Profile Likelihood Information Criteria for Parametric and Semiparametric Models

Yuichi Hirose[1*]
Victoria University of Wellington

September 13, 2011

Summary

In this paper we propose an information criteria using the quadratic expansion of the profile likelihood. We call this information criteria the profile likelihood information criteria (PLIC). The profile Akaike information criteria in Xu, Vaida and Harrington (2009) and Claeskens and Carroll (2007) assumed the true model is in the chosen parametric family. Our information criteria does not require the assumption.

Key words: Information criterion; Semi-parametric model; Profile likelihood; Efficient semi-parametric estimation.

[1] School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, New Zealand. e-mail: Yuichi.Hirose@msor.vuw.ac.nz

1 Introduction

Information criteria is a popular tool for a model selection, since it can be used to compare non-nested models as well as nested models. Generally, there are two assumptions under which we derive information criteria: (A) the true distribution may not be in the chosen parametric family; (B) the true distribution is in the chosen parametric family. In model selection problem, usually we do not know the true distribution is in the chosen parametric family. Therefore, the assumption (A) may be a more practical assumption.

Some authors considered model selection using the profile likelihood under the assumption (B). For example Xu, Vaida and Harrington (2009) used profile likelihood to derive the profile Akaike information criteria. Claeskens and Carroll (2007) used the profile Akaike information criteria to apply their method of focused information criteria (Claeskens and Hjort (2003)) in the context of semi-parametric models. In this paper we derive an information criteria using the profile likelihood under the assumption (A) for parametric and semi-parametric models.

We consider parametric and semi-parametric models of the form

$$\mathcal{P} = \{f(x;\theta,\eta) : \ \theta \in \Theta,\, \eta \in \mathcal{H}\} \tag{1}$$

where $f(x;\theta,\eta)$ denotes a density with two parameters: θ is a finite-dimensional parameter of interest, and η is a nuisance parameter, which is finite-dimensional for a parametric model and infinite-dimensional for a semi-parametric model.

Let $G(x)$ and $g(x)$ be the true cdf and pdf for which the data $X_1, \ldots, X_n$ are generated. The expectation of a function $\phi(X)$ of X with respect to G is denoted by $E\{\phi(X)\}$. Let us define (θ_0, η_0) as the maximizer of the expected log of density

$$E\{\log f(X;\theta_0,\eta_0)\} = \max_{\theta,\eta} E\{\log f(X;\theta,\eta)\}.$$

Then the assumption (A) is equivalent to the situation that "we may have $g(x) \neq f(x;\theta,\eta)$ for all $\theta \in \Theta$ and $\eta \in \mathcal{H}$" and the assumption (B) is "we have $g(x) = f(x;\theta_0,\eta_0)$".

2 Profile Likelihood Information Criteria

In this section we derive information criteria for parametric and semi-parametric models with nuisance parameters. We assume the assumption (A) that the true distribution (density) $g(x)$ may not be in the chosen parametric family. The information criteria derived here will be called the profile likelihood information criteria (PLIC) since we use the profile likelihood for the derivation.

Suppose the model $\mathcal{P}$ in (1) may be a semi-parametric model and a function $\hat{\eta}_\theta$ is the maximizer of the expected log of density given a value of the parameter θ:

$$\hat{\eta}_\theta = \text{argmax}_\eta E\{\log f(X;\theta,\eta)\}. \tag{2}$$

Then the profile log-likelihood is given by

$$\int \log f(x;\theta,\hat{\eta}_\theta) dG_n(x)$$

and the profile likelihood score equation is

$$\frac{\partial}{\partial\theta}\int \log f(x;\theta,\hat{\eta}_\theta) dG_n(x) = 0. \tag{3}$$

Note that the function $\hat{\eta}_\theta$ satisfies $\hat{\eta}_{\theta_0} = \eta_0$.

Let

$$\tilde{\ell}_1(x;\theta) = \frac{\partial}{\partial\theta}\log f(x;\theta,\hat{\eta}_\theta),\ \tilde{\ell}_{11}(x;\theta) = \frac{\partial^2}{\partial\theta\partial\theta^T}\log f(x;\theta,\hat{\eta}_\theta), \tag{4}$$

and

$$\tilde{J}_{11} = -E\{\tilde{\ell}_{11}(X;\theta_0)\}. \tag{5}$$

Under the mild regularity conditions the standard Taylor's expansion argument can show that the solution $\hat{\theta}$ to the profile likelihood score equation (3) is asymptotically linear estimator such that

$$n^{1/2}(\hat{\theta}-\theta_0) = n^{-1/2}\sum_{i=1}^{n}\tilde{J}_{11}^{-1}\tilde{\ell}_1(X_i;\theta_0) + o_P(1), \tag{6}$$

(cf. Murphy and van der Vaart (2000), Hirose (2011)). The function $\tilde{\ell}_1(x;\theta_0)$ is called the efficient score function.

For the $\hat{\theta}$ and $\hat{\eta}_\theta$ given above, we will derive an information criteria based on the Kullback–Leibler distance between $g(\cdot)$ and $f(\cdot;\hat{\theta},\hat{\eta}_{\hat{\theta}})$:

$$I\{g(\cdot), f(\cdot;\hat{\theta},\hat{\eta}_{\hat{\theta}})\} \ = \ \int \log g(x) dG(x) - \int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta}}) dG(x).$$

Since the first term in the right hand side is constant, the Kullback–Leibler distance is determined by the second term.

If we use the integral $\int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta}}) dG_n(x)$ as an estimator of the second term, the bias of the estimator is

$$\begin{aligned}\text{bias} &= E\left\{\int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta}}) dG_n(x) - \int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta}}) dG(x)\right\} \\ &= E\left\{\int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta}}) d(G_n - G)(x)\right\}\end{aligned}$$

By Taylor's expansion

$$\log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta}}) \ = \ \log f(x;\theta_0,\eta_0) + \tilde{\ell}_1(x;\theta_0)^T(\hat{\theta}-\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)^T\tilde{\ell}_{11}(x;\theta^*)(\hat{\theta}-\theta_0)$$

where θ^* lies between $\hat{\theta}$ and θ_0. It follows that the bias is

$$\begin{aligned}\text{bias} &= E\left\{\int \log f(x;\theta_0,\eta_0) d(G_n - G)(x)\right\} \\ &\quad + n^{-1}E\left\{n^{1/2}\int \tilde{\ell}_1(x;\theta_0,\eta_0)^T d(G_n - G)(x) n^{1/2}(\hat{\theta}-\theta_0)\right\} \\ &\quad + (2n)^{-1}E\left\{n^{1/2}(\hat{\theta}-\theta_0)^T\int \tilde{\ell}_{11}(x;\theta^*) d(G_n - G)(x) n^{1/2}(\hat{\theta}-\theta_0)\right\}.\end{aligned}$$

In the right hand side, the first term is zero. The equation (6) implies $n^{1/2}(\hat{\theta}-\theta_0) = O_P(1)$. It follows that the third term is $o_P(n^{-1})$ due to the uniform law of large numbers (Glivenko-Canteli theorem). We look closely at the second term in the right hand side. Using $\int \tilde{\ell}_1(x;\theta_0)dG(x) = 0$ and (6), this term is equal to

$$\begin{aligned} & n^{-1}E\left\{n^{1/2}\int \tilde{\ell}_1(x;\theta_0)^T dG_n(x) n^{1/2}(\hat{\theta}-\theta_0)\right\} \\ = \; & n^{-1}E\left\{\left(n^{1/2}\int \tilde{\ell}_1 dG_n\right)^T \tilde{J}_{11}^{-1}\left(n^{1/2}\int \tilde{\ell}_1 dG_n\right)\right\} + o(n^{-1}) \\ = \; & n^{-1}\mathrm{tr}\left\{\tilde{J}_{11}^{-1}\mathrm{var}\left(n^{1/2}\int \tilde{\ell}_1 dG_n\right)\right\} + o(n^{-1}) \\ = \; & n^{-1}\mathrm{tr}\left(\tilde{J}_{11}^{-1}\tilde{I}_{11}\right) + o(n^{-1}) \end{aligned} \tag{7}$$

where

$$\tilde{I}_{11} = E\{\tilde{\ell}_1(X;\theta_0)^{\otimes 2}\}.$$

Therefore

$$\mathrm{bias} = n^{-1}\mathrm{tr}\left(\tilde{J}_{11}^{-1}\tilde{I}_{11}\right) + o(n^{-1}).$$

Since the form of information criteria using the profile likelihood is

$$-2n \times \left\{\int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta}}) dG_n(x) - \mathrm{bias}\right\},$$

we have the profile likelihood information criteria (PLIC)

$$\mathrm{PLIC} \;\; = \;\; -2\sum_{i=1}^{n} \log f(X_i;\hat{\theta},\hat{\eta}_{\hat{\theta}}) + 2\mathrm{tr}(\tilde{J}_{11}^{-1}\tilde{I}_{11}) \tag{8}$$

where we ignored the $o(1)$ term.

When the assumption (B): $g(x) = f(x;\theta_0,\eta_0)$, is true, we have $\tilde{I}_{11} = \tilde{J}_{11}$ and

$$\mathrm{tr}(\tilde{J}_{11}^{-1}\tilde{I}_{11}) = \mathrm{tr}(\tilde{I}_{11}^{-1}\tilde{I}_{11}) = p,$$

where p is the dimension of the parameter θ. Then the PLIC become the profile Akaike information criteria in Xu, Vaida and Harrington (2009)

$$-2\sum_{i=1}^{n} \log f(X_i;\hat{\theta},\hat{\eta}_{\hat{\theta}}) + 2p.$$

2.1 Alternative formula to compute PLIC

The PLIC given in (8) requires the function $\hat{\eta}_\theta$ which is defined by (2). In practice, it is sometimes computationally difficult to deal with the function $\hat{\eta}_\theta$. Therefore it is nice have an alternative expression for the PLIC without using this function. In this section we will derive such an expression for the PLIC under the assumption (A) that we may have $g(x) \neq f(x;\theta,\eta)$ for all θ and η.

First we assume that η is a finite-dimensional parameter so that the model $\mathcal{P}$ given in (1) is a parametric model. Let us denote the score functions

$$\begin{pmatrix} \dot{\ell}_1(x;\theta,\eta) \\ \dot{\ell}_2(x;\theta,\eta) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial\theta}\log f(x;\theta,\eta) \\ \frac{\partial}{\partial\eta}\log f(x;\theta,\eta) \end{pmatrix} \tag{9}$$

and the second derivatives

$$\begin{pmatrix} \ddot{\ell}_{11}(x;\theta,\eta), & \ddot{\ell}_{12}(x;\theta,\eta) \\ \ddot{\ell}_{21}(x;\theta,\eta), & \ddot{\ell}_{22}(x;\theta,\eta) \end{pmatrix} = \begin{pmatrix} \frac{\partial^2}{\partial\theta\,\partial\theta^T}\log f(x;\theta,\eta), & \frac{\partial^2}{\partial\theta\,\partial\eta^T}\log f(x;\theta,\eta) \\ \frac{\partial^2}{\partial\eta\,\partial\theta^T}\log f(x;\theta,\eta), & \frac{\partial^2}{\partial\eta\,\partial\eta^T}\log f(x;\theta,\eta) \end{pmatrix}. \tag{10}$$

Also let

$$\begin{pmatrix} J_{11}, & J_{12} \\ J_{21}, & J_{22} \end{pmatrix} = -\begin{pmatrix} E\{\ddot{\ell}_{11}(X;\theta_0,\eta_0)\}, & E\{\ddot{\ell}_{12}(X;\theta_0,\eta_0)\} \\ E\{\ddot{\ell}_{21}(X;\theta_0,\eta_0)\}, & E\{\ddot{\ell}_{22}(X;\theta_0,\eta_0)\} \end{pmatrix}.$$

LEMMA 1. *For the efficient score function $\tilde{\ell}_1(x;\theta_0)$ and the efficient information matrix $\tilde{J}_{11}$ given by (4) and (5), we have*

$$\tilde{\ell}_1(x;\theta_0) = \dot{\ell}_1(x;\theta_0,\eta_0) - J_{12}J_{22}^{-1}\dot{\ell}_2(x;\theta_0,\eta_0), \tag{11}$$

and

$$\tilde{J}_{11} = J_{11} - J_{12}J_{22}^{-1}J_{21}. \tag{12}$$

PROOF. Note that the function $\hat{\eta}_\theta$ given by (2) satisfies $\hat{\eta}_{\theta_0} = \eta_0$ and

$$E\{\dot{\ell}_2(X;\theta,\hat{\eta}_\theta)\} = 0 \quad \text{for all } \theta,$$

where $\dot{\ell}_2(x;\theta,\eta)$ is the score function for η defined by (9). By differentiating this equality with respect to θ, we get

$$E\{\ddot{\ell}_{12}(X;\theta,\hat{\eta}_\theta)\} + \left(\frac{\partial}{\partial\theta^T}\hat{\eta}_\theta\right) E\{\ddot{\ell}_{22}(X;\theta,\hat{\eta}_\theta)\} = 0 \quad \text{for all } \theta.$$

It follows that, at θ_0, we have

$$\frac{\partial}{\partial\theta^T}\hat{\eta}_\theta = -J_{12}J_{22}^{-1}. \tag{13}$$

Hence the efficient score function

$$\tilde{\ell}_1(x;\theta_0) = \frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0}\log f(x;\theta,\hat{\eta}_\theta) = \dot{\ell}_1(x;\theta_0,\eta_0) + \left(\frac{\partial}{\partial\theta^T}\hat{\eta}_\theta\right)\dot{\ell}_2(x;\theta_0,\eta_0)$$

is given by (11). Now, the second derivative is

$$\begin{aligned} \frac{\partial^2}{\partial\theta\partial\theta^T}\bigg|_{\theta=\theta_0}\log f(x;\theta,\hat{\eta}_\theta) &= \ddot{\ell}_{11}(x;\theta_0,\eta_0) + 2\left(\frac{\partial}{\partial\theta^T}\hat{\eta}_{\theta_0}\right)\ddot{\ell}_{12}(x;\theta_0,\eta_0) \\ &\quad + \left(\frac{\partial}{\partial\theta^T}\hat{\eta}_{\theta_0}\right)\ddot{\ell}_{22}(x;\theta_0,\eta_0)\left(\frac{\partial}{\partial\theta}\hat{\eta}_{\theta_0}\right) + \left(\frac{\partial^2}{\partial\theta\partial\theta^T}\hat{\eta}_{\theta_0}\right)\dot{\ell}_2(x;\theta_0,\eta_0). \end{aligned}$$

By taking expectations of these functions, it follows with (13) that (12) holds. □

Equations (11) and (12) give us a formula to compute the PLIC: let

$$\hat{J}^{11} = (\hat{J}_{11} - \hat{J}_{12}\hat{J}_{22}^{-1}\hat{J}_{21})^{-1} \tag{14}$$

where

$$\hat{J}_{ij} = \frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}_{ij}(X_i; \hat{\theta}, \hat{\eta})$$

and let

$$\hat{I}_{11} = \frac{1}{n}\sum_{i=1}^{n} \{\dot{\ell}_1(X_i; \hat{\theta}, \hat{\eta}) - \hat{J}_{12}\hat{J}_{22}^{-1}\dot{\ell}_2(X_i; \hat{\theta}, \hat{\eta})\}^{\otimes 2}.$$

The PLIC is computed by

$$\widehat{\text{PLIC}} \;=\; -2\sum_{i=1}^{n} \log f(X_i; \hat{\theta}, \hat{\eta}) + 2\text{tr}(\hat{J}^{11}\hat{I}_{11}). \tag{15}$$

This formula does not require the function $\hat{\eta}_\theta$. It only requires the score functions (9) and the second derivatives of the log of density (10) in addition to the MLE $(\hat{\theta}, \hat{\eta})$.

Suppose η is an infinite dimensional parameter so that the model (1) is a semi-parametric model. In the case of nonparametric maximum likelihood estimator $\hat{\eta}$ of η, the estimator is in a finite dimensional space. We can have score functions for θ and η as if it were a parametric model. So the formula (15) is applicable in these semi-parametric models.

2.2 Example: Semi-parametric stratified sampling model

Suppose the underlying data generating process on the sample space $\mathcal{Y} \times \mathcal{X}$ is a model

$$\mathcal{Q} = \{f(y, x; \theta, \eta) = f(y|x; \theta)\eta(x) : \; \theta \in \Theta, \, \eta \in \mathcal{H}\}. \tag{16}$$

Here $f(y|x;\theta)$ is a conditional density of Y given X which depends on a finite dimensional parameter θ, $\eta(x)$ is an unspecified density of X which is an infinite-dimensional nuisance parameter. For a partition of the sample space $\mathcal{Y} \times \mathcal{X} = \cup_{s=1}^{S} \mathcal{S}_s$, define

$$Q_s(x; \theta) = \int f(y|x; \theta)\, 1_{(y,x)\in\mathcal{S}_s}\, dy,$$

and let

$$Q_s(\theta, \eta) = \int Q_s(x; \theta)\eta(x)\, dx$$

be the probability of (Y, X) belonging to stratum $\mathcal{S}_s$. In standard stratified sampling, for each $s = 1, \ldots, S$, a random sample of size n_s, $(Y_{s1}, X_{s1}), \ldots, (Y_{sn_s}, X_{sn_s})$, is taken from the conditional distribution

$$f_s(y, x; \theta, \eta) = \frac{f(y|x; \theta)\eta(x)1_{(y,x)\in\mathcal{S}_s}}{Q_s(\theta, \eta)} \tag{17}$$

of (Y, X) given stratum $\mathcal{S}_s$. We aim to find the maximum likelihood estimators for θ and η based on the data from the stratified sampling.

To find a nonparametric maximum likelihood estimator $\hat{\eta}$ of η, we assume that the support of the distribution of X is only at the observed values:

$$\text{SUPP}(X) = \{v_1, \ldots, v_K\}.$$

Let $(\eta_1, \ldots, \eta_K) = \{\eta(v_1), \ldots, \eta(v_K)\}$, then $\eta_K = 1 - \sum_{k=1}^{K-1} \eta_k$ and $\log \eta(x)$ and $Q_s(\theta, \eta)$ can be expressed as

$$\log \eta(x) = \sum_{k=1}^{K} 1_{\{x=v_k\}} \log \eta_k$$

and

$$Q_s(\theta, \eta) = \sum_{k=1}^{K} Q_s(v_k; \theta)\eta_k.$$

The log of density is

$$\log f_s(y, x; \theta, \eta) \quad = \quad \log f(y|x; \theta) + \sum_{k=1}^{K} 1_{\{x=v_k\}} \log \eta_k + \log \left\{ \sum_{k=1}^{K} Q_s(v_k; \theta)\eta_k \right\}.$$

The derivatives of the log of density are

$$\dot{\ell}_1(s, x; \theta, \eta) \quad = \quad \frac{\partial}{\partial \theta} \log f_s(y, x; \theta, \eta) = \frac{\frac{\partial}{\partial \theta} f(y|x; \theta)}{f(y|x; \theta)} + \frac{\sum_{k=1}^{K} \frac{\partial}{\partial \theta} Q_s(v_k; \theta)\eta_k}{\sum_{k=1}^{K} Q_s(v_k; \theta)\eta_k}$$

and

$$\begin{aligned} \dot{\ell}_2(s, x; \theta, \eta) \quad &= \quad \left(\frac{\partial}{\partial \eta_k} \log f_s(y, x; \theta, \eta) : k = 1, \ldots, K-1 \right) \\ &= \quad \left(\frac{1_{\{x=v_k\}}}{\eta_k} - \frac{1_{\{x=v_K\}}}{\eta_K} + \frac{Q_s(v_k; \theta) - Q_s(v_K; \theta)}{\sum_{k=1}^{K} Q_s(v_k; \theta)\eta_k} : k = 1, \ldots, K-1 \right). \end{aligned}$$

The second derivatives $\dot{\ell}_{11}$, $\dot{\ell}_{12}$, $\dot{\ell}_{21}$, $\dot{\ell}_{22}$ can be calculated similarly.

Let $(\hat{\theta}, \hat{\eta})$ be the solution to the score equations $\sum_{s=1}^{S} \sum_{i=1}^{n_s} \dot{\ell}_1(s, Y_{si}, X_{si}; \hat{\theta}, \hat{\eta}) = 0$ and $\sum_{s=1}^{S} \sum_{i=1}^{n_s} \dot{\ell}_2(s, Y_{si}, X_{si}; \hat{\theta}, \hat{\eta}) = 0$. With the log-likelihood $\log L_n(\theta, \eta) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \log f_s(Y_{si}, X_{si}; \theta, \eta)$, the PLIC is calculated by

$$\widehat{\text{PLIC}} \quad = \quad -2 \log L_n(\hat{\theta}, \hat{\eta}) + 2\text{tr}(\hat{J}^{11} \hat{I}_{11}) \tag{18}$$

where $\hat{J}^{11}$ is given by (30) with $\hat{J}_{ij} = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \ddot{\ell}_{ij}(s, Y_{si} X_{si}; \hat{\theta}, \hat{\eta})$ and $\hat{I}_{11} = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \{\dot{\ell}_1(s, Y_{si}, X_{si}; \hat{\theta}, \hat{\eta}) - \hat{J}_{12} \hat{J}_{22}^{-1} \dot{\ell}_2(s, Y_{si}, X_{si}; \hat{\theta}, \hat{\eta})\}^{\otimes 2}$.

3 Method of re-parametrization

Scott and Wild (1997, 2001) proposed a method of re-parametrization of profile-likelihood so that the log-likelihood is an explicitly defined function in terms of the parameters in the re-parametrized model. It turns out that their estimator is efficient. Motivated by their work, under the assumption (B) that the true distribution is in the chosen parametric family, Hirose

and Lee (2011) showed conditions under which re-parametrization gives efficient estimation in a context of semi-parametric multiple-sample model. In this section, we extend the results to the more general situation (A) that the true distribution may not be in the chosen parametric family.

The reason why we consider this method is that if the estimators of the parameter of interest are the same in the original semi-parametric model and the re-parametrized model, then the information criteria for these two models should coincide. We will show that the PLIC given above has this property.

In the multi-sample model, we observe S independent samples

$$X_{s1}, \ldots, X_{sn_s}, \quad s = 1, \ldots, S,$$

where each sample $X_{s1}, \ldots, X_{sn_s}$ is independently and identically distributed according to the true cdf G_s. Let $n = \sum_{s=1}^{S} n_s$. We assume $(\frac{n_1}{n}, \ldots, \frac{n_S}{n}) \to (w_1, \ldots, w_S)$ where $w_s > 0$ and $\sum_{s=1}^{S} w_s = 1$.

Suppose we choose an S-vector of semi-parametric models $(\mathcal{P}_1, \ldots, \mathcal{P}_S)$ where, for each $s = 1, \ldots, S$,

$$\mathcal{P}_s = \{f_s(x; \theta, \eta) : \ \theta \in \Theta, \eta \in \mathcal{H}\}$$

is a probability model on the sample space $\mathcal{X}_s$ with the parameter of interest θ, a finite-dimensional parameter, and the nuisance parameter η, which is an infinite-dimensional parameter.

Let (θ_0, η_0) be the maximizer of the expected log of density:

$$(\theta_0, \eta_0) = \text{argmax}_{\theta,\eta} \sum_{s=1}^{S} w_s E_s\{\log f_s(X; \theta, \eta)\},$$

here E_s denotes the expectation with respect to the cdf G_s.

We assume that the function $\hat{\eta}_\theta$ satisfies

$$\left.\frac{\partial}{\partial \eta}\right|_{\eta=\hat{\eta}_\theta} \sum_{s=1}^{S} w_s E_s\{\log f_s(X; \theta, \eta)\} = 0 \quad \text{for all } \theta \in \Theta.$$

Then the the efficient score function in the multi-sample model is given by

$$\tilde{\ell}_1(s, x; \theta_0) = \left.\frac{\partial}{\partial \theta}\right|_{\theta=\theta_0} \log f_s(x; \theta, \hat{\eta}_\theta). \tag{19}$$

Define

$$\tilde{I}_{11} = \sum_{s=1}^{S} w_s E_s \left\{\tilde{\ell}_1(s, X; \theta_0)^{\otimes 2}\right\} \tag{20}$$

and

$$\tilde{J}_{11} = -\sum_{s=1}^{S} w_s E_s \left\{ \left.\frac{\partial^2}{\partial \theta \partial \theta^T}\right|_{\theta=\theta_0} \log f_s(X; \theta, \hat{\eta}_\theta) \right\}. \tag{21}$$

By the derivation of PLIC in *Section 2*, the PLIC for the multi-sample semi-parametric model is

$$\text{PLIC} \quad = \quad -2\sum_{s=1}^{S}\sum_{i=1}^{n_s}\log f_s(X_{si};\hat{\theta},\hat{\eta}_{\hat{\theta}}) + 2\text{tr}(\tilde{J}_{11}^{-1}\tilde{I}_{11}). \tag{22}$$

In the method of re-parametrization, we assume that the density for the least favorable submodel is of the form

$$f_s(x;\theta,\hat{\eta}_\theta) = f'_s(x;\theta,q_\theta), \quad \text{for } \theta\in\Theta, \quad s=1,\ldots,S, \tag{23}$$

where q_θ is a function of θ whose values are in a finite dimensional space, the function $f'_s(x;\theta,q)$ is twice continuously differentiable with respect to (θ,q) and q is a finite dimensional parameter. Let $q_0 = q_{\theta_0}$, then, since $\eta_0 = \hat{\eta}_{\theta_0}$, we have $f_s(x;\theta_0,\eta_0) = f'_s(x;\theta_0,q_0)$. With an appropriate neighborhood D_q of q_0 in the Euclidean space, the model

$$\mathcal{P}'_s = \{f'_s(x;\theta,q) : \theta\in\Theta, q\in D_q\}, \quad s=1,\ldots,S$$

is called the re-parametrized model. Further, suppose

$$\left.\frac{\partial}{\partial q}\right|_{q=q_\theta}\sum_{s=1}^{S} w_s E_s\{\log f'_s(x;\theta,q)\} = 0 \text{ for } \theta\in\Theta. \tag{24}$$

Then the the efficient score function in the re-parametrized model is given by

$$\tilde{\ell}'_1(s,x;\theta_0) = \left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta_0}\log f'_s(x;\theta,q_\theta). \tag{25}$$

Let

$$\tilde{I}'_{11} = \sum_{s=1}^{S} w_s E_s\left\{\tilde{\ell}'_1(s,X;\theta_0)^{\otimes 2}\right\} \tag{26}$$

and

$$\tilde{J}'_{11} = -\sum_{s=1}^{S} w_s E_s\left\{\left.\frac{\partial^2}{\partial\theta\partial\theta^T}\right|_{\theta=\theta_0}\log f'_s(X;\theta,q_\theta)\right\}. \tag{27}$$

Again, by the derivation of PLIC in *Section 2*, the PLIC for the re-parametrized model is

$$\text{PLIC} \quad = \quad -2\sum_{s=1}^{S}\sum_{i=1}^{n_s}\log f'_s(X_{si};\hat{\theta},q_{\hat{\theta}}) + 2\text{tr}\{(\tilde{J}'_{11})^{-1}\tilde{I}'_{11}\}. \tag{28}$$

From the assumption (23), it is immediate that the PLICs (22) and (28) are the same information criteria with different expressions.

Let

$$\dot{\ell}'_1(s,x;\theta,q) = \frac{\partial}{\partial\theta}\log f'_s(x;\theta,q) \quad \text{and} \quad \dot{\ell}'_2(s,x;\theta,q) = \frac{\partial}{\partial q}\log f'_s(x;\theta,q)$$

be the score functions for θ and q in the re-parametrized model, respectively. Also denote the second derivatives

$$\ddot{\ell}'_{12}(s,x;\theta,q) = \frac{\partial^2}{\partial\theta\partial q^T}\log f'_s(x;\theta,q) \quad \text{and} \quad \ddot{\ell}'_{22}(s,x;\theta,q) = \frac{\partial^2}{\partial q\partial q^T}\log f'_s(x;\theta,q).$$

By the derivation of the formula (15), we compute the PLIC (28) in the re-parametrized model by

$$\widehat{\text{PLIC}} \quad = \quad -2\sum_{s=1}^{S}\sum_{i=1}^{n_s}\log f'_s(X_{si};\hat{\theta},\hat{q}) + 2\text{tr}\{(\hat{J}')^{11}\hat{I}'_{11}\}. \tag{29}$$

where

$$(\hat{J}')^{11} = (\hat{J}'_{11} - \hat{J}'_{12}(\hat{J}'_{22})^{-1}\hat{J}'_{21})^{-1}, \tag{30}$$

$$\hat{J}'_{ij} = n^{-1}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\ddot{\ell}'_{ij}(s,X_{si};\hat{\theta},\hat{q})$$

and

$$\hat{I}'_{11} = n^{-1}\sum_{s=1}^{S}\sum_{i=1}^{n}\{\dot{\ell}'_1(s,X_{si};\hat{\theta},\hat{q}) - \hat{J}'_{12}(\hat{J}'_{22})^{-1}\dot{\ell}'_2(s,X_{si};\hat{\theta},\hat{q})\}^{\otimes 2}.$$

In the next, we apply the result to the example of stratified sampling example given in Section 2.2.

3.1 Example: Semi-parametric stratified sampling model continued

This example is a continuation of the semi-parametric stratified sampling model which we discussed in Section 2.2. For each $s = 1,\ldots,S$, let G_s be the true cumulative distribution function (cdf) which we aim to approximate by the chosen model $f_s(y,x;\theta,\eta)$. Let w_s, $s = 1,\ldots,S$, be the weight probabilities, i.e., $w_s > 0$ for all s and $\sum_s w_s = 1$. The expected log likelihood with the weight probabilities w_s and the cdfs G_s is

$$\sum_{s=1}^{S} w_s \int \log f_s(y,x;\theta,\eta)dG_s \quad = \quad \sum_{s=1}^{S} w_s\left[\int\{\log f(y|x;\theta) + \log\eta(x)\}\,dG_s - \log Q_s(\theta,\eta)\right].$$

Again we assume that the support of the distribution of X is finite:

$$\text{SUPP}(X) = \{v_1,\ldots,v_K\}.$$

To find the maximizer $(\eta_1,\ldots,\eta_K)$ of the expected log-likelihood at θ, differentiate the expected likelihood with respect to η_k and set the derivative equal to zero,

$$\frac{\partial}{\partial\eta_k}\sum_{s=1}^{S} w_s \int \log f_s(y,x;\theta,\eta)dG_s \quad = \quad \sum_{s=1}^{S} w_s\left\{\frac{\int 1_{x=v_k}dG_s}{\eta_k} - \frac{Q_{s|X}(v_k;\theta)}{Q_s(\theta,\eta)}\right\} = 0.$$

The solution η_k to the equation is

$$\hat{\eta}_\theta(v_k) = \eta_k \quad = \quad \frac{\sum_{s=1}^{S} w_s \int 1_{x=v_k}dG_s}{\sum_{s=1}^{S} w_s \frac{Q_{s|X}(v_k;\theta)}{Q_s(\theta,\eta)}}.$$

This form motivate us to work with a continuous extension of the solution: Suppose $g_s(y,x)$ is the density funciton corresponds to the distribution function G_s, then the continuous extension of the solution can be written as

$$\hat{\eta}_\theta(x) = \hat{\eta}(x,\theta,\hat{Q}(\theta)) = \frac{g^*(x)}{\sum_{s=1}^S w_s \frac{Q_{s|X}(x;\theta)}{\hat{Q}_s(\theta)}},$$

where

$$g^*(x) = \sum_{s=1}^S w_s \int g_s(y,x)dy,$$

and

$$\hat{Q}_s(\theta) = \int Q_{s|X}(x;\theta)\hat{\eta}(x,\theta,\hat{Q}(\theta))dx, \quad s=1,\ldots,S.$$

Now let

$$f_s(y,x;\theta,\hat{\eta}_\theta) = \frac{f(y|x;\theta)1_{(y,s)\in\mathcal{S}_s}\hat{\eta}(x,\theta,\hat{Q}(\theta))}{\hat{Q}_s(\theta)}, \quad s=1,\ldots,S, \tag{31}$$

be the model we wish to re-parametrize. From this it is immediate that condition (23) is satisfied (with $\hat{Q}(\theta) = (\hat{Q}_1(\theta),\ldots,\hat{Q}_S(\theta))$ as the function q_θ).

By replacing $\hat{Q}(\theta) = (\hat{Q}_1(\theta),\ldots,\hat{Q}_{S-1}(\theta),\hat{Q}_S(\theta))$ with $q = (q_1,\ldots,q_{S-1},1)$, we consider a re-parametrized model of the form

$$f'_s(y,x;\theta,q) = \frac{f(y|x;\theta)1_{(y,s)\in\mathcal{S}_s}\hat{\eta}(x,\theta,q)}{q_s}, \quad s=1,\ldots,S,$$

where

$$\hat{\eta}(x,\theta,q) = \frac{g^*(x)}{\sum_{s=1}^S w_s \frac{Q_{s|X}(x;\theta)}{q_s}}.$$

The true value of (θ,q) is

$$(\theta_0,q_0) = \left(\theta_0, \left(\frac{Q_1(\theta_0,g_0)}{Q_S(\theta_0,g_0)},\ldots,\frac{Q_{S-1}(\theta_0,g_0)}{Q_S(\theta_0,g_0)},1\right)\right).$$

For $j=1,\ldots,S-1$, the derivative is

$$\begin{aligned}\frac{\partial}{\partial q_j}\sum_{s=1}^S w_s E_s\{\log f'_s(y,x;\theta,q)\} &= -\frac{\partial}{\partial q_j}\sum_{s=1}^S w_s E_s\left\{\log\sum_{s'=1}^S w_{s'}\frac{Q_{s'|X}(x;\theta)}{q_{s'}} + \log q_s\right\} \\ &= \frac{w_j}{q_j^2}\left\{\int Q_{j|X}(x;\theta)\hat{\eta}(x,\theta,q)dx - q_j\right\}.\end{aligned}$$

It follows that, for all $\theta\in\Theta$, we have

$$\frac{\partial}{\partial q}\Big|_{q=\hat{Q}(\theta)}\sum_{s=1}^S w_s E_s\{\log f'_s(y,x;\theta,q)\} = 0.$$

This verifies the condition (24). Therefore the formula (29) is applicable to this example.

The log of density in the re-parametrized model is

$$\log f'_s(y,x;\theta,q) \;=\; \log f(y|x;\theta) + \log g^*(x) - \log \sum_{s'=1}^{S} w_{s'} \frac{Q_{s'|X}(x;\theta)}{q_{s'}} - \log q_s.$$

The score functions in the re-parametrized model are

$$\begin{aligned}\dot{\ell}'_1(s,y,x;\theta,q) &= \frac{\partial}{\partial\theta}\log f'_s(y,x;\theta,q) \\ &= \frac{\frac{\partial}{\partial\theta}f(y|x;\theta)}{f(y|x;\theta)} - \frac{\sum_{s'=1}^{S} w_{s'} \frac{\frac{\partial}{\partial\theta}Q_{s'|X}(x;\theta)}{q_{s'}}}{\sum_{s'=1}^{S} w_{s'} \frac{Q_{s'|X}(x;\theta)}{q_{s'}}}\end{aligned}$$

and

$$\begin{aligned}\dot{\ell}'_2(s,y,x;\theta,q) &= \left(\frac{\partial}{\partial q_k}\log f'_s(y,x;\theta,q) : k=1,\ldots,S-1\right) \\ &= \left(\frac{w_k \frac{Q_{k|X}(x;\theta)}{q_k^2}}{\sum_{s'=1}^{S} w_{s'} \frac{Q_{s'|X}(x;\theta)}{q_{s'}}} - \frac{1_{k=s}}{q_k} : k=1,\ldots,S-1\right)\end{aligned}$$

The second derivatives $\ddot{\ell}'_{11}$, $\ddot{\ell}'_{12}$, $\ddot{\ell}'_{21}$, $\ddot{\ell}'_{22}$ can be calculated similarly.

Then PLIC in this model is calculated by (29) with obvious modification. Since PLICs in the original semi-parametric model and the re-parametrized model coincides, this value must be close to the one calculated by (18).

4 An extension of PLIC

We have developed PLIC using Kullback–Leibler distance between g and $f(x;\hat{\theta},\hat{\eta}_{\hat{\theta}})$ where the function $\hat{\eta}_\theta$ is given by (2). For the empirical cdf G_n, define

$$\hat{\eta}_{\theta,G_n} = \mathrm{argmax}_\eta \int \log f(x;\theta,\eta) dG_n(x). \tag{32}$$

In this section, we show that the reason why it is not useful to have a PLIC based on the Kullback–Leibler distance between g and $f(x;\hat{\theta},\hat{\eta}_{\hat{\theta},G_n})$:

$$I\{g(\cdot), f(\cdot;\hat{\theta},\hat{\eta}_{\hat{\theta},G_n})\} \;=\; \int \log g(x) dG(x) - \int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta},G_n}) dG(x) \tag{33}$$

We assume that the parameter η is a finite dimensional parameter so that the model (1) is a parametric model.

The bias of $\int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta},G_n}) dG_n(x)$ as an estimator of the second term in (33) is

$$\begin{aligned}\text{bias} &= E\left\{\int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta},G_n}) dG_n(x) - \int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta},G_n}) dG(x)\right\} \\ &= E\left\{\int \log f(x;\hat{\theta},\hat{\eta}_{\hat{\theta},G_n}) d(G_n - G)(x)\right\}\end{aligned}$$

By Taylor's expansion, for some θ^* between $\hat{\theta}$ and θ_0, and for some η^* between $\hat{\eta}_{\theta_0,G_n}$ and η_0, the bias is expanded as

$$\begin{aligned}
\text{bias} &= E\left\{\int \log f(x;\theta_0,\eta_0)d(G_n-G)(x)\right\} \\
&\quad +n^{-1}E\left\{n^{1/2}\int \tilde{\ell}_1(x;\theta_0,G)^T d(G_n-G)(x)n^{1/2}(\hat{\theta}-\theta_0)\right\} \\
&\quad +n^{-1}E\left\{n^{1/2}\int \dot{\ell}_2(x;\theta_0,\eta_0)^T d(G_n-G)(x)n^{1/2}(\hat{\eta}_{\theta_0,G_n}-\eta_0)\right\} \\
&\quad +(2n)^{-1}E\left\{n^{1/2}(\hat{\theta}-\theta_0)^T\int \tilde{\ell}_{11}(x;\theta^*,G_n)d(G_n-G)(x)n^{1/2}(\hat{\theta}-\theta_0)\right\} \\
&\quad +(2n)^{-1}E\left\{n^{1/2}(\hat{\eta}_{\theta_0,G_n}-\eta_0)^T\int \ddot{\ell}_{22}(x;\theta_0,\eta^*)d(G_n-G)(x)n^{1/2}(\hat{\eta}_{\theta_0,G_n}-\eta_0)\right\} \\
&\quad +n^{-1}E\left\{n^{1/2}(\hat{\theta}-\theta_0)^T n^{1/2}\int\{\tilde{\ell}_1(x;\theta_0,G_n)-\tilde{\ell}_1(x;\theta_0,G)\}d(G_n-G)(x)\right\}.
\end{aligned}$$

The first term in the right hand side is zero. Due to (35) in Lemma 2 given below, the last term in the right hand side is $o_P(n^{-1})$. By (37), $n^{1/2}(\hat{\theta}-\theta_0)=O_P(1)$. Under the mild regularity conditions we have the uniform weak law of large numbers: $\sup_\theta \int \tilde{\ell}_{11}(x;\theta,G_n)d(G_n-G)(x)=o_P(1)$. These imply the forth term in the right hand side is $o_P(n^{-1})$. Similarly, the fifth term in the right hand side is $o_P(n^{-1})$. We look closely the second and third terms in the right hand side.

Using (37) and (7), the second term in right hand side is equal to

$$n^{-1}\text{tr}\left[\tilde{J}_{11}^{-1}\tilde{I}_{11}\right]+o(n^{-1}).$$

Similarly, using (34) and $\int \dot{\ell}_2(x;\theta_0,\eta_0)dG(x)=0$, the third term in right hand side is equal to

$$\begin{aligned}
& n^{-1}E\left\{n^{1/2}\int \dot{\ell}_2(x;\theta_0,\eta_0)^T d(G_n-G)(x)n^{1/2}(\hat{\eta}_{\theta_0,G_n}-\eta_0)\right\} \\
= \;& n^{-1}E\left\{\left(n^{1/2}\int \dot{\ell}_2 dG_n\right)^T J_{22}^{-1}\left(n^{1/2}\int \dot{\ell}_2 dG_n\right)\right\}+o(n^{-1}) \\
= \;& n^{-1}\text{tr}\left\{J_{22}^{-1}\text{var}\left(n^{1/2}\int \dot{\ell}_2 dG_n\right)\right\}+o(n^{-1}) \\
= \;& n^{-1}\text{tr}(J_{22}^{-1}I_{22})+o(n^{-1}),
\end{aligned}$$

where $I_{22}=E(\dot{\ell}_2\dot{\ell}_2^T)$.

If we combine all of these, the bias can be written as

$$\text{bias}=n^{-1}\text{tr}(\tilde{J}_{11}^{-1}\tilde{I}_{11})+n^{-1}\text{tr}(J_{22}^{-1}I_{22})+o(n^{-1}).$$

The extended profile likelihood information criteria (PLIC_{ext}) is given by

$$\text{PLIC}_{\text{ext}} = -2\sum_{i=1}^{n}\log f(X_i;\hat{\theta},\hat{\eta}_{\hat{\theta},G_n})+2\text{tr}(\tilde{J}_{11}^{-1}\tilde{I}_{11})+2\text{tr}(J_{22}^{-1}I_{22})$$

The term $\mathrm{tr}(J_{22}^{-1}I_{22})$ depends on the choice of parametrization for the nuisance parameter η. For example, in the stratified sampling example, when $g(x) = f(x;\theta_0,\eta_0)$ we have $J_{22} = I_{22}$ and, $\mathrm{tr}(J_{22}^{-1}I_{22}) \approx n$ for the original semi-parametric model and $\mathrm{tr}(J_{22}^{-1}I_{22}) = S-1$ for the re-parametrized model. The $\mathrm{PLIC}_{\mathrm{ext}}$ is not useful to compare parametric/semi-parametric models with nuisance parameters.

The next lemma gives asymptotic linear expansion of the maximum profile likelihood estimator $\hat{\theta}$ and $\hat{\eta}_{\theta_0,G_n}$.

LEMMA 2.

(a) For the function $\hat{\eta}_{\theta,G_n}$ given by (32), we have

$$n^{1/2}(\hat{\eta}_{\theta_0,G_n} - \eta_0) = n^{1/2}\int J_{22}^{-1}\dot{\ell}_2(x;\theta_0,\eta_0)dG_n(x) + o_P(1) \tag{34}$$

where $J_{22} = -E = \{\frac{\partial^2}{\partial\eta\partial\eta^T}|_{\theta=\theta_0,\eta=\eta_0}\log f(x;\theta,\eta)\}$.

(b) For the function $\tilde{\ell}_1(x;\theta,G_n) = \frac{\partial}{\partial\theta}\log f(x;\theta,\hat{\eta}_{\theta,G_n})$, we have

$$n^{1/2}\int\{\tilde{\ell}_1(x;\theta_0,G_n) - \tilde{\ell}_1(x;\theta_0,G)\}dG_n(x) = o_P(1). \tag{35}$$

(c) A solution $\hat{\theta}$ to the profile likelihood score equation

$$n^{1/2}\int\tilde{\ell}_1(x;\hat{\theta},G_n)dG_n(x) = 0 \tag{36}$$

is an asymptotically linear estimator such that

$$n^{1/2}(\hat{\theta}-\theta_0) = n^{1/2}\int\tilde{J}_{11}^{-1}\tilde{\ell}_1(x;\theta_0,G)dG_n(x) + o_P(1). \tag{37}$$

PROOF. (a) From (32), the function $\hat{\eta}_{\theta_0,G_n}$ is solution to

$$n^{1/2}\int\dot{\ell}_2(x;\theta_0,\hat{\eta}_{\theta_0,G_n})dG_n(x) = 0$$

where $\dot{\ell}_2(x;\theta,\eta) = \frac{\partial}{\partial\eta}\log f(x;\theta,\eta)$ is the score function for η. By usual Taylor's expansion argument, it follows that $n^{1/2}(\hat{\eta}_{\theta_0,G_n} - \eta_0)$ has an asymptotic linear expansion (34).

(b) Using

$$\tilde{\ell}_1(x;\theta,G) = \dot{\ell}_1(x;\theta,\hat{\eta}_{\theta,G}) + \left(\frac{\partial}{\partial\theta^T}\hat{\eta}_{\theta,G}\right)\dot{\ell}_2(x;\theta,\hat{\eta}_{\theta,G})$$

and Taylor's expansion, there are η^* and η^{**} in between $\hat{\eta}_{\theta_0,G_n}$ and $\hat{\eta}_{\theta_0,G} = \eta_0$ such that

$$\begin{aligned}
& n^{1/2}\int\{\tilde{\ell}_1(x;\theta_0,G_n) - \tilde{\ell}_1(x;\theta_0,G)\}dG_n(x) \\
= & \int\ddot{\ell}_{12}(x;\theta_0,\eta^*)dG_n(x)n^{1/2}(\hat{\eta}_{\theta_0,G_n} - \eta_0) \\
& + \left(\frac{\partial}{\partial\theta^T}\hat{\eta}_{\theta_0,G_n}\right)\int\ddot{\ell}_{22}(x;\theta_0,\eta^{**})dG_n(x)n^{1/2}(\hat{\eta}_{\theta_0,G_n} - \eta_0) \\
& + \left\{\frac{\partial}{\partial\theta^T}n^{1/2}(\hat{\eta}_{\theta_0,G_n} - \hat{\eta}_{\theta_0,G})\right\}\int\dot{\ell}_2(x;\theta_0,\eta_0)dG_n(x).
\end{aligned}$$

Since $G_n \xrightarrow{P} G$, we have $\hat{\eta}_{\theta_0,G_n} \xrightarrow{P} \hat{\eta}_{\theta_0,G} = \eta_0$ and it follows that $\eta^*, \eta^{**} \xrightarrow{P} \eta_0$.

Then under the mild regularity conditions, we have that

$$\frac{\partial}{\partial \theta^T}\hat{\eta}_{\theta_0,G_n} \xrightarrow{P} \frac{\partial}{\partial \theta^T}\hat{\eta}_{\theta_0,G} = -J_{12}J_{22}^{-1},$$

$$\int \ddot{\ell}_{12}(x;\theta_0,\eta^*)dG_n(x) \xrightarrow{P} \int \ddot{\ell}_{12}(x;\theta_0,\eta_0)dG(x) = -J_{12},$$

$$\int \ddot{\ell}_{22}(x;\theta_0,\eta^{**})dG_n(x) \xrightarrow{P} \int \ddot{\ell}_{22}(x;\theta_0,\eta_0)dG(x) = -J_{22}$$

and

$$\left\{\frac{\partial}{\partial \theta^T} n^{1/2}(\hat{\eta}_{\theta_0,G_n} - \hat{\eta}_{\theta_0,G})\right\} = O_P(1).$$

The claim follows from these.

(c) By (35), the solution $\hat{\theta}$ to the equation (36) also satisfies

$$n^{1/2}\int \tilde{\ell}_1(x;\hat{\theta},G)dG_n(x) = o_P(1).$$

Then by usual Taylor's expansion argument, we have (37). □

References

Claeskens G. and Carroll R. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94** 249–265.

Claeskens G. and Hjort N.L. (2003). Focused information Criterion (with Discussion). *J. Am. Statist. Assoc.* **98** 900–945.

Hirose, Y. (2011). Efficiency of profile likelihood in semi-parametric models, *Ann. Inst. Statist. Math.* DOI 10.1007/s10463-010-0280-y.

Hirose, Y. and Lee A.(2011). Reparametrization of the Least Favorable Submodel in Semi-Parametric Multi-Sample Models, *Bernoulli* To appear.

Hjort N.L. and Claeskens G. (2003). Frequentist model average estimators (with Discussion). *J. Am. Statist. Assoc.* **98** 879–899.

Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood (with discussion). *J. Amer. Statist. Assoc.* **95** 449–485.

Xu R., Vaida F. and Harrington D.P. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazard mixed models. *Statistica Sinica* **19** 819–842.