# An Asymptotic Theory for Model Selection Inference in General Semiparametric Problems

By Gerda Claeskens

*ORSTAT, K.U. 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, TX 77843-3143, U.S.A.*

carroll@stat.tamu.edu

**Abstract**

Recently, Hjort and Claeskens (2003) developed an asymptotic theory for model selection, model averaging and post-model selection/averaging inference using likelihood methods in parametric models, along with associated confidence statements. In this paper, 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 and 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. The results also describe the behavior of semiparametric model estimates when the parametric component is misspecified, and have implications as well for pointwise consistent model selectors.

**KEY WORDS:** Akaike Information Criterion; Bayes Information Criterion; Efficient semiparametric estimation; Frequentist model averaging; Model averaging; Model selection; Profile likelihood; Semiparametric model.

**Short title: Model Selection Inference in Semiparametric Models**

# 1 Introduction

We consider semiparametric models where the responses $Y$ are 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 $Y_i = Z_i^{\mathrm{T}}\beta + \theta(X_i) + \varepsilon_i$. In generalized linear models, or in general likelihood problems, we start with a likelihood function

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

where the value of $\beta_{\mathrm{true}}$ as well as the function $\theta_{\mathrm{true}}$ 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 the partially linear models. Shi and Tsai (1999) use B-splines to estimate the nonparametric function $\theta(\cdot)$ and develop a small sample adjustment to Akaike's information criterion AIC. Recently, Fan and 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 and Tsai (1999) developed an improvement to the AIC for variable selection in semiparametric and additive models. Naik and Tsai (2001) developped an AIC-type information criterion for use in single-index models, with extension to partially linear models. None of these papers, however, 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 paper, in particular, we go further than model selection by extending the frequentist model averaging results of Hjort and Claeskens (2003) to semiparametric models. By appropriate use of profile likelihood methods, we show that their results continue to hold, provided parametric likelihood ratio methods are replaced by semiparametric profile likelihood methods, and in addition that quantities related to the Fisher information matrix in

parametric models are replaced by a suitable profile information matrix, namely the semi-parametric information bound. Our methods of proof employ Le Cam's contiguity lemmas, leading to transparent results.

Definitions and notation are given in Section 2. The asymptotic results are split in two parts. First, we focus on the nonparametric part of the model, for which we use local linear estimators. The parametric model part is dealt with in Section 3.2. All technical details, as well as regularity conditions, are collected in the appendix. The distributions of estimators in submodels there obtained are combined in Section 4 to give the main results on the distribution of model averaged estimators. The distribution of estimators post-model selection is a special case. The applicability of the method is illustrated in a simulation study in Section 5, where we also show that BIC has poor behavior, despite it being a consistent model selector. Final comments, along with a brief discussion of the problems with pointwise consistent model selectors, are given in Section 6.

# 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 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: (1) a reduced model where $\beta_{\text{red}}^{\text{T}} = (\alpha^{\text{T}}, 0_q^{\text{T}})$, and (2) a full model where $\beta_{\text{full}}^{\text{T}} = (\alpha^{\text{T}}, \gamma^{\text{T}})$. As in Hjort and Claeskens (2003) we make the local misspecification assumption: 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.

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 log-likelihood function $\mathcal{L}\{Y, Z, \beta, \theta(X)\}$. The true values are $\beta_{\text{true}}$ and $\theta_{\text{true}}(\cdot)$. Partial derivatives of the log-likelihood function are denoted

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

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

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

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

The key assumptions that will hold in likelihood problems are that

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

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

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

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

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

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

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

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:

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

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

$$0 = n^{-1}\sum_{i=1}^{n}\frac{d}{d\beta}\mathcal{L}\{Y_i,Z_i,\beta_S,\widehat{\theta}(X_i,\beta_S)\}$$

$$= \quad n^{-1}\sum_{i=1}^{n}\left[\mathcal{L}_{\beta}\{Y_i, Z_i, \beta_S, \widehat{\theta}(X_i, \beta_S)\} + \mathcal{L}_{\theta}\{Y_i, Z_i, \beta_S, \widehat{\theta}(X_i, \beta_S)\}\frac{\partial}{\partial \beta_S}\widehat{\theta}(X_i, \beta_S)\right].$$

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, as well as its inverse, is partitioned as

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

with

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

$G_{\theta\beta} = G_{\beta\theta}^{\mathrm{T}}$, and $G^{\beta\beta} = (G_{\beta\beta} - G_{\beta\theta}G_{\theta\theta}^{-1}G_{\beta\theta}^{\mathrm{T}})^{-1}$. 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 do not know whether all of the components of $\beta$ are needed. Our interest is in the distribution of $\mu(\widehat{\beta})$ where $\widehat{\beta}$ is obtained via a model selection procedure. 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. Via some lemmas we arrive at the distribution of the profile likelihood estimator $\widehat{\beta}$ in both reduced and full models, under the local misspecification assumptions. Technical details are given in an appendix.

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

**Lemma 3.1** *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 differentiating (4) with respect to $\beta$ and solving the resulting equation. ■

## 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 as well as its inverse in the following way,

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

We give a basic expansion of the profile kernel method, first in the full model, then in the reduced model. Recall that $\beta_{\text{true}} = (\alpha_{\text{true}}^{\text{T}}, \delta^{\text{T}}/\sqrt{n})^{\text{T}}$.

**Theorem 3.1** *Under the local misspecification assumption and when working in the full model, assuming conditions (C1)–(C4),*

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

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

**Theorem 3.2** *If the reduced model holds, that is $\gamma = 0_q$,*

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

*Moreover, $n^{1/2}(\widehat{\alpha}_{\text{red}} - \alpha_{\text{true}}) \Rightarrow$ Normal$(0, S_{\alpha\alpha}^{-1})$.*

The proof of the first statement is very similar to the proof of Theorem 3.1. 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 estimate when the full model holds, and the other describing the relationship between the full and reduced model estimates in this case.

**Theorem 3.3** *If the local misspecified model holds, that is* $\gamma_{\text{true}} = n^{-1/2}\delta$,

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

**Theorem 3.4** *Under the local misspecification assumption,*

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

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

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

**Theorem 3.5** *Under the local misspecification assumption,*

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

**Proof**. Follows immediately via the delta method, and Theorems 3.1 and 3.3. ■

# 4 Model Averaging and Inference

## 4.1 Limit Results and Confidence Sets

Theorem 3.5 is the main ingredient to obtain the distribution of estimators after model selection. Here we follow the approach leading to Theorem 4.1 of Hjort and Claeskens (2003), with the approach to inference following their Section 4.3.

We consider cases where model selection and model averaging are based on weights depending on $\widehat{\delta}_{\text{full}} = n^{1/2}\widehat{\gamma}_{\text{full}} \Rightarrow D = \text{Normal}(\delta, S^{\gamma\gamma})$, see below for more discussion. Estimators after model selection take the form

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

where the data-driven weights $c(\widehat{\delta}_{\text{full}})$ are between zero and one. We then immediately have the following result.

**Theorem 4.1** *Recall that $D$ is the limiting distribution of $n^{1/2}\widehat{\gamma}_{\text{full}}$, and that $\Lambda_{\text{full}}$ and $\Lambda_{\text{red}}$ are described in Theorem 3.5. Then, under the local misspecification assumption, $n^{1/2}\{\widehat{\mu} - \mu(\beta_{\text{true}})\} \Rightarrow c(D)\Lambda_{\text{full}} + \{1 - c(D)\}\Lambda_{\text{red}}$.*

We can combine Theorem 4.1 with the methods in Section 4.3 of Hjort and Claeskens to develop asymptotically correct confidence limits for $\mu(\beta_{\text{true}})$. This is simply a matter of making identifications of notation. In our case, let $\mu_\alpha(\beta_{\text{true}}) = \{\partial\mu(\beta_{\text{true}})\}/\partial\alpha$ and let $\mu_\gamma(\beta_{\text{true}}) = \{\partial\mu(\beta_{\text{true}})\}/\partial\gamma$. Define $\tau_0^2 = \mu_\alpha^{\mathrm{T}}(\beta_{\text{true}})S_{\alpha\alpha}^{-1}\mu_\alpha(\beta_{\text{true}})$, $\omega = S_{\gamma\alpha}S_{\alpha\alpha}^{-1}\mu_\alpha(\beta_{\text{true}}) - \mu_\gamma(\beta_{\text{true}})$, $\kappa = (\tau_0^2 + \omega^{\mathrm{T}}S^{\gamma\gamma}\omega)^{1/2}$ and replace their $\widehat{\delta}_n(D)$ by $Q(D) = c(D)D$. Then their equation (4.8) gives asymptotically correct confidence statements for $\mu(\beta_{\text{true}})$, when estimates are substituted at $\widehat{\beta}_{\text{full}}$.

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

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

see for example equation (6) of Murphy and van der Vaart (2000).

Post-model selection estimators take indicator functions as weights, pointing to the selected model. The theory also applies with more general weighting schemes, allowing to average estimators across models.

# 5 Simulation Example

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

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

where $Z_i = (Z_{i1}, Z_{i2})^{\mathrm{T}}$, $\epsilon_i = \text{Normal}(0, \sigma^2)$, $\mathcal{B} = (\mathcal{B}_1, \mathcal{B}_2)$, $\beta = (\sigma^2, \mathcal{B}^{\mathrm{T}})$, $\alpha = (\sigma^2, \mathcal{B}_1)$ 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 $c = 0.0, 0.5, 1.0, \ldots 10.0$. The experiment was repeated $2,000$ times in each configuration. We use 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 DPI bandwidth selection method of Ruppert, Sheather and Wand (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 get a preliminary estimate $\widehat{\beta}_{\text{start}}$ of $\beta$. Following this, we regressed $Y - Z^{\mathrm{T}}\widehat{\beta}_{\text{start}}$ on $X$ to get a final common bandwidth, and then reestimated $\beta$.

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

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

and the information bound then becomes

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

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

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 fairly close to being the same as 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. BIC in particular had significant bias for estimating $\mathcal{B}_1$.

# 6 Discussion

In this paper, we have computed the limit distributions and asymptotic expansion for general semiparametric models with misspecified parametric components, results that are summarized in Theorem 3.3 and Theorem 3.5. Our method of argument is to exploit the contiguity of locally misspecified models. In effect, we show that the results are the same as what one would expect in fully parametric models, as described by Hjort and Claeskens (2003), but with the Fisher information matrix for parametric models replaced by the semiparametric information bound for semiparametric models, and with maximum likelihood estimators for parametric models replaced by semiparametric efficient profile estimators. These results form the model misspecification and model selection analogue of the correct model profile likelihood results of Murphy and van der Vaart (2000).

Our work has focused on the case that $X$ 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, e.g., the partially linear additive model with mean $Z^{\mathrm{T}}\beta + \sum_{j=1}^{m} \theta_j(X_j)$.

Finally, in our simulation we found that BIC estimates and confidence intervals had bias 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 hold for other so-called consistent model selectors.

# Acknowledgments

Carroll's research was supported by a grant from the National Cancer Institute (CA–57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30–ES09106).

# Appendix

## Regularity conditions

(C1) The bandwidth sequence $h_n \to 0$ as $n \to \infty$, in such a way that $nh_n/\log(n) \to \infty$ and $h_n \geq \{\log(n)/n\}^{1-2/\lambda}$ for $\lambda$ as in condition (C4).

(C2) The kernel function $K$ is a symmetric, continuously differentiable pdf on $[-1, 1]$ taking on 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 2 continuous derivatives on $B$ and is also twice differentiable with respect to $\beta$.

(C3) 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 4th partial derivative exists for almost all $(y, z)$. Further, mixed partial derivatives $\frac{\partial^{r+s}}{\partial\beta^r\partial v^s}\mathcal{L}\{y, z, \beta, v)\}|_{v=\theta(x)}$, with $0 \leq r, s \leq 4, r + s \leq 4$ exist for almost all $(y, z)$ and $E\{\sup_\beta \sup_v \left|\frac{\partial^{r+s}}{\partial\beta^r\partial v^s}\mathcal{L}\{y, z, \beta, v)\}\right|^2\} < \infty$. The Fisher information, $G(x)$, possesses a continuous derivative and $\inf_{x\in B} G(x) > 0$.

(C4) There exists a neighborhood $\mathcal{N}\{\beta_{\text{true}}, \theta_{\text{true}}(x)\}$ 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 $X = x$. Further,

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

**Asymptotic Theory For The Nonparametric Part of the Model**

For each fixed value of $\beta$, the local linear estimator $\widehat{\theta}(x,\beta)$ 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 and Van Keilegom (2003); precise regularity conditions are formulated above. We summarize the strong uniform consistency result in the first part of Lemma A.1, and add a result about the derivatives with respect to the parameters $\beta$.

**Lemma A.1** *As $n \to \infty$, and under regularity conditions (C1)–(C4) on the kernel, bandwidth and likelihood function, $\widehat{\theta}(x,\beta)$ and $\widehat{\theta}_1(x,\beta)$ exist and $\sup_x |\widehat{\theta}(x,\beta) - \theta(x,\beta)| = O[\{nh/log(n)\}^{-1/2} + h^2]$ almost surely. For the estimator of the derivative of the curve it follows that $\sup_x |\widehat{\theta}_1(x,\beta) - \frac{\partial}{\partial x}\theta(x,\beta)| = O(\{nh^3/log(n)\}^{-1/2} + h^2)$ almost surely. Furthermore, $\frac{\partial}{\partial\beta}\widehat{\theta}(x,\beta)$ exists, is strongly consistent and $\sup_x |\frac{\partial}{\partial\beta}\widehat{\theta}(x,\beta) - \frac{\partial}{\partial\beta}\theta(x,\beta)| = O_P[\{nh/log(n)\}^{-1/2} + h^2]$ and for some $\delta > 0$, $\sup_x |\frac{\partial^2}{\partial x\partial\beta}\widehat{\theta}(x,\beta) - \frac{\partial^2}{\partial x\partial\beta}\theta(x,\beta)| = o_p(n^{-\delta})$.*

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

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

By the first part of this lemma, $\widehat{\theta}_1(x,\beta_S)$ is a strongly consistent estimator of $\theta_1(x,\beta_S)$. Since by assumption (C3) the Fisher information matrix is positive definite, and the design density $f_X(x) > 0$ (C2), the implicit function theorem implies that the function $\beta_S \to \widehat{\theta}_0(x,\beta_S)$ is a $C^1$ function. As a consequence there exists a neighborhood of $\widehat{\beta}_S$ such that for all $\beta_S$ in this neighborhood,

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

It follows that

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

where

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

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 and Van Keilegom (2003),

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

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

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

$$\frac{d}{d\beta}\mathcal{L}\{Y,Z,\beta,\theta(X,\beta)\} = \mathcal{L}_\beta\{Y,Z,\beta,\theta(X,\beta)\} + \mathcal{L}_\theta\{Y,Z,\beta,\theta(X,\beta)\}\frac{\partial}{\partial\beta}\theta(X,\beta).$$

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

$$E[\frac{d}{d\beta}\mathcal{L}\{Y,Z,\beta_{\text{true}},\theta^*(X,\beta_{\text{true}})\}\frac{d}{d\beta}\mathcal{L}\{Y,Z,\beta_{\text{true}},\theta^*(X,\beta_{\text{true}})\}^{\text{T}}|X] \tag{A.1}$$

is minimal. In other words, $-\mathcal{L}_\theta\{Y,Z,\beta_{\text{true}},\theta^*(X,\beta_{\text{true}})\}\frac{\partial}{\partial\beta}\theta^*(X,\beta_{\text{true}})$ is the projection of $\mathcal{L}_\beta\{Y,Z,\beta_{\text{true}},\theta^*(X,\beta_{\text{true}})\}$ onto the space spanned by $\mathcal{L}_\theta\{Y,Z,\beta_{\text{true}},\theta^*(X,\beta_{\text{true}})\}$, as implied by (A.1).

**Lemma A.2** *The local linear estimator, defined as the maximizer of (5), is a consistent estimator of the least favorable curve which minimizes (A.1).*

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

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

Bartlett's identities together with Lemma 3.1 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 A.1. ■

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

**Asymptotic Theory For The Parametric Part of the Model**

**Lemma A.3** *Assume that regularity conditions (C1)–(C4) hold. The (generalized) profile likelihood estimator of* $\beta_{\text{true}}$ *in the full model is consistent.*

**Proof**.

This follows from Lemmas 3.1, A.1 and A.2, which show that for the local linear likelihood estimator the Severini-Wong conditions of their Proposition 1 hold. ■

**Proof of Theorem 3.1**. Via a Taylor expansion we obtain that

$$\begin{aligned} 0 &= \frac{d}{d\beta}\mathcal{L}\{Y_i, Z_i, \widehat{\beta}, \widehat{\theta}(X_i,\widehat{\beta})\} \\ &= \frac{d}{d\beta}\mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \widehat{\theta}(X_i,\beta_{\text{true}})\} + \frac{d^2}{d\beta d\beta^{\text{T}}}\mathcal{L}\{Y_i, Z_i, \widehat{\beta}^*, \widehat{\theta}(X_i,\widehat{\beta}^*)\}(\widehat{\beta}_{\text{full}} - \beta_{\text{true}}), \end{aligned}$$

where $\widehat{\beta}^*$ lies in between $\widehat{\beta}$ and $\beta_{\text{true}}$. Lemma A.3 implies that $\widehat{\beta}^* \to \beta_{\text{true}}$ in probability as $n \to \infty$. Using assumption (2) it follows that the total score function satisfies

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

This implies that

$$\mathcal{S}(\beta_{\text{true}}) = -E[\frac{d^2}{d\beta d\beta^{\text{T}}}\mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta(X_i,\beta_{\text{true}})\}].$$

The theorem is proven if the following equations hold

$$n^{-1/2}\frac{d}{d\beta}\mathcal{L}\{Y, Z, \beta_{\text{true}}, \widehat{\theta}(X,\beta_{\text{true}})\} = n^{-1/2}\frac{d}{d\beta}\mathcal{L}\{Y, Z, \beta_{\text{true}}, \theta(X,\beta_{\text{true}})\} + o_P(1)$$

$$\sup_{\beta}\left| n^{-1}\frac{d^2}{d\beta d\beta^{\text{T}}}\mathcal{L}\{Y_i, Z_i, \beta, \widehat{\theta}(X_i,\beta)\} - n^{-1}\frac{d^2}{d\beta d\beta^{\text{T}}}\mathcal{L}\{Y_i, Z_i, \beta, \theta(X_i,\beta)\}\right| + o_P(1).$$

This follows by the same line of arguments as in Proposition 2 of Severini and Wong (1992).

■

The asymptotic distributions of the estimators $\widehat{\beta}_{\text{full}}$ and $\widehat{\beta}_{\text{red}}$ 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 $\sigma^2_{LC}$, as $n \to \infty$,

$$\sum_{i=1}^{n} \Big[\mathcal{L}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\} - \mathcal{L}\{Y_i, Z_i, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X_i)\}\Big] \Rightarrow \text{Normal}(-\frac{1}{2}\sigma^2_{LC} \ , \ \sigma^2_{LC}) \quad \text{(A.2)}$$

**Lemma A.4** *Equation (A.2) holds with* $\sigma^2_{LC} = \delta^{\text{T}} E\{\mathcal{L}_\gamma(Y, X, (\alpha_{\text{true}}, 0_q), \theta_{\text{true}}(X)\}\delta$.

**Proof**. Via a Taylor series expansion

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

The first term above converges in distribution to

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

while the second term converges in probability to

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

which equals $-\frac{1}{2}\sigma^2_{LC}$ under the likelihood assumptions. ■

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

**Lemma A.5** *The vector* $n^{1/2}(\widehat{\alpha}_{\text{red}} - \alpha_{\text{true}})$ *and the log-likelihood difference in (A.2) are jointly asymptotically normal under the reduced model. The limiting distribution has mean vector* $(0, -\frac{1}{2}\sigma^2_{LC})$ *and covariance matrix*

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

**Proof.** Via the Cramér-Wold theorem it remains to compute the covariance matrix. We use the asymptotic expansion in the proof of (A.2) together with Lemma 3.1 applied to the reduced model to yield the result. ■

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

**Proof of Theorem 3.3**. The convergence in distribution follows from Le Cam's third Lemma, using Lemma A.5. Theorem 3.2 together with a Taylor series expansion give that

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

Lemma 3.1 applied to the reduced model gives that $\frac{\partial}{\partial\alpha}\theta(X_i, \beta_{\text{true}}) = -G_{\alpha\theta}/G_{\theta\theta}$. We use this result to show that the sum of the last three terms in the above expansion converge in probability to

$$\delta E[\mathcal{L}_{\alpha\gamma}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\}] + \delta E\Big[\tfrac{\partial}{\partial\alpha}\theta(X_i, \beta_{\text{true}}) \mathcal{L}^{\text{T}}_{\theta\gamma}\{Y_i, Z_i, \beta_{\text{true}}, \theta_{\text{true}}(X_i)\} = -\delta S_{\alpha\gamma} \blacksquare$$

**Proof of Theorem 3.4**. We start from the expansion in Theorem 3.1 which is in matrix notation equal to

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

It now follows that

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

Since $\{S^{\alpha\alpha}(\beta_{\text{true}}) - S^{\alpha\gamma}(\beta_{\text{true}})\{S^{\gamma\gamma}(\beta_{\text{true}})\}^{-1}S^{\gamma\alpha}(\beta_{\text{true}})\} = S^{-1}_{\alpha\alpha}$, the first result follows after application of Theorem 3.3. The correlation is computed as $(S^{\gamma\alpha}S_{\alpha\alpha} + S^{\gamma\gamma}S_{\gamma\alpha})S^{-1}_{\alpha\alpha}$ and equals zero by definition of $S^{-1}$. ■

# References


Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics*, **32**, 898–927.

Claeskens, G. & Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, **31**, 1852–1884.

Fan, J. & Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, **99**, 710-723.

Foutz, V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*, **72**, 147–148.

Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators (with discussion). *Journal of the American Statistical Association*, **98**, 879–899.

Leeb, H. & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, **21**, 21-59.

Murphy, S. A. & van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, **95**, 449–485.

Naik, P. A. & Tsai, C.-L. (2001). Single-index model selections. *Biometrika*, **88**, 821–832.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression (Corr: 96V91 p1380). *Journal of the American Statistical Association*, **90**, 1257–1270.

Severini, T. A. & Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics*, **20**, 1768–1802.

Shi, P. & Tsai, C.-L. (1999). Semiparametric regression model selections. *Journal of Statistical Planning & Inference*, **77**, 119–139.

Simonoff, J.S. & Tsai, C.-L. (1999). Semiparametric and additive model selection using an improved Akaike information criterion. *Journal of Computational and Graphical Statistics*, **8**, 22–40.