working correlation. The QIF approach approximates the inverse of $\mathbf{R}$ by a linear combination of some basis matrixes, that is,

$$\mathbf{R}^{-1} \approx a_0\mathbf{I} + a_1\mathbf{M}_1 + \cdots + a_K\mathbf{M}_K,$$

where $\mathbf{I}$ is the identity and $\mathbf{M}_i$ are known symmetric basis matrices. For example, if $\mathbf{R}$ is of compound symmetric structure with correlation ρ, then $\mathbf{R}^{-1}$ can be represented as $a_0\mathbf{I} + a_1\mathbf{M}_1$ with $\mathbf{M}_1$ being a matrix with 0 on the diagonal and 1 off the diagonal, $a_0 = -\{(m-2)\rho + 1\}/k_1$, and $a_1 = \rho/k_1$, where $k_1 = (m-1)\rho^2 - (n-2)\rho - 1$ and m is the dimension of $\mathbf{R}$. The linear representation of $\mathbf{R}^{-1}$ is also applicable for the AR-1 and the block diagonal correlation structures. The advantage of the QIF approach is that it does not require the estimation of the linear coefficients a_i's associated with the working correlation matrix, which are treated as nuisance parameters here.

For any $\mathbf{x} = (x_1, \ldots, x_d)^T$, let $\mathbf{B}(\mathbf{x}) = (1, B_{11}(x_1), \ldots, B_{1J_n}(x_1), \ldots, B_{d1}(x_d), \ldots, B_{dJ_n}(x_d))^T$. Denote $\mathbf{B}_i = (\mathbf{B}(\mathbf{x}_{i1}), \ldots, \mathbf{B}(\mathbf{x}_{im_i}))^T$ for $i = 1, \ldots, n$. Define

$$\begin{aligned}\mathbf{G}_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\beta}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \mathbf{B}_i^T\boldsymbol{\Delta}_i^T(\boldsymbol{\beta})\mathbf{A}_i^{-1}\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} \\ \mathbf{B}_i^T\boldsymbol{\Delta}_i^T(\boldsymbol{\beta})\mathbf{A}_i^{-1/2}\mathbf{M}_1\mathbf{A}_i^{-1/2}\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} \\ \vdots \\ \mathbf{B}_i^T\boldsymbol{\Delta}_i^T(\boldsymbol{\beta})\mathbf{A}_i^{-1/2}\mathbf{M}_K\mathbf{A}_i^{-1/2}\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} \end{pmatrix}\end{aligned}$$

with $\boldsymbol{\Delta}_i(\boldsymbol{\beta}) = \operatorname{diag}\{\dot{\mu}_{i1}(\boldsymbol{\beta}), \ldots, \dot{\mu}_{im_i}(\boldsymbol{\beta})\}$ and $\dot{\mu}_{ij}(\boldsymbol{\beta})$ being the first derivative of g^{-1} evaluated at $\mathbf{B}^T(\mathbf{x}_{ij})\boldsymbol{\beta}$; and $\mathbf{A}_i = \operatorname{diag}\{V(\mu_{i1}), \ldots, V(\mu_{im_i})\}$. Because there are more estimation equations than the number of unknown parameters, the QIF approach estimates $\boldsymbol{\beta}$ by setting $\mathbf{G}_n$ as close to 0 as possible, in the sense of minimizing the quadratic inference function $Q_n(\boldsymbol{\beta})$, that is,

$$\widetilde{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}) = \operatorname*{argmin}_{\boldsymbol{\beta}} n\mathbf{G}_n^T(\boldsymbol{\beta})\mathbf{C}_n^{-1}(\boldsymbol{\beta})\mathbf{G}_n(\boldsymbol{\beta}), \qquad (2)$$

where

$$\mathbf{C}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i^T(\boldsymbol{\beta}).$$

As a result, for any $\mathbf{x} \in [0, 1]^d$, the estimator of the unknown components in (1) is given as

$$\widetilde{\alpha}_0 = \widetilde{\beta}_0,$$

$$\widetilde{\alpha}_l\big(x^{(l)}\big) = \sum_{j=1}^{J_n} \widetilde{\beta}_{lj} B_{lj}\big(x^{(l)}\big), \qquad l = 1, \ldots, d,$$

and

$$\widetilde{\alpha}(\mathbf{x}) = \widetilde{\beta}_0 + \sum_{l=1}^{d} \widetilde{\alpha}_l\big(x^{(l)}\big). \qquad (3)$$

Using a spline basis, we convert a problem in the continuum to one that is governed by only a finite number of parameters. For any given set of spline bases, $\widetilde{\boldsymbol{\beta}}$ can be obtained using the Newton–Raphson algorithm developed by Qu, Lindsay, and Li (2000). QIF estimation through spline basis expansion produces functional estimators that achieve optimal nonparametric properties on the rate of convergence, as shown in Theorem 1. This can be used in its own right, or as the initial value for an interactive scheme as developed in the next section for simultaneous variable selection and functional estimation.

To study the rate of convergence for $\widetilde{\alpha}_0, \widetilde{\alpha}_l$, and $\widetilde{\alpha}$, we first introduce some notation and present regularity conditions. We assume equal cluster sizes ($m_i = m < \infty$), and $(\mathbf{Y}_i, \mathbf{X}_i), i = 1, \ldots, n$, are iid copies of $(\mathbf{Y}, \mathbf{X})$ with $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$, and $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_m^T)^T$. The asymptotic results still hold for data with unequal cluster sizes, m_i, using a cluster-specific transformation, as discussed later in Section 3.5. For any matrix $\mathbf{A}$, $\|\mathbf{A}\|$ denotes the modulus of the largest singular value of $\mathbf{A}$. To prove the theoretical arguments, we need the following assumptions:

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

(C1) The covariates $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_m^T)^T$ are compactly supported, and without loss of generality, we assume that each $\mathbf{X}_j$ has support $\chi = [0,1]^d$. The density of $\mathbf{X}_j$, denoted by $f_j(\mathbf{x})$, is absolutely continuous, and there exist constants c_1 and c_2 such that $0 < c_1 \le \min_{\mathbf{x}\in\chi} f_j(\mathbf{x}) \le \max_{\mathbf{x}\in\chi} f_j(\mathbf{x}) \le c_2 < \infty$ for all $j = 1, \ldots, m$.

(C2) For each $l = 1, \ldots, d$, $\alpha_l(\cdot)$ has r continuous derivatives for some $r \ge 2$.

(C3) Let $\mathbf{e} = \mathbf{Y} - \boldsymbol{\mu}_0(\mathbf{X})$. Then $\boldsymbol{\Sigma} = E\mathbf{e}\mathbf{e}^T$ is positive definite, and for some $\delta > 0$, $E\|\mathbf{e}\|^{2+\delta} < +\infty$.

(C4) The knots sequences $\upsilon_l = \{0 = \upsilon_{l,0} \le \upsilon_{l,1} \le \cdots \le \upsilon_{l,N_n} \le \upsilon_{l,N_n+1} = 1\}$ for $l = 1, \ldots, d$, are quasi-uniform, that is, there exists $c_3 > 0$, such that

$$\max_{l=1,\ldots,d} \frac{\max(\upsilon_{l,j+1} - \upsilon_{l,j}, j = 0, \ldots, N_n)}{\min(\upsilon_{l,j+1} - \upsilon_{l,j}, j = 0, \ldots, N_n)} \le c_3.$$

Furthermore, the number of interior knots $N_n \asymp n^{1/(2r+1)}$, where $\asymp$ means that both sides have the same order. In particular, let $h = 1/N_n \asymp n^{-1/(2r+1)}$.

(C5) Let $\boldsymbol{\Gamma}_0^{(k)} = \boldsymbol{\Delta}_0^T \mathbf{V}_0^{(k)} \boldsymbol{\Sigma} \mathbf{V}_0^{(k)} \boldsymbol{\Delta}_0$ where $\mathbf{V}_0^{(k)} = \mathbf{A}_0^{-1/2} \mathbf{M}_k \times \mathbf{A}_0^{-1/2}$ and $\boldsymbol{\Delta}_0$, $\mathbf{A}_0$ are evaluated at $\boldsymbol{\mu} = \boldsymbol{\mu}^0(\mathbf{X})$. Then for sufficiently large n, the eigenvalues of $E(\boldsymbol{\Gamma}_0^{(k)})$ are bounded away from 0 and infinity for any $k = 0, \ldots, K$.

(C6) There exists a positive constant c_4 such that $0 < c_4 \le \inf_{i,j} V(\mu_{ij}) \le \sup_{i,j} V(\mu_{ij}) < \infty$. The functions $V(\cdot)$ and $g^{-1}(\cdot)$ have bounded second derivatives.

(C7) Let $\mathbf{M} = (\mathbf{M}_0^T, \ldots, \mathbf{M}_K^T)^T$. Assume that the modular of the singular value of $\mathbf{M}$ is bounded away from 0 and infinity.

These conditions are common in the polynomial spline estimation literature. Assumptions similar to (C1)–(C4) were also considered by Huang (1998), Huang (2003), and Xue (2009). In particular, the smoothness condition in (C2) determines the rate of convergence of the spline estimators $\widetilde{\alpha}_0, \widetilde{\alpha}_l$, and $\widetilde{\alpha}$. Conditions (C5) and (C6) are similar to assumptions (A3) and (A4) of He, Fung, and Zhu (2005), and can be easily verified for a broad family of distributions. Condition (C7) is needed for the weighting matrix $\mathbf{C}_n$ in (2), so it maintains positive definite asymptotically. This holds for the choice of basis matrices of exchangeable and AR-1 correlation structures discussed in this section.

Theorem 1. Under conditions (C1)–(C7), there exists a local minimizer of (2) such that

$$|\widetilde{\alpha}_0 - \alpha_0| = O_p\big(n^{-r/(2r+1)}\big),$$

$$\max_{1\le l\le d} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \big\{\widetilde{\alpha}_l\big(x_{ij}^{(l)}\big) - \alpha_l\big(x_{ij}^{(l)}\big)\big\}^2 = O_p\big(n^{-2r/(2r+1)}\big).$$

For $m_i = m = 1$, this reduces to a special case where the responses are iid. The rate of convergence given here is the same optimal rate as that obtained for polynomial spline regression for independent data (Huang 1998; Xue 2009). The main advantage of the QIF approach is that it incorporates within-cluster correlation by optimally combining estimating equations without estimating the correlation parameters. However, when there are redundant predictor variables, it fails to produce efficient and accurate estimators due to model complexity. In the next section we propose the penalized QIF method, which automatically sets small estimated functions to 0 and selects a parsimonious model.

3.2 Penalized Quadratic Inference Function for Marginal Generalized Additive Models

We propose a new variable selection approach by penalizing the QIF in (2). This enables one to perform simultaneous estimation and variable selection in the generalized additive model. The proposed method is able to shrink small components of estimated functions to 0, thus performing variable selection. In addition, it produces efficient estimators of the nonzero components by taking the within-cluster correlation into consideration. The penalized quadratic inference function estimator is defined as

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left\{ Q_n(\boldsymbol{\beta}) + n \sum_{l=1}^{d} p_{\lambda_n}\big(\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l}\big) \right\}, \tag{4}$$

in which $p_{\lambda_n}(\cdot)$ is a given penalty function depending on a tuning parameter λ_n, and $\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l}^2 = \boldsymbol{\beta}_l^T \mathbf{K}_l \boldsymbol{\beta}_l$, where $\mathbf{K}_l = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbf{B}_l(\mathbf{X}_{ij}^{(l)}) \mathbf{B}_l^T(\mathbf{X}_{ij}^{(l)})$ with $\mathbf{B}_l(\cdot) = (\mathbf{B}_{l1}(\cdot), \ldots, \mathbf{B}_{lJ_n}(\cdot))^T$. Note that

$$\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} s_l^2\big(\mathbf{X}_{ij}^{(l)}\big) \right)^{1/2} = \|s_l\|_n,$$

which is the empirical norm of the spline function s_l. The reason that we choose to penalize on $\|s_l\|_n$ is that it does not depend on a particular choice of spline bases, and intuitively, shrinking $\|s_l\|_n$ to 0 entails s_l is 0 almost surely. Note that the proposed penalization strategy achieves the same effect as the groupwise model selection approach (Yuan and Lin 2006), because each additive component containing many coefficients is treated as a whole group in model selection.

There are a number of choices for the penalty function $p_{\lambda_n}(\cdot)$. For the linear models, the L_1 penalty $p_{\lambda_n}(|\cdot|) = \lambda_n|\cdot|$ provides a LASSO-type of estimator, and the L_2 penalty $p_{\lambda_n}(|\cdot|) = \lambda_n|\cdot|^2$ gives a ridge-type estimator. Here we choose the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) penalty, which is defined by the derivative of p_{λ_n},

$$p'_{\lambda_n}(\theta) = \lambda_n \left\{ I(\theta \le \lambda_n) + \frac{(a\lambda_n - \theta)_+}{(a-1)\lambda_n} I(\theta > \lambda_n) \right\},$$

where a is a constant and is usually set to be $a = 3.7$ (Fan and Li 2001; Xue 2009), and $\lambda_n > 0$ is a tuning parameter. After obtaining the estimator $\widehat{\boldsymbol{\beta}}$ through penalization in (4), for any given $\mathbf{x} \in [0,1]^d$, an estimator of the unknown components in (1) is given as

$$\widehat{\alpha}_0 = \widehat{\beta}_0,$$

$$\widehat{\alpha}_l\big(x^{(l)}\big) = \sum_{j=1}^{J_n} \widehat{\beta}_{lj} B_{lj}\big(x^{(l)}\big), \qquad l = 1, \ldots, d.$$

Fan and Li (2001) established the asymptotic property of the regression parameter estimator using SCAD for the linear model. In what follows, we establish the asymptotic properties of the estimators of the nonparametric components for the marginal generalized additive model. We denote $\eta^0(\mathbf{x}_{ij}) = \alpha_0 +$

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

$\sum_{l=1}^{d}\alpha_l(x_{ij}^{(l)}) = \alpha_0 + \sum_{l=1}^{s}\alpha_l(x_{ij}^{(l)}) + \sum_{l=s+1}^{d}\alpha_l(x_{ij}^{(l)})$, where, without loss of generality, $\alpha_l = 0$ almost surely for $l = s+1,\ldots,d$, and s is the total number of nonzero function components. We first show that the penalized QIF estimators $\{\widehat{\alpha}_l\}_{l=1}^{d}$ achieve the same rate of convergence as the unpenalized ones $\{\widetilde{\alpha}_l\}_{l=1}^{d}$. Furthermore, we prove that the penalized estimators $\{\widehat{\alpha}_l\}_{l=1}^{d}$ in Theorem 2 have the sparsity property, that is, $\widehat{\alpha}_l = 0$ almost surely for $l = s+1,\ldots,d$. The sparsity property ensures that the proposed model selection is consistent, that is, it selects the correct variables with probability tending to 1 as the sample size goes to infinity.

Theorem 2. Under the same assumptions of Theorem 1, and if the tuning parameter $\lambda_n \to 0$, then there exists a local minimizer of (4) such that

$$|\widehat{\alpha}_0 - \alpha_0| = O_p\big(n^{-r/(2r+1)}\big),$$

$$\max_{1\le l\le d}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\big\{\widehat{\alpha}_l\big(x_{ij}^{(l)}\big) - \alpha_l\big(x_{ij}^{(l)}\big)\big\}^2 = O_p\big(n^{-2r/(2r+1)}\big).$$

Theorem 3. Under the same assumptions of Theorem 1, and if the tuning parameter $\lambda_n \to 0$, and $\lambda_n n^{r/(2r+1)} \to +\infty$, then, with probability approaching 1, $\widehat{\alpha}_l = 0$ almost surely for $l = s+1,\ldots,d$.

Theorem 3 also implies that the foregoing generalized additive model selection has the consistency property. The results in Theorems 2 and 3 are similar to those for penalized least squares established by Xue (2009); however, the theoretical proof is very different from the penalized least squares approach due to the nonlinear structure of the problem. Therefore, a more thorough study of the properties of the quadratic inference functions for the infinite-dimension case is indeed necessary here to overcome the difficulties.

3.3 An Algorithm Using Local Quadratic Approximation

In this section we provide an algorithm to minimize the penalized quadratic inference function defined as in (4) using the local quadratic approximation (Fan and Li 2001). Let $\boldsymbol{\beta}^0$ be an initial value that is close to the true minimizer of (4). For example, we could take $\widetilde{\boldsymbol{\beta}}$, the QIF estimator, as the initial value. Denote $\boldsymbol{\beta}^k = (\boldsymbol{\beta}_0^k, \boldsymbol{\beta}_1^{kT},\ldots,\boldsymbol{\beta}_d^{kT})^T$ as the value at the kth iteration. If $\boldsymbol{\beta}_l^k$ is close to 0 in the sense that $\|\boldsymbol{\beta}_l^k\|_{\mathbf{K}_l} \le \epsilon$, for some small threshold value ϵ, set $\boldsymbol{\beta}_l^{k+1}$ to $\mathbf{0}$. In the implementation, we have used $\epsilon = 10^{-6}$. Without loss of generality, suppose that $\boldsymbol{\beta}_l^{k+1} = 0$, for $l = d_k+1,\ldots,d$, and write $\boldsymbol{\beta}^{k+1} = (\beta_0^{k+1}, (\boldsymbol{\beta}_1^{k+1})^T,\ldots,(\boldsymbol{\beta}_{d_k}^{k+1})^T, (\boldsymbol{\beta}_{d_k+1}^{k+1})^T,\ldots, (\boldsymbol{\beta}_d^{k+1})^T)^T = (\beta_0^{k+1}, (\boldsymbol{\beta}_{11}^{k+1})^T, (\boldsymbol{\beta}_{22}^{k+1})^T)^T$, with $\boldsymbol{\beta}_{22}^{k+1}$ containing the $d-d_k$ zero components. Let $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_{11}^T, \boldsymbol{\beta}_{22}^T)^T$ be the same partition of any $\boldsymbol{\beta}$ with the same length as $\boldsymbol{\beta}^{k+1}$.

We obtain a value for the nonzero component $\boldsymbol{\beta}_{11}^{k+1}$ using the following quadratic approximation. For $\|\boldsymbol{\beta}_l^k\|_{\mathbf{K}_l} > \epsilon$, one can locally approximate the penalty function by

$$\begin{aligned} p_{\lambda_n}\big(\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l}\big) &\approx p_{\lambda_n}\big(\|\boldsymbol{\beta}_l^k\|_{\mathbf{K}_l}\big) \\ &\quad + \tfrac{1}{2}p'_{\lambda_n}\big(\|\boldsymbol{\beta}_l^k\|_{\mathbf{K}_l}\big)\|\boldsymbol{\beta}_l^k\|_{\mathbf{K}_l}^{-1}\boldsymbol{\beta}^{kT}\mathbf{K}_l(\boldsymbol{\beta}_l - \boldsymbol{\beta}_l^k) \\ &\approx p_{\lambda_n}\big(\|\boldsymbol{\beta}_l^k\|_{\mathbf{K}_l}\big) \\ &\quad + \tfrac{1}{2}p'_{\lambda_n}\big(\|\boldsymbol{\beta}_l^k\|_{\mathbf{K}_l}\big)\|\boldsymbol{\beta}_l^k\|_{\mathbf{K}_l}^{-1}(\boldsymbol{\beta}_l^T\mathbf{K}_l\boldsymbol{\beta}_l - \boldsymbol{\beta}_l^{kT}\mathbf{K}_l\boldsymbol{\beta}_l^k), \end{aligned}$$

where p'_{λ_n} is the first-order derivative of p_{λ_n}. Therefore, the objective function in (4) can be locally approximated (up to a constant) by a quadratic function

$$\begin{aligned} &Q_n(\boldsymbol{\beta}^k) + \nabla Q_n(\boldsymbol{\beta}^k)^T(\boldsymbol{\beta}_{11} - \boldsymbol{\beta}_{11}^k) \\ &\quad + \tfrac{1}{2}(\boldsymbol{\beta}_{11} - \boldsymbol{\beta}_{11}^k)^T\nabla^2 Q_n(\boldsymbol{\beta}^k)(\boldsymbol{\beta}_{11} - \boldsymbol{\beta}_{11}^k) \\ &\quad + \tfrac{1}{2}n\boldsymbol{\beta}_{11}^T\Lambda(\beta^k)\boldsymbol{\beta}_{11}, \end{aligned}$$

where $\nabla Q_n(\boldsymbol{\beta}^k) = \frac{\partial Q_n(\boldsymbol{\beta}^k)}{\partial \boldsymbol{\beta}_{11}}$, $\nabla^2 Q_n(\boldsymbol{\beta}^k) = \frac{\partial^2 Q_n(\boldsymbol{\beta}^k)}{\partial \boldsymbol{\beta}_{11}\partial \boldsymbol{\beta}_{11}^T}$, and

$$\begin{aligned} \Lambda(\beta^k) = \operatorname{diag}\big\{&p'_{\lambda_n}\big(\|\boldsymbol{\beta}_1^k\|_{\mathbf{K}_l}\big)\|\boldsymbol{\beta}_1^k\|_{\mathbf{K}_l}^{-1},\ldots, \\ &p'_{\lambda_n}\big(\|\boldsymbol{\beta}_{d_k}^k\|_{\mathbf{K}_l}\big)\|\boldsymbol{\beta}_{d_k}^k\|_{\mathbf{K}_l}^{-1}\big\}. \end{aligned} \quad (5)$$

$\boldsymbol{\beta}_{11}^{k+1}$ can be obtained by minimizing the foregoing quadratic function. Specifically,

$$\begin{aligned} \boldsymbol{\beta}_{11}^{k+1} = \boldsymbol{\beta}_{11}^k - &\{\nabla^2 Q_n(\boldsymbol{\beta}^k) + n\Lambda(\beta^k)\}^{-1} \\ &\times \{\nabla Q_n(\boldsymbol{\beta}^k) + n\Lambda(\beta^k)\boldsymbol{\beta}_{11}^k\}. \end{aligned}$$

The foregoing iteration process can be repeated until convergence is reached. Here we use the convergence criterion such that $\sqrt{(\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k)^T(\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k)} \le 10^{-6}$. In our experience, the algorithm is very stable and fast to compute. It usually reaches a reasonable convergence tolerance within a few iterations. However, the computational time increases as the number of covariates increases. In one of our numerical examples in Section 4, it takes about 25 iterations on average to converge when there are 100 covariates in the model.

3.4 Tuning Parameter Selection

It is important to select sensible tuning parameters in the implementation, because the performance of our proposed method depends critically on the choice of tuning parameters. The QIF method in Section 3.1 requires an appropriate choice of the knot sequences $\{\upsilon_l\}_{l=1}^{d}$ for the spline approximation. For the penalized QIF method in Section 3.2, in addition to $\{\upsilon_l\}_{l=1}^{d}$, one also needs to choose tuning parameter λ_n in the SCAD penalty function.

For knot selection in the QIF method, we use equally spaced knots and select only the number of interior knots, N_n. A similar strategy was described by Huang, Wu, and Zhou (2004), Xue and Yang (2006), and Xue (2009). For the penalized QIF, we use the same knot sequences selected in the QIF procedure and select only the tuning parameter λ_n for simplicity. For any given N_n, denote the solution that minimizes (2) by $\widetilde{\boldsymbol{\beta}}_{N_n}$. Similarly, for any given λ_n, denote the estimator that minimizes (4) by $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. Motivated by Qu and Li (2006) and Wang, Li, and Tsai (2008), we use a consistent BIC procedure to select the optimal tuning parameters. Because the quadratic inference function Q_n behaves equivalently to minus twice the log-likelihood function (Qu, Lindsay, and Li 2000), we define the BIC in the QIF procedure as

$$\mathrm{BIC}_1(N_n) = Q_n(\widetilde{\boldsymbol{\beta}}_{N_n}) + \log(n)\,\mathrm{DF}_{N_n},$$

where $\mathrm{DF}_{N_n} = 1 + dJ_n$ is the total number of unknown parameters in (2). We then select the optimal knot number $\widehat{N}_n =$

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

$\text{argmin}_{N_n} \text{BIC}_1(N_n)$. In our experience, only a small number of knots are needed. In the numerical examples in Section 4, the selected the optimal knot numbers are in the range of 2 to 5. Similarly, for the penalized QIF procedure, we define

$$\text{BIC}_2(\lambda_n) = Q_n\big(\widehat{\boldsymbol{\beta}}_{\lambda_n}\big) + \log(n)\,\text{DF}_{\lambda_n}.$$

The effective degrees of freedom is defined as

$$\text{DF}_{\lambda_n} = \text{trace}\big\{\big(\nabla^2 Q_n\big(\widehat{\boldsymbol{\beta}}_{\lambda_n}\big) + n\Lambda\big(\widehat{\boldsymbol{\beta}}_{\lambda_n}\big)\big)^{-1}\nabla^2 Q_n\big(\widehat{\boldsymbol{\beta}}_{\lambda_n}\big)\big\},$$

where $\nabla^2 Q_n(\cdot)$ and $\Lambda(\cdot)$ are defined similarly as in (5) for the nonzero components in $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. Consequently, we choose the optimal λ_n such that the BIC value reaches the minimum, that is, $\widehat{\lambda}_n = \text{argmin}_{\lambda_n} \text{BIC}_2(\lambda_n)$.

3.5 Unequal Cluster Sizes

In the cases where the clusters have unequal cluster sizes, we use cluster-specific transformation matrices to reformulate the data with unequal cluster size. That is, for the ith cluster with cluster size m_i, the $m \times m_i$ transformation matrix T_i is defined by deleting the corresponding columns (indexed by the missing observations in the cluster) of the $m \times m$ identity matrix, where m is the largest size of a fully observed cluster and is assumed to be bounded. Using the transformation matrices, we define $Y_i^* = T_i Y_i$, $\Delta_i^*(\beta) = T_i\Delta_i(\beta)$, $B_i^* = T_i B_i$, $\mu_i^*(\beta) = T_i\mu_i(\beta)$, and $A_i^* = T_i A_i$ for the ith cluster. Note that the entries in Y_i^*, $\Delta_i^*(\beta)$, B_i^*, $\mu_i^*(\beta)$, and A_i^*, either can be obtained if the measurements are observed or can be 0 if the corresponding observations are missing. We substitute Y_i, $\Delta_i(\beta)$, B_i, $\mu_i(\beta)$, and A_i with Y_i^*, $\Delta_i^*(\beta)$, B_i^*, $\mu_i^*(\beta)$, and A_i^*, respectively, in the previous procedures and in Lemmas 10 and 11 (provided in the supplemental material). Parameter estimation and variable selection also could be carried out for data with unequal cluster sizes, assuming an additional condition that the cluster size is m_i. Following the arguments in the proofs of the theorems, it can be shown that the asymptotic results of Theorems 1, 2, and 3 still hold for correlated data with unequal cluster sizes for parameter estimation and variable selection procedures.

4. SIMULATION STUDIES

In this section we conduct simulation studies for both continuous and binary outcomes. We evaluate the total averaged integrated squared error (TAISE) to assess estimation efficiency. Let $\widehat{\alpha}^{(r)}$ be the estimator of a nonparametric function, α, in the rth ($1 \le r \le R$) replication and let $\{x_m\}_{m=1}^{n_{\text{grid}}}$ be the grid points where $\widehat{\alpha}^{(r)}$ are evaluated. We define

$$\text{AISE}(\widehat{\alpha}) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_{\text{grid}}}\sum_{m=1}^{n_{\text{grid}}}\big\{\alpha(x_m) - \widehat{\alpha}^{(r)}(x_m)\big\}^2$$

and $\text{TAISE} = \sum_{l=1}^{d}\text{AISE}(\widehat{\alpha}_l)$. Let $\mathcal{S}$ and $\mathcal{S}_0$ be the selected and true index set containing relevant variables, respectively. We say that $\mathcal{S}$ is correct if $\mathcal{S} = \mathcal{S}_0$, $\mathcal{S}$ overfits if $\mathcal{S}_0 \subset \mathcal{S}$ and $\mathcal{S}_0 \neq \mathcal{S}$, and $\mathcal{S}$ underfits if $\mathcal{S}_0 \not\subset \mathcal{S}$. In all of the simulation studies, the number of replications is 100.

4.1 Example 1: Continuous Response and Moderate Dimension of Covariates

In this example, the continuous responses $\{Y_{ij}\}$ are generated from

$$Y_{ij} = \sum_{l=1}^{d}\alpha_l\big(X_{ij}^{(l)}\big) + \varepsilon_{ij}, \qquad i = 1,\ldots,n, j = 1,\ldots,5, \quad (6)$$

where $d = 10$ and the number of clusters is $n = 100, 250,$ or 500. The additive functions are

$$\alpha_1(x) = 2x - 1, \qquad \alpha_2(x) = 8(x - 0.5)^3,$$
$$\alpha_3(x) = \sin(2\pi x),$$

and $\alpha_l(x) = 0$ for $l = 4,\ldots,10$. Thus the last seven variables in this model are null variables and do not contribute to the model. The covariates $\mathbf{X}_{ij} = (X_{ij}^{(1)},\ldots,X_{ij}^{(10)})^T$ are generated independently from Uniform$([0,1]^{10})$. The error $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1},\ldots,\varepsilon_{i5})^T$ follows a multivariate normal distribution with mean 0, a common marginal variance $\sigma^2 = 1$, and an exchangeable correlation structure with correlation $\rho = 0.7$.

We apply the penalized QIF with SCAD penalty and fit the linear splines ($p = 1$) and cubic splines ($p = 3$). To illustrate the effect incorporating within-cluster correlation on estimation efficiency, we compare the estimation efficiency of using basis matrices from different working correlation structures: exchangeable (EC), AR-1, and independent (IND). In addition, we compare the penalized QIF approach with the QIF estimations of a full model (FULL) and an oracle model (ORACLE). Here the full model consists of all 10 variables and the oracle model contains only the first three relevant variables. The oracle model is available only in simulation studies where the true information is known.

Table 1 summarizes the estimation results of all procedures. It shows that the estimation procedures with a correct EC working correlation have the smallest TAISEs, and thus the estimators are more efficient than their counterparts with IND working correlation that ignore within-cluster correlation. For example, the efficiency gained by incorporating correlation could be doubled in the cubic splines approach. Estimation based on a misspecified AR-1 correlation structure will lead to some efficiency loss compared with using the true EC structure, but it is still more efficient than assuming independent structure. In general, estimation using cubic splines is more efficient than that using linear splines. Furthermore, with the same type of working correlation, ORACLE has the smallest TAISEs compared with the SCAD and FULL model approaches. This is not surprising, because ORACLE takes advantage of the true information from the data generation and contains only three relevant variables. However, the efficiency of the SCAD is closer to that of ORACLE, with much smaller TAISEs than those of the FULL model.

From one selected simulated data set with $n = 250$, Figure 1 plots the first four estimated functional components from the SCAD, FULL, and ORACLE models using linear spline and exchangeable working correlation. Note that for the fourth variable, both the true and estimated component functions from SCAD are 0. This clearly indicates that the proposed method estimates unknown functions reasonably well.

Table 1. Example 1: Continuous response and moderate dimension of covariates. The TAISEs (standard errors in parentheses) of SCAD, ORACLE, and FULL with exchangeable (EC), AR-1 or independent (IND) working correlation using linear or cubic splines. The number of replications is 100

	n	Method	EC	AR-1	IND
Linear spline	100	SCAD	$0.0385_{(0.0017)}$	$0.0468_{(0.0019)}$	$0.0496_{(0.0018)}$
		ORACLE	$0.0263_{(0.0004)}$	$0.0311_{(0.0005)}$	$0.0377_{(0.0007)}$
		FULL	$0.0522_{(0.0010)}$	$0.0649_{(0.0012)}$	$0.0776_{(0.0015)}$
	250	SCAD	$0.0121_{(0.0002)}$	$0.0143_{(0.0004)}$	$0.0173_{(0.0006)}$
		ORACLE	$0.0103_{(0.0001)}$	$0.0117_{(0.0002)}$	$0.0164_{(0.0004)}$
		FULL	$0.0205_{(0.0004)}$	$0.0247_{(0.0005)}$	$0.0385_{(0.0007)}$
	500	SCAD	$0.0093_{(0.0001)}$	$0.0095_{(0.0001)}$	$0.0119_{(0.0002)}$
		ORACLE	$0.0085_{(0.0001)}$	$0.0089_{(0.0001)}$	$0.0117_{(0.0002)}$
		FULL	$0.0134_{(0.0001)}$	$0.0150_{(0.0002)}$	$0.0218_{(0.0003)}$
Cubic spline	100	SCAD	$0.0378_{(0.0042)}$	$0.0422_{(0.0032)}$	$0.0454_{(0.0038)}$
		ORACLE	$0.0172_{(0.0009)}$	$0.0217_{(0.0010)}$	$0.0262_{(0.0011)}$
		FULL	$0.0778_{(0.0022)}$	$0.0843_{(0.0024)}$	$0.0867_{(0.0044)}$
	250	SCAD	$0.0062_{(0.0003)}$	$0.0083_{(0.0004)}$	$0.0099_{(0.0004)}$
		ORACLE	$0.0052_{(0.0003)}$	$0.0067_{(0.0003)}$	$0.0099_{(0.0004)}$
		FULL	$0.0220_{(0.0007)}$	$0.0284_{(0.0008)}$	$0.0362_{(0.0008)}$
	500	SCAD	$0.0038_{(0.0001)}$	$0.0043_{(0.0001)}$	$0.0075_{(0.0002)}$
		ORACLE	$0.0026_{(0.0001)}$	$0.0032_{(0.0001)}$	$0.0056_{(0.0002)}$
		FULL	$0.0075_{(0.0001)}$	$0.0096_{(0.0002)}$	$0.0174_{(0.0003)}$

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

Table 2 provides variable selection results for the SCAD procedures. It shows that for the EC, AR-1, or IND type of working correlation, the percentage of correct model fitting goes to 1 quickly as the sample size increases. This confirms the consistency theorem of variable selection provided in Section 3.2. Furthermore, for the cubic spline approach, the correct model-fitting percentage using the EC correlation structure is noticeably higher than that using the IND when the sample size is

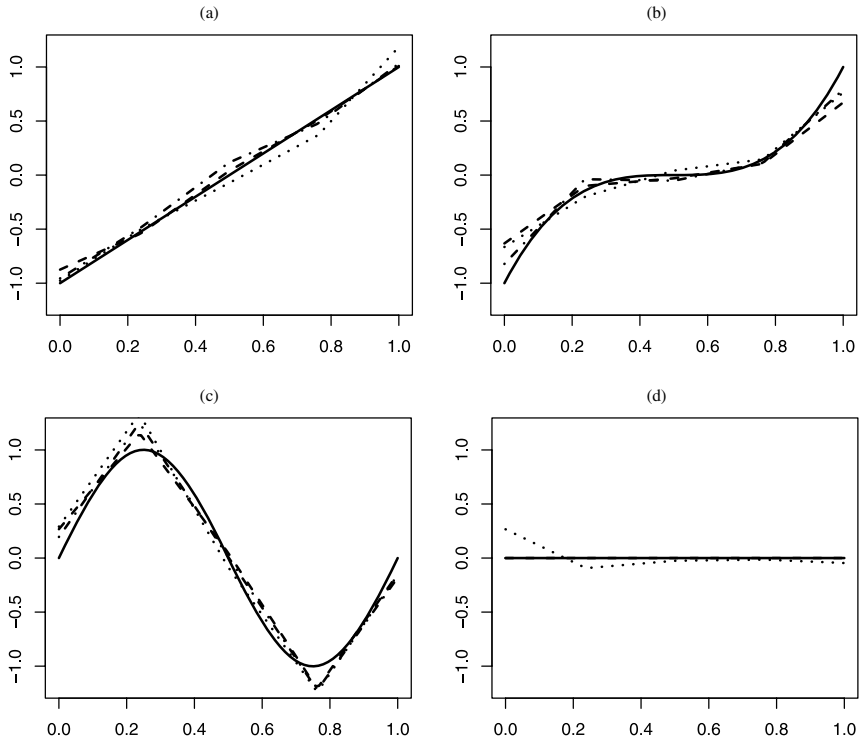

Figure 1. Example 1: Continuous response and moderate dimension of covariates. The estimated component functions of (a) $\alpha_1(x) = 2x - 1$, (b) $\alpha_2(x) = 8(x - 0.5)^3$, (c) $\alpha_3(x) = \sin(2\pi x)$, and (d) $\alpha_4(x) = 0$ from SCAD (dot–dash), FULL (dot), and ORACLE (dash) using linear spline and exchangeable working correlation. The true functions are plotted in solid lines.

Table 2. Example 1: Continuous response and moderate dimension of covariates. Variable selection results of SCAD with exchangeable, AR-1 or independent working correlation and linear or cubic splines. The columns of C, O, and U give the percentage of correct fitting, overfitting, and underfitting from 100 replications

		EC			AR-1			IND		
	n	C	O	U	C	O	U	C	O	U
Linear spline	100	0.93	0.00	0.07	0.93	0.01	0.06	0.91	0.01	0.08
	250	0.99	0.00	0.01	0.98	0.00	0.02	0.98	0.00	0.02
	500	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
Cubic spline	100	0.81	0.09	0.10	0.79	0.17	0.04	0.73	0.18	0.09
	250	0.99	0.00	0.01	0.99	0.00	0.01	0.95	0.00	0.05
	500	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

moderate. Finally, we give the computing time of the algorithm proposed in Section 3.3. For SCAD with $p = 1$ and EC working correlation, completing one run of the simulation for $n = 100, 250$, and 500 takes 0.5, 1.2, and 3.8 seconds, respectively. The computing time for cubic spline and other correlation structures are similar.

4.2 Example 2: Continuous Response With High Dimension of Covariates

To assess our method in more challenging cases for high-dimensional data, we consider a model with the dimension of functional components $d = 100$ in (6). However, only the first three variables are relevant and take the same functional forms as in Example 1. We consider the model (6) with number of clusters $n = 200$ and cluster size 5, with errors $\{\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i5})^T\}_{i=1}^{200}$ generated independently from a multivariate normal distribution with mean 0, and an AR-1 correlation structure with $\text{corr}(\varepsilon_{ij_1}, \varepsilon_{ij_2}) = 0.7^{|j_1 - j_2|}$ for $1 \le j_1, j_2 \le 5$.

We apply the linear spline QIF to estimate the full (FULL) and oracle (ORACLE) models with working independent (IND), exchangeable (EC), or AR-1 working correlation. For variable selection, we consider the penalized linear spline QIF with SCAD penalty with basis matrices from IND, EC, or AR-1 working correlations. Table 3 reports TAISEs for FULL, ORACLE, and SCAD and variable selection results on correct, overfit, and underfit percentages of the SCAD approach for both IND and AR-1 working correlations. Table 3 clearly indicates that the improvement from incorporating within-cluster correlation is very significant. In particular, the estimation procedures with a correctly specified AR-1 structure always give smaller TAISEs than those with a misspecified EC or IND working correlation. For variable selection, the SCAD with an AR-1 working correlation also performs noticeably better than the one with EC or independent working correlation. Furthermore, Table 3 also shows that the SCAD procedure dramatically improves the estimation accuracy for this high-dimensional case, with TAISEs from the SCAD less than one-tenth of the TAISEs from the FULL model. Finally, for computing time, it takes 19.8 seconds to compute SCAD with $p = 1$ and EC working correlation in one run of simulation. The computing times for other correlation structures are similar. However, compared with Example 1, the computation burden increases dramatically as the number of covariates increases.

4.3 Example 3: Binary Response

To assess the performance of our method for binary outcomes, we generate a random sample of 250 clusters in each simulation run. Within each cluster, binary responses $\{Y_{ij}\}_{j=1}^{20}$ are generated from a marginal logit model

$$\text{logit}\, P(Y_{ij} = 1 | \mathbf{X}_{ij} = \mathbf{x}_{ij}) = \sum_{l=1}^{10} \alpha_l\big(x_{ij}^{(l)}\big),$$

where $\alpha_1(x) = [\exp(x+1) - (\exp(2) - \exp(1))]/16$, $\alpha_2(x) = \cos(2\pi x)/4$, $\alpha_3(x) = x(1-x) - 1/6$, $\alpha_4(x) = 2(x - 0.5)^3$, and the remaining 6 covariates are null variables with $\alpha_l(x) = 0$ for $l = 5, \ldots, 10$. The covariates $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(10)})^T$ are generated independently from Uniform$([0,1]^{10})$. The algorithm provided by Macke et al. (2008) is applied to generate correlated binary responses with exchangeable correlation structure with a correlation parameter of 0.5.

We applied the penalized QIF with a SCAD penalty and linear spline (SCAD) for the variable selection, and the linear spline QIF for estimation of the full (FULL) and oracle (ORACLE) models. To illustrate how different working correlations could affect our estimation and variable selection results, we consider minimizing (2) and (4) using AR-1 (AR-1) working correlation and independent (IND) structures, in addition to the true exchangeable (EC) correlation structure.

Table 4 gives the TAISEs for the SCAD, ORACLE, and FULL models with three different working correlations. Sim-

Table 3. Example 2: Continuous response and high dimension of covariates. The TAISEs (standard errors in parentheses), and percentages of correct fitting (C), underfitting (U), and overfitting (O) from 100 replications

	SCAD	ORACLE	FULL	C	O	U
AR-1	$0.03185_{(0.0007)}$	$0.0189_{(0.0005)}$	$0.4125_{(0.0005)}$	0.92	0.02	0.06
EC	$0.03544_{(0.0008)}$	$0.0214_{(0.0006)}$	$0.5241_{(0.0006)}$	0.89	0.05	0.06
IND	$0.04499_{(0.0009)}$	$0.0237_{(0.0006)}$	$0.7278_{(0.0006)}$	0.85	0.04	0.11

Table 4. Example 3: Binary response. The TAISEs (standard errors in parentheses) from 100 simulations

	SCAD	ORACLE	FULL
EC	$0.0062_{(0.0004)}$	$0.0068_{(0.0003)}$	$0.0191_{(0.0008)}$
AR-1	$0.0089_{(0.0005)}$	$0.0083_{(0.0004)}$	$0.0229_{(0.0008)}$
IND	$0.0123_{(0.0006)}$	$0.0112_{(0.0005)}$	$0.0276_{(0.0010)}$

ilar to the previous continuous simulation study, the estimation based on correctly specified exchangeable correlation structure has the smallest TAISEs. For the SCAD approach, the efficiency is almost doubled if the correct correlation information is incorporated. Estimation based on misspecified AR-1 correlation structure will lead to some efficiency loss compared to using the true structure, but it is still more efficient than assuming independent structure. However, this could be a different case for the GEE if the true exchangeable correlation were misspecified as AR-1, since GEE requires one to estimate the correlation ρ for misspecified AR-1, and the estimator of ρ may not be valid. Consequently, the estimator from the misspecified correlation structure could be less efficient than assuming independence.

Furthermore, similar to the previous study, TAISEs calculated based on the SCAD approach are also shown to be close to the TAISEs from ORACLE, and much smaller than those from the FULL model. The relative efficiency ratio between the SCAD and FULL model approaches is close to 2. This shows that the SCAD approach can gain significant estimation accuracy by effectively removing the zero component variables. Table 5 gives the frequency of appearance for each of the 10 variables in the selected model. Overall, the SCAD procedures work reasonably well, and the SCAD with EC and AR-1 working correlation structures provides better variable selection results than the SCAD with IND working structure. In addition, the SCAD approach with EC working correlation performs the best in selecting non-zero component variables $(X_1, \ldots, X_4)$. Note that the model selection assuming IND working structure performs poorly for selecting variable X_4.

For one simulation run, Figure 2 plots the first four estimated functional components from the SCAD approach with the three different working correlations. For the fourth functional component, the SCAD with IND structure does not estimate the nonzero component correctly, because it apparently shrinks all coefficients of the function form to 0. In general, estimation assuming IND structure provides poorer results than estimation assuming EC or AR-1 working correlations. For the EC and AR-1 correlation structures, the SCAD approach provides reasonable estimations of the unknown functions. Finally, for computing time, it takes 9.7 seconds to compute SCAD with $p = 1$ and EC working correlation in one run of simulation. Overall, the proposed algorithm is fast to compute.

Table 5. Example 3: Binary response. The appearance frequency of the variables in the selected model from 100 simulations

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
EC	100	100	90	98	0	0	0	2	0	0
AR-1	96	100	79	94	0	1	0	0	0	0
IND	89	100	75	66	0	1	0	1	0	0

5. REAL DATA ANALYSIS

To illustrate the proposed method, we apply it to an observational study of periodontal disease (Stoner 2000). The partial data set consists of 528 patients with chronic periodontal disease who had an initial periodontal exam during 1988–1992. Each patient was followed annually for 5 years after the initial examination. One of the study goals is to identify risk factors that influence tooth loss to improve the treatment of periodontal disease. Let a binary response be $Y_{ij} = 1$ if patient i in jth year has at least one tooth extraction and $Y_{ij} = 0$ otherwise. The covariates of interest include age at time of initial exam (Age), number of teeth present at time of initial exam (Teeth), number of diseased sites (Sites), mean pocket depth in diseased sites (Pddis), mean pocket depth in all sites (Pdall), number of nonsurgical periodontal procedures in a year (Nonsurg), number of non-periodontal dental treatments in a year (Dent), and date of initial exam in fractional years (Entry). It is obvious that the last variable (Entry) is not related to the model prediction, but it is included here so we can verify whether the model selection procedure will exclude this variable as expected.

To model the binary responses Y_{ij}, we consider a marginal generalized additive model with a logit link in (1). We apply the proposed penalized QIF with SCAD penalty and fit both linear and cubic splines to select relevant variables, in addition to the QIF generalized additive model with linear spline based on the full model with all eight covariates. In both procedures, we use exchangeable working correlation. Procedures with other types of working correlation are also considered and lead to similar conclusions, thus they are not reported here.

Figure 3 plots estimated function components from the SCAD approach with linear spline (dashed) and cubic spline (dotdash), and the QIF estimation using the full model (solid line). Figure 3 shows that for both SCAD procedures, the selected relevant covariates are Age, Sites, Pdall, and Dent. The others are not selected since their function components are shrunken to 0 by the penalized QIF procedures. Note that the variable Entry is not selected, as expected. Furthermore, we also calculate the 95% pointwise confidence intervals of the estimated component functions using the QIF procedure with linear spline based on 500 bootstrapped samples. Notice that for the four variables that are not selected, their 95% bootstrap confidence intervals contain the zero line completely. This confirms our findings on variable selection based on the penalized QIF procedure.

6. DISCUSSION

We propose efficient estimation and model selection procedures for generalized additive models when the responses are correlated. We provide a SCAD penalty in model selection for the additive functional forms, and we are able to select the functional components groupwise. We incorporate correlation from the clustered data through optimally combining moment conditions that contain correlation information from the correlated data. The advantage of this approach is that we do not require explicit estimation of the correlation parameters. This improves estimation efficiency significantly as well as saving computational cost. We found that it may be quite computationally intensive in high-dimensional variable selection settings, because

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

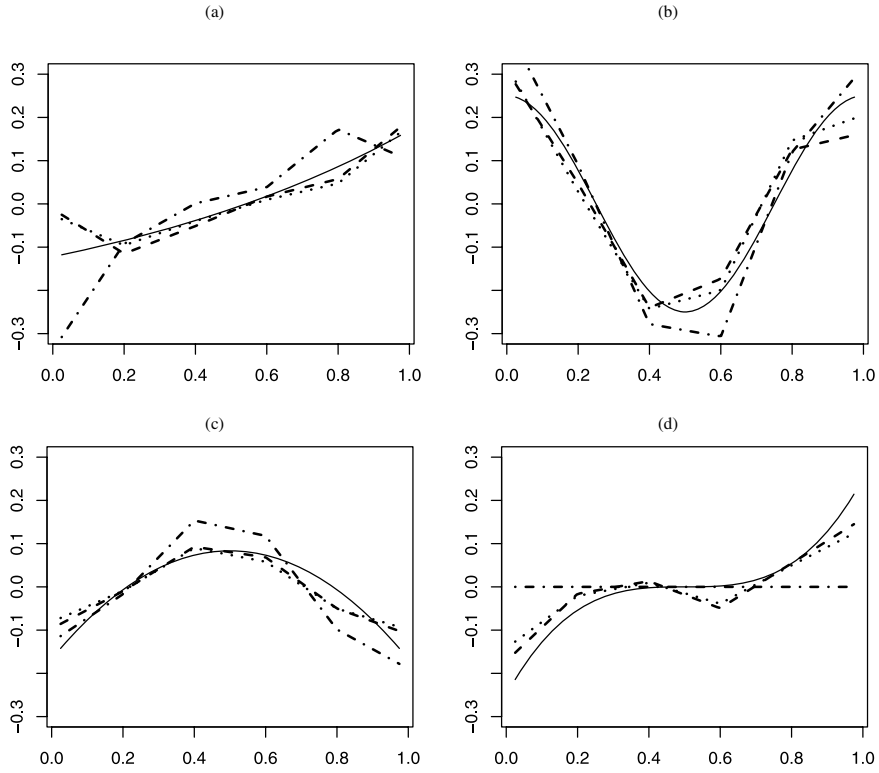

Figure 2. Example 3: Binary response. The estimated component functions of (a) $\alpha_1(x) = (e^{x+1} - e^2 + e)/16$, (b) $\alpha_2(x) = \cos(2\pi x)/4$, (c) $\alpha_3(x) = x(1-x) - 1/6$, and (d) $\alpha_4(x) = 2(x-0.5)^3$ from linear spline SCAD with exchangeable (dot), AR (dash), and independent (dot–dash) working correlation. The true functions are plotted in solid lines.

the dimension of the parameters involved in the nonparametric forms increases significantly compared with parametric model selection settings.

We show that the nonparametric estimator is consistent with an optimal L_2-norm convergence rate of $n^{-r/(2r+1)}$, which is the same optimal rate of L_2-norm convergence as in nonparametric models for independent and continuous data. In addition, we show that the model selection approach is consistent, that is, with probability tending to 1, it selects the correct model with nonzero functional forms converging to 1 almost surely.

Another advantage of our approach is that estimation and model selection are achieved simultaneously, in contrast to other model selection procedures such as AIC or BIC. Our simulation studies show that the proposed estimator recovers a significant amount of efficiency by performing model selection. This is reflected in that the total average integrated squared error from the SCAD model selection approach is much closer to the true ORACLE model, and is smaller than the squared error from the full model. In addition, we also show that estimation is more accurate when the true correlation structure is taken into consideration.

Finally, the theoretical techniques in deriving asymptotic properties of the proposed estimator are quite different from the existing nonparametric approaches of which we are aware. We should also mention that because there is no closed form of the likelihood function, the numerical integration technique to obtain a maximum likelihood estimator for a setting with a small number of parameters is not applicable here, because there are infinite dimensions of parameters involved in each functional form. In addition, the conventional penalized least squares approach for continuous data is not applicable for correlated categorical response data because it could result in biased estimation and inconsistent model selection.

APPENDIX: PROOFS OF THEOREMS

The necessary lemmas for the following proofs are given in the supplemental materials. We first introduce some notation.

For any real-valued function f on $[0, 1]$, let $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$. Let $\|\cdot\|_2$ be the usual L_2 norm for functions and vectors. Let $\mathcal{L}_2([0, 1])$ be the space of all square integrable functions defined on $[0, 1]$. For any $\alpha_1, \alpha_2 \in \mathcal{L}_2([0, 1])$, let

$$\langle \alpha_1, \alpha_2 \rangle = \int_0^1 \alpha_1(x)\alpha_2(x)\,dx \qquad \text{and} \qquad \|\alpha\|_2^2 = \int_0^1 \alpha^2(x)\,dx. \tag{A.1}$$

Let $\mathcal{H}_0$ be the space of square-integrable constant functions on $[0, 1]$, and let $\mathcal{L}_2^0$ be the space of square-integrable functions on $[0, 1]$, which is orthogonal to $\mathcal{H}_0$. That is, let $\mathcal{L}_2^0 = \{\alpha : \langle \alpha, 1\rangle = 0, \alpha \in \mathcal{L}_2\}$, where 1 is the identity function defined on $[0, 1]$. Define the additive model space $\mathcal{M}$ as

$$\mathcal{M} = \left\{ \alpha(\mathbf{x}) = \alpha_0 + \sum_{l=1}^{d} \alpha_l(x_l); \alpha_0 \in \mathcal{H}_0, \alpha_l \in \mathcal{L}_2^0 \right\}.$$

Then the regression function $\eta(\mathbf{x})$ in (1) of the main text is modeled as an element of $\mathcal{M}$. Define the polynomial spline approximation space

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

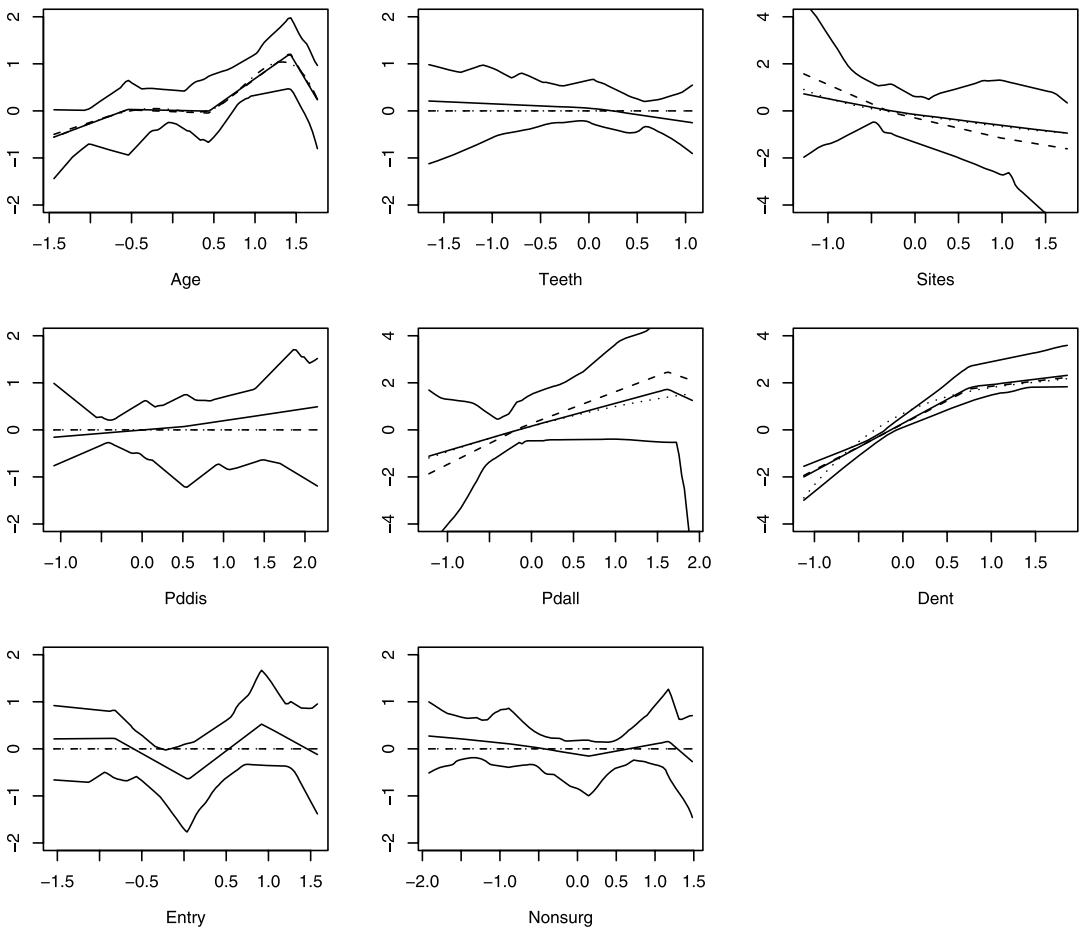

Figure 3. Periodontal disease application. The estimated component functions from the SCAD procedure with linear spline (dashed line) and cubic spline (dot–dash), and the QIF of the full model (solid line) using linear spline and the exchangeable working correlation. Two SCAD procedures give exactly zero estimates for variables: Teeth, Pddis, Entry, and Nonsurg. Also plotted are 95% bootstrap confidence intervals of the component functions from the QIF procedure.

$\mathcal{M}_n$, as

$$\mathcal{M}_n = \left\{ s(\mathbf{x}) = s_0 + \sum_{l=1}^{d} s_l(x_l); s_0 \in \mathcal{H}_0, s_l \in \varphi_l^{0,n} \right\},$$

in which $\varphi_l^{0,n} = \{s_l(\cdot) : s_l \in \varphi_l, \langle s_l, \mathbf{1} \rangle = 0\}$, the centered polynomial spline space. Note that the definitions of $\mathcal{M}$ and $\mathcal{M}_n$ do not depend on the constraints that $\alpha_l \in \mathcal{L}_2^0$, $s_l \in \varphi_l^{0,n}$. We impose those constraints to ensure a unique additive decomposition of the functions in $\mathcal{M}$ and $\mathcal{M}_n$ almost surely. Furthermore, for any $\alpha \in \mathcal{M}$, define two norms that will be used later: $\|\alpha\|^2 = E\{\alpha^T(\mathbf{X})\alpha(\mathbf{X})\}$ and $\|\alpha\|_n^2 = \frac{1}{n}\sum_{i=1}^n \alpha^T(\mathbf{X}_i)\alpha(\mathbf{X}_i)$, where $\alpha(\mathbf{X}) = (\alpha(\mathbf{X}_1), \ldots, \alpha(\mathbf{X}_m))^T$ for $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_m^T)^T$.

Proof of Theorem 1

Lemma A.7 in the supplement and triangular inequality given that for each $l = 1, \ldots, d$,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\big[\widetilde{\alpha}_l\big(x_{ij}^{(l)}\big) - \alpha_l\big(x_{ij}^{(l)}\big)\big]^2 \le \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\big[\mathbf{B}_l^T\big(x_{ij}^{(l)}\big)\big(\widetilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^0\big)\big]^2 + ch^{2r}.$$

Therefore, it is sufficient to show that

$$\|\mathbf{B}_l^T(\widetilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^0)\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\big[\mathbf{B}_l^T\big(x_{ij}^{(l)}\big)(\widetilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^0)\big]^2 = O_p((nh)^{-1}). \tag{A.2}$$

Note that Lemmas A.10 and A.11 in the supplement entail that for any $\epsilon > 0$, there exists a sufficiently large $C > 0$ such that, as $n \to \infty$,

$$P\Big\{ \inf_{\boldsymbol{\beta}:\|\mathbf{B}^T(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_n = C(nh)^{-1/2}} Q_n(\boldsymbol{\beta}) > Q_n(\boldsymbol{\beta}_0) \Big\} > 1 - \epsilon.$$

Therefore,

$$P\big\{\|\mathbf{B}^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_n \le C(nh)^{-1/2}\big\} > 1 - \epsilon,$$

which entails that $\|\mathbf{B}^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_n = O_p((nh)^{-1/2})$. Furthermore, Lemma A.5 in the supplement entails that for each $l = 1, \ldots, d$, there exists a constant $C > 0$, such that

$$\|\mathbf{B}_l^T(\widetilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^0)\|_n \le C\|\mathbf{B}^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_n = O_p\big((nh)^{-1/2}\big).$$

Therefore, (A.2) is proven.

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012

Proof of Theorem 2

Let $L_n(\boldsymbol{\beta}) = Q_n(\boldsymbol{\beta}) + n\sum_{l=1}^{d} p_{\lambda_n}(\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l})$ be the object function in (4). Let

$$\widehat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}=(\beta_0,\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_s^T,\mathbf{0}^T,\ldots,\mathbf{0}^T)^T}{\operatorname{argmin}} Q_n(\boldsymbol{\beta}),$$

which leads to the spline QIF estimator of the first s function components, knowing that the rest are zero terms. As a special case of Theorem 1, we have $\|\mathbf{B}^T(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0)\|_n = O_p(1/\sqrt{nh})$. We want to show that for large n and any $\varepsilon > 0$, there exists a constant C large enough such that

$$P\Big(\inf_{\boldsymbol{\beta}:\|\mathbf{B}^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}^*)\|_n = C(nh)^{-1/2}} L_n(\boldsymbol{\beta}) > L_n(\widehat{\boldsymbol{\beta}}^*)\Big) \geq 1-\varepsilon. \tag{A.3}$$

As a result, this implies that $L_n(\cdot)$ has a local minimum in the ball $\{\boldsymbol{\beta}:\|\mathbf{B}^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)\|_n \leq C(nh)^{-1/2}\}$. Thus $\|\mathbf{B}^T(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^*)\|_n = O_p(1/\sqrt{nh})$. Further, the triangular inequality gives $\|\mathbf{B}^T\widehat{\boldsymbol{\beta}} - \alpha\|_n \leq \|\mathbf{B}^T(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^*)\|_n + \|\mathbf{B}^T(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0)\|_n + \|\mathbf{B}^T\boldsymbol{\beta}_0 - \alpha\|_n = O_p(1/\sqrt{nh} + h^r)$. It proves the theorem by noting that $h = n^{-1/(2r+1)}$.

To show (A.3), using $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\cdot) \geq 0$, we have

$$\begin{aligned} &L_n(\boldsymbol{\beta}) - L_n(\widehat{\boldsymbol{\beta}}^*) \\ &\quad \geq Q_n(\boldsymbol{\beta}) - Q_n(\widehat{\boldsymbol{\beta}}^*) + \sum_{l=1}^{s} n\big\{p_{\lambda_n}\big(\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l}\big) - p_{\lambda_n}\big(\|\widehat{\boldsymbol{\beta}}_l^*\|_{\mathbf{K}_l}\big)\big\}. \end{aligned}$$

By similar arguments as in the proof of theorem 2 of Xue (2009), if $\lambda_n \to 0$, then for any $\boldsymbol{\beta}$ with $\|\mathbf{B}^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)\|_n = C(nh)^{-1/2}$, we have $\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l} \geq a\lambda_n$, and $\|\widehat{\boldsymbol{\beta}}_l^*\|_{\mathbf{K}_l} \geq a\lambda_n$ for each $l = 1,\ldots,s$, when n is sufficiently large. Therefore, when n is sufficiently large,

$$\sum_{l=1}^{s}\big\{p_{\lambda_n}\big(\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l}\big) - p_{\lambda_n}\big(\|\widehat{\boldsymbol{\beta}}_l^*\|_{\mathbf{K}_l}\big)\big\} = 0,$$

by the definition of the SCAD penalty function. Furthermore,

$$\begin{aligned} &Q_n(\boldsymbol{\beta}) - Q_n(\widehat{\boldsymbol{\beta}}^*) \\ &\quad = (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)^T\dot{Q}_n(\widehat{\boldsymbol{\beta}}^*) \\ &\qquad + \tfrac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)^T\ddot{Q}_n(\widehat{\boldsymbol{\beta}}^*)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)\{1 + o_p(1)\} \end{aligned} \tag{A.4}$$

with $\dot{Q}_n$ and $\ddot{Q}_n$ being the gradient vector and Hessian matrix of Q_n, respectively. Following Qu, Lindsay, and Li (2000) and Lemma A.9 in the supplement, for any $\boldsymbol{\beta}$, with $\|\mathbf{B}^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)\|_n = C(nh)^{-1/2}$, we have

$$\begin{aligned} &(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)^T\dot{Q}_n(\widehat{\boldsymbol{\beta}}^*) \\ &\quad = n(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)^T\mathbf{G}_n^T(\widehat{\boldsymbol{\beta}}^*)\mathbf{C}_n^{-1}(\widehat{\boldsymbol{\beta}}^*)\mathbf{G}_n(\widehat{\boldsymbol{\beta}}^*)\{1 + o_p(1)\} \\ &\quad \asymp Ch^{-1} \end{aligned}$$

and

$$\begin{aligned} &(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)^T\ddot{Q}_n(\widehat{\boldsymbol{\beta}}^*)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*) \\ &\quad = n(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)^T\dot{\mathbf{G}}_n^T(\widehat{\boldsymbol{\beta}}^*)\mathbf{C}_n^{-1}(\widehat{\boldsymbol{\beta}}^*)\dot{\mathbf{G}}_n(\widehat{\boldsymbol{\beta}}^*)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^*)\{1 + o_p(1)\} \\ &\quad \asymp C^2h^{-1}, \end{aligned}$$

where $\dot{\mathbf{G}}_n$ is the first-order derivative of $\mathbf{G}_n$. Therefore, by choosing C sufficiently large, the second term on the right side of (A.4) dominates its first term. Therefore, (A.3) holds when C and n are sufficiently large. This completes the proof of Theorem 2.

Proof of Theorem 3

Let $\Theta_1 = \{\boldsymbol{\beta}:\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_s^T, \mathbf{0}^T,\ldots,\mathbf{0}^T)^T, \|\mathbf{B}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_n = O_p(1/\sqrt{nh})\}$. For $l = s+1,\ldots,d$, define $\Theta_l = \{\boldsymbol{\beta}:\boldsymbol{\beta} = (0,\mathbf{0},\ldots,\mathbf{0},\boldsymbol{\beta}_l^T,\mathbf{0},\ldots,\mathbf{0})^T, \|\mathbf{B}^T\boldsymbol{\beta}\|_n = O_p(1/\sqrt{nh})\}$.

It is sufficient to show that uniformly for any $\boldsymbol{\beta} \in \Theta_1$ and $\boldsymbol{\beta}_l^* \in \Theta_l$, $L_n(\boldsymbol{\beta}) \leq L_n(\boldsymbol{\beta} + \boldsymbol{\beta}_l^*)$, with probability tending to 1 as $n \to \infty$. Note that, similar to the proof of Theorem 2,

$$\begin{aligned} &L_n(\boldsymbol{\beta} + \boldsymbol{\beta}_l^*) - L_n(\boldsymbol{\beta}) \\ &\quad = Q_n(\boldsymbol{\beta} + \boldsymbol{\beta}_l^*) - Q_n(\boldsymbol{\beta}) + np_{\lambda_n}\big(\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l}\big) \\ &\quad = \boldsymbol{\beta}_l^{*T}\dot{Q}_n(\widehat{\boldsymbol{\beta}}^*) + \frac{1}{2}\boldsymbol{\beta}_l^{*T}\ddot{Q}_n(\widehat{\boldsymbol{\beta}}^*)\boldsymbol{\beta}_l^*\{1 + o_p(1)\} + np_{\lambda_n}\big(\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l}\big) \\ &\quad = n\lambda_n\|\mathbf{B}^T\boldsymbol{\beta}_l\|_n\left\{\frac{R_n}{\lambda_n} + \frac{p'_{\lambda_n}(w)}{\lambda_n}\right\}\{1 + o_p(1)\}, \end{aligned}$$

where

$$R_n = \frac{\boldsymbol{\beta}_l^{*T}\dot{Q}_n(\widehat{\boldsymbol{\beta}}^*) + (1/2)\boldsymbol{\beta}_l^{*T}\ddot{Q}_n(\widehat{\boldsymbol{\beta}}^*)\boldsymbol{\beta}_l^*}{n\|\mathbf{B}^T\boldsymbol{\beta}_l\|_n} = O_p(1/\sqrt{nh}),$$

and w is a value between 0 and $\|\boldsymbol{\beta}_l\|_{\mathbf{K}_l}$. We complete the proof by observing that $\operatorname{plim}_{n\to\infty} R_n/\lambda_n = 0$, whereas

$$\liminf_{n\to\infty}\liminf_{w\to 0^+}\frac{p'_{\lambda_n}(w)}{\lambda_n} = 1.$$

SUPPLEMENTAL MATERIALS

Web Appendix: The Web appendix provides necessary lemmas and their proofs to support the proofs of the theorem in the Appendix. (WebAppendix.pdf)

[Received March 2010. Revised June 2010.]

REFERENCES

Berhane, K., and Tibshirani, R. J. (1998), "Generalized Additive Models for Longitudinal Data," *Canadian Journal Statistics*, 26, 517–535. [1518,1519]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1519,1521,1522]

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman & Hall. [1518]

He, X., Fung, W. K., and Zhu, Z. (2005), "Robust Estimation in Generalized Partial Linear Models for Clustered Data," *Journal of the American Statistical Association*, 100, 1176–1184. [1521]

Horowitz, J. L. (2001), "Nonparametric Estimation and a Generalized Additive Model With an Unknown Link Function," *Econometrica*, 69, 499–513. [1519]

Horowitz, J. L., and Mammen, E. (2007), "Rate-Optimal Estimation for a General Class of Nonparametric Regression Models With Unknown Link Functions," *The Annals of Statistics*, 35, 2589–2619. [1518]

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [1518]

Huang, J. Z. (1998), "Projection Estimation in Multiple Regression With Application to Functional ANOVA Models," *The Annals of Statistics*, 26, 242–272. [1519,1521]

——— (2003), "Local Asymptotics for Polynomial Spline Regression," *The Annals of Statistics*, 31, 1600–1635. [1521]

Huang, J. Z., and Yang, L. (2004), "Identification of Nonlinear Additive Autoregressive Models," *Journal of the Royal Statistical Society, Ser. B*, 66, 463–477. [1518]

Huang, J. Z., Wu, C. O., and Zhou, L. (2004), "Polynomial Spline Estimation and Inference for Varying Coefficient Models With Longitudinal Data," *Statistica Sinica*, 14, 763–788. [1522]

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22. [1519]

Linton, O. B., and Härdle, W. (1996), "Estimation of Additive Regression Models With Known Links," *Biometrika*, 83, 529–540. [1518]

Macke, J. H., Berens, P., Ecker, A. S., Tolias, A. S., and Bethge, M. (2008), "Generating Spike-Trains With Specified Correlation-Coefficients," *Neural Computation*, 21, 397–423. [1525]

Meier, L., van de Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society, Ser. B*, 70, 53–71. [1518]

——— (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [1518]

Qu, A., and Li, R. (2006), "Quadratic Inference Functions for Varying-Coefficient Models With Longitudinal Data," *Biometrics*, 62, 379–391. [1522]

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving Generalised Estimating Equations Using Quadratic Inference Functions," *Biometrika*, 87, 823–836. [1519,1520,1522,1529]

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [1518]

Stoner, J. A. (2000), "Analysis of Clustered Data: A Combined Estimating Equations Approach," Ph.D. thesis, University of Washington. [1526]

Wang, H., Li, R., and Tsai, C. (2008), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [1522]

Wang, L., Li, H., and Huang, J. Z. (2008), "Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements," *Journal of the American Statistical Association*, 103, 1556–1569. [1519]

Wang, N. (2003), "Marginal Nonparametric Kernel Regression Accounting for Within-Subject Correlation," *Biometrika*, 90, 43–52. [1518]

Xue, L. (2009), "Variable Selection in Additive Models," *Statistica Sinica*, 19, 1281–1296. [1518,1519,1521,1522,1529]

Xue, L., and Yang, L. (2006), "Additive Coefficient Modeling via Polynomial Spline," *Statistica Sinica*, 16, 1423–1446. [1519,1522]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Ser. B*, 68, 49–67. [1521]

Zhu, Z., Fung, W. K., and He, X. (2008), "On the Asymptotics of Marginal Regression Splines With Longitudinal Data," *Biometrika*, 95, 907–917. [1518]

Downloaded by [University of Wisconsin - Madison] at 10:29 02 March 2012