FOCUSED INFORMATION CRITERION AND MODEL AVERAGING FOR GENERALIZED ADDITIVE PARTIAL LINEAR MODELS

BY XINYU ZHANG[1] AND HUA LIANG[2]

Chinese Academy of Sciences and University of Rochester

We study model selection and model averaging in generalized additive partial linear models (GAPLMs). Polynomial spline is used to approximate nonparametric functions. The corresponding estimators of the linear parameters are shown to be asymptotically normal. We then develop a focused information criterion (FIC) and a frequentist model average (FMA) estimator on the basis of the quasi-likelihood principle and examine theoretical properties of the FIC and FMA. The major advantages of the proposed procedures over the existing ones are their computational expediency and theoretical reliability. Simulation experiments have provided evidence of the superiority of the proposed procedures. The approach is further applied to a real-world data example.

1. Introduction. Generalized additive models, which are a generalization of the generalized models and involve a summand of one-dimensional nonparametric functions instead of a summand of linear components, have been widely used to explore the complicated relationships between a response to treatment and predictors of interest [Hastie and Tibshirani (1990)]. Various attempts are still being made to balance the interpretation of generalized linear models and the flexibility of generalized additive models such as generalized additive partial linear models (GAPLMs), in which some of the additive component functions are linear, while the remaining ones are modeled nonparametrically [Härdle et al. (2004a, 2004b)]. A special case of a GAPLM with a single nonparametric component, the generalized partial linear model (GPLM), has been well studied in the literature; see, for example, Severini and Staniswalis (1994), Lin and Carroll (2001), Hunsberger (1994), Hunsberger et al. (2002) and Liang (2008). The profile quasi-likelihood procedure has generally been used, that is, the estimation of GPLM is made computationally feasible by the idea that estimates of the parameters can be found for a known nonparametric function, and an estimate of the nonparametric function can

be found for the estimated parameters. Severini and Staniswalis (1994) showed that the resulting estimators of the parameter are asymptotically normal and that estimators of the nonparametric functions are consistent in supremum norm. The computational algorithm involves searching for maxima of global and local likelihoods simultaneously. It is worthwhile to point out that studying GPLM is easier than studying GAPLMs, partly because there is only one nonparametric term in GPLM. Correspondingly, implementation of the estimation for GPLM is simpler than for GAPLMs. Nevertheless, the GAPLMs are more flexible and useful than GPLM because the former allow several nonparametric terms for some covariates and parametric terms for others, and thus it is possible to explore more complex relationships between the response variables and covariates. For example, Shiboski (1998) used a GAPLM to study AIDS clinical trial data and Müller and Rönz (2000) used a GAPLM to carry out credit scoring. However, few theoretical results are available for GAPLMs, due to their general flexibility. In this article, we shall study estimation of GAPLMs using polynomial spline, establish asymptotic normality for the estimators of the linear parameters and develop a focused information criterion (FIC) for model selection and a frequentist model averaging (FMA) procedure in construction of the confidence intervals for the focus parameters with improved coverage probability.

We know that traditional model selection methods such as the Akaike information criterion [AIC, Akaike (1973)] and the Bayesian information criterion [BIC, Schwarz (1978)] aim to select a model with good overall properties, but the selected model is not necessarily good for estimating a specific parameter under consideration, which may be a function of the model parameters; see an inspiring example in Section 4.4 of Claeskens and Hjort (2003). Exploring the data set from the Wisconsin epidemiologic study of diabetic retinopathy, Claeskens, Croux and van Kerckhoven (2006) also noted that different models are suitable for different patient groups. This occurrence has been confirmed by Hand and Vinciotti (2003) and Hansen (2005). Motivated by this concern, Claeskens and Hjort (2003) proposed a new model selection criterion, FIC, which is an unbiased estimate of the limiting risk for the limit distribution of an estimator of the focus parameter, and systematically developed a general asymptotic theory for the proposed criterion. More recently, FIC has been studied in several models. Hjort and Claeskens (2006) developed the FIC for the Cox hazard regression model and applied it to a study of skin cancer; Claeskens, Croux and van Kerckhoven (2007) introduced the FIC for autoregressive models and used it to predict the net number of new personal life insurance policies for a large insurance company.

The existing model selection methods may arrive at a model which is thought to be able to capture the main information of the data, and to be decided in advance in data analysis. Such an approach may lead to the ignoring of uncertainty introduced by model selection. Thus, the reported confidence intervals are too narrow or shift away from the correct location, and the corresponding coverage probabilities of the resulting confidence intervals can substantially deviate from the nominal

level [Danilov and Magnus (2004) and Shen, Huang and Ye (2004)]. Model averaging, as an alternative to model selection, not only provides a kind of insurance against selecting a very poor model, but can also avoid model selection instability [Yang (2001) and Leung and Barron (2006)] by weighting/smoothing estimators across several models, instead of relying entirely on a single model selected by some model selection criterion. As a consequence, analysis of the distribution of model averaging estimators can improve coverage probabilities. This strategy has been adopted and studied in the literature, for example, Draper (1995), Buckland, Burnham and Augustin (1997), Burnham and Anderson (2002), Danilov and Magnus (2004) and Leeb and Pöstcher (2006). A seminal work, Hjort and Claeskens (2003), developed asymptotic distribution theories for estimation and inference after model selection and model averaging across parametric models. See Claeskens and Hjort (2008) for a comprehensive survey on FIC and model averaging.

FIC and FMA have been well studied for parametric models. However, few efforts have been made to study FIC and FMA for semiparametric models. To the best of our knowledge, only Claeskens and Carroll (2007) studied FMA in semiparametric partial linear models with a univariate nonparametric component. The existing results are hard to extend directly to GAPLMs, for the following reasons: (i) there exist nonparametric components in GAPLMs, so the ordinary likelihood method cannot be directly used in estimation for GAPLMs; (ii) unlike the semiparametric partial linear models in Claeskens and Carroll (2007), GAPLMs allow for multivariate covariate consideration in nonparametric components and also allow for the mean of the response variable to be connected to the covariates by a link function, which means that the binary/count response variable can be considered in the model. Thus, to develop FIC and FMA procedures for GAPLMs and to establish asymptotic properties for these procedures are by no means straightforward to achieve. Aiming at these two goals, we first need to appropriately estimate the coefficients of the parametric components (hereafter, we call these coefficients "linear parameters").

There are two commonly used estimation approaches for GAPLMs: the first is local scoring backfitting, proposed by Buja, Hastie and Tibshirani (1989); the second is an application of the marginal integration approach on the nonparametric component [Linton and Nielsen (1995)]. However, theoretical properties of the former are not well understood since it is only defined implicitly as the limit of a complicated iterative algorithm, while the latter suffers from the *curse of dimensionality* [Härdle et al. (2004a)], which may lead to an increase in the computational burden and which also conflicts with the purpose of using a GAPLM, that is, dimension reduction. Therefore, in this article, we apply polynomial spline to approximate nonparametric functions in GAPLMs. After the spline basis is chosen, the nonparametric components are replaced by a linear combination of spline basis, then the coefficients can be estimated by an efficient one-step maximizing procedure. Since the polynomial-spline-based method solves much smaller systems of equations than kernel-based methods that solve larger systems (which may

lead to identifiability problems), our polynomial-spline-based procedures can substantially reduce the computational burden. See a similar discussion about this computational issue in Yu, Park and Mammen (2008), in the generalized additive models context.

The use of polynomial spline in generalized nonparametric models can be traced back to Stone (1986), where the rate of convergence of the polynomial spline estimates for the generalized additive model were first obtained. Stone (1994) and Huang (1998) investigated the polynomial spline estimation for the generalized functional ANOVA model. In a widely discussed paper, Stone et al. (1997) presented a completely theoretical setting of polynomial spline approximation, with applications to a wide array of statistical problems, ranging from least-squares regression, density and conditional density estimation, and generalized regression such as logistic and Poisson regression, to polychotomous regression and hazard regression. Recently, Xue and Yang (2006) studied estimation in the additive coefficient model with continuous response using polynomial spline to approximate the coefficient functions. Sun, Kopciuk and Lu (2008) used polynomial spline in partially linear single-index proportional hazards regression models. Fan, Feng and Song (2009) applied polynomial spline to develop nonparametric independence screening in sparse ultra-high-dimensional additive models. Few attempts have been made to study polynomial spline for GAPLMs, due to the extreme technical difficulties involved.

The remainder of this article is organized as follows. Section 2 sets out the model framework and provides the polynomial spline estimation and asymptotic normality of estimators. Section 3 introduces the FIC and FMA procedures and constructs confidence intervals for the focus parameters on a basis of FMA estimators. A simulation study and real-world data analysis are presented in Sections 4 and 5, respectively. Regularity conditions and technical proofs are presented in the Appendix.

2. Model framework and estimation. We consider a GAPLM where the response Y is related to covariates $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in R^p$ and $\mathbf{Z} = (Z_1, \ldots, Z_d)^{\mathrm{T}} \in R^d$. Let the unknown mean response $\mathbf{u}(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ and the conditional variance function be defined by a known positive function V, $\mathrm{var}(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = V\{\mathbf{u}(\mathbf{x}, \mathbf{z})\}$. In this article, the mean function $\mathbf{u}$ is defined via a known link function g by an additive linear function

$$(2.1) \qquad g\{\mathbf{u}(\mathbf{x}, \mathbf{z})\} = \sum_{\alpha=1}^{p} \eta_\alpha(x_\alpha) + \mathbf{z}^{\mathrm{T}}\beta,$$

where x_α is the αth element of x, $\boldsymbol{\beta}$ is a d-dimensional regression parameter and the η_α's are unknown smooth functions. To ensure identifiability, we assume that $E\{\eta_\alpha(X_\alpha)\} = 0$ for $1 \le \alpha \le p$.

Let $\beta = (\beta_c^{\mathrm{T}}, \beta_u^{\mathrm{T}})^{\mathrm{T}}$ be a vector with $d = d_c + d_u$ components, where β_c consists of the first d_c parameters of β (which we certainly wish to be in the selected

model) and β_u consists of the remaining d_u parameters (for which we are unsure whether or not they should be included in the selected model). In what follows, we call the elements of $\mathbf{z}$ corresponding to β_c and β_u the *certain* and *exploratory* variables, respectively. As in the literature on FIC, we consider a local misspecification framework where the true value of the parameter vector β is $\beta_0 = (\beta_{c,0}^{\mathrm{T}}, \delta^{\mathrm{T}}/\sqrt{n})^{\mathrm{T}}$, with δ being a $d_u \times 1$ vector; that is, the true model is away from the deduced model with a distance $O(1/\sqrt{n})$. This framework indicates that squared model biases and estimator variances are both of size $O(1/n)$, the most possible large-sample approximations. Some arguments related to this framework appear in Hjort and Claeskens (2003, 2006).

Denote by $\beta_S = (\beta_c^{\mathrm{T}}, \beta_{u,S}^{\mathrm{T}})^{\mathrm{T}}$ the parameter vector in the Sth submodel, in the same sense as β, with $\beta_{u,S}$ being a $d_{u,S}$-subvector of β_u. Let π_S be the projection matrix of size $d_{u,S} \times d_u$ mapping β_u to $\beta_{u,S}$. With d_u exploratory covariates, our setup allows 2^{d_u} extended models to choose among. However, it is not necessary to deal with all 2^{d_u} possible models and one is free to consider only a few relevant submodels (unnecessarily nested or ordered) to be used in the model selection or averaging. A special example is the James–Stein-type estimator studied by Kim and White (2001), which is a weighted summand of the estimators based on the reduced model ($d_{u,S} = 0$) and the full model ($d_{u,S} = d_u$). So, the covariates in the Sth submodel are $\mathbf{X}$ and $\Pi_S \mathbf{Z}$, where $\Pi_S = \mathrm{diag}(I_{d_c}, \pi_S)$. To save space, we generally ignore the dimensions of zero vectors/matrices and identity matrices, simply denoting them by 0 and I, respectively. If necessary, we will write their dimensions explicitly. In the remainder of this section, we shall investigate polynomial spline estimation for $(\beta_{c,0}^{\mathrm{T}}, 0)$ based on the Sth submodel and establish a theoretical property for the resulting estimators.

Let $\eta_0 = \sum_{\alpha=1}^p \eta_{0,\alpha}(x_\alpha)$ be the true additive function and the covariate X_α be distributed on a compact interval $[a_\alpha, b_\alpha]$. Without loss of generality, we take all intervals $[a_\alpha, b_\alpha] = [0, 1]$ for $\alpha = 1, \ldots, p$. Noting (A.7) in Appendix A.2, under some smoothness assumptions in Appendix A.1, η_0 can be well approximated by spline functions. Let $\mathcal{S}_n$ be the space of polynomial splines on $[0, 1]$ of degree $\varrho \geq 1$. We introduce a knot sequence with J interior knots, $k_{-\varrho} = \cdots = k_{-1} = k_0 = 0 < k_1 < \cdots < k_J < 1 = k_{J+1} = \cdots = k_{J+\varrho+1}$, where $J \equiv J_n$ increases when sample size n increases and the precise order is given in condition (C6). Then, $\mathcal{S}_n$ consists of functions ς satisfying the following:

(i) ς is a polynomial of degree ϱ on each of the subintervals $[k_j, k_{j+1})$, $j = 0, \ldots, J_n - 1$, and the last subinterval is $[k_{J_n}, 1]$;
(ii) for $\varrho \geq 2$, ς is $(\varrho - 1)$-times continuously differentiable on $[0, 1]$.

For simplicity of proof, equally spaced knots are used. Let $h = 1/(J_n + 1)$ be the distance between two consecutive knots.

Let $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, be independent copies of $(Y, \mathbf{X}, \mathbf{Z})$. In the Sth submodel, we consider the additive spline estimates of η_0 based on the independent random sample $(Y_i, \mathbf{X}_i, \Pi_S \mathbf{Z}_i)$, $i = 1, \ldots, n$. Let $\mathcal{G}_n$ be the collection of functions

η with the additive form $\eta(\mathbf{x}) = \sum_{\alpha=1}^{p} \eta_\alpha(x_\alpha)$, where each component function $\eta_\alpha \in \mathcal{S}_n$.

We would like to find a function $\eta \in \mathcal{G}_n$ and a value of $\boldsymbol{\beta}_S$ that maximize the quasi-likelihood function

$$(2.2) \qquad L(\eta, \boldsymbol{\beta}_S) = \frac{1}{n} \sum_{i=1}^{n} Q[g^{-1}\{\eta(\mathbf{X}_i) + (\Pi_S \mathbf{Z}_i)^{\mathrm{T}} \boldsymbol{\beta}_S\}, Y_i], \qquad \eta \in \mathcal{G}_n,$$

where $Q(m, y)$ is the quasi-likelihood function satisfying $\frac{\partial Q(m,y)}{\partial m} = \frac{y-m}{V(m)}$.

For the αth covariate x_α, let $b_{j,\alpha}(x_\alpha)$ be the B-spline basis function of degree ϱ. For any $\eta \in \mathcal{G}_n$, one can write $\eta(\mathbf{x}) = \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{b}(\mathbf{x})$, where $\mathbf{b}(\mathbf{x}) = \{b_{j,\alpha}(x_\alpha), j = -\varrho, \ldots, J_n, \alpha = 1, \ldots, p\}^{\mathrm{T}}$ are the spline basis functions and $\boldsymbol{\gamma} = \{\gamma_{j,\alpha}, j = -\varrho, \ldots, J_n, \alpha = 1, \ldots, p\}^{\mathrm{T}}$ is the spline coefficient vector. Thus, the maximization problem in (2.2) is equivalent to finding values of $\boldsymbol{\beta}_S^*$ and $\boldsymbol{\gamma}^*$ that maximize

$$(2.3) \qquad \frac{1}{n} \sum_{i=1}^{n} Q[g^{-1}\{\boldsymbol{\gamma}^{*\mathrm{T}} \mathbf{b}(\mathbf{X}_i) + (\Pi_S \mathbf{Z}_i)^{\mathrm{T}} \boldsymbol{\beta}_S^*\}, Y_i].$$

We denote the maximizers as $\widehat{\boldsymbol{\beta}}_S^*$ and $\widehat{\boldsymbol{\gamma}}_S^* = \{\widehat{\gamma}_{S,j,\alpha}^*, j = -\varrho, \ldots, J_n, \alpha = 1, \ldots, p\}^{\mathrm{T}}$. The spline estimator of η_0 is then $\widehat{\eta}_S^* = \widehat{\boldsymbol{\gamma}}_S^{*\mathrm{T}} \mathbf{b}(\mathbf{x})$ and the centered spline estimators of each component function are

$$\widehat{\eta}_{S,\alpha}^*(x_\alpha) = \sum_{j=-\varrho}^{J_n} \widehat{\gamma}_{S,j,\alpha}^* b_{j,\alpha}(x_\alpha) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=-\varrho}^{J_n} \widehat{\gamma}_{S,j,\alpha}^* b_{j,\alpha}(X_{i\alpha}), \qquad \alpha = 1, \ldots, p.$$

The above estimation approach can be easily implemented with commonly used statistical software since the resulting model is a generalized linear model.

For any measurable functions φ_1, φ_2 on $[0, 1]^p$, define the empirical inner product and the corresponding norm as

$$\langle \varphi_1, \varphi_2 \rangle_n = n^{-1} \sum_{i=1}^{n} \{\varphi_1(\mathbf{X}_i)\varphi_2(\mathbf{X}_i)\}, \qquad \|\varphi\|_n^2 = n^{-1} \sum_{i=1}^{n} \varphi^2(\mathbf{X}_i).$$

If φ_1 and φ_2 are L^2-integrable, define the theoretical inner product and the corresponding norm as $\langle \varphi_1, \varphi_2 \rangle = E\{\varphi_1(\mathbf{X})\varphi_2(\mathbf{X})\}$, $\|\varphi\|_2^2 = E\varphi^2(\mathbf{X})$, respectively. Let $\|\varphi\|_{n\alpha}^2$ and $\|\varphi\|_{2\alpha}^2$ be the empirical and theoretical norms, respectively, of a function φ on $[0, 1]$, that is,

$$\|\varphi\|_{n\alpha}^2 = n^{-1} \sum_{i=1}^{n} \varphi^2(X_{i\alpha}), \qquad \|\varphi\|_{2\alpha}^2 = E\varphi^2(X_\alpha) = \int_0^1 \varphi^2(x_\alpha) f_\alpha(x_\alpha) \, dx_\alpha,$$

where $f_\alpha(x_\alpha)$ is the density function of X_α.

Define the centered version spline basis for any $\alpha = 1, \ldots, p$ and $j = -\varrho + 1, \ldots, J_n$, $b_{j,\alpha}^*(x_\alpha) = b_{j,\alpha}(x_\alpha) - \|b_{j,\alpha}\|_{2\alpha} / \|b_{j-1,\alpha}\|_{2\alpha} b_{j-1,\alpha}(x_\alpha)$, with the standardized version given by

$$\tag{2.4} B_{j,\alpha}(x_\alpha) = \frac{b_{j,\alpha}^*(x_\alpha)}{\|b_{j,\alpha}^*\|_{2\alpha}}.$$

Note that to find (γ^*, β_S^*) that maximizes (2.3) is mathematically equivalent to finding (γ, β_S) that maximizes

$$\tag{2.5} \ell(\boldsymbol{\gamma}, \boldsymbol{\beta}_S) = \frac{1}{n} \sum_{i=1}^{n} Q[g^{-1}\{\boldsymbol{\gamma}^{\mathrm{T}} \mathbf{B}(\mathbf{X}_i) + (\Pi_S \mathbf{Z}_i)^{\mathrm{T}} \boldsymbol{\beta}_S\}, Y_i],$$

where $\mathbf{B}(\mathbf{x}) = \{B_{j,\alpha}(x_\alpha), j = -\varrho + 1, \ldots, J_n, \alpha = 1, \ldots, p\}^{\mathrm{T}}$. Similarly to $\widehat{\boldsymbol{\beta}}_S^*$, $\widehat{\boldsymbol{\gamma}}_S^*$, $\widehat{\eta}_S^*$ and $\widehat{\eta}_{S,\alpha}^*$, we can define $\widehat{\boldsymbol{\beta}}_S$, $\widehat{\boldsymbol{\gamma}}_S$, $\widehat{\eta}_S$ and the centered spline estimators of each component function $\widehat{\eta}_{S,\alpha}(x_\alpha)$. In practice, the basis $\{b_{j,\alpha}(x_\alpha), j = -\varrho, \ldots, J_n, \alpha = 1, \ldots, p\}^{\mathrm{T}}$ is used for data analytic implementation and the mathematically equivalent expression (2.4) is convenient for asymptotic derivation.

Let $\rho_l(m) = \{\frac{dg^{-1}(m)}{dm}\}^l / V\{g^{-1}(m)\}$, $l = 1, 2$. Write $\mathbf{T} = (\mathbf{X}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$, $m_0(\mathbf{T}) = \eta_0(\mathbf{X}) + \mathbf{Z}^{\mathrm{T}} \boldsymbol{\beta}_0$ and $\varepsilon = Y - g^{-1}\{m_0(\mathbf{T})\}$. $\mathbf{T}_i$, $m_0(\mathbf{T}_i)$ and ε_i are defined in the same way after replacing $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{T}$ by $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{T}_i$, respectively. Write

$$\boldsymbol{\Gamma}(\mathbf{x}) = \frac{E[\mathbf{Z}\rho_1\{m_0(\mathbf{T})\}|\mathbf{X} = \mathbf{x}]}{E[\rho_1\{m_0(\mathbf{T})\}|\mathbf{X} = \mathbf{x}]}, \qquad \psi(\mathbf{T}) = \mathbf{Z} - \boldsymbol{\Gamma}(\mathbf{X}),$$

$$\mathbf{G}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \rho_1\{m_0(\mathbf{T}_i)\} \psi(\mathbf{T}_i), \qquad \mathbf{D} = E[\rho_1\{m_0(\mathbf{T})\} \psi(\mathbf{T})\{\psi(\mathbf{T})\}^{\mathrm{T}}]$$

and $\boldsymbol{\Sigma} = E[\rho_1^2\{m_0(\mathbf{T})\}\varepsilon^2 \psi(\mathbf{T})\{\psi(\mathbf{T})\}^{\mathrm{T}}]$.

The following theorem shows that the estimators $\widehat{\beta}_S$ on the basis of the Sth submodel are asymptotically normal.

THEOREM 1. *Under the local misspecification framework and conditions* (C1)–(C11) *in the Appendix,*

$$\sqrt{n}\{\widehat{\beta}_S - (\beta_{c,0}^{\mathrm{T}}, 0)^{\mathrm{T}}\}$$

$$= -(\Pi_S \mathbf{D} \Pi_S^{\mathrm{T}})^{-1} \Pi_S \mathbf{G}_n + (\Pi_S \mathbf{D} \Pi_S^{\mathrm{T}})^{-1} \Pi_S \mathbf{D} \begin{pmatrix} 0 \\ \delta \end{pmatrix} + o_p(1)$$

$$\xrightarrow{d} -(\Pi_S \mathbf{D} \Pi_S^{\mathrm{T}})^{-1} \Pi_S \mathbf{G} + (\Pi_S \mathbf{D} \Pi_S^{\mathrm{T}})^{-1} \Pi_S \mathbf{D} \begin{pmatrix} 0 \\ \delta \end{pmatrix}$$

with $\mathbf{G}_n \xrightarrow{d} \mathbf{G} \sim N(0, \boldsymbol{\Sigma})$, *where* "$\xrightarrow{d}$" *denotes convergence in distribution.*

REMARK 1. If the link function g is identical and there is only one nonparametric component (i.e., $p = 1$), then the result of Theorem 1 will simplify to those of Theorems 3.1–3.4 of Claeskens and Carroll (2007) under the corresponding submodels.

REMARK 2. Assume that $d_u = 0$. Theorem 1 indicates that the polynomial-spline-based estimators of the linear parameters are asymptotically normal. This is the first explicitly theoretical result on asymptotic normality for estimation of the linear parameters in GAPLMs and is of independent interest and importance. This theorem also indicates that although there are several nonparametric functions and their polynomial approximation deduces biases for the estimators of each nonparametric component, these biases do not make the estimators of β biased under condition (C6) imposed on the number of knots.

3. Focused information criterion and frequentist model averaging. In this section, based on the asymptotic result in Section 2, we develop an FIC model selection for GAPLMs, an FMA estimator, and propose a proper confidence interval for the focus parameters.

3.1. *Focused information criterion.* Let $\mu_0 = \mu(\beta_0) = \mu(\beta_{c,0}, \delta/\sqrt{n})$ be a focus parameter. Assume that the partial derivatives of $\mu(\beta_0)$ are continuous in a neighborhood of $\beta_{c,0}$. Note that, in the Sth submodel, μ_0 can be estimated by $\widehat{\mu}_S = \mu([I_{d_c}, 0_{d_c \times d_u}]\Pi_S^{\mathrm{T}}\widehat{\beta}_S, [0_{d_u \times d_c}, I_{d_u}]\Pi_S^{\mathrm{T}}\widehat{\beta}_S)$. We now show the asymptotic normality of $\widehat{\mu}_S$. Write $\mathbf{R}_S = \Pi_S^{\mathrm{T}}(\Pi_S \mathbf{D} \Pi_S^{\mathrm{T}})^{-1}\Pi_S$, $\mu_c = \frac{\partial \mu(\beta_c, \beta_u)}{\partial \beta_c}|_{\beta_c = \beta_{c,0}, \beta_u = 0}$, $\mu_u = \frac{\partial \mu(\beta_c, \beta_u)}{\partial \beta_u}|_{\beta_c = \beta_{c,0}, \beta_u = 0}$ and $\mu_\beta = (\mu_c^{\mathrm{T}}, \mu_u^{\mathrm{T}})^{\mathrm{T}}$.

THEOREM 2. *Under the local misspecification framework and conditions* (C1)–(C11) *in the Appendix, we have*

$$\sqrt{n}(\widehat{\mu}_S - \mu_0) = -\mu_\beta^{\mathrm{T}} \mathbf{R}_S \mathbf{G}_n + \mu_\beta^{\mathrm{T}}(\mathbf{R}_S \mathbf{D} - I)\begin{pmatrix} 0 \\ \delta \end{pmatrix} + o_p(1)$$

$$\xrightarrow{d} \Lambda_S \equiv -\mu_\beta^{\mathrm{T}} \mathbf{R}_S \mathbf{G} + \mu_\beta^{\mathrm{T}}(\mathbf{R}_S \mathbf{D} - I)\begin{pmatrix} 0 \\ \delta \end{pmatrix}.$$

Recall $\mathbf{G} \sim N(0, \Sigma)$. A direct calculation yields

$$(3.1) \quad E(\Lambda_S^2) = \mu_\beta^{\mathrm{T}} \left\{ \mathbf{R}_S \Sigma \mathbf{R}_S + (\mathbf{R}_S \mathbf{D} - I)\begin{pmatrix} 0 \\ \delta \end{pmatrix}\begin{pmatrix} 0 \\ \delta \end{pmatrix}^{\mathrm{T}} (\mathbf{R}_S \mathbf{D} - I)^{\mathrm{T}} \right\} \mu_\beta.$$

Let $\widehat{\delta}$ be the estimator of δ by the full model. Then, from Theorem 1, we know that

$$\widehat{\delta} = -[0, I]\mathbf{D}^{-1}\mathbf{G}_n + \delta + o_p(1).$$

If we define $\Delta = -[0, I]\mathbf{D}^{-1}\mathbf{G} + \delta \sim N(\delta, [0, I]\mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}[0, I]^{\mathrm{T}})$, then $\widehat{\delta} \xrightarrow{d} \Delta$. Following Claeskens and Hjort (2003) and (3.1), we define the FIC of the Sth submodel as

$$
\begin{aligned}
\mathrm{FIC}_S = \mu_\beta^{\mathrm{T}} \Bigg\{ &\mathbf{R}_S \mathbf{\Sigma} \mathbf{R}_S + (\mathbf{R}_S \mathbf{D} - I) \begin{pmatrix} 0 \\ \delta \end{pmatrix} \begin{pmatrix} 0 \\ \delta \end{pmatrix}^{\mathrm{T}} (\mathbf{R}_S \mathbf{D} - I)^{\mathrm{T}} \\
&- (\mathbf{R}_S \mathbf{D} - I) \begin{pmatrix} 0 & 0 \\ 0 & I_{d_u} \end{pmatrix} \mathbf{D}^{-1} \mathbf{\Sigma} \mathbf{D}^{-1} \begin{pmatrix} 0 & 0 \\ 0 & I_{d_u} \end{pmatrix} (\mathbf{R}_S \mathbf{D} - I)^{\mathrm{T}} \Bigg\} \mu_\beta,
\end{aligned}
$$

(3.2)

which is an approximately unbiased estimator of the mean squared error when $\sqrt{n}\mu_0$ is estimated by $\sqrt{n}\widehat{\mu}_S$. This FIC can be used for choosing a proper submodel relying on the parameter of interest.

3.2. *Frequentist model averaging.* As mentioned previously, an average estimator is an alternative to a model selection estimator. There are at least two advantages to the use of an average estimator. First, an average estimator often reduces mean square error in estimation because it avoids ignoring useful information from the form of the relationship between response and covariates and it provides a kind of insurance against selecting a very poor submodel. Second, model averaging procedures can be more stable than model selection, for which small changes in the data often lead to a significant change in model choice. Similar discussions of this issue appear in Bates and Granger (1969) and Leung and Barron (2006).

By choosing a submodel with the minimum value of FIC, the FIC estimators of μ can be written as $\widehat{\mu}_{\mathrm{FIC}} = \sum_S \mathbf{I}(\mathrm{FIC}$ selects the Sth submodel$)\widehat{\mu}_S$, where $\mathbf{I}(\cdot)$, an indicator function, can be thought of as a weight function depending on the data via $\widehat{\delta}$, yet it just takes value either 0 or 1. To smooth estimators across submodels, we may formulate the model average estimator of μ as

$$
\widehat{\mu} = \sum_S w(S|\widehat{\delta})\widehat{\mu}_S, \tag{3.3}
$$

where the weights $w(S|\widehat{\delta})$ take values in the interval $[0, 1]$ and their sum equals 1. It is readily seen that smoothed AIC, BIC and FIC estimators investigated in Hjort and Claeskens (2003) and Claeskens and Carroll (2007) share this form. The following theorem shows an asymptotic property for the general model average estimators $\widehat{\mu}$ defined in (3.3) under certain conditions.

THEOREM 3. *Under the local misspecification framework and conditions* (C1)–(C11) *in the* Appendix, *if the weight functions have at most a countable number of discontinuities, then*

$$
\begin{aligned}
\sqrt{n}(\widehat{\mu} - \mu_0) &= -\mu_\beta^{\mathrm{T}} \mathbf{D}^{-1} \mathbf{G}_n + \mu_\beta^{\mathrm{T}} \left\{ Q(\widehat{\delta}) \begin{pmatrix} 0 \\ \widehat{\delta} \end{pmatrix} - \begin{pmatrix} 0 \\ \widehat{\delta} \end{pmatrix} \right\} + o_p(1) \\
&\xrightarrow{d} \Lambda \equiv -\mu_\beta^{\mathrm{T}} \mathbf{D}^{-1} \mathbf{G} + \mu_\beta^{\mathrm{T}} \left\{ Q(\Delta) \begin{pmatrix} 0 \\ \Delta \end{pmatrix} - \begin{pmatrix} 0 \\ \Delta \end{pmatrix} \right\},
\end{aligned}
$$

where $Q(\cdot) = \sum_S w(s|\cdot) \mathbf{R}_S \mathbf{D}$ and Δ is defined in Section 3.1.

Referring to the above theorems, we construct a confidence interval for μ based on the model average estimator $\hat{\mu}$, as follows. Assume that $\hat{\kappa}^2$ is a consistent estimator of $\mu_\beta^{\mathrm{T}} \mathbf{D}^{-1} \boldsymbol{\Sigma} \mathbf{D}^{-1} \mu_\beta$. It is easily seen that

$$\left[\sqrt{n}(\hat{\mu} - \mu_0) - \mu_\beta^{\mathrm{T}} \left\{ Q(\hat{\delta}) \begin{pmatrix} 0 \\ \hat{\delta} \end{pmatrix} - \begin{pmatrix} 0 \\ \hat{\delta} \end{pmatrix} \right\} \right] / \hat{\kappa} \xrightarrow{d} N(0, 1).$$

If we define the lower bound (low_n) and upper bound (up_n) by

$$(3.4) \qquad \hat{\mu} - \mu_\beta^{\mathrm{T}} \left\{ Q(\hat{\delta}) \begin{pmatrix} 0 \\ \hat{\delta} \end{pmatrix} - \begin{pmatrix} 0 \\ \hat{\delta} \end{pmatrix} \right\} / \sqrt{n} \mp z_J \hat{\kappa} / \sqrt{n},$$

where z_J is the Jth standard normal quantile, then we have $\Pr\{\mu_0 \in (low_n, up_n)\} \to 2\Phi(z_J) - 1$, where $\Phi(\cdot)$ is a standard normal distribution function. Therefore, the interval (low_n, up_n) can be used as a confidence interval for μ_0 with asymptotic level $2\Phi(z_J) - 1$.

REMARK 3. Note that the limit distribution of $\sqrt{n}(\hat{\mu} - \mu_0)$ is a nonlinear mixture of several normal variables. As argued in Hjort and Claeskens (2006), a direct construction of a confidence interval based on Theorem 3 may not be easy. The confidence interval based on (3.4) is better in terms of coverage probability and computational simplicity, as promoted in Hjort and Claeskens (2003) and advocated by Claeskens and Carroll (2007).

REMARK 4. A referee has asked whether the focus parameter can depend on the nonparametric function η_0. Our answer is "yes." For instance, we consider a general focus parameter, $\eta_0(\mathbf{x}) + \mu_0$, a summand of μ_0, which we have studied, and a nonparametric value at $\mathbf{x}$. We may continue to get an estimator of $\eta_0(\mathbf{x}) + \mu_0$ by minimizing (3.2) and then model-averaging estimators by weighting the estimators of μ_0 and η_0 as in (3.3). However, the underlying FMA estimators are not root-n consistent because the bias of these estimators is proportional to the bias of the estimators of η_0, which is larger than $n^{-1/2}$, whereas we can establish their rates of convergence using easier arguments than those employed in the proof of Theorem 3. Even though the focus parameters generally depend on μ_0 and η_0 of form $H(\mu_0, \eta_0)$ for a given function $H(\cdot, \cdot)$, the proposed method can be still applied. However, to develop asymptotic properties for the corresponding FMA estimators depends on the form of $H(\cdot, \cdot)$ and will require further investigation. We omit the details. Our numerical studies below follow these proposals when the focus parameters are related to the nonparametric functions.

4. Simulation study. We generated 1000 data sets consisting of $n = 200$ and 400 observations from the GAPLM

$$\text{logit}\{\Pr(Y_i = 1)\} = \eta_1(\mathbf{X}_{i,1}) + \eta_2(\mathbf{X}_{i,2}) + \mathbf{Z}_i^\mathrm{T}\beta$$
$$= \sin(2\pi\mathbf{X}_{i,1}) + 5\mathbf{X}_{i,2}^4 + 3\mathbf{X}_{i,2}^2 - 2 + \mathbf{Z}_i^\mathrm{T}\beta, \qquad i = 1, \ldots, n,$$

where: the true parameter $\beta = \{1.5, 2, r_0(2, 1, 3)/\sqrt{n}\}^\mathrm{T}$; $\mathbf{X}_{i,1}$ and $\mathbf{X}_{i,2}$ are independently uniformly distributed on $[0, 1]$; $\mathbf{Z}_{i,1}, \ldots, \mathbf{Z}_{i,5}$ are normally distributed with mean 0 and variance 1; when $\hbar_1 \neq \hbar_2$, the correlation between $\mathbf{Z}_{i,\hbar_1}$ and $\mathbf{Z}_{i,\hbar_2}$ is $\varpi^{|\hbar_1 - \hbar_2|}$ with $\varpi = 0$ or $\varpi = 0.5$; $\mathbf{Z}_i$ is independent of $\mathbf{X}_{i,1}$ and $\mathbf{X}_{i,2}$. We set the first two components of β to be in all submodels. The other three may or may not be present, so we have $2^3 = 8$ submodels to be selected or averaged across. r_0 varies from 1 or 4 to 7. Our focus parameters are (i) $\mu_1 = \beta_1$, (ii) $\mu_2 = \beta_2$, (iii) $\mu_3 = 0.75\beta_1 + 0.05\beta_2 - 0.3\beta_3 + 0.1\beta_4 - 0.06\beta_5$ and (iv) $\mu_4 = \eta_1(0.86) + \eta_2(0.53) + 0.32\beta_1 - 0.87\beta_2 - 0.33\beta_3 - 0.15\beta_4 + 0.13\beta_5$.

The cubic B-splines have been used to approximate the two nonparametric functions. We propose to select J_n using a BIC procedure. Based on condition (C6), the optimal order of J_n can be found in the range $(n^{1/(2\upsilon)}, n^{1/3})$. Thus, we propose to choose the optimal knot number, J_n, from a neighborhood of $n^{1/5.5}$. For our numerical examples, we have used $[2/3N_r, 4/3N_r]$, where $N_r = \text{ceiling}(n^{1/5.5})$ and the function ceiling($\cdot$) returns the smallest integer not less than the corresponding element. Under the full model, let the log-likelihood function be $l_n(N_n)$. The optimal knot number, N_n^{opt}, is then the one which minimizes the BIC value. That is,

$$(4.1) \qquad N_n^{\text{opt}} = \arg\min_{N_n \in [2/3N_r, 4/3N_r]} \{-2l_n(N_n) + q_n \log n\},$$

where q_n is the total number of parameters.

Four model selection or model averaging methods are compared in this simulation: AIC, BIC, FIC and the smoothed FIC (S-FIC). The smoothed FIC weights we have used are

$$w(S|\hat{\delta}) = \exp\left(-\frac{\text{FIC}_S}{\mu_\beta^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}\mu_\beta}\right) \Big/ \sum_{\text{all } S'} \exp\left(-\frac{\text{FIC}_{S'}}{\mu_\beta^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}\mu_\beta}\right),$$

a case of expression (5.4) in Hjort and Claeskens (2003). When using the FIC or S-FIC method, we estimate $\mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}$ by the covariance matrix of $\widehat{\beta}_{\text{full}}$ and estimate $\mathbf{D}$ by its sample mean, as advocated by Hjort and Claeskens (2003) and Claeskens and Carroll (2007). Thus, $\mathbf{\Sigma}$ can be calculated straightforwardly. Note that the subscript "*full*" denotes the estimator using the full model.

In this simulation, one of our purposes is to see whether the traditional selection methods like AIC and BIC lead to an overly optimistic coverage probability (CP) of a claimed confidence interval (CI). We consider a claimed 95% confidence interval (CI). The other purpose is to check the accuracy of estimators in terms of their

mean squared errors (MSE) $1/1000 \sum_j (\widehat{\mu}_a^{(j)} - \mu_a)^2$ for $a = 1, \ldots, 4$, where j denotes the jth replication. Our results are listed in Table 1.

These results indicate that the performance of both the FIC and S-FIC, especially the latter, is superior to that of AIC and BIC in terms of CP and mean squared error (MSE), regardless of whether the focus parameter depends on the nonparametric components or not. The CPs based on FIC and S-FIC are generally close to the nominal level. When the smallest CPs based on S-FIC and FIC are respectively 0.921 and 0.914, the corresponding CPs of AIC and BIC are only 0.860 and 0.843, respectively, which are much lower than the level 95%. The CPs of both S-FIC and FIC are higher than those from full models, but close to the nominal level, whereas

TABLE 1

Simulation results. Full: *using all variables*; *CP*: *coverage probability*; *MSE*: *mean squared error*

| | | | μ_1 | | | | μ_2 | | | | μ_3 | | | | μ_4 | | | |
| | | | $\varpi = 0$ | | $\varpi = 0.5$ | | $\varpi = 0$ | | $\varpi = 0.5$ | | $\varpi = 0$ | | $\varpi = 0.5$ | | $\varpi = 0$ | | $\varpi = 0.5$ | |
n	r_0	Method	CP	MSE	CP	MSE	CP	MSE	CP	MSE	CP	MSE	CP	MSE	CP	MSE	CP	MSE
200	1	Full	0.9	0.33	0.9	0.49	0.89	0.49	0.88	0.8	0.9	0.22	0.9	0.32	0.92	2.25	0.91	2.92
		AIC	0.91	0.31	0.9	0.46	0.9	0.45	0.87	0.76	0.89	0.21	0.88	0.3	0.91	2.15	0.9	2.77
		BIC	0.92	0.28	0.9	0.4	0.91	0.39	0.88	0.71	0.9	0.19	0.88	0.26	0.92	1.98	0.9	2.66
		FIC	0.92	0.28	0.93	0.39	0.92	0.33	0.91	0.79	0.92	0.19	0.92	0.25		2		2.66
		S-FIC	0.93	0.28	0.93	0.41	0.93	0.4	0.92	0.68	0.93	0.19	0.92	0.26		2		2.61
	4	Full	0.89	0.35	0.9	0.73	0.9	0.48	0.88	1.16	0.9	0.19	0.91	0.35	0.94	1.79	0.91	3.42
		AIC	0.89	0.34	0.9	0.69	0.9	0.47	0.86	1.16	0.89	0.19	0.89	0.35	0.94	1.75	0.9	3.39
		BIC	0.9	0.31	0.91	0.63	0.91	0.42	0.84	1.17	0.87	0.19	0.87	0.35	0.94	1.67	0.89	3.4
		FIC	0.95	0.19	0.95	0.34	0.94	0.33	0.93	0.79	0.93	0.14	0.95	0.24		1.52		2.7
		S-FIC	0.97	0.17	0.97	0.32	0.97	0.22	0.97	0.68	0.96	0.13	0.97	0.22		1.32		2.47
	7	Full	0.89	0.46	0.9	1.02	0.89	0.66	0.87	2.04	0.9	0.2	0.92	0.41	0.92	2.26	0.92	5.32
		AIC	0.89	0.46	0.9	1	0.89	0.65	0.86	2.04	0.9	0.2	0.91	0.41	0.92	2.24	0.91	5.28
		BIC	0.89	0.44	0.91	0.93	0.9	0.62	0.86	1.92	0.9	0.2	0.88	0.41	0.92	2.18	0.91	4.79
		FIC	0.94	0.21	0.97	0.36	0.94	0.33	0.95	0.79	0.95	0.12	0.97	0.19		1.87		2.98
		S-FIC	0.97	0.12	0.98	0.22	0.97	0.16	0.98	0.63	0.98	0.09	0.98	0.15		1.24		2.57
400	1	Full	0.93	0.07	0.92	0.1	0.93	0.11	0.93	0.15	0.93	0.05	0.92	0.07	0.94	0.52	0.94	0.67
		AIC	0.94	0.07	0.92	0.1	0.93	0.1	0.91	0.14	0.93	0.04	0.91	0.07	0.94	0.51	0.93	0.66
		BIC	0.94	0.06	0.93	0.09	0.94	0.09	0.91	0.14	0.93	0.04	0.91	0.06	0.94	0.5	0.93	0.65
		FIC	0.94	0.06	0.93	0.09	0.94	0.09	0.94	0.15	0.94	0.04	0.93	0.06		0.5		0.65
		S-FIC	0.95	0.06	0.93	0.09	0.94	0.1	0.93	0.14	0.94	0.04	0.94	0.06		0.51		0.64
	4	Full	0.94	0.07	0.91	0.12	0.93	0.11	0.9	0.19	0.94	0.04	0.91	0.08	0.94	0.54	0.93	0.78
		AIC	0.94	0.07	0.92	0.12	0.93	0.11	0.89	0.2	0.94	0.04	0.87	0.08	0.94	0.53	0.92	0.79
		BIC	0.94	0.07	0.92	0.12	0.94	0.1	0.88	0.22	0.92	0.05	0.87	0.09	0.94	0.52	0.9	0.83
		FIC	0.95	0.05	0.93	0.09	0.95	0.09	0.92	0.15	0.96	0.04	0.94	0.06		0.49		0.72
		S-FIC	0.97	0.05	0.95	0.09	0.97	0.07	0.95	0.16	0.97	0.04	0.94	0.06		0.46		0.69
	7	Full	0.92	0.08	0.9	0.14	0.93	0.11	0.91	0.21	0.94	0.04	0.91	0.08	0.94	0.52	0.93	0.82
		AIC	0.92	0.08	0.9	0.14	0.92	0.11	0.91	0.21	0.94	0.04	0.89	0.08	0.94	0.51	0.93	0.81
		BIC	0.93	0.08	0.91	0.13	0.93	0.11	0.89	0.22	0.94	0.04	0.86	0.09	0.94	0.5	0.92	0.82
		FIC	0.94	0.06	0.92	0.1	0.93	0.09	0.93	0.15	0.94	0.04	0.93	0.07		0.47		0.68
		S-FIC	0.95	0.05	0.96	0.07	0.95	0.06	0.96	0.12	0.96	0.03	0.96	0.05		0.38		0.6

the intervals of FIC and S-FIC have the same length as those from the full models because we estimate the unknown quantities in (3.4) under the full model.

When r_0 gets bigger, the MSEs based on S-FIC are substantially smaller than those obtained from other criteria. It is worth mentioning that in Tables 1 and 2, we do not report the CPs corresponding to FIC and S-FIC for μ_4 because we do not derive an asymptotic distribution for the proposed estimators of this focus parameter.

As suggested by a referee, we now numerically examine the effects of the number of knots on the performance of these criteria. We generalize the data and conduct the simulation in the same way as above, but oversmoothing and undersmoothing nonparametric terms by letting $N_r = \text{ceiling}(n^{1/3})$ and $N_r = \text{ceiling}(n^{1/10})$, respectively. The results corresponding to undersmoothing show a similar pattern as in Table 1. Note that derivatives of all orders of functions $\eta_1(X_{i,1})$ and $\eta_2(X_{i,2})$ exist and satisfy the Lipschitz condition. $N_r = \text{ceiling}(n^{1/10})$ is still in the range $(n^{1/(2v)}, n^{1/3})$, so this similarity is not surprising and supports our theory. However, oversmoothing of the nonparametric functions causes significant changes and generally produces larger MSEs but lower CPs, while all of the results show a preference for the S-FIC and FIC. To save space, we report the results with $n = 400$ and $r_0 = 4$ in Table 2, but omit other results, which show similar features to those reported in Table 2.

5. Real-world data analysis. In this section, we apply our methods to a data set from a Pima Indian diabetes study and perform some model selection and averaging procedures. The data set is obtained from the UCI Repository of Machine

TABLE 2
Simulation results of overfitting with $n = 400$ and $r_0 = 4$

	μ_1				μ_2			
	$\varpi = 0$		$\varpi = 0.5$		$\varpi = 0$		$\varpi = 0.5$	
Method	CP	MSE	CP	MSE	CP	MSE	CP	MSE
Full	0.864	0.131	0.852	0.232	0.852	0.211	0.840	0.365
AIC	0.869	0.129	0.863	0.226	0.851	0.207	0.805	0.381
BIC	0.884	0.117	0.872	0.210	0.863	0.186	0.770	0.409
FIC	0.942	0.086	0.917	0.154	0.922	0.131	0.874	0.300
S-FIC	0.952	0.081	0.932	0.149	0.946	0.123	0.916	0.300
	μ_3				μ_4			
Full	0.884	0.073	0.863	0.138	0.928	1.055	0.910	1.548
AIC	0.874	0.073	0.813	0.142	0.931	1.053	0.909	1.571
BIC	0.863	0.077	0.782	0.152	0.929	1.028	0.897	1.606
FIC	0.914	0.060	0.915	0.107		0.967		1.443
S-FIC	0.949	0.064	0.921	0.110		0.910		1.361

Learning Databases and selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. The patients under consideration are Pima Indian women at least 21 years old and living near Phoenix, Arizona. The response variable, Y, taking the value of 0 or 1, indicates a positive or negative test for diabetes. The eight covariates are *PGC* (plasma glucose concentration after two hours in an oral glucose tolerance test), *DPF* (diabetes pedigree function), *DBP* [diastolic blood pressure (mm Hg)], *NumPreg* (the number of times pregnant), *SI* [two-hour serum insulin (mu U/ml)], *TSFT* [triceps skin fold thickness (mm)], *BMI* (body mass index [weight in kg/(height in m)^2]) and *AGE* (years). We then consider the following GAPLM for this data analysis:

$$\text{logit}\{\Pr(Y=1)\} = \eta_1(BMI) + \eta_2(AGE) + \beta_1 PGC + \beta_2 DPF$$
$$+ \beta_3 DBP + \beta_4 NumPreg + \beta_5 SI + \beta_6 TSFT,$$

where *AGE* and *BMI* are set in nonparametric components and the following Figure 1 confirms that the effects of these two covariates on the log odd are nonlinear. All covariates have been centralized by sample mean and standardized by sample standard error.

We first fit the model with all covariates using the polynomial spline method introduced in Section 2. The cubic B-splines have been used to approximate the two nonparametric functions. The number of knots was chosen using the BIC, presented in (4.1). The fitted curves of the two nonparametric components $\eta_1(BMI)$ and $\eta_2(AGE)$ are depicted in Figure 1. The estimated values of the β_i's, their standard error (SE) and corresponding z-values are listed in Table 3. The results indicate that *PGC* and *DPF* are very significant, while the other four seem not to be, so we run model selection and averaging on these four covariates. Accordingly, there are $2^4 = 16$ submodels.
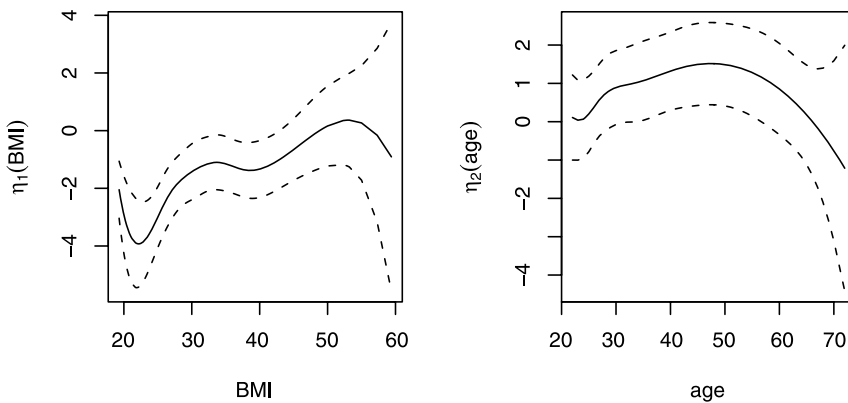

FIG. 1. *The patterns of the nonparametric functions of BMI and AGE (solid lines) with ±SE (broken lines).*

TABLE 3

Results for the diabetes study: *estimated values, associated standard errors and P-values obtained using the full model*

	Estimated value	Standard error	*P*-value
PGC	1.1698	0.1236	0.0000
DPF	0.3323	0.1029	0.0012
DBP	−0.2662	0.1040	0.0104
NumPreg	0.1887	0.1209	0.1184
SI	−0.1511	0.1078	0.1610
TSFT	0.0179	0.1135	0.8749

We now consider four focus parameters: $\mu_1 = \beta_1$, $\mu_2 = \beta_2$, $\mu_3 = \eta_1(-1.501) + \eta_2(0.585) + 0.028\beta_1 - 0.899\beta_2 - 1.570\beta_3 + 1.087\beta_4 - 0.223\beta_5 - 0.707\beta_6$ and $\mu_4 = \eta_1(-0.059) + \eta_2(1.363) + 0.994\beta_1 + 0.423\beta_2 + 0.645\beta_3 + 1.117\beta_4 - 0.221\beta_5 + 0.055\beta_6$. The first two are just the single coefficients of *PGC* and *DPF*, the so-called two most significant linear components. The second two are related to the nonparametric terms. Specifically speaking, μ_3 represents the log odd at *BMI* = 22.2, the lowest point of the estimated curve in the left panel of Figure 1, and the corresponding means of other predictors when *BMI* = 22.2, while μ_4 represents the log odd at *AGE* = 49, the highest point of the estimated curve in the right panel of Figure 1, and the corresponding means of other predictors when *AGE* = 49. We label the potential 16 submodels "0," "3," "4," "5," "6,"..., "3456" corresponding to a submodel which includes (or not) *DBP*, *NumPreg*, *SI* and *TSFT*. The results based on AIC, BIC and FIC methods are presented in Table 4. Regardless of focus parameter, the AIC and BIC select submodels "345" and "3," respectively. On the other hand, the FIC prefers submodels "3," "34," "345" and "5" when the focus is on μ_1, μ_2, μ_3 and μ_4, respectively. It is noticeable that submodel "36" is also competitive for μ_1. We are inclined to use submodel "3" since it has fewer parameters.

We further examine the predictive power of above model selection and averaging methods through a cross-validation experiment. For each patient in the data set, we use the AIC, BIC, FIC and S-FIC to carry out estimations based on all of the other patients as a training sample, and then predict the left-out observation. The prediction error ratios (the ratio of the number of mistaken predictions to the sample size) corresponding to AIC, BIC, FIC and S-FIC are 0.228, 0.225, 0.221 and 0.221, respectively. Both FIC and S-FIC show smaller prediction errors than those of AIC and BIC, although the differences among these errors are not substantial. These results indicate the superiority of the FIC and S-FIC to the AIC and BIC.

6. Discussion. We have proposed an effective procedure using the polynomial spline technique along with the model average principle to improve accuracy of

TABLE 4
Results for the diabetes study: AIC, BIC and FIC values, and estimators of focus parameters

	0	3	4	5	6	34	35	36	45	46	56	345	346	356	456	3456
AIC	717.2	712.1	716.7	716.5	718.3	711.5	711.8	713.8	716.1	717.8	718.4	711.3★	713.2	713.7	718.0	713.3
BIC	791.3	790.7★	795.4	795.2	797.0	794.7	795.0	797.1	799.4	801.1	801.7	799.2	801.1	801.6	806.0	805.8
μ_1-FIC	11.58	9.86★	13.69	11.07	11.4	10.97	11.74	9.86	11.63	13.41	11.36	11.16	10.96	12.17	12.08	11.54
$\hat{\mu}_1$	1.09	1.11	1.09	1.15	1.092	1.11	1.17	1.11	1.15	1.09	1.15	1.17	1.11	1.17	1.15	1.17
μ_2-FIC	7.87	7.83	7.66	7.77	7.95	7.58★	7.77	8.07	7.79	7.93	7.97	7.80	8.00	7.98	8.01	7.99
$\hat{\mu}_2$	0.31	0.31	0.31	0.33	0.32	0.32	0.33	0.32	0.33	0.33	0.33	0.33	0.32	0.33	0.34	0.33
μ_3-FIC	261.5	51.9	143.7	245.8	219.1	38.0	48.7	47.9	144.7	122.5	235.1	37.7★	38.8	51.0	140.7	38.7
$\hat{\mu}_3$	−2.62	−2.23	−2.57	−2.70	−2.66	−2.17	−2.30	−2.26	−2.65	−2.61	−2.70	−2.24	−2.19	−2.29	−2.65	−2.23
μ_4-FIC	10.56	53.98	24.17	4.22★	9.82	30.70	30.08	51.82	35.28	23.98	6.02	31.38	30.71	30.10	35.43	31.93
$\hat{\mu}_4$	1.63	1.59	1.66	1.73	1.61	1.62	1.68	1.58	1.75	1.64	1.71	1.71	1.61	1.69	1.74	1.71

★ denotes the minimal AIC, BIC or FIC values of the corresponding row.

estimation in GAPLMs when uncertainty potentially appears. Our method avoids any iterative algorithms and reduces computational challenges, therefore its computational gain is remarkable. Most importantly, the estimators of the linear components we have developed are still asymptotically normal. Both theoretical and numerical studies show promise for the proposed methods.

GAPLMs are generally enough to cover a variety of semiparametric models such as partially linear additive models [Liang et al. (2008)] and generalized partially linear models [Severini and Staniswalis (1994)]. It is worth pointing out that GAPLMs do not involve any interaction between nonparametric components (which may appear in a particular issue) and thus our current methods do not deal with this situation. We conjecture that our procedure can be applied when the interactions may also be included in the model search through tensor polynomial spline approximation, but this extension poses additional challenges. How to develop model selection and model averaging procedures in such a complex structure warrants further investigation.

APPENDIX

Let $\| \cdot \|$ be the Euclidean norm and $\|\varphi\|_\infty = \sup_m |\varphi(m)|$ be the supremum norm of a function φ on $[0, 1]$. As in Carroll et al. (1997), we let $q_l(m, y) = \frac{\partial^l Q\{g^{-1}(m), y\}}{\partial m^l}$, then $q_1(m, y) = \partial Q\{g^{-1}(m), y\}/\partial m = \{y - g^{-1}(m)\}\rho_1(m)$ and $q_2(m, y) = \partial^2 Q\{g^{-1}(m), y\}/\partial m^2 = \{y - g^{-1}(m)\}\rho_1'(m) - \rho_2(m)$.

A.1. Conditions. Let r be a positive integer and $\nu \in (0, 1]$ be such that $\upsilon = r + \nu > 1.5$. Let $\mathcal{H}$ be the collection of functions f on $[0, 1]$ whose rth derivative, $f^{(r)}$, exists and satisfies the Lipschitz condition of order ν; that is,

$$\left| f^{(r)}(m^*) - f^{(r)}(m) \right| \le C_1 |m^* - m|^\nu \qquad \text{for } 0 \le m^*, m \le 1,$$

where C_1 is a generic positive constant. In what follows, c, C, $c.$, $C.$ and $C_.^*$ are all generic positive constants. The following are the conditions needed to obtain Theorems 1–3:

(C1) each component function $\eta_{0,\alpha} \in \mathcal{H}$, $\alpha = 1, \ldots, p$;

(C2) $q_2(m, y) < 0$ and $c_q < |q_2(m, y)| < C_q$ for $m \in R$ and y in the range of the response variable;

(C3) the function $\eta_0''(\cdot)$ is continuous;

(C4) the distribution of $\mathbf{X}$ is absolutely continuous and its density f is bounded away from zero and infinity on $[0, 1]^p$;

(C5) $E(\mathbf{Z}\mathbf{Z}^T | \mathbf{X} = \mathbf{x})$ exists and $\mathbf{A} = E[\rho_2\{m_0(\mathbf{T})\}\mathbf{Z}\mathbf{Z}^T]$ is invertible, almost surely;

(C6) the number of interior knots $n^{1/(2\upsilon)} \ll J_n \ll n^{1/3}$;

(C7) $\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \mathbf{B}(\mathbf{X}_i)\mathbf{B}^T(\mathbf{X}_i) & \mathbf{B}(\mathbf{X}_i)\mathbf{Z}_i^T \\ \mathbf{Z}_i\mathbf{B}^T(\mathbf{X}_i) & \mathbf{Z}_i\mathbf{Z}_i^T \end{pmatrix}$ exists and is nonsingular;

(C8) for ρ_1 introduced in Section 2, $|\rho_1(m_0)| \leq C_\rho$ and

$$|\rho_1(m) - \rho_1(m_0)| \leq C_\rho^*|m - m_0| \qquad \text{for all } |m - m_0| \leq C_m;$$

(C9) the matrix $\mathbf{D}$ is invertible almost surely;

(C10) the link function g in model (2.1) satisfies $|\frac{d}{dm}g(m)|_{m=m_0}| \leq C_g$ and

$$\left| \frac{d}{dm}g^{-1}(m) - \frac{d}{dm}g^{-1}(m) \right|_{m=m_0} \left|\right.$$
$$\leq C_g^*|m - m_0| \qquad \text{for all } |m - m_0| \leq C_m^*;$$

(C11) there exists a positive constant C_ε such that $E(\varepsilon^2|\mathbf{T} = \mathbf{t}) \leq C_\varepsilon$ almost surely.

A.2. Technical lemmas. In the following, for any probability measure P, we define $L_2(P) = \{f : \int f^2\,dP < \infty\}$. Let $\mathcal{F}$ be a subclass of $L_2(P)$. The bracketing number $\mathcal{N}_{[]}(\tau, \mathcal{F}, L_2(P))$ of $\mathcal{F}$ is defined as the smallest value of N for which there exist pairs of functions $\{[f_j^L, f_j^U]\}_{j=1}^N$ with $\|f_j^U - f_j^L\| \leq \tau$, such that for each $f \in \mathcal{F}$, there exists a $j \in \{1, \ldots, N\}$ such that $f_j^L \leq f \leq f_j^U$. Define the entropy integral $J_{[]}(\tau, \mathcal{F}, L_2(P)) = \int_0^\tau \sqrt{1 + \log \mathcal{N}_{[]}(\iota, \mathcal{F}, L_2(P))}\,d\iota$. Let P_n be the empirical measure of P. Define $G_n = \sqrt{n}(P_n - P)$ and $\|G_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |G_n f|$ for any measurable class of functions $\mathcal{F}$.

We state or prove several preliminary lemmas first. Lemmas A.1–A.3 will be used to prove the remaining lemmas. Lemmas A.4–A.5 are used to prove Theorem 1. Theorems 2–3 are obtained from Theorem 1.

LEMMA A.1 [Lemma 3.4.2 of van der Vaart and Wellner (1996)]. *Let M_0 be a finite positive constant. Let $\mathcal{F}$ be a uniformly bounded class of measurable functions such that $Pf^2 < \tau^2$ and $\|f\|_\infty < M_0$. Then*

$$E_P\|G_n\|_{\mathcal{F}} \leq C_0 J_{[]}(\tau, \mathcal{F}, L_2(P))\left\{1 + \frac{J_{[]}(\tau, \mathcal{F}, L_2(P))}{\tau^2\sqrt{n}} M_0\right\},$$

where C_0 is a finite constant not dependent on n.

LEMMA A.2 [Lemma A.2 of Huang (1999)]. *For any $\tau > 0$, let $\Theta_n = \{\eta(\mathbf{x}) + \mathbf{z}^{\mathrm{T}}\beta; \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \tau, \eta \in \mathcal{G}_n, \|\eta - \eta_0\|_2 \leq \tau\}$. Then, for any $\iota \leq \tau$, $\log \mathcal{N}_{[]}(\iota, \Theta_n, L_2(P)) \leq c_0(J_n + \varrho)\log \tau/\iota$, where c_0 is a finite constant not dependent on n.*

Referring to the result of de Boor [(2001), page 149], for any function $f \in \mathcal{H}$ and $n \geq 1$, there exists a function $\widetilde{f} \in \mathcal{S}_n$ such that $\|\widetilde{f} - f\|_\infty \leq Ch^\upsilon$, where C is some fixed positive constant. From condition (C1), we can find $\widetilde{\boldsymbol{\gamma}}_S = \{\widetilde{\gamma}_{S,j,\alpha}, j =$

$-\varrho + 1, \ldots, J_n, \alpha = 1, \ldots, p\}^{\mathrm{T}}$ and an additive spline function $\widetilde{\eta}_S = \widetilde{\boldsymbol{\gamma}}_S^{\mathrm{T}} \mathbf{B}(\mathbf{x}) \in \mathcal{G}_n$ such that

$$(\text{A.1}) \qquad\qquad \|\widetilde{\eta}_S - \eta_0\|_{\infty} = O(h^{\upsilon}).$$

Let $\widetilde{\beta}_S = \arg\max \frac{1}{n} \sum_{i=1}^{n} Q[g^{-1}\{\widetilde{\eta}_S(\mathbf{X}_i) + (\Pi_S \mathbf{Z}_i)^{\mathrm{T}} \beta_S\}, Y_i]$, $m_{0,i} = m_0(\mathbf{T}_i) = \eta_0(\mathbf{X}_i) + \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_0$ and $\widetilde{m}_{S,i} = \widetilde{m}_S(\mathbf{T}_i) = \widetilde{\eta}_S(\mathbf{X}_i) + \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_0 = \widetilde{\gamma}_S^{\mathrm{T}} \mathbf{B}(\mathbf{X}_i) + \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_0$.

LEMMA A.3. *Under the local misspecification framework and conditions* (C1)–(C6),

$$(\text{A.2}) \qquad \sqrt{n} \Pi_S^{\mathrm{T}} (\widetilde{\beta}_S - \Pi_S \boldsymbol{\beta}_0) - \bar{\Pi}_S^{\mathrm{T}} \bar{\delta}_S \xrightarrow{d} N(0, \mathbf{A}^{-1} \boldsymbol{\Sigma}_1 \mathbf{A}^{-1}),$$

where $\bar{\delta}$ *consists of the elements of* δ *that are not in the Sth submodel,* $\bar{\pi}_S$ *is the project matrix mapping* δ *to* $\bar{\delta}$, $\bar{\Pi}_S = [0_{(d_u - d_{u,S}) \times d_c}, \bar{\pi}_S]$ *and* $\boldsymbol{\Sigma}_1 = E[q_1^2\{m_0(\mathbf{T})\} \mathbf{Z} \mathbf{Z}^{\mathrm{T}}]$.

PROOF. Let $\vartheta = \sqrt{n} \Pi_S^{\mathrm{T}} (\beta_S - \Pi_S \boldsymbol{\beta}_0) - \bar{\Pi}_S^{\mathrm{T}} \bar{\delta}_S$ and $\widetilde{\vartheta} = \sqrt{n} \Pi_S^{\mathrm{T}} (\widetilde{\beta}_S - \Pi_S \boldsymbol{\beta}_0) - \bar{\Pi}_S^{\mathrm{T}} \bar{\delta}_S$. Note that $\widetilde{\beta}_S$ maximizes $\frac{1}{n} \sum_{i=1}^{n} Q[g^{-1}\{\widetilde{\eta}_S(\mathbf{X}_i) + (\Pi_S \mathbf{Z}_i)^{\mathrm{T}} \beta_S\}, Y_i]$, so $\widetilde{\vartheta}$ maximizes

$$\widetilde{\ell}_n(\vartheta) = \sum_{i=1}^{n} [Q\{g^{-1}(\widetilde{m}_{S,i} + n^{-1/2} \vartheta^{\mathrm{T}} \mathbf{Z}_i), Y_i\} - Q\{g^{-1}(\widetilde{m}_{S,i}), Y_i\}].$$

By Taylor expansion, one has $\widetilde{\ell}_n(\vartheta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q_1(\widetilde{m}_{S,i}, Y_i) \vartheta^{\mathrm{T}} \mathbf{Z}_i + \frac{1}{2} \vartheta^{\mathrm{T}} \mathbf{A}_n \boldsymbol{\vartheta}$, where $\mathbf{A}_n = \frac{1}{n} \sum_{i=1}^{n} \{Y_i \rho_1'(\widetilde{m}_{S,i} + \zeta_{ni}) - \rho_3(\widetilde{m}_{0i} + \zeta_{ni}^*)\} \mathbf{Z}_i \mathbf{Z}_i^{\mathrm{T}}$ with ζ_{ni} and ζ_{ni}^* both lying between 0 and $n^{-1/2} \vartheta^{\mathrm{T}} \mathbf{Z}_i$, and $\rho_3(m) = g^{-1}(m) \rho_1'(m) - \rho_2(m)$. From the proof of Theorem 2 in Carroll et al. (1997), $\mathbf{A}_n = -E[\rho_2\{m_0(\mathbf{T})\} \mathbf{Z} \mathbf{Z}^{\mathrm{T}}] + o_p(1) = -\mathbf{A} + o_p(1)$ and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} q_1(\widetilde{m}_{S,i}, Y_i) \mathbf{Z}_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q_1(m_{0,i}, Y_i) \mathbf{Z}_i$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q_2(m_{0,i}, Y_i) \{\widetilde{\eta}_S(\mathbf{X}_i) - \eta_0(\mathbf{X}_i)\} \mathbf{Z}_i$$

$$+ O_p(n^{1/2} \|\widetilde{\eta}_S - \eta_0\|_{\infty}^2).$$

In addition, by (A.1) and conditions (C2), (C5) and (C6), we have

$$n^{-1/2} \sum_{i=1}^{n} q_2(m_{0,i}, Y_i) \mathbf{Z}_i \{\widetilde{\eta}_S(\mathbf{X}_i) - \eta_0(\mathbf{X}_i)\} = O_p(n^{1/2} h^{\upsilon}) = o_p(1).$$

Therefore, by the convexity lemma of Pollard (1991) and condition (C5), one has $\widetilde{\vartheta} = \mathbf{A}^{-1} n^{-1/2} \sum_{i=1}^{n} q_1(m_{0,i}, Y_i) \mathbf{Z}_i + o_p(1)$ and $\text{var}\{q_1\{m_0(\mathbf{T}), Y\} \mathbf{Z}\} = E[q_1^2\{m_0(\mathbf{T}), Y\} \mathbf{Z} \mathbf{Z}^{\mathrm{T}}] = \Sigma_1$, so (A.2) holds. $\square$

Define $a_{n,h} = h^{\upsilon} + (n^{-1} \log n)^{1/2}$, $\theta_S = (\gamma^{\mathrm{T}}, \beta_S^{\mathrm{T}})^{\mathrm{T}}$, $\widetilde{\theta}_S = (\widetilde{\gamma}_S^{\mathrm{T}}, \widetilde{\beta}_S^{\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\theta}_S = (\widehat{\gamma}_S^{\mathrm{T}}, \widehat{\beta}_S^{\mathrm{T}})^{\mathrm{T}}$.

LEMMA A.4. *Under the local misspecification framework and conditions* (C1)–(C8), *one has* $\|\widehat{\theta}_S - \widetilde{\theta}_S\| = O_p(J_n^{1/2} a_{n,h})$.

PROOF. Note that

$$(A.3) \qquad \left.\frac{\partial \ell_n(\theta_S)}{\partial \theta_S}\right|_{\theta_S = \widehat{\theta}_S} - \left.\frac{\partial \ell_n(\theta_S)}{\partial \theta_S}\right|_{\theta_S = \widetilde{\theta}_S} = \left.\frac{\partial^2 \ell_n(\theta_S)}{\partial \theta_S \, \partial \theta_S^{\mathrm{T}}}\right|_{\theta_S = \bar{\theta}_S} (\widehat{\theta}_S - \widetilde{\theta}_S),$$

with $\bar{\theta}_S$ lying between $\widehat{\theta}_S$ and $\widetilde{\theta}_S$. Recalling the equation (2.5), one has

$$\left.\frac{\partial \ell_n(\theta_S)}{\partial \theta_S}\right|_{\theta_S = \widetilde{\theta}_S} = \left\{\left(\frac{\partial \ell_n(\theta_S)}{\partial \gamma}\right)^{\mathrm{T}}, \left(\frac{\partial \ell_n(\theta_S)}{\partial \beta_S}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}\Bigg|_{\theta_S = \widetilde{\theta}_S},$$

where

$$\begin{aligned}
\left.\frac{\partial \ell_n(\theta_S)}{\partial \gamma}\right|_{\theta_S = \widetilde{\theta}_S} &= \frac{1}{n} \sum_{i=1}^{n} q_1(m_{0,i}, Y_i) \mathbf{B}(\mathbf{X}_i) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} q_2(\xi_i, Y_i)\{\widetilde{\eta}(\mathbf{X}_i) - \eta_0(\mathbf{X}_i)\} \mathbf{B}(\mathbf{X}_i) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} q_2(\xi_i, Y_i)\{\Pi_S^{\mathrm{T}}(\widetilde{\beta}_S - \Pi_S \beta_0) - \bar{\Pi}_S^{\mathrm{T}} \bar{\delta}_S / \sqrt{n}\}^{\mathrm{T}} \mathbf{Z}_i \mathbf{B}(\mathbf{X}_i)
\end{aligned}$$

and

$$\begin{aligned}
\left.\frac{\partial \ell_n(\theta_S)}{\partial \beta_S}\right|_{\theta_S = \widetilde{\theta}_S} &= \frac{1}{n} \sum_{i=1}^{n} q_1(m_{0,i}, Y_i) \Pi_S \mathbf{Z}_i \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} q_2(\xi_i^*, Y_i)\{\widetilde{\eta}(\mathbf{X}_i) - \eta_0(\mathbf{X}_i)\} \Pi_S \mathbf{Z}_i \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} q_2(\xi_i^*, Y_i)\{\Pi_S^{\mathrm{T}}(\widetilde{\beta}_S - \Pi_S \beta_0) - \bar{\Pi}_S^{\mathrm{T}} \bar{\delta}_S / \sqrt{n}\}^{\mathrm{T}} \mathbf{Z}_i \Pi_S \mathbf{Z}_i,
\end{aligned}$$

with ξ_i and ξ_i^* both lying between $m_{0,i}$ and $\widetilde{m}_{S,i}$. According to the Bernstein inequality and condition (C8),

$$\begin{aligned}
\left\|\frac{1}{n} \sum_{i=1}^{n} q_1(m_{0,i}, Y_i) \mathbf{B}(\mathbf{X}_i)\right\|_{\infty} &= \max_{-\varrho+1 \leq j \leq J, 1 \leq \alpha \leq p} \frac{1}{n} \left|\sum_{i=1}^{n} \rho_1(m_{0,i}) B_{j,\alpha}(X_{i\alpha}) \varepsilon_i\right| \\
&= O_p\{(n^{-1} \log n)^{1/2}\}.
\end{aligned}$$

And, by (A.1), Lemma A.3 and condition (C2), one has

$$\frac{1}{n} \sum_{i=1}^{n} \|q_2(\xi_i, Y_i)\{\widetilde{\eta}_S(\mathbf{X}_i) - \eta_0(\mathbf{X}_i)\}\mathbf{B}(\mathbf{X}_i)\|_\infty = O_p(h^\upsilon)$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \|q_2(\xi_i, Y_i)\{\Pi_S^{\mathrm{T}}(\widetilde{\beta}_S - \Pi_S \beta_0) - \bar{\Pi}_S^{\mathrm{T}} \bar{\delta}_S / \sqrt{n}\}^{\mathrm{T}} \mathbf{Z}_i \mathbf{B}(\mathbf{X}_i)\|_\infty = O_p(n^{-1/2}).$$

Therefore, $\|\partial \ell_n(\theta_S)/\partial \gamma|_{\theta_S = \widetilde{\theta}_S}\|_\infty = O_p(a_{n,h})$. Similarly, we can prove

$$\left\| \frac{\partial \ell_n(\theta_S)}{\partial \beta_S} \Big|_{\theta_S = \widetilde{\theta}_S} \right\|_\infty = O_p(h^\upsilon + (n^{-1} \log n)^{1/2}).$$

Thus,

$$(A.4) \qquad \left\| \frac{\partial \ell_n(\theta_S)}{\partial \theta_S} \Big|_{\theta_S = \widetilde{\theta}_S} \right\|_\infty = O_p(a_{n,h}).$$

Let $\bar{m}_{S,i} = \bar{m}_S(\mathbf{T}_i) = \bar{\theta}^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}(\mathbf{X}_i), (\Pi_S \mathbf{Z}_i)^{\mathrm{T}})^{\mathrm{T}}$. For the second order derivative, one has

$$\frac{\partial^2 \ell_n(\theta_S)}{\partial \theta_S \, \partial \theta_S^{\mathrm{T}}} \Big|_{\theta_S = \bar{\theta}_S} = \left(\begin{matrix} \dfrac{\partial^2 \ell_n(\theta_S)}{\partial \gamma \, \partial \gamma^{\mathrm{T}}} & \dfrac{\partial^2 \ell_n(\theta_S)}{\partial \gamma \, \partial \beta_S^{\mathrm{T}}} \\[2mm] \dfrac{\partial^2 \ell_n(\theta_S)}{\partial \beta_S \, \partial \gamma^{\mathrm{T}}} & \dfrac{\partial^2 \ell_n(\theta_S)}{\partial \beta_S \, \partial \beta_S^{\mathrm{T}}} \end{matrix} \right) \Bigg|_{\theta_S = \bar{\theta}_S}$$

$$= \frac{1}{n} \sum_{i=1}^{n} q_2(\bar{m}_{S,i}, Y_i) \left\{ \left(\begin{matrix} \mathbf{B}(\mathbf{X}_i)\mathbf{B}^{\mathrm{T}}(\mathbf{X}_i) & \mathbf{B}(\mathbf{X}_i)\mathbf{Z}_i^{\mathrm{T}} \\ \mathbf{Z}_i \mathbf{B}^{\mathrm{T}}(\mathbf{X}_i) & \mathbf{Z}_i \mathbf{Z}_i^{\mathrm{T}} \end{matrix} \right) \right\},$$

by which, along with conditions (C2) and (C7), we know that the matrix $\frac{\partial^2 \ell_n(\theta_S)}{\partial \theta_S \, \partial \theta_S^{\mathrm{T}}}|_{\theta_S = \bar{\theta}_S}$ is nonsingular in probability. So, according to (A.3) and (A.4), we have completed the proof.   □

Define $\mathcal{M}_n = \{m(\mathbf{x}, \mathbf{z}) = \eta(\mathbf{x}) + \mathbf{z}^{\mathrm{T}} \beta : \eta \in \mathcal{G}_n\}$ and a class of functions $\mathcal{A}(\tau) = \{\rho_1(m(\mathbf{t}))\psi(\mathbf{t}) : m \in \mathcal{M}_n, \|m - m_0\| \le \tau\}$.

LEMMA A.5. *Under the local misspecification framework and conditions* (C1)–(C8), *we have*

$$(A.5) \qquad \frac{1}{n} \sum_{i=1}^{n} \{\widehat{\eta}_S(\mathbf{X}_i) - \eta_0(\mathbf{X}_i)\}\rho_1(m_{0,i})\psi(\mathbf{T}_i) = o_p(n^{-1/2}),$$

$$(A.6) \qquad \frac{1}{n} \sum_{i=1}^{n} \rho_1(m_{0,i})\psi(\mathbf{T}_i)\Gamma(\mathbf{X}_i)^{\mathrm{T}} \Pi_S^{\mathrm{T}}(\widehat{\beta}_S - \Pi_S \beta_0) = o_p(n^{-1/2}).$$

PROOF. Noting that ψ and ρ_1 are fixed bounded functions under condition (C8), by Lemma A.2, similar to the proof of Corollary A.1 in Huang (1999), we can show, for any $\iota \leq \tau$, $\log \mathcal{N}_{[]}(\iota, \mathcal{A}(\tau), \| \cdot \|) \leq c_0((J_n + \varrho) \log(\tau/\iota) + \log(\iota^{-1}))$, so the corresponding entropy integral satisfies $J_{[]}(\tau, \mathcal{A}(\tau), \| \cdot \|) \leq c_0 \tau \{(J_n + \varrho)^{1/2} + (\log \tau^{-1})^{1/2}\}$. According to Lemma A.4, $\|\widehat{\eta}_S - \widetilde{\eta}_S\|_2^2 = (\widehat{\gamma}_S - \widetilde{\gamma}_S)^{\mathrm{T}} \sum_{i=1}^n E\{\mathbf{B}(\mathbf{X}_i)\mathbf{B}^{\mathrm{T}}(\mathbf{X}_i)\}(\widehat{\gamma}_S - \widetilde{\gamma}_S)/n \leq C_7 \|\widehat{\gamma}_S - \widetilde{\gamma}_S\|_2^2$, thus $\|\widehat{\eta}_S - \widetilde{\eta}_S\|_2 = O_p(J_n^{1/2} a_{n,h})$ and $\|\widehat{\eta}_S - \eta_0\|_2 \leq \|\widehat{\eta}_S - \widetilde{\eta}_S\|_2 + \|\widetilde{\eta}_S - \eta_0\|_2 = O_p(J_n^{1/2} a_{n,h})$. Now, by Lemma 7 of Stone (1986),

$$(A.7) \qquad \|\widehat{\eta}_S - \eta_0\|_\infty \leq C_8 J_n^{1/2} \|\widehat{\eta}_S - \eta_0\|_2 = O_p(J_n a_{n,h}).$$

Thus, by Lemma A.1, together with conditions (C1) and (C6), we have

$$E\left| \frac{1}{n} \sum_{i=1}^n \{\widehat{\eta}_S(\mathbf{X}_i) - \eta_0(\mathbf{X}_i)\} \rho_1(m_{0,i}) \psi(\mathbf{T}_i) \right.$$

$$\left. - E[\{\widehat{\eta}_S(\mathbf{X}) - \eta_0(\mathbf{X})\} \rho_1\{m_{0,i} \psi(\mathbf{T})\}] \right| = o(n^{-1/2}).$$

In addition, by the definition of ψ, $E[\phi(\mathbf{X})\rho_1\{m_0(\mathbf{T})\}\psi(\mathbf{T})] = 0$ for any measurable function ϕ. Hence (A.5) holds. Similarly, (A.6) follows from Lemmas A.1–A.4. $\square$

A.3. Proof of Theorem 1. Let $\widehat{m}_{S,i} = \widehat{m}_S(\mathbf{T}_i) = \widehat{\eta}_S(\mathbf{X}_i) + \widehat{\beta}_S^{\mathrm{T}} \Pi_S \mathbf{Z}_i$. For any $\mathbf{v} \in R^{d_c+d_{u,S}}$, define $\widehat{m}_S(\mathbf{v}) = \widehat{m}_S(\mathbf{x}, \Pi_S\mathbf{z}) + \mathbf{v}^{\mathrm{T}}\{\Pi_S\mathbf{z} - \Pi_S\Gamma(\mathbf{x})\} = \widehat{m}_S(\mathbf{x}, \Pi_S\mathbf{z}) + \mathbf{v}^{\mathrm{T}}\Pi_S\psi(\mathbf{t})$. Note that when $\mathbf{v} = 0$, $\widehat{m}_S(\mathbf{v})$ maximizes $1/n \sum_{i=1}^n Q[g^{-1}\{m_S(\mathbf{T}_i)\}, Y_i]$ for all $m_S \in \{m_S(\mathbf{x}, \mathbf{z}) = \eta(\mathbf{x}) + (\Pi_S\mathbf{z})^{\mathrm{T}}\beta_S : \eta \in \mathcal{G}_n\}$, by which

$$0 = \frac{\partial}{\partial \mathbf{v}} \ell_n(\widehat{m}_S(\mathbf{v}))\bigg|_{\mathbf{v}=0}$$

$$= \frac{1}{n} \sum_{i=1}^n \{Y_i - g^{-1}(\widehat{m}_{S,i})\}\rho_1(\widehat{m}_{S,i})\Pi_S\psi(\mathbf{T}_i).$$

$$(A.8) \quad = \frac{1}{n} \sum_{i=1}^n q_1(m_{0,i}, Y_i)\Pi_S\psi(\mathbf{T}_i) + \frac{1}{n} \sum_{i=1}^n \varepsilon_i\{\rho_1(\widehat{m}_{S,i}) - \rho_1(m_{0,i})\}\Pi_S\psi(\mathbf{T}_i)$$

$$- \frac{1}{n} \sum_{i=1}^n \{g^{-1}(\widehat{m}_{S,i}) - g^{-1}(m_{0,i})\}\rho_1(\widehat{m}_{S,i})\Pi_S\psi(\mathbf{T}_i)$$

$$\equiv I + II - III.$$

Note that for the second term $E[\varepsilon_i\{\rho_1(\widehat{m}_{S,i}) - \rho_1(m_{0,i})\}\Pi_S\psi(\mathbf{T}_i)] = 0$. From Lemma A.3, (A.6) and (A.7), we have $\|\widehat{m}_S - m_0\|_\infty = O_p(J_n^{1/2} a_{n,h})$, so, by condi-

tion (C8), $\|\rho_1(\widehat{m}_S) - \rho_1(m_0)\|_\infty = O_p(J_n^{1/2} a_{n,h})$. Now, by the Bernstein inequality, under condition (C11), we show that

$$(A.9) \qquad II = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \{\rho_1(\widehat{m}_{S,i}) - \rho_1(m_{0,i})\} \Pi_S \psi(\mathbf{T}_i) = o_p(n^{-1/2}).$$

Express the third term as

$$III = \frac{1}{n} \sum_{i=1}^{n} \{g^{-1}(\widehat{m}_{S,i}) - g^{-1}(m_{0,i})\} \rho_1(\widehat{m}_{S,i}) \Pi_S \psi(\mathbf{T}_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} (\widehat{m}_{S,i} - m_{0,i}) \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \{g^{-1}(\widehat{m}_{S,i}) - g^{-1}(m_{0,i}) - (\widehat{m}_{S,i} - m_{0,i})\} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \{g^{-1}(\widehat{m}_{S,i}) - g^{-1}(m_{0,i})\} \{\rho_1(\widehat{m}_{S,i}) - \rho_1(m_{0,i})\} \Pi_S \psi(\mathbf{T}_i)$$

$$\equiv III_1 + III_2 + III_3.$$

From Lemma A.5, a direct simplification yields

$$III_1 = \frac{1}{n} \sum_{i=1}^{n} \{\widehat{\eta}_S(\mathbf{X}_i) + \widehat{\beta}_S^T \Pi_S \mathbf{Z}_i - \eta_0(\mathbf{X}_i) - \beta_0^T \mathbf{Z}_i\} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \{\widehat{\eta}_S(\mathbf{X}_i) - \eta_0(\mathbf{X}_i) + (\Pi_S^T \widehat{\beta}_S - \beta_0)^T \psi(\mathbf{T}_i)$$

$$+ (\Pi_S^T \widehat{\beta}_S - \beta_0)^T \Gamma(\mathbf{X}_i)\} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \{\widehat{\eta}_S(\mathbf{X}_i) - \eta_0(\mathbf{X}_i)\} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i) \psi(\mathbf{T}_i)^T \left\{ \Pi_S^T \widehat{\beta}_S - \begin{pmatrix} \beta_{c,0} \\ 0 \end{pmatrix} \right\}$$

$$- \frac{1}{n} \sum_{i=1}^{n} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i) \psi(\mathbf{T}_i)^T [0, I]^T \delta / \sqrt{n}$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i) \Gamma(\mathbf{X}_i)^T \Pi_S^T (\widehat{\beta}_S - \Pi_S \beta_0)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i) \Gamma(\mathbf{X}_i)^T \bar{\Pi}_S^T (-\bar{\delta}_S / \sqrt{n})$$

$$= \frac{1}{n} \sum_{i=1}^{n} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i) \psi(\mathbf{T}_i)^{\mathrm{T}} \Pi_S^{\mathrm{T}} \left\{ \widehat{\beta}_S - \binom{\beta_{c,0}}{0} \right\}$$

$$- \frac{1}{\sqrt{n}} \frac{1}{n} \sum_{i=1}^{n} \rho_1(m_{0,i}) \Pi_S \psi(\mathbf{T}_i) \psi(\mathbf{T}_i)^{\mathrm{T}} [0, I]^{\mathrm{T}} \delta + o_p(n^{-1/2}).$$

In addition, from conditions (C8) and (C10), referring to the proof of (A.5), we have $III_2 = o_p(n^{-1/2})$ and $III_3 = o_p(n^{-1/2})$. Therefore,

$$III = [E\{\rho_1(m_0) \Pi_S \psi(\mathbf{T}) \psi(\mathbf{T})^{\mathrm{T}} \Pi_S^{\mathrm{T}}\} + o_p(1)] \left\{ \widehat{\beta}_S - \binom{\beta_{c,0}}{0} \right\}$$

(A.10)

$$- \frac{1}{\sqrt{n}} [E\{\rho_1(m_0) \Pi_S \psi(\mathbf{T}) \psi(\mathbf{T})^{\mathrm{T}} [0, I]^{\mathrm{T}}\} + o_p(1)] \delta + o_p(n^{-1/2}).$$

Thus, by combining (A.8), (A.9), (A.10) and condition (C9), the desired distribution of $\widehat{\beta}_S$ follows.

A.4. Proof of Theorem 2. By the Taylor expansion, $\mu_0 = \mu(\beta_{c,0}, \delta/\sqrt{n}) = \mu(\beta_{c,0}, 0) + \mu_u^{\mathrm{T}} \delta/\sqrt{n} + o(n^{-1/2})$ and

$$\widehat{\mu}_S = \mu([I, 0]\Pi_S^{\mathrm{T}} \widehat{\beta}_S, [0, I]\Pi_S^{\mathrm{T}} \widehat{\beta}_S)$$

$$= \mu(\beta_{c,0}, 0) + \mu_\beta^{\mathrm{T}} \left\{ \Pi_S^{\mathrm{T}} \widehat{\beta}_S - \binom{\beta_{c,0}}{0} \right\} + o_p(n^{-1/2}),$$

where the second equation follows from the asymptotic normality of $\widehat{\beta}_S$. Thus, by Theorem 1,

$$\sqrt{n}(\widehat{\mu}_S - \mu_0) = \mu_\beta^{\mathrm{T}} \left\{ \Pi_S^{\mathrm{T}} \widehat{\beta}_S - \binom{\beta_{c,0}}{0} \right\} - \mu_u^{\mathrm{T}} \delta + o_p(1)$$

$$= -\mu_\beta^{\mathrm{T}} \mathbf{R}_S \mathbf{G}_n + \mu_\beta^{\mathrm{T}} \mathbf{R}_S \mathbf{D} \binom{0}{\delta} - \mu_u^{\mathrm{T}} \delta + o_p(1)$$

$$\xrightarrow{d} -\mu_\beta^{\mathrm{T}} \mathbf{R}_S \mathbf{G} + \mu_\beta^{\mathrm{T}} (\mathbf{R}_S \mathbf{D} - I) \binom{0}{\delta}.$$

Thus, the proof is complete.

A.5. Proof of Theorem 3. Recalling the definitions of Π_S and $\mathbf{R}_S$, we have

$$\mathbf{R}_S \mathbf{D} \begin{pmatrix} I & 0 \\ 0 & 0_{d_u \times d_u} \end{pmatrix} = \mathbf{R}_S \mathbf{D} \Pi_S^{\mathrm{T}} \begin{pmatrix} I & 0 \\ 0 & 0_{d_{u,s} \times d_u} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 0_{d_u \times d_u} \end{pmatrix},$$

which, along with the definition of $\widehat{\delta}$ and Theorem 2, indicates that

$$\sqrt{n}(\widehat{\mu} - \mu_0) = \sum_S w(S|\widehat{\delta}) \sqrt{n}(\widehat{\mu}_S - \mu_0)$$

$$= \sum_S w(S|\widehat{\delta}) \left\{ -\mu_\beta^{\mathrm{T}} \mathbf{R}_S \mathbf{G}_n + \mu_\beta^{\mathrm{T}} \mathbf{R}_S \mathbf{D} \binom{0}{\delta} - \mu_u^{\mathrm{T}} \delta + o_p(1) \right\}$$

$$
\begin{aligned}
&= \mu_\beta^{\mathrm{T}} \sum_S w(S|\widehat{\delta}) \mathbf{R}_S \mathbf{D} \begin{pmatrix} -[I,0]\mathbf{D}^{-1}\mathbf{G}_n \\ -[0,I]\mathbf{D}^{-1}\mathbf{G}_n + \delta \end{pmatrix} - \mu_u^{\mathrm{T}}\delta + o_p(1) \\
&= -\mu_\beta^{\mathrm{T}} \sum_S w(S|\widehat{\delta}) \mathbf{R}_S \mathbf{D} \begin{pmatrix} I & 0 \\ 0 & 0_{d_u \times d_u} \end{pmatrix} \mathbf{D}^{-1}\mathbf{G}_n \\
&\quad + \mu_\beta^{\mathrm{T}} \sum_S w(S|\widehat{\delta}) \mathbf{R}_S \mathbf{D} \begin{pmatrix} 0 \\ \widehat{\delta} \end{pmatrix} - \mu_u^{\mathrm{T}}(\widehat{\delta} + [0,I]\mathbf{D}^{-1}\mathbf{G}_n) + o_p(1) \\
&= -\mu_\beta^{\mathrm{T}} \mathbf{D}^{-1}\mathbf{G}_n + \mu_\beta^{\mathrm{T}} \left\{ Q(\widehat{\delta}) \begin{pmatrix} 0 \\ \widehat{\delta} \end{pmatrix} - \begin{pmatrix} 0 \\ \widehat{\delta} \end{pmatrix} \right\} + o_p(1) \\
&\xrightarrow{d} -\mu_\beta^{\mathrm{T}} \mathbf{D}^{-1}\mathbf{G} + \mu_\beta^{\mathrm{T}} \left\{ Q(\Delta) \begin{pmatrix} 0 \\ \Delta \end{pmatrix} - \begin{pmatrix} 0 \\ \Delta \end{pmatrix} \right\}
\end{aligned}
$$

and thus the proof is complete.

REFERENCES

AKAIKE, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **22** 203–217. MR0326953

BATES, J. M. and GRANGER, C. M. J. (1969). The combination of forecasts. *Operations Res. Quart.* **20** 451–468. MR0295497

BUCKLAND, S. T., BURNHAM, K. P. and AUGUSTIN, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53** 603–618.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555. MR0994249

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach*, 2nd ed. Springer, New York. MR1919620

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489. MR1467842

CLAESKENS, G. and CARROLL, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94** 249–265. MR2331485

CLAESKENS, G., CROUX, C. and VAN KERCKHOVEN, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* **62** 972–979. MR2297667

CLAESKENS, G., CROUX, C. and VAN KERCKHOVEN, J. (2007). Prediction-focused model selection for autoregressive models. *Aus. J. Statist.* **49** 359–379. MR2413576

CLAESKENS, G. and HJORT, N. L. (2003). The focused information criterion (with discussion). *J. Amer. Statist. Assoc.* **98** 900–916. MR2041482

CLAESKENS, G. and HJORT, N. L. (2008). *Model Selection and Model Averaging*. Cambridge Univ. Press, Cambridge. MR2431297

DANILOV, D. and MAGNUS, J. R. (2004). On the harm that ignoring pretesting can cause. *J. Econometrics* **122** 27–46. MR2082531

DE BOOR, C. (2001). *A Practical Guide to Splines*. Springer, New York. MR1900298

DRAPER, D. (1995). Assessment and propagation of model uncertainty. *J. R. Stat. Soc. Ser. B* **57** 45–70. MR1325378

FAN, J., FENG, Y. and SONG, R. (2009). Nonparametric independence screening in sparse ultra-high dimensional additive models. Technical report, Dept. Operations Research and Financial Engineering, Princeton Univ.

HAND, D. J. and VINCIOTTI, V. (2003). Local versus global models for classification problems: Fitting models where it matters. *Amer. Statist.* **57** 124–131. MR1977118

HANSEN, B. E. (2005). Challenges for econometric model selection. *Econ. Theory* **21** 60–68. MR2161958

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London. MR1082147

HÄRDLE, W., HUET, S., MAMMEN, E. and SPERLICH, S. (2004a). Bootstrap inference in semi-parametric generalized additive models. *Econom. Theory* **20** 265–300. MR2044272

HÄRDLE, W., MÜLLER, M., SPERLICH, S. and WERWATZ, A. (2004b). *Nonparametric and Semi-parametric Models*. Springer, New York.

HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators (with discussion). *J. Amer. Statist. Assoc.* **98** 879–945. MR2041481

HJORT, N. L. and CLAESKENS, G. (2006). Focussed information criteria and model averaging for Cox's hazard regression model. *J. Amer. Statist. Assoc.* **101** 1449–1464. MR2279471

HUANG, J. (1998). Functional ANOVA models for generalized regression. *J. Multivariate Anal.* **67** 49–71. MR1659096

HUANG, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27** 1536–1563. MR1742499

HUNSBERGER, S. (1994). Semiparametric regression in likelihood-based models. *J. Amer. Statist. Assoc.* **89** 1354–1365. MR1310226

HUNSBERGER, S., ALBERT, P., FOLLMANN, D. and SUH, E. (2002). Parametric and semiparametric approaches to testing for seasonal trend in serial count data. *Biostatistics* **3** 289–298.

KIM, T. H. and WHITE, H. (2001). James–Stein-type estimators in large samples with application to the least absolute deviations estimator. *J. Amer. Statist. Assoc.* **96** 697–705. MR1946435

LEEB, H. and PÖTSCHER, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34** 2554–2591. MR2291510

LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410. MR2242356

LIANG, H. (2008). Generalized partially linear models with missing covariates. *J. Multivariate Anal.* **99** 880–895. MR2405096

LIANG, H., THURSTON, S., RUPPERT, D., APANASOVICH, T. and HAUSER, R. (2008). Additive partial linear models with measurement errors. *Biometrika* **95** 667–678.

LIN, X. H. and CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96** 1045–1056. MR1947252

LINTON, O. B. and NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–101. MR1332841

MÜLLER, M. and RÖNZ, B. (2000). Credit scoring using semiparametric methods. In *Measuring Risk in Complex Stochastic Systems. Lecture Notes in Statistics* (J. Franks, W. Härdle and G. Stahl, eds.) **147** 83–98. Springer, Berlin.

POLLARD, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econom. Theory* **7** 186–199. MR1128411

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SEVERINI, T. A. and STANISWALIS, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501–511. MR1294076

SHEN, X., HUANG, H. C. and YE, J. (2004). Inference after model selection. *J. Amer. Statist. Assoc.* **99** 751–762. MR2090908

SHIBOSKI, C. S. (1998). Generalized additive models for current status data. *Lifetime Data Anal.* **4** 29–50.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606. MR0840516

STONE, C. J. (1994). The use of polynomial spline and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184. MR1272079

STONE, C. J., HANSEN, M. H., KOOPERBERG, C. and TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** 1371–1470. MR1463561

SUN, J., KOPCIUK, A. K. and LU, X. (2008). Polynomial spline estimation of partially linear single-index proportional hazards regression models. *Comput. Statist. Data Anal.* **53** 176–188. MR2528601

VAN DER VAART, A. and WELLNER, J. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

XUE, L. and YANG, L. (2006). Additive coefficient modeling via polynomial spline. *Statist. Sinica* **16** 1423–1446. MR2327498

YANG, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96** 574–586. MR1946426

YU, K., PARK, B. U. and MAMMEN, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970

INSTITUTE OF SYSTEMS SCIENCE
ACADEMY OF MATHEMATICS AND SYSTEM SCIENCE
CHINESE ACADEMY OF SCIENCES
BEIJING, 100190
CHINA
E-MAIL: xinyu@amss.ac.cn

DEPARTMENT OF BIOSTATISTICS
    AND COMPUTATIONAL BIOLOGY
UNIVERSITY OF ROCHESTER
ROCHESTER, NEW YORK 14642
USA
E-MAIL: hliang@bst.rochester.edu