

# An asymptotic theory for model selection inference in general semiparametric problems

BY GERDA CLAESKENS

*Operations Research & Business Statistics and University Center for Statistics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*  
 gerda.claeskens@econ.kuleuven.be

AND RAYMOND J. CARROLL

*Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.*  
 carroll@stat.tamu.edu

## SUMMARY

Hjort & Claeskens (2003) developed an asymptotic theory for model selection, model averaging and subsequent inference using likelihood methods in parametric models, along with associated confidence statements. In this article, we consider a semiparametric version of this problem, wherein the likelihood depends on parameters and an unknown function, and model selection/averaging is to be applied to the parametric parts of the model. We show that all the results of Hjort & Claeskens hold in the semiparametric context, if the Fisher information matrix for parametric models is replaced by the semiparametric information bound for semiparametric models, and if maximum likelihood estimators for parametric models are replaced by semiparametric efficient profile estimators. Our methods of proof employ Le Cam's contiguity lemmas, leading to transparent results. The results also describe the behaviour of semiparametric model estimators when the parametric component is misspecified, and also have implications for pointwise-consistent model selectors.

*Some key words:* Akaike information criterion; Bayes information criterion; Efficient semiparametric estimation; Frequentist model averaging; Model averaging; Model selection; Profile likelihood; Semiparametric model.

## 1. INTRODUCTION

We consider semiparametric models where the response  $Y$  is related to a vector of covariates  $Z$ , and where at the same time there is an unknown nonlinear relationship to a covariate  $X$ . Thus the model has a parametric component in  $Z$  and  $\beta$  and a nonparametric component  $\theta(X)$ . With normal errors, a typical example is a partially linear model where  $Y_i = Z_i^T \beta + \theta(X_i) + \varepsilon_i$ . In generalized linear models or in general likelihood problems, we start with a loglikelihood function

$$\sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\}, \quad (1)$$

where the value of  $\beta_{\text{true}}$  and the function  $\theta_{\text{true}}$  are unknown.

Our goal is to perform variable selection in the parametric part of the model, without assuming the nonparametric part to be known, and to obtain correct inference in the selected model.

Most other results in semiparametric model selection only consider partially linear models. Shi & Tsai (1999) use  $B$ -splines to estimate the nonparametric function  $\theta(\cdot)$  and develop a small-sample adjustment to Akaike's information criterion AIC. Fan & Li (2004) use local polynomial estimators in a longitudinal data setting and select the variables of the parametric part of the partially linear model by means of a penalized least squares criterion. Simonoff & Tsai (1999) developed an improvement to the AIC for variable selection in semiparametric and additive models. Naik & Tsai (2001) developed an AIC-type information criterion for use in single-index models, with extension to partially linear models. However, none of these articles deals with inference in the selected model. An exception is Bunea (2004), who studies post-model-selection inference in, again, partially linear regression models using penalized least squares estimation in combination with a construction of sieves.

In this article we go further than model selection by extending the frequentist model-averaging results of Hjort & Claeskens (2003) to semiparametric models.

## 2. DEFINITIONS AND MODEL ASSUMPTIONS

The true model (1) contains the parameter vector  $\beta_{\text{true}}$ , of which some components might be zero, and the unknown curve  $\theta_{\text{true}}(\cdot)$ . Since it is unsure whether or not all components of  $\beta$  are needed in the model, a model-selection criterion is applied. For simplicity we consider the case of two models of interest, namely a reduced model where  $\beta_{\text{red}}^T = (\alpha^T, 0_q^T)$  and a full model where  $\beta_{\text{full}}^T = (\alpha^T, \gamma^T)$ . As in Hjort & Claeskens (2003) we make the local misspecification assumption that the  $q$ -dimensional vector  $\gamma_{\text{true}} = \delta/\sqrt{n}$ . This implies that the true model is a distance  $O(1/\sqrt{n})$  away from the reduced model.

In the general setting, a model is chosen from amongst the  $2^q$  submodels of the full model. A particular submodel includes the full vector of coefficients  $\alpha$ , but only part of the coefficient vector  $\gamma$ ; the excluded components of  $\gamma$  are set to zero. The reduced model mentioned above is the simplest one, obtained by setting all of the components of  $\gamma$  equal to zero. Of course, one is free to allow only a few relevant submodels to participate in the model selection procedure, as opposed to dealing with all  $2^q$  possible models.

Under the full model, we have a set of responses  $Y$  and covariates  $Z$ , other covariates  $X$ , a parameter  $\beta$  and a function  $\theta(\cdot)$ , with a loglikelihood function  $\mathcal{L}\{Y, Z, \beta, \theta(X)\}$ . The true values are  $\beta_{\text{true}}$  and  $\theta_{\text{true}}(\cdot)$ . Partial derivatives of the loglikelihood function are denoted by

$$\begin{aligned}\mathcal{L}_\theta\{Y, Z, \beta, \theta(X)\} &= \left. \frac{\partial}{\partial v} \mathcal{L}(Y, Z, \beta, v) \right|_{v=\theta(x)}, \\ \mathcal{L}_\beta\{Y, Z, \beta, \theta(X)\} &= \frac{\partial}{\partial \beta} \mathcal{L}\{Y, Z, \beta, \theta(X)\}.\end{aligned}$$

The second derivatives are denoted by  $\mathcal{L}_{\beta\beta}(\cdot)$ , and so on.

In general, for any function  $F$ , we will use the following notation for partial and total derivatives:

$$\frac{\partial}{\partial \beta} F\{\beta, \theta(x, \beta)\} = \frac{\partial}{\partial u} F\{u, \theta(x, \beta)\}_{u=\beta} = F_\beta\{\beta, \theta(x, \beta)\},$$

$$\begin{aligned} \frac{d}{d\beta} F\{\beta, \theta(x, \beta)\} &= \frac{\partial}{\partial u} F\{u, \theta(x, \beta)\}_{u=\beta} + \frac{\partial}{\partial v} F\{\beta, v\}_{v=\theta(x, \beta)} \frac{\partial}{\partial \beta} \theta(x, \beta) \\ &= F_\beta\{\beta, \theta(x, \beta)\} + F_\theta\{\beta, \theta(x, \beta)\} \frac{\partial}{\partial \beta} \theta(x, \beta). \end{aligned}$$

The key assumptions that will hold in likelihood problems are

$$0 = E[\mathcal{L}_\theta\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(X)\} | X, Z], \quad (2)$$

$$0 = E[\mathcal{L}_\beta\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(X)\} | X, Z]. \quad (3)$$

Here and elsewhere in the article, the expectation is with respect to the true distribution of the data  $Y$ . Assumption (2) implies that  $E[\mathcal{L}_\theta\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(X)\} | X] = 0$ . For each fixed  $\beta$ , define  $\theta(x, \beta)$  as the solution to

$$E[\mathcal{L}_\theta\{Y, Z, \beta, \theta(X, \beta)\} | X = x] = 0. \quad (4)$$

Of course,  $\theta(\cdot, \beta_{\text{true}}) = \theta_{\text{true}}(\cdot)$ .

Let the subscript  $S$  refer either to the reduced model, where  $\gamma = 0_q$ , or to the full model that includes all  $q$   $\gamma$ -components. We define  $\hat{\theta}(x, \beta_S)$  as the local linear estimator of  $\theta(\cdot)$  at location  $x$ , when  $\beta = \beta_S$ . To be specific,  $\{\hat{\theta}(x; \beta_S), \hat{\theta}_1(x; \beta_S)\}$  is the maximizer, with respect to  $(\psi_0, \psi_1)$ , of

$$n^{-1} \sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \beta_S, \psi_0 + \psi_1(X_i - x)\} K_h(X_i - x), \quad (5)$$

where, for a kernel function  $K$  and bandwidth  $h$ ,  $K_h(\cdot) = K(\cdot/h)/h$ . If the first partial derivatives of the likelihood exist, we have the following set of estimating equations in the semiparametric model:

$$0 = n^{-1} \sum_{i=1}^n \mathcal{L}_\theta\{Y_i, Z_i, \beta_S, \psi_0 + \psi_1(X_i - x)\} K_h(X_i - x) (1, X_i - x)^\top.$$

The covariate  $X$  has density function  $f_X(\cdot)$ . Given the estimator  $\hat{\theta}(x, \beta_S)$ , we define the generalized profile likelihood estimator  $\hat{\beta}_S$  as the solution to

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n \frac{d}{d\beta} \mathcal{L}\{Y_i, Z_i, \beta_S, \hat{\theta}(X_i, \beta_S)\} \\ &= n^{-1} \sum_{i=1}^n \left[ \mathcal{L}_\beta\{Y_i, Z_i, \beta_S, \hat{\theta}(X_i, \beta_S)\} + \mathcal{L}_\theta\{Y_i, Z_i, \beta_S, \hat{\theta}(X_i, \beta_S)\} \frac{\partial}{\partial \beta_S} \hat{\theta}(X_i, \beta_S) \right]. \end{aligned}$$

For any given  $X$ , were  $\theta_{\text{true}}(\cdot)$  known, the Fisher information matrix would be calculated as follows. The matrix of conditional expected values of second derivatives given  $X$  is denoted by  $G(X)$ . This matrix and its inverse are partitioned as

$$G = G(X) = \begin{pmatrix} G_{\beta\beta} & G_{\beta\theta} \\ G_{\theta\beta} & G_{\theta\theta} \end{pmatrix}, \quad G^{-1} = G^{-1}(X) = \begin{pmatrix} G^{\beta\beta} & G^{\beta\theta} \\ G^{\theta\beta} & G^{\theta\theta} \end{pmatrix},$$

with

$$G_{\beta\beta} = E[\mathcal{L}_{\beta\beta}\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(x) | X],$$

$$G_{\beta\theta} = E[\mathcal{L}_{\beta\theta}\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(x)|X],$$

$$G_{\theta\theta} = E[\mathcal{L}_{\theta\theta}\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(x)|X],$$

$G_{\theta\beta} = G_{\beta\theta}^T$ , and  $G^{\beta\beta} = (G_{\beta\beta} - G_{\beta\theta}G_{\theta\theta}^{-1}G_{\beta\theta}^T)^{-1}$ . In parametric likelihood models in  $\beta$  induced by distributions given  $X$ ,  $-G(X)$  is the Fisher information matrix.

### 3. ASYMPTOTIC RESULTS

#### 3.1. Introduction

The reason for considering model selection is that we wish to estimate a specific function  $\mu(\beta)$ , though we do not know whether or not all of the components of  $\beta$  are needed. Our interest is in the distribution of  $\mu(\hat{\beta})$ , where  $\hat{\beta}$  is obtained through a model-selection procedure. The function  $\mu$  is assumed to possess continuous partial derivatives in a neighbourhood of the true parameter values. We obtain this distribution in several steps. First we study the nonparametric part of the model since an estimator of  $\theta_{\text{true}}(\cdot)$  is necessary to define the profile likelihood function. Next we continue with the parametric part. With the help of some lemmas we arrive at the distribution of the profile likelihood estimator  $\hat{\beta}$  in both reduced and full models, under the local misspecification assumptions. Technical details are given in the Appendix.

Our study of the profile likelihood estimator  $\hat{\beta}$  will make frequent use of the derivative of the curve  $\theta(x, \beta)$  with respect to  $\beta$ , for which we prove the following result.

LEMMA 1. *The derivative of the curve  $\theta(x, \beta)$  satisfies*

$$\frac{\partial}{\partial \beta} \theta(x, \beta_{\text{true}}) = \mathcal{G}(x) = -G_{\beta\theta}(x)/G_{\theta\theta}(x).$$

*Proof.* The lemma follows by differentiation of (4) with respect to  $\beta$  and solution of the resulting equation.  $\square$

#### 3.2. Main results

Our main results are stated as a series of Theorems. We first define the semiparametric information bound  $\mathcal{S}(\beta) = \text{cov}[\frac{d}{d\beta} \mathcal{L}\{Y, Z, \beta_{\text{true}}, \theta(X, \beta_{\text{true}})\}]$  and partition this matrix and its inverse as

$$\mathcal{S}(\beta) = \begin{pmatrix} S_{\alpha\alpha}(\beta) & S_{\alpha\gamma}(\beta) \\ S_{\gamma\alpha}(\beta) & S_{\gamma\gamma}(\beta) \end{pmatrix}, \quad \mathcal{S}^{-1}(\beta) = \begin{pmatrix} S^{\alpha\alpha}(\beta) & S^{\alpha\gamma}(\beta) \\ S^{\gamma\alpha}(\beta) & S^{\gamma\gamma}(\beta) \end{pmatrix}.$$

We give a basic expansion of the profile kernel method, first in the full model and then in the reduced model which sets all  $q$  components of  $\gamma$  equal to zero. The general situation, with more than two models to choose from and with only some of the components of  $\gamma$  set equal to zero, requires the same method of proof as for our simple reduced model, only with the notation becoming more cumbersome. For each model considered, there is a corresponding limiting distribution similar to the one given in Theorem 2 below. For example, the limiting covariance matrix is adjusted to the specific situation, selecting the correct part of the matrix  $\mathcal{S}(\beta)$ , so as to include only those rows and columns for which the corresponding component of  $\gamma$  is included in the model considered at that time. For a similar construction and the required notational issues in a parametric regression setting, see Hjort & Claeskens (2003). Recall that  $\beta_{\text{true}} = (\alpha_{\text{true}}^T, \delta^T/\sqrt{n})^T$ .

**THEOREM 1.** *Under the local misspecification assumption and when working in the full model, assuming Conditions A1–A4 in the Appendix, we have*

$$n^{1/2}(\hat{\beta}_{\text{full}} - \beta_{\text{true}}) = S^{-1}(\beta_{\text{true}})n^{-1/2} \sum_{i=1}^n \frac{d}{d\beta} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta(X_i, \beta_{\text{true}})\} + o_P(1).$$

*The limiting distribution of  $\hat{\beta}_{\text{full}}$  can now be constructed immediately:  $n^{1/2}(\hat{\beta}_{\text{full}} - \beta_{\text{true}}) \rightarrow N\{0, S^{-1}(\beta_{\text{true}})\}$ , in distribution.*

**THEOREM 2.** *If the reduced model holds, that is  $\gamma = 0_q$ , for the estimator  $\hat{\alpha}_{\text{red}}$  obtained by fitting the reduced model it holds that*

$$\begin{aligned} n^{1/2}(\hat{\alpha}_{\text{red}} - \alpha_{\text{true}}) &= S_{\alpha\alpha}^{-1}(\alpha_{\text{true}}, 0_q)n^{-1/2} \\ &\quad \times \sum_{i=1}^n \frac{d}{d\alpha} \mathcal{L}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta(X_i, \alpha_{\text{true}}, 0_q)\} + o_P(1). \end{aligned}$$

*Moreover,  $n^{1/2}(\hat{\alpha}_{\text{red}} - \alpha_{\text{true}}) \rightarrow N(0, S_{\alpha\alpha}^{-1})$ , in distribution.*

The proof of the first statement is very similar to the proof of Theorem 1. The second part follows immediately from the central limit theorem.

We now state two results describing what happens under the local model misspecification, one concerning the reduced model estimator when the full model holds, and the other describing the relationship between the full and reduced model estimators in this case.

**THEOREM 3.** *If the local misspecified model holds, that is  $\gamma_{\text{true}} = n^{-1/2}\delta$ , for the estimator  $\hat{\alpha}_{\text{red}}$  obtained by fitting the reduced model it holds that*

$$\begin{aligned} n^{1/2}(\hat{\alpha}_{\text{red}} - \alpha_{\text{true}}) &= S_{\alpha\alpha}^{-1}(\beta_{\text{true}})S_{\alpha\gamma}(\beta_{\text{true}})\delta \\ &\quad + n^{-1/2} \sum_{i=1}^n S_{\alpha\alpha}^{-1}(\beta_{\text{true}}) \frac{d}{d\alpha} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta(X_i, \beta_{\text{true}})\} + o_P(1) \\ &\rightarrow N\{S_{\alpha\alpha}^{-1}(\beta_{\text{true}})S_{\alpha\gamma}(\beta_{\text{true}})\delta, S_{\alpha\alpha}^{-1}(\beta_{\text{true}})\}, \end{aligned}$$

*in distribution.*

**THEOREM 4.** *Under the local misspecification assumption,*

$$\begin{aligned} n^{1/2}(\hat{\alpha}_{\text{full}} - \alpha_{\text{true}}) &= n^{1/2}(\hat{\alpha}_{\text{red}} - \alpha_{\text{true}}) - S_{\alpha\alpha}^{-1}(\beta_{\text{true}})S_{\alpha\gamma}(\beta_{\text{true}})\delta \\ &\quad + S^{\alpha\gamma}(\beta_{\text{true}})\{S^{\gamma\gamma}(\beta_{\text{true}})\}^{-1}n^{1/2}(\hat{\gamma}_{\text{full}} - \gamma_{\text{true}}) + o_P(1), \end{aligned}$$

*and the estimators  $\hat{\gamma}_{\text{full}}$  and  $\hat{\alpha}_{\text{red}}$  are asymptotically uncorrelated.*

The above discussion is summarized in the following theorem, which describes what happens to estimators of functions of the parameters under local model misspecification.

**THEOREM 5.** *Under the local misspecification assumption, in distribution,*

$$\begin{aligned} n^{1/2}\{\mu(\hat{\beta}_{\text{full}}) - \mu(\beta_{\text{true}})\} &\rightarrow \Lambda_{\text{full}} = \frac{\partial\mu}{\partial\beta} N\{0, S^{-1}(\beta_{\text{true}})\}, \\ n^{1/2}\{\mu(\hat{\beta}_{\text{red}}) - \mu(\beta_{\text{true}})\} &\rightarrow \Lambda_{\text{red}} = \frac{\partial\mu}{\partial\alpha} N\{S_{\alpha\alpha}^{-1}(\beta_{\text{true}})S_{\alpha\gamma}(\beta_{\text{true}})\delta, S_{\alpha\alpha}^{-1}(\beta_{\text{true}})\} - \frac{\partial\mu}{\partial\gamma}\delta. \end{aligned}$$

*Proof.* The results follow immediately via the delta method, and Theorems 1 and 3.  $\square$

When more than two models are considered, each of the models gives rise to its own limiting normal distribution. Selecting a model implies that we do not decide beforehand which of the models will be used; hence the distributions given above are only conditional on the model for which they are obtained. The distribution of the estimator in the selected model needs to take the model selection process into account. This is dealt with via model averaging, as explained in the next section.

#### 4. MODEL AVERAGING AND INFERENCE

##### 4.1. Model selection weights

Model selection leads to the simplest possible weighted estimator. The estimator after model selection is a weighted sum of the estimators in all of the considered models, where the estimator in the selected model receives weight one and all other estimators receive weight zero. In other words, we consider the estimator in the selected model only.

Let  $S$  be a subset of the index set  $\{1, \dots, q\}$ , and let  $\emptyset$  denote the smallest model, with no extra variable  $\gamma_j$ . We will use the words ‘model  $S$ ’ to indicate the model with additional variables  $\gamma_j$  for which  $j$  belongs to the set  $S$ . We consider explicitly the example of the AIC. We define AIC using the profile loglikelihood,

$$\sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \beta_S, \hat{\theta}(X_i, \beta_S)\};$$

see for example equation (6) of Murphy & van der Vaart (2000), who give a Taylor series expansion for the profile loglikelihood function, similar to that for the ‘full’ loglikelihood. The AIC is now defined as the penalized profile loglikelihood,

$$2 \sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \beta_S, \hat{\theta}(X_i, \beta_S)\} - 2|S|,$$

where  $|S|$  denotes the number of variables in the set  $S$ . The model with the largest value of AIC is selected. Let  $\hat{S}_{\text{aic}}$  denote the index set selected by AIC, and let  $\hat{\mu}(S)$  denote the estimator of  $\mu$  in model  $S$ . The final estimator after AIC model selection is

$$\hat{\mu} = \hat{\mu}(\hat{S}_{\text{aic}}) = \sum_{\text{all } S} I(\text{AIC selects model } S) \hat{\mu}(S).$$

We now show that the indicator value  $I(\text{AIC selects model } S)$  can be written as  $c(\hat{\delta}_{\text{full}}) + o_p(1)$ . For simplicity of derivation, we take the case of only two models, a reduced model and a full model. Exactly the same computation is needed to obtain the result in the case of more than two models.

Denote the semiparametric score by

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \frac{d}{d\beta} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \hat{\theta}(X_i, \beta_{\text{true}})\} \\ &= \begin{pmatrix} \bar{U}_{\alpha, n} \\ \bar{U}_{\gamma, n} \end{pmatrix} = \begin{pmatrix} n^{-1} \sum_{i=1}^n \frac{d}{d\alpha} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \hat{\theta}(X_i, \beta_{\text{true}})\} \\ n^{-1} \sum_{i=1}^n \frac{d}{d\gamma} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \hat{\theta}(X_i, \beta_{\text{true}})\} \end{pmatrix}. \end{aligned}$$

For the reduced model only the first component  $\bar{U}_{\alpha,n}$  is needed. From Theorem 1, we have

$$\hat{\beta}_{\text{full}} - \beta_{\text{true}} = \mathcal{S}^{-1} \begin{pmatrix} \bar{U}_{\alpha,n} \\ \bar{U}_{\gamma,n} \end{pmatrix} + o_P(n^{-1/2}). \quad (6)$$

Since  $\beta_{\text{full}} = (\alpha, \gamma)$ , working out the matrix product leads to the representation

$$\hat{\gamma}_{\text{full}} - \gamma_{\text{true}} = S^{\gamma\gamma}(\beta_{\text{true}}) \{ \bar{U}_{\gamma,n} - S_{\gamma\alpha}(\beta_{\text{true}}) S_{\alpha\alpha}^{-1} \bar{U}_{\alpha,n} \} + o_P(n^{-1/2}). \quad (7)$$

Using Theorem 3, we have

$$\hat{\beta}_{\text{red}} - \beta_{\text{true}} = S_{\alpha\alpha}^{-1}(\beta_{\text{true}}) S_{\alpha,\gamma}(\beta_{\text{true}}) \delta / n^{1/2} + S_{\alpha\alpha}^{-1}(\beta_{\text{true}}) \bar{U}_{\alpha,n} + o_P(n^{-1/2}). \quad (8)$$

Consider the AIC difference between the full and reduced models,

$$\begin{aligned} \Delta_n &= 2 \sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \hat{\beta}_{\text{full}}, \hat{\theta}(X_i, \hat{\beta}_{\text{full}})\} - 2 \sum_{i=1}^n \mathcal{L}\{Y_i, Z_i, \hat{\beta}_{\emptyset}, \hat{\theta}(X_i, \hat{\beta}_{\emptyset})\} - 2q \\ &= 2 \sum_{i=1}^n [\mathcal{L}\{Y_i, Z_i, \hat{\beta}_{\text{full}}, \hat{\theta}(X_i, \hat{\beta}_{\text{full}})\} - \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \hat{\theta}(X_i, \beta_{\text{true}})\}] \\ &\quad - 2 \sum_{i=1}^n [\mathcal{L}\{Y_i, Z_i, \hat{\alpha}_{\text{red}}, 0_q, \hat{\theta}(X_i, \hat{\alpha}_{\text{red}}, 0_q)\} - \mathcal{L}\{Y_i, Z_i, \alpha_{\text{true}}, 0_q, \hat{\theta}(X_i, \alpha_{\text{red}}, 0_q)\}] \\ &\quad + 2 \sum_{i=1}^n [\mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \hat{\theta}(X_i, \beta_{\text{true}})\} - \mathcal{L}\{Y_i, Z_i, \alpha_{\text{true}}, 0_q, \hat{\theta}(X_i, \alpha_{\text{red}}, 0_q)\}] - 2q. \end{aligned}$$

By a standard two-term Taylor expansion based on the total derivative ‘ $d$ ’ and not the partial derivative, we obtain

$$\begin{aligned} \Delta_n &= n \left\{ (\hat{\beta}_{\text{full}} - \beta_{\text{true}})^T \mathcal{S}(\hat{\beta}_{\text{full}} - \beta_{\text{true}}) - (\hat{\alpha}_{\text{red}} - \alpha_{\text{true}}) S_{\alpha\alpha} (\hat{\alpha}_{\text{red}} - \alpha_{\text{true}}) + 2(\delta/\sqrt{n})^T \bar{U}_{\gamma,n} \right\} \\ &\quad + \delta^T S_{\gamma\gamma} \delta - 2q + o_P(1), \end{aligned}$$

since, for example,

$$\begin{aligned} &2 \sum_{i=1}^n [\mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \hat{\theta}(X_i, \beta_{\text{true}})\} - \mathcal{L}\{Y_i, Z_i, \alpha_{\text{true}}, 0_q, \hat{\theta}(X_i, \alpha_{\text{red}}, 0_q)\}] \\ &= 2n\gamma_{\text{full}}^T \bar{U}_{\gamma,n} + n\gamma_{\text{full}} S_{\gamma\gamma} \gamma_{\text{full}} + o_P(1) = 2n(\delta/n^{1/2})^T \bar{U}_{\gamma,n} + \delta^T S_{\gamma\gamma} \delta + o_P(1). \end{aligned}$$

Using (6) and (8) leads to

$$\begin{aligned} \Delta_n &= n \begin{pmatrix} \bar{U}_{\alpha,n} & \bar{U}_{\gamma,n} \end{pmatrix} \mathcal{S}^{-1} \begin{pmatrix} \bar{U}_{\alpha,n} & \bar{U}_{\gamma,n} \end{pmatrix}^T + o_P(1) \\ &\quad + n \{ -(S_{\alpha\alpha}^{-1} S_{\alpha\gamma} \delta / \sqrt{n} + S_{\alpha\alpha}^{-1} \bar{U}_{\alpha,n})^T S_{\alpha\alpha} (S_{\alpha\alpha}^{-1} S_{\alpha\gamma} \delta / \sqrt{n} + S_{\alpha\alpha}^{-1} \bar{U}_{\alpha,n}) + 2(\delta/\sqrt{n})^T \bar{U}_{\gamma,n} \} \\ &\quad + \delta^T S_{\gamma\gamma} \delta - 2q \\ &= n \{ \bar{U}_{\alpha,n}^T (S^{\alpha\alpha} - S_{\alpha\alpha}^{-1}) \bar{U}_{\alpha,n} + 2\bar{U}_{\alpha,n}^T S^{\alpha\gamma} \bar{U}_{\gamma,n} + \bar{U}_{\gamma,n}^T S^{\gamma\gamma} \bar{U}_{\gamma,n} \\ &\quad + 2(\delta/\sqrt{n})^T \bar{U}_{\gamma,n} - 2(\delta/\sqrt{n})^T S_{\gamma\alpha} S_{\alpha\alpha}^{-1} \bar{U}_{\alpha,n} \} \\ &\quad + \delta^T (S_{\gamma\gamma} - S_{\gamma\alpha} S_{\alpha\alpha}^{-1} S_{\alpha\gamma}) \delta - 2q + o_P(1). \end{aligned}$$

Now use the expansion (7) for  $\hat{\gamma}_{\text{full}} - \gamma_{\text{true}}$ , along with the fact that

$$(S^{\gamma\gamma})^{-1} = S_{\gamma\gamma} - S_{\gamma\alpha} S_{\alpha\alpha}^{-1} S_{\alpha\gamma},$$

to see that we have

$$\begin{aligned} \Delta_n &= n \left\{ \bar{U}_{\alpha n}^T (S^{\alpha\alpha} - S_{\alpha\alpha}^{-1}) \bar{U}_{\alpha,n} + 2 \bar{U}_{\alpha,n}^T S^{\alpha\gamma} \bar{U}_{\gamma,n} + \bar{U}_{\gamma,n}^T S^{\gamma\gamma} \bar{U}_{\gamma,n} \right. \\ &\quad \left. + 2(\delta/\sqrt{n})^T (S^{\gamma\gamma})^{-1} (\hat{\gamma}_{\text{full}} - \gamma_{\text{true}}) \right\} + \delta^T (S^{\gamma\gamma})^{-1} \delta - 2q + o_P(1) \\ &= n \left\{ (\bar{U}_{\gamma,n} - S_{\gamma\alpha} S_{\alpha\alpha}^{-1} \bar{U}_{\alpha,n})^T S^{\gamma\gamma} (\bar{U}_{\gamma,n} - S_{\gamma\alpha} S_{\alpha\alpha}^{-1} \bar{U}_{\alpha,n}) \right. \\ &\quad \left. + 2(\delta/\sqrt{n})^T (S^{\gamma\gamma})^{-1} (\hat{\gamma}_{\text{full}} - \gamma_{\text{true}}) \right\} + \delta^T (S^{\gamma\gamma})^{-1} \delta - 2q + o_P(1) \\ &= n (\hat{\gamma}_{\text{full}} - \gamma_{\text{true}} + \delta/\sqrt{n})^T (S^{\gamma\gamma})^{-1} (\hat{\gamma}_{\text{full}} - \gamma_{\text{true}} + \delta/\sqrt{n}) - 2q + o_P(1) \\ &= n \hat{\gamma}_{\text{full}}^T (S^{\gamma\gamma})^{-1} \hat{\gamma}_{\text{full}} - 2q + o_P(1). \end{aligned}$$

The criterion AIC selects the full model if the AIC value for the full model is larger than the value for the reduced model, that is, when  $\Delta_n > 0$ . This is equivalent to selecting the full model when  $\hat{\delta}_{\text{full}}^T (S^{\gamma\gamma})^{-1} \hat{\delta}_{\text{full}} > 2q$ . Therefore, for AIC model selection, the weight  $c(\hat{\delta}_{\text{full}})$  is

$$c(\hat{\delta}_{\text{full}}) = I\{\hat{\delta}_{\text{full}}^T (S^{\gamma\gamma})^{-1} \hat{\delta}_{\text{full}} > 2q\},$$

and the AIC-selected estimator equals

$$\hat{\mu} = c(\hat{\delta}_{\text{full}}) \mu(\hat{\beta}_{\text{full}}) + \{1 - c(\hat{\delta}_{\text{full}})\} \mu(\hat{\beta}_{\text{red}}).$$

The above statement shows that under local misspecification the probability that AIC selects the full model, with similar statements holding for models other than the full one, is the probability that a noncentral chi-squared variable exceeds a certain threshold, in this case equal to  $2q$ . Indeed, since  $\hat{\delta}_{\text{full}} = n^{1/2} \hat{\gamma}_{\text{full}} \rightarrow D = N(\delta, S^{\gamma\gamma})$ , in distribution, it follows that

$$\hat{\delta}_{\text{full}}^T (S^{\gamma\gamma})^{-1} \hat{\delta}_{\text{full}} \rightarrow \chi_q^2\{\delta^T (S^{\gamma\gamma})^{-1} \delta\},$$

in distribution. If there is no local misspecification, Woodroffe (1982) obtains, using central chi-squared variables, the generalized arc-sine laws, which give the probabilities that AIC selects a certain model order in a sequence of nested models.

#### 4.2. Limit results and confidence sets

Theorem 5 is the main ingredient for obtaining the distribution of estimators after model selection. Here we follow the approach leading to Theorem 4.1 of Hjort & Claeskens (2003), with the approach to inference following their §4.3.

Estimators after model selection are viewed as weighted sums of the estimators in the separate models under consideration. The simplest type of weight functions are indicator functions, pointing to the selected model; an example based on the AIC is given in §4.1. Since the choice of the model is data-dependent, the weights are random and depend on the data. Instead of zero/one weights, other weight functions, with values between zero and one, can be chosen. For example, the AIC model-averaged estimator assigns weights

$$\frac{\exp\{(AIC_S - AIC_{\emptyset})/2\}}{\sum_{\text{all } S'} \exp\{(AIC_{S'} - AIC_{\emptyset})/2\}}$$

to the model  $S$ . The denominator takes the sum over all considered models, and hence guarantees that the total weight equals one. In particular, for the simple case of only two models, the weight for the model indexed by  $S$  equals  $\exp\{(\text{AIC}_S - \text{AIC}_{\text{red}})/2\}/[1 + \exp\{(\text{AIC}_{\text{full}} - \text{AIC}_{\text{red}})/2\}]$ . For the BIC, similar model averaging weights are defined,

$$\frac{\exp\{(\text{BIC}_S - \text{BIC}_{\emptyset})/2\}}{\sum_{\text{all } S'} \exp\{(\text{BIC}_{S'} - \text{BIC}_{\emptyset})/2\}},$$

see Burnham & Anderson (2002) for examples of their use.

We consider cases where model selection and model averaging are based on weights depending on  $\hat{\delta}_{\text{full}} = n^{1/2}\hat{\gamma}_{\text{full}} \rightarrow D = N(\delta, S^{\gamma\gamma})$ , in distribution; see §4.1 for more discussion and a derivation in the case of the AIC. In fact, the calculations above show that any model selection procedure for which the random part can be written as a function of the likelihood ratio statistic depends on the data through  $\hat{\delta}_{\text{full}}$  only.

We then immediately have the following result.

**THEOREM 6.** *Recall that  $D$  is the limiting distribution of  $n^{1/2}\hat{\gamma}_{\text{full}}$ , and that  $\Lambda_{\text{full}}$  and  $\Lambda_{\text{red}}$  are described in Theorem 5. Then, under the local misspecification assumption,  $n^{1/2}\{\hat{\mu} - \mu(\beta_{\text{true}})\} \rightarrow \Lambda = c(D)\Lambda_{\text{full}} + \{1 - c(D)\}\Lambda_{\text{red}}$ , in distribution.*

It is obvious from these calculations that the weights need only equal  $c(\hat{\delta}_{\text{full}}) + o_p(1)$ . Thus, for example, these results apply if one uses AIC or BIC based on the semiparametric profile loglikelihood.

We can combine Theorem 6 with the methods in §4.3 of Hjort & Claeskens to develop asymptotically correct confidence limits for  $\mu(\beta_{\text{true}})$ . Let  $\mu_{\alpha} = \mu_{\alpha}(\beta_{\text{true}}) = \{\partial\mu(\beta_{\text{true}})\}/\partial\alpha$  and let  $\mu_{\gamma} = \mu_{\gamma}(\beta_{\text{true}}) = \{\partial\mu(\beta_{\text{true}})\}/\partial\gamma$ .

Using Theorem 5, we have

$$\Lambda_{\text{full}} = \begin{pmatrix} \mu_{\alpha} \\ \mu_{\gamma} \end{pmatrix}^{\text{T}} \begin{pmatrix} S^{\alpha\alpha} & S^{\alpha\gamma} \\ S^{\gamma\alpha} & S^{\gamma\gamma} \end{pmatrix} \begin{pmatrix} M_{\alpha} \\ M_{\gamma} \end{pmatrix},$$

where  $(M_{\alpha}, M_{\gamma})^{\text{T}} \sim N(0, S)$ . Also, define  $W = S^{\gamma\alpha}M_{\alpha} + S^{\gamma\gamma}M_{\gamma}$ . The random variables  $M_{\alpha}$  and  $W$  are stochastically independent and  $\hat{\delta}_{\text{full}} \rightarrow D = \delta + W$ , in distribution. We rewrite  $\Lambda_{\text{full}}$  as

$$\begin{aligned} \Lambda_{\text{full}} &= (\mu_{\alpha}^{\text{T}}S^{\alpha\alpha} + \mu_{\gamma}^{\text{T}}S^{\gamma\alpha})M_{\alpha} + (\mu_{\alpha}^{\text{T}}S^{\alpha\gamma} + \mu_{\gamma}^{\text{T}}S^{\gamma\gamma})M_{\gamma} \\ &= \mu_{\alpha}^{\text{T}}(S_{\alpha\alpha}^{-1} - S_{\alpha\alpha}^{-1}S_{\alpha\gamma}S^{\gamma\alpha})M_{\alpha} + \mu_{\gamma}^{\text{T}}S^{\gamma\alpha}M_{\alpha} + (-\mu_{\alpha}^{\text{T}}S_{\alpha\alpha}^{-1}S_{\alpha\gamma}S^{\gamma\gamma} + \mu_{\gamma}^{\text{T}}S^{\gamma\gamma})M_{\gamma} \\ &= \mu_{\alpha}^{\text{T}}S_{\alpha\alpha}^{-1}M_{\alpha} + (S_{\gamma\alpha}S_{\alpha\alpha}^{-1}\mu_{\alpha} - \mu_{\gamma})^{\text{T}}(-S^{\gamma\alpha}M_{\alpha} - S^{\gamma\gamma}M_{\gamma}) \\ &= \mu_{\alpha}^{\text{T}}S_{\alpha\alpha}^{-1}M_{\alpha} + (S_{\gamma\alpha}S_{\alpha\alpha}^{-1}\mu_{\alpha} - \mu_{\gamma})^{\text{T}}(\delta - D). \end{aligned}$$

Furthermore, from Theorem 5,  $\Lambda_{\text{red}} = \mu_{\alpha}^{\text{T}}S_{\alpha\alpha}^{-1}M_{\alpha} + (S_{\gamma\alpha}S_{\alpha\alpha}^{-1}\mu_{\alpha} - \mu_{\gamma})^{\text{T}}\delta$ . With these expressions it is easier to derive the mean and variance of  $\Lambda$  which are needed for confidence interval construction. We use  $\omega$  as an abbreviation for  $S_{\gamma\alpha}S_{\alpha\alpha}^{-1}\mu_{\alpha} - \mu_{\gamma}$ . It follows that

$$\begin{aligned} E(\Lambda) &= \omega^{\text{T}}[\delta - E\{c(D)D\}], \\ \text{var}(\Lambda) &= \mu_{\alpha}^{\text{T}}S_{\alpha\alpha}^{-1}\mu_{\alpha} + \omega^{\text{T}}\text{var}\{c(D)D\}\omega. \end{aligned}$$

For confidence limits we consider

$$\hat{\mu} - \hat{\omega}^{\text{T}}\{[1 - c(\hat{\delta}_{\text{full}})]\hat{\delta}_{\text{full}}\}/n^{1/2} \pm z_u\hat{\kappa}/n^{1/2},$$

where  $z_u$  is the  $u$ th standard normal quantile and  $\hat{\kappa}^2$  is a consistent estimator of the variance in the full model,  $\kappa^2 = \mu_\alpha^\top S_{\alpha\alpha}^{-1} \mu_\alpha + \omega^\top S^{\gamma\gamma} \omega$ . We substituted the unknown parameters by  $\hat{\beta}_{\text{full}}$ . The constructed confidence interval has asymptotically the correct coverage probability since, by arguments similar to those used to prove Theorem 6, it follows that, in distribution,

$$(n^{1/2}(\hat{\mu} - \mu_{\text{true}}), \hat{\delta}_{\text{full}}) \rightarrow (\mu_\alpha^\top S_{\alpha\alpha}^{-1} M_\alpha + \omega^\top \{\delta - c(D)D\}, D),$$

and also that, in distribution,

$$T_n = [n^{1/2}(\hat{\mu} - \mu_{\text{true}}) - \hat{\omega}\{\hat{\delta}_{\text{full}} - c(\hat{\delta}_{\text{full}})\hat{\delta}_{\text{full}}\}]/[\hat{\kappa} \rightarrow \mu_\alpha^\top S_{\alpha\alpha}^{-1} M_\alpha + \omega^\top \{\delta - c(D)D\}]/\kappa,$$

which follows a standard normal distribution.

## 5. SIMULATION EXAMPLE

We performed a small simulation study for the partially linear Gaussian model

$$Y_i = Z_i^\top \mathcal{B} + \theta(X_i) + \epsilon_i,$$

where  $Z_i = (Z_{i1}, Z_{i2})^\top$ ,  $\epsilon_i \sim N(0, \sigma^2)$ ,  $\mathcal{B} = (\mathcal{B}_1, \mathcal{B}_2)$ ,  $\beta = (\sigma^2, \mathcal{B}^\top)$ ,  $\alpha = (\sigma^2, \mathcal{B}_1)$  and  $\gamma = \mathcal{B}_2$ . In the simulation, we took  $\sigma^2 = 0.20$ ,  $\mathcal{B}_1 = 1$ ,  $n = 100, 200$ ,  $\theta(x) = \sin(8x - 2)$ ,  $X_i$  uniform on  $[0, 1]$  and  $Z_i$  bivariate normal with mean zero, variances  $1/12$  and correlation  $0.70$ . We varied  $\mathcal{B}_2 = cn^{-1/2}$  for  $c = 0.0, 0.5, 1.0, \dots, 10.0$ . The experiment was repeated 2000 times in each configuration, and we used the Epanechnikov kernel function. To cut down on Monte-Carlo variability, the same random numbers were used for each value of  $c$ .

In our calculations, we estimated the bandwidth as follows. First, we regressed  $Y$ ,  $Z_1$  and  $Z_2$  separately on  $X$ , using the ‘Direct Plug-In’ bandwidth selection method of Ruppert et al. (1995) to form different estimated bandwidths on each. We then calculated the residuals from these fits, and regressed the residual in  $Y$  on the residual in  $(Z_1, Z_2)$  to obtain a preliminary estimate  $\hat{\beta}_{\text{start}}$  of  $\beta$ . Following this, we regressed  $Y - Z^\top \hat{\beta}_{\text{start}}$  on  $X$  to obtain a common bandwidth, then undersmoothed it by multiplication by  $n^{-2/15}$  to obtain a bandwidth of order  $n^{-1/3}$  to eliminate bias, and then reestimated  $\beta$  and  $\theta(\cdot)$ . This algorithm was repeated for the reduced model.

The calculations are relatively straightforward. It is readily seen that the profile loglikelihood is  $\mathcal{L}(\beta) = -(1/2)\log(\sigma^2) - (2\sigma^2)^{-1}(R_y - R_z^\top \mathcal{B})^2$ , where  $R_y = Y - E(Y|X)$  and  $R_z = Z - E(Z|X)$ . The score then is

$$\begin{bmatrix} -(2\sigma^2)^{-1} + (2\sigma^4)^{-1}(R_y - R_z^\top \mathcal{B})^2 \\ (\sigma^2)^{-1} R_z (R_y - R_z^\top \mathcal{B}) \end{bmatrix},$$

and the information bound then becomes

$$\begin{bmatrix} (2\sigma^4)^{-1} & 0 \\ 0 & \sigma^{-2} E(R_z R_z^\top) \end{bmatrix} = \begin{bmatrix} (2\sigma^4)^{-1} & 0 \\ 0 & \sigma^{-2} \Omega \end{bmatrix}.$$

Our goal is to estimate  $\mathcal{B}_1 = (0, 1, 0)\beta$ . This means that  $\mu_\gamma(\beta_{\text{true}}) = 0$  and that  $\mu_\alpha(\beta_{\text{true}}) = (0, 1)^\top$ .

The Hjort & Claeskens confidence interval is the following. Its lower and upper values for 95% coverage are

$$\begin{aligned} \text{low} &= \hat{\mu} - \hat{\omega}^\top \{[1 - c(\hat{\delta}_{\text{full}})]\hat{\delta}_{\text{full}}\} / \sqrt{n} - 1.96\hat{\kappa} / \sqrt{n}, \\ \text{high} &= \hat{\mu} - \hat{\omega}^\top \{[1 - c(\hat{\delta}_{\text{full}})]\hat{\delta}_{\text{full}}\} / \sqrt{n} + 1.96\hat{\kappa} / \sqrt{n}, \end{aligned}$$

where  $\hat{\mu} = \hat{\mathcal{B}}_1$ , and the other terms are defined as follows:  $\hat{\delta}_{\text{full}} = n^{1/2}\hat{\mathcal{B}}_2$ ,  $c(\hat{\delta}_{\text{full}})$  is the weight for the full model,

$$\hat{\Omega} = n^{-1} \sum_{i=1}^n R_{zi} R_{zi}^T = \begin{bmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{12} & \hat{\Omega}_{22} \end{bmatrix},$$

$\mu_\alpha = (0, 1)^T$ ,  $\mu_\gamma = 0$ , leading to  $\hat{\omega} = \hat{\Omega}_{12}/\hat{\Omega}_{11}$ , and with  $c = \hat{\Omega}_{11}\hat{\Omega}_{22} - \hat{\Omega}_{12}^2$  it follows that  $\hat{\kappa} = (\hat{\sigma}^2\hat{\Omega}_{22}/c)^{1/2}$ .

In summary, if the weight attached to the full model is  $\hat{c}_n$ , then the confidence interval has endpoints

$$\begin{aligned} \text{low} &= \hat{\mathcal{B}}_1 - \hat{\omega}(1 - \hat{c}_n)\hat{\mathcal{B}}_2 - 1.96\hat{\kappa}/\sqrt{n}, \\ \text{high} &= \hat{\mathcal{B}}_1 - \hat{\omega}(1 - \hat{c}_n)\hat{\mathcal{B}}_2 + 1.96\hat{\kappa}/\sqrt{n}. \end{aligned}$$

The AIC and BIC weights for model selection and model averaging are computed exactly as described in §4.

When we used the model-averaged AIC estimator, the coverage properties were quite good. In all situations, for both  $n = 100$  and  $n = 200$ , the actual coverage of the nominal 90% intervals ranged between 0.88 and 0.89, while the actual coverage of the nominal 95% intervals ranged between 0.935 and 0.940. These intervals were very similar to intervals based on fitting the full model only. In contrast, when we selected the model and then used the standard errors from that selected model, neither AIC nor BIC performed well. The former had minimum coverage of 0.71 for a nominal 95% interval, while the latter's coverage had minimum value 0.46. The method based on BIC in particular had significant bias for estimating  $\mathcal{B}_1$ . For the 95% intervals, the mean lengths for the confidence intervals are 0.875 for the interval constructed as described in §4.2, while the intervals based on the naive method, without using the limiting distribution  $\Lambda$ , for AIC had mean length 0.710 and for BIC the mean length was 0.651.

The confidence intervals using the correct procedure are indeed wider, leading to higher coverage. The BIC-selected confidence intervals are badly biased, and combined with the smallest length this leads to the lowest coverage in this comparison.

## 6. DISCUSSION

Our work has focused on the case in which  $X$  is scalar, although because of the contiguity argument employed we expect the results to hold when  $X$  is multivariate. Other special cases await further development, such as the partially linear additive model with mean  $Z^T\beta + \sum_{j=1}^m \theta_j(X_j)$ .

In our simulation we found that BIC estimates and confidence intervals had biases and very poor coverage probabilities, as low as 46% for a nominal 95% interval. This may seem somewhat surprising, given that BIC is known to be a consistent model selector. As Leeb & Pötscher (2005) point out in parametric problems, however, and as our results verify in semiparametric problems, BIC is not a uniformly consistent model selector; that is, for fixed misspecification, BIC can consistently distinguish between models, but, for local misspecification, it cannot consistently distinguish between models. This lack of uniform consistency translates into the bias and poor coverage that we observe for BIC. Of course,

this problem is not restricted to BIC, and can be shown using our asymptotic theory on a case-by-case basis to obtain for other so-called consistent model selectors.

#### ACKNOWLEDGEMENT

The authors thank the editor, associate editor and referee for their careful reading and comments, which improved the readability of our manuscript. Carroll's research was supported by a grant from the U.S. National Cancer Institute, and by the Texas A&M Center for Environmental and Rural Health via a grant from the U.S. National Institute of Environmental Health Sciences. The research of Claeskens was partly supported by the Fund for Scientific Research Flanders.

#### APPENDIX

##### *Technical details*

*Regularity conditions.* We require the following conditions.

*Condition A1.* The bandwidth sequence  $h_n \rightarrow 0$  as  $n \rightarrow \infty$ , in such a way that  $nh_n/\log(n) \rightarrow \infty$  and  $h_n \geq \{\log(n)/n\}^{1-2/\lambda}$  for  $\lambda$  as in Condition A4.

*Condition A2.* The kernel function  $K$  is a symmetric, continuously differentiable probability density function on  $[-1, 1]$  taking the value zero at the boundaries. The design density  $f_X$  is differentiable on  $B = [b_1, b_2]$ , the derivative is continuous, and  $\inf_{x \in B} f_X(x) > 0$ . The function  $\theta(\cdot, \beta)$  has two continuous derivatives on  $B$  and is also twice differentiable with respect to  $\beta$ .

*Condition A3.* For  $\beta \neq \beta'$ , the Kullback-Leibler distance between  $\mathcal{L}\{\cdot, \cdot, \beta, \theta(\cdot, \beta)\}$  and  $\mathcal{L}\{\cdot, \cdot, \beta', \theta(\cdot, \beta')\}$  is strictly positive. For every  $(y, z)$ , third partial derivatives of  $\mathcal{L}\{y, z, \beta, \theta(x)\}$  with respect to  $\beta$  exist and are continuous in  $\beta$ . The fourth partial derivative exists for almost all  $(y, z)$ . Furthermore, mixed partial derivatives  $(\partial^{r+s}/\partial\beta^r\partial v^s)\mathcal{L}\{y, z, \beta, v\}|_{v=\theta(x)}$ , with  $0 \leq r, s \leq 4$ ,  $r+s \leq 4$ , exist for almost all  $(y, z)$  and  $E\left[\sup_{\beta} \sup_v |(\partial^{r+s}/\partial\beta^r\partial v^s)\mathcal{L}\{y, z, \beta, v\}|^2\right] < \infty$ . The Fisher information,  $G(x)$ , possesses a continuous derivative and  $\inf_{x \in B} G(x) > 0$ .

*Condition A4.* There exists a neighbourhood  $\mathcal{N}\{\beta_{\text{true}}, \theta_{\text{true}}(x)\}$  such that

$$\max_{k=1,2} \sup_{x \in B} \left\| \sup_{(\beta, \theta) \in \mathcal{N}\{\beta_{\text{true}}, \theta_{\text{true}}(x)\}} \left| \frac{\partial^k}{\partial \theta^k} \log\{\mathcal{L}(Y, Z, \beta, \theta)\} \right| \right\|_{\lambda, x} < \infty$$

for some  $\lambda \in (2, \infty]$ , where  $\|\cdot\|_{\lambda, x}$  is the  $L^\lambda$ -norm, conditional on  $X = x$ . Furthermore,

$$\sup_{x \in B} E_x \left[ \sup_{(\beta, \theta) \in \mathcal{N}\{\beta_{\text{true}}, \theta_{\text{true}}(x)\}} \left| \frac{\partial^3}{\partial \theta^3} \log\{\mathcal{L}(Y, Z, \beta, \theta)\} \right| \right] < \infty.$$

The regularity conditions stated above are the same as those used in a local likelihood setting where one wishes to obtain strong uniform consistency of the local likelihood estimators. This is needed in Lemma's A1–A3. Condition A3 requires the fourth partial derivative of the log profile likelihood to have a bounded second moment, and it further requires the Fisher information matrix to be invertible and to be differentiable with respect to  $x$ . Condition A4 requires a bound on the first and second derivatives of the log profile likelihood and on the first moment of the third partial derivative, in a neighbourhood of the true parameter values.

*Asymptotic theory for the nonparametric part of the model.* For each fixed value of  $\beta$ , the local linear estimator  $\hat{\theta}(x, \beta)$  exists and is a strongly consistent estimator of  $\theta(x, \beta)$  defined in (4). This follows from local likelihood calculations. See for example Theorem 2.1 in Claeskens & Van

Keilegom (2003); precise regularity conditions are formulated above. We summarize the strong uniform consistency result in the first part of Lemma A1, and add a result about the derivatives with respect to the parameters  $\beta$ .

**LEMMA A1.** *As  $n \rightarrow \infty$ , and under Conditions A1–A4 on the kernel, bandwidth and likelihood function,  $\hat{\theta}(x, \beta)$  and  $\hat{\theta}_1(x, \beta)$  exist and  $\sup_x |\hat{\theta}(x, \beta) - \theta(x, \beta)| = O[\{nh/\log(n)\}^{-1/2} + h^2]$  almost surely. For the estimator of the derivative of the curve it follows that  $\sup_x |\hat{\theta}_1(x, \beta) - (\partial/\partial x)\theta(x, \beta)| = O[\{nh^3/\log(n)\}^{-1/2} + h^2]$  almost surely. Furthermore,  $(\partial/\partial \beta)\hat{\theta}(x, \beta)$  exists, is strongly consistent and  $\sup_x |(\partial/\partial \beta)\hat{\theta}(x, \beta) - (\partial/\partial \beta)\theta(x, \beta)| = O_P[\{nh/\log(n)\}^{-1/2} + h^2]$ . For some  $\delta > 0$ ,  $\sup_x |(\partial^2/\partial x \partial \beta)\hat{\theta}(x, \beta) - (\partial^2/\partial x \partial \beta)\theta(x, \beta)| = o_P(n^{-\delta})$ .*

*Proof.* The first part of the lemma has been shown in Theorem 2.1 in Claeskens & Van Keilegom (2003). For the part about the derivatives with respect to  $\beta$ , define, for fixed  $x$ , the function

$$u(\beta_S, \psi_0) = n^{-1} \sum_{i=1}^n \mathcal{L}_\theta\{Y_i, Z_i, \beta_S, \psi_0 + \hat{\theta}_1(x, \beta_S)(X_i - x)\} K_h(X_i - x).$$

By the first part of this lemma,  $\hat{\theta}_1(x, \beta_S)$  is a strongly consistent estimator of  $\theta_1(x, \beta_S)$ . Since by Condition A3 the Fisher information matrix is positive definite, and the design density  $f_X(x) > 0$  by Condition A2, the implicit function theorem implies that the function  $\beta_S \rightarrow \hat{\theta}_0(x, \beta_S)$  is a  $C^1$  function. As a consequence there exists a neighbourhood of  $\hat{\beta}_S$  such that, for all  $\beta_S$  in this neighbourhood,

$$0 = \frac{d}{d\beta} u\{\beta_S, \hat{\theta}_0(x, \beta_S)\} = \frac{\partial}{\partial \beta} u\{\beta_S, \hat{\theta}_0(x, \beta_S)\} + \frac{\partial}{\partial \theta} u\{\beta_S, \hat{\theta}_0(x, \beta_S)\} \frac{\partial}{\partial \beta} \hat{\theta}_0(x, \beta_S).$$

It follows that

$$\frac{\partial}{\partial \beta} \hat{\theta}_0(x, \beta_S) = -G_{n,\theta\theta}^{-1} G_{n,\beta\theta},$$

where

$$G_{n,\theta\theta}(x) = n^{-1} \sum_{i=1}^n \mathcal{L}_{\theta\theta}\{Y_i, Z_i, \beta_S, \hat{\theta}_0(x, \beta_S) + \hat{\theta}_1(x, \beta_S)(X_i - x)\} K_h(X_i - x),$$

$$G_{n,\beta\theta}(x) = n^{-1} \sum_{i=1}^n \mathcal{L}_{\beta\theta}\{Y_i, Z_i, \beta_S, \hat{\theta}_0(x, \beta_S) + \hat{\theta}_1(x, \beta_S)(X_i - x)\} K_h(X_i - x).$$

Application of the inverse function theorem, for example as in Foutz (1977), yields strong consistency of the estimator. Using the proof of Corollary 2.1 of Claeskens & Van Keilegom (2003), we have

$$\begin{aligned} \sup_x |G_{n,\theta\theta}(x) - G_{\theta\theta}(x)f_X(x)| &= O_P[\sqrt{\{\log(n)/(nh)\} + h^2}], \\ &= \sup_x |G_{n,\beta\theta}(x) - G_{\beta\theta}(x)f_X(x)| = O_P[\sqrt{\{\log(n)/(nh)\} + h^2}]. \end{aligned}$$

This proves the statement about  $(\partial/\partial \beta)\hat{\theta}_0(x, \beta_S)$ . A similar proof can be constructed for  $(\partial^2/\partial x \partial \beta)\hat{\theta}_0(x, \beta_S)$ .  $\square$

Inference on the parametric part in a semiparametric model via local profile likelihood estimation involves the concept of a least favourable curve. Define the score function for  $\beta$  as

$$\begin{aligned} \sum_{i=1}^n \frac{d}{d\beta} \mathcal{L}\{Y_i, Z_i, \beta, \theta(X_i, \beta)\} &= \sum_{i=1}^n [\mathcal{L}_\beta\{Y_i, Z_i, \beta, \theta(X_i, \beta)\} \\ &\quad + \mathcal{L}_\theta\{Y_i, Z_i, \beta, \theta(X_i, \beta)\} \frac{\partial}{\partial \beta} \theta(X_i, \beta)]. \end{aligned}$$

The least favourable curve  $\theta^*(\cdot, \beta)$  is the curve for which

$$E\left[\frac{d}{d\beta}\mathcal{L}\{Y, Z, \beta_{\text{true}}, \theta^*(X, \beta_{\text{true}})\}(d/d\beta)\mathcal{L}\{Y, Z, \beta_{\text{true}}, \theta^*(X, \beta_{\text{true}})\}^T|X\right] \quad (\text{A1})$$

is minimal. In other words,  $-\mathcal{L}_\theta\{Y, Z, \beta_{\text{true}}, \theta^*(X, \beta_{\text{true}})\}\partial/\partial\beta\theta^*(X, \beta_{\text{true}})$  is the projection of  $\mathcal{L}_\beta\{Y, Z, \beta_{\text{true}}, \theta^*(X, \beta_{\text{true}})\}$  on to the space spanned by  $\mathcal{L}_\theta\{Y, Z, \beta_{\text{true}}, \theta^*(X, \beta_{\text{true}})\}$ , as implied by (A1).

**LEMMA A2.** *The local linear estimator, defined as the maximizer of (5), is a consistent estimator of the least favourable curve which minimizes (A1).*

*Proof.* By the projection interpretation it follows immediately that, for a least favourable curve  $\theta^*(\cdot, \beta_{\text{true}})$ ,

$$0 = E[\mathcal{L}_\beta\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(X)\} + \mathcal{L}_\theta\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(X)\}\frac{\partial}{\partial\beta}\theta^*(X, \beta_{\text{true}})] \\ \times \mathcal{L}_\theta\{Y, Z, \beta_{\text{true}}, \theta_{\text{true}}(X)\}|X).$$

Bartlett's identities together with Lemma 1 show that

$$\frac{\partial}{\partial\beta}\theta^*(X, \beta_{\text{true}}) = \frac{\partial}{\partial\beta}\theta(X, \beta_{\text{true}}).$$

The proof ends by application of Lemma A1.  $\square$

We have now shown that the conditions NP of Severini & Wong (1992) hold.

*Asymptotic theory for the parametric part of the model.*

**LEMMA A3.** *Assume that Conditions A1–A4 hold. The generalized profile likelihood estimator of  $\beta_{\text{true}}$  in the full model is consistent.*

*Proof.* This follows from Lemmas 1, A1 and A2, which show that for the local linear likelihood estimator the Severini-Wong conditions of their Proposition 1 hold.  $\square$

*Proof of Theorem 1.* By a Taylor expansion we obtain that

$$0 = \sum_{i=1}^n \frac{d}{d\beta}\mathcal{L}\{Y_i, Z_i, \hat{\beta}, \hat{\theta}(X_i, \hat{\beta})\} \\ = \sum_{i=1}^n \frac{d}{d\beta}\mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \hat{\theta}(X_i, \beta_{\text{true}})\} + \sum_{i=1}^n \frac{d^2}{d\beta d\beta^T}\mathcal{L}\{Y_i, Z_i, \hat{\beta}^*, \hat{\theta}(X_i, \hat{\beta}^*)\}(\hat{\beta}_{\text{full}} - \beta_{\text{true}}),$$

where  $\hat{\beta}^*$  lies between  $\hat{\beta}$  and  $\beta_{\text{true}}$ . Lemma A3 implies that  $\hat{\beta}^* \rightarrow \beta_{\text{true}}$  in probability as  $n \rightarrow \infty$ . Using assumption (2) we obtain that the total score function satisfies

$$E\left[\frac{d}{d\beta}\mathcal{L}\{Y, Z, \beta_{\text{true}}, \theta(X, \beta_{\text{true}})\}|X, Z\right] = 0.$$

This implies that

$$\mathcal{S}(\beta_{\text{true}}) = -E\left[\frac{d^2}{d\beta d\beta^T}\mathcal{L}\{Y, Z, \beta_{\text{true}}, \theta(X, \beta_{\text{true}})\}\right].$$

The theorem is proven if the following equations hold:

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^n \frac{d}{d\beta} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \hat{\theta}(X_i, \beta_{\text{true}})\} \\ &= n^{-1/2} \sum_{i=1}^n \frac{d}{d\beta} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta(X_i, \beta_{\text{true}})\} + o_P(1), \\ & \sup_{\beta} \left| n^{-1} \sum_{i=1}^n \left[ \frac{d^2}{d\beta d\beta^T} \mathcal{L}\{Y_i, Z_i, \beta, \hat{\theta}(X_i, \beta)\} - \frac{d^2}{d\beta d\beta^T} \mathcal{L}\{Y_i, Z_i, \beta, \theta(X_i, \beta)\} \right] \right| = o_P(1). \end{aligned}$$

This follows by the same line of argument as in Proposition 2 of Severini & Wong (1992).  $\square$

The asymptotic distributions of the estimators  $\hat{\beta}_{\text{full}}$  and  $\hat{\beta}_{\text{red}}$  will be derived under the misspecification assumption by showing that the distributions are contiguous.

Contiguity follows from Le Cam's first lemma provided we can show that, under the reduced model, for some positive value  $\sigma_{LC}^2$ , as  $n \rightarrow \infty$ , in distribution,

$$\sum_{i=1}^n \left[ \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\} - \mathcal{L}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X_i)\} \right] \rightarrow N\left(-\frac{1}{2}\sigma_{LC}^2, \sigma_{LC}^2\right). \quad (\text{A2})$$

LEMMA A4. Equation (A2) holds with  $\sigma_{LC}^2 = \delta^T E[-\mathcal{L}_{\gamma\gamma}\{Y, X, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X)\}]\delta$ .

*Proof.* By a Taylor series expansion,

$$\begin{aligned} & \sum_{i=1}^n \left[ \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\} - \mathcal{L}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X_i)\} \right] \\ &= n^{-1/2} \delta^T \sum_{i=1}^n \mathcal{L}_{\gamma}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X_i)\} \\ & \quad + \frac{1}{2} n^{-1} \sum_{i=1}^n \delta^T \mathcal{L}_{\gamma\gamma}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}\} \delta + o_P(1). \end{aligned}$$

The first term above converges in distribution to

$$N\left(0, \delta^T E\left[\mathcal{L}_{\gamma}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X_i)\} \mathcal{L}_{\gamma}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X_i)\}^T\right] \delta\right),$$

while the second term converges in probability to

$$\frac{1}{2} \delta^T E\left[\mathcal{L}_{\gamma\gamma}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X_i)\} \right] \delta,$$

which equals  $-\frac{1}{2}\sigma_{LC}^2$  under the likelihood assumptions.  $\square$

We shall apply Le Cam's third lemma to derive the distribution of the estimator  $\hat{\alpha}_{\text{red}}$  under the full model. To establish this result we first show the following lemma.

LEMMA A5. The vector  $n^{1/2}(\hat{\alpha}_{\text{red}} - \alpha_{\text{true}})$  and the loglikelihood difference in (A2) are jointly asymptotically normal under the reduced model. The limiting distribution has mean vector  $(0, -\frac{1}{2}\sigma_{LC}^2)^T$  and covariance matrix

$$\begin{pmatrix} S_{\alpha\alpha}^{-1}(\alpha_{\text{true}}, 0_q) & S_{\alpha\alpha}^{-1}(\beta_{\text{true}}) S_{\alpha\gamma}(\beta_{\text{true}}) \delta \\ \delta^T S_{\alpha\gamma}(\beta_{\text{true}}) S_{\alpha\alpha}^{-1}(\beta_{\text{true}}) & \sigma_{LC}^2 \end{pmatrix}.$$

*Proof.* From the Cramér-Wold theorem it remains to compute the covariance matrix. We use the asymptotic expansion in the proof of (A2) together with Lemma 1 applied to the reduced model to yield the result.  $\square$

Le Cam's third Lemma immediately yields the distribution of  $\hat{\alpha}_{\text{red}}$  under the local misspecification assumption.

*Proof of Theorem 2.* The convergence in distribution follows from Le Cam's third Lemma, using Lemma A5. Theorem 2 together with a Taylor series expansion give that

$$\begin{aligned} & \mathcal{S}_{\alpha\alpha}(\alpha_{\text{true}}, 0_q)n^{1/2}(\hat{\alpha}_{\text{red}} - \alpha_{\text{true}}) \\ &= n^{-1/2} \sum_{i=1}^n \frac{d}{d\alpha} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta(X_i, \beta_{\text{true}})\} + n^{-1} \sum_{i=1}^n \mathcal{L}_{\alpha\gamma}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\}\delta \\ & \quad + n^{-1} \sum_{i=1}^n \frac{\partial}{\partial\alpha} \theta(X_i, \beta_{\text{true}}) \mathcal{L}_{\theta\gamma}^{\text{T}}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\}\delta \\ & \quad + n^{-1} \sum_{i=1}^n \mathcal{L}_{\theta} \{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\} \frac{\partial^2}{\partial\alpha\partial\gamma} \theta(X_i, \beta_{\text{true}})\delta + o_P(1). \end{aligned}$$

Lemma 1 applied to the reduced model gives that  $(\partial/\partial\alpha)\theta(X_i, \beta_{\text{true}}) = -G_{\alpha\theta}/G_{\theta\theta}$ . We use this result to show that the sum of the last three terms in the above expansion converges in probability to

$$\delta E[\mathcal{L}_{\alpha\gamma}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\}] + \delta E\left[\frac{\partial}{\partial\alpha} \theta(X_i, \beta_{\text{true}}) \mathcal{L}_{\theta\gamma}^{\text{T}}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\}\right] = -\delta S_{\alpha\gamma}. \quad \square$$

*Proof of Theorem 3.* We start from the expansion in Theorem 1 which, in matrix notation, is equal to

$$\begin{aligned} & \begin{pmatrix} n^{1/2}(\hat{\alpha}_{\text{full}} - \alpha_{\text{true}}) \\ n^{1/2}(\hat{\gamma}_{\text{full}} - \gamma_{\text{true}}) \end{pmatrix} \\ &= n^{-1} \begin{pmatrix} S^{\alpha\alpha}(\beta_{\text{true}}) & S^{\alpha\gamma}(\beta_{\text{true}}) \\ S^{\gamma\alpha}(\beta_{\text{true}}) & S^{\gamma\gamma}(\beta_{\text{true}}) \end{pmatrix} \sum_{i=1}^n \begin{pmatrix} \frac{d}{d\alpha} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta(X_i, \beta_{\text{true}})\} \\ \frac{d}{d\gamma} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta(X_i, \beta_{\text{true}})\} \end{pmatrix} + o_P(1). \end{aligned}$$

It now follows that

$$\begin{aligned} & n^{1/2}(\hat{\alpha}_{\text{full}} - \alpha_{\text{true}}) - S^{\alpha\gamma}(\beta_{\text{true}})\{S^{\gamma\gamma}(\beta_{\text{true}})\}^{-1}n^{1/2}(\hat{\gamma}_{\text{full}} - \gamma_{\text{true}}) \\ &= (I \quad S^{\alpha\gamma}(\beta_{\text{true}})\{S^{\gamma\gamma}(\beta_{\text{true}})\}^{-1})n^{1/2}(\hat{\beta}_{\text{full}} - \beta_{\text{true}}) \\ &= \{S^{\alpha\alpha}(\beta_{\text{true}}) - S^{\alpha\gamma}(\beta_{\text{true}})\{S^{\gamma\gamma}(\beta_{\text{true}})\}^{-1}S^{\gamma\alpha}(\beta_{\text{true}})\}n^{-1} \\ & \quad \times \sum_{i=1}^n \frac{d}{d\alpha} \mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta(X_i, \beta_{\text{true}})\}. \end{aligned}$$

Since  $\{S^{\alpha\alpha}(\beta_{\text{true}}) - S^{\alpha\gamma}(\beta_{\text{true}})\{S^{\gamma\gamma}(\beta_{\text{true}})\}^{-1}S^{\gamma\alpha}(\beta_{\text{true}})\} = S_{\alpha\alpha}^{-1}$ , the first result follows after application of Theorem 3. The correlation is computed as  $(S^{\gamma\alpha}S_{\alpha\alpha}^{-1} + S^{\gamma\gamma}S_{\gamma\alpha})S_{\alpha\alpha}^{-1}$  and equals zero by definition of  $S^{-1}$ .  $\square$

## REFERENCES

- BUNEA, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.* **32**, 898–927.
- BURNHAM, K. P. & ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*, 2nd ed. New York: Springer-Verlag.
- CLAESKENS, G. & VAN KEILEGOM, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31**, 1852–84.
- FAN, J. & LI, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Am. Statist. Assoc.* **99**, 710–23.
- FOUTZ, V. (1977). On the unique consistent solution to the likelihood equations. *J. Am. Statist. Assoc.* **72**, 147–8.

- HJORT, N. L. & CLAESKENS, G. (2003). Frequentist model average estimators (with Discussion). *J. Am. Statist. Assoc.* **98**, 879–99.
- LEEB, H. & PÖTSCHER, B. M. (2005). Model selection and inference: facts and fiction. *Economet. Theory* **21**, 21–59.
- MURPHY, S. A. & VAN DER VAART, A. W. (2000). On profile likelihood (with Discussion). *J. Am. Statist. Assoc.* **95**, 449–85.
- NAIK, P. A. & TSAI, C.-L. (2001). Single-index model selections. *Biometrika* **88**, 821–32.
- RUPPERT, D., SHEATHER, S. J. & WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.* **90**, 1257–70, Correction (1996) **91**, 1380.
- SEVERINI, T. A. & WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768–802.
- SHI, P. & TSAI, C.-L. (1999). Semiparametric regression model selections. *J. Statist. Plan. Infer.* **77**, 119–39.
- SIMONOFF, J. S. & TSAI, C.-L. (1999). Semiparametric and additive model selection using an improved Akaike information criterion. *J. Comp. Graph. Statist.* **8**, 22–40.
- WOODROOFE, M. (1982). On model selection and the arc sine laws. *Ann. Statist.* **10**, 1182–94.

[Received May 2005. Revised October 2006]