Working Paper Series, N. 17, December 2005

Department of Statistical Sciences
University of Padua
Italy

Likelihood theory, prediction, model selection: asymptotic connections

Luigi Pace
Department of Statistics
University of Udine
Italy

Alessandra Salvan and Laura Ventura
Department of Statistical Sciences
University of Padua
Italy

Abstract: Plug-in estimation and corresponding refinements involving penalisation have been considered in various areas of parametric statistical inference. One major example is adjustment of the profile likelihood for inference in the presence of nuisance parameters. Another important setting is prediction, where improved estimative predictive densities have been recently developed. A third related setting is model selection, where information criteria based on penalisation of maximised likelihood have been proposed starting from the pioneering contribution of Akaike. The seminal contributions in the last setting predate those introducing the former two classes of procedures, and pertinent portions of literature seem to have evolved quite independently. The aim of this paper is to establish some simple asymptotic connections among these classes of procedures. In particular, all the three kinds of penalisations involved can be viewed as bias corrections of plug-in estimates of theoretical target criteria which are shown to be very closely connected. As a by-product, we obtain adjusted profile likelihoods from optimal predictive densities. Links between adjusted procedures in likelihood theory and model selection procedures are also briefly enquired throuh some simulation studies.

Keywords: Akaike's information criterion, Likelihood asymptotics, Model selection, Nuisance parameter, Predictive density, Profile likelihood.

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Laura Ventura
tel: +39 049 827 4177
ventura@stat.unipd.it
http://www.stat.unipd.it/˜ventura

Likelihood theory, prediction, model selection: asymptotic connections

Luigi Pace
Department of Statistics
University of Udine
Italy

Alessandra Salvan and Laura Ventura
Department of Statistical Sciences
University of Padua
Italy

Abstract: Plug-in estimation and corresponding refinements involving penalisation have been considered in various areas of parametric statistical inference. One major example is adjustment of the profile likelihood for inference in the presence of nuisance parameters. Another important setting is prediction, where improved estimative predictive densities have been recently developed. A third related setting is model selection, where information criteria based on penalisation of maximised likelihood have been proposed starting from the pioneering contribution of Akaike. The seminal contributions in the last setting predate those introducing the former two classes of procedures, and pertinent portions of literature seem to have evolved quite independently. The aim of this paper is to establish some simple asymptotic connections among these classes of procedures. In particular, all the three kinds of penalisations involved can be viewed as bias corrections of plug-in estimates of theoretical target criteria which are shown to be very closely connected. As a by-product, we obtain adjusted profile likelihoods from optimal predictive densities. Links between adjusted procedures in likelihood theory and model selection procedures are also briefly enquired throuh some simulation studies.

Keywords: Akaike's information criterion, Likelihood asymptotics, Model selection, Nuisance parameter, Predictive density, Profile likelihood.

1 Introduction

In statistical modelling, penalisation is a general idea to pursue a balance between goodness of fit, as description of the available data, and appropriate behaviour under repeated sampling. Techniques based on penalisation of a maximized likelihood, or of the maximum likelihood estimate of a density, have been developed in different areas of parametric statistical inference. Here we focus on (*i*) adjustments of the profile loglikelihood for inference on a parameter of interest in likelihood asymptotics, (*ii*) improvements of the estimative density in prediction, and (*iii*) model selection criteria, such as the well-known Akaike's AIC. The seminal contributions in the last setting predate those introducing the former two classes of procedures. Our aim is to show that, under suitable assumptions, these three classes of penalised procedures are more closely interrelated than is currently suggested within the literature.

A unifying aspect is the need to take properly into account uncertainty introduced by sampling variability of the maximum likelihood estimates of the nuisance components and to accomodate for the ensuing bias. The intended nuisance components are nuisance parameters in the likelihood setting and the whole model parameter in prediction as well as in model selection. Each of these settings has of course its own specific inferential objective. Accordingly, there are various ways, aiming at different theoretical target criteria, to adjust for the substitution of nuisance components with their empirical surrogates. Within a unified formalisation, under the assumption that the overall statistical model is not misspecified, it turns out that there exist some simple asymptotic connections among classes (*i*), (*ii*) and (*iii*) above.

A unified notation and background material is given in Sections 2 and 3, referring to purely estimative and adjusted methods, respectively. Section 4 discusses theoretical target criteria which lead to different penalisations in the three settings we are considering. Estimative counterparts of the theoretical objectives presented in Section 4 are discussed in Section 5. Section 6 is devoted to the relation between adjustments of the profile likelihood and optimal predictive densities.

2 Background: plug-in methods

Let us denote by $y = (y_1, \ldots, y_n)$ the available data, considered for simplicity as a random sample of size n, i.e. as a realisation of a random variable $Y = (Y_1, \ldots, Y_n)$, having independent and identically distributed components. While relaxing independence usually requires care, non-identically distributed components of Y are easily coped with in the theory below.

Let us consider first the typical setting of likelihood theory for inference on a q-dimensional parameter of interest ψ in the presence of an m-dimensional nuisance parameter λ. Let $p(y;\theta) = p(y;\psi,\lambda) = \prod_{i=1}^n p_{Y_i}(y_i;\psi,\lambda)$ denote the density of Y, where $\theta = (\psi,\lambda) \in \Theta \subseteq \mathrm{IR}^d$, with $d = q + m$. Let us denote by $\ell(\theta) = \ell(\psi,\lambda) = \ell(\psi,\lambda;y) = \log p(y;\theta)$ the loglikelihood function based on y and by $\hat\theta = (\hat\psi,\hat\lambda)$ the maximum likelihood estimate (MLE) of $\theta = (\psi,\lambda)$. Moreover, let $\hat\lambda_\psi$ be the constrained MLE of λ for a given value of ψ and $\hat\theta_\psi = (\psi,\hat\lambda_\psi)$. In the presence of a nuisance parameter, likelihood inference is often based on a pseudolikelihood. This is a function of ψ and y ideally having properties similar to those of a genuine likelihood for ψ, such as a marginal or conditional likelihood for ψ. When no exact reduction by marginalisation or conditioning is available, the most commonly used pseudologlikelihood for inference about ψ is the profile loglikelihood

$$\ell_P(\psi) = \ell_P(\psi;y) = \ell(\hat\theta_\psi) = \ell(\psi,\hat\lambda_\psi)\ . \tag{1}$$

As is well known, (1) is not a genuine loglikelihood for ψ, but shares most first-order properties of a genuine loglikelihood for ψ (see Barndorff-Nielsen and Cox, 1994, Section 3.4).

Elimination of nuisance parameters through maximum likelihood estimates is widely considered in prediction as well. In the simplest instance of this setting, the object of inference is a future or as yet unobserved random vector $X = (X_1, \dots, X_h)$, $h \geq 1$, independent of $Y = (Y_1, \dots, Y_n)$, and having independent and identically distributed components where X_1 has the same distribution as Y_1. Let us denote by $p_X(x;\theta)$ the density of X. Formal links with the previous setting are recovered by letting $\theta = (\psi, \lambda)$, with $\theta \in \Theta \subseteq \mathrm{I\!R}^{q+m}$, and treating ψ as known, while λ is considered as nuisance. The simplest frequentist approach to prediction of X, on the basis of the observed y from Y, consists in using the estimative predictive density function

$$p_e(x;\psi) = p_X(x;\psi,\hat{\lambda}_\psi) \ , \tag{2}$$

obtained by substituting, for the given ψ, the unknown λ with its MLE based on y, $\hat{\lambda}_\psi = \hat{\lambda}_\psi(y)$.

Maximum likelihood estimation of nuisance parameters plays an important role also in model selection. Here we have to compare several competing models in order to choose one that most suitably describes the data generation process. A selection procedure consists of a rule that associates to data y one model among those available. Suppose that k plausible parametric models M_j, $j = 1, \dots, k$, may be used to analyse y, with $k \geq 2$. Let us write the probability density functions specified within model M_j, $j = 1, \dots, k$, in the form $p_j(y;\lambda_{(j)})$, with $\lambda_{(j)} \in \Lambda_j \subseteq \mathrm{I\!R}^{m_j}$. Such a notation underlines the fact that parameters pertaining to different models may belong to different parameter spaces. Model selection may be recast as a problem of inference about a parameter of interest ψ, which is the index j of the competing models. The overall model M for y is then $M = \cup_{j=1}^k M_j$ and the probability density functions specified within M may be written in the form $p(y;\theta)$, with $\theta = (\psi, \lambda_{(\psi)})$, where $\psi \in \Psi = \{1, \dots, k\}$ and $\lambda_{(\psi)} \in \Lambda_\psi \subseteq \mathrm{I\!R}^{m_\psi}$. The specific value of $\lambda_{(\psi)}$ is not of primary interest, so $\lambda_{(\psi)}$ is a nuisance parameter. One major difference with respect to the standard setting of likelihood theory is that here Ψ is finite, with k elements. Moreover, the overall parameter space $\Theta = \{(\psi, \lambda_{(\psi)}) : \psi \in \Psi, \ \lambda_{(\psi)} \in \Lambda_\psi\}$ is usually not of form $\Psi \times \Lambda$ for some Λ. In frequentist model selection, the nuisance parameter $\lambda_{(\psi)}$ can be eliminated through a suitable pseudolikelihood which only depends on the index parameter ψ, such as the marginal loglikelihood (Quesenberry, 1985; Pace *et al.*, 2005) or the profile loglikelihood $\ell_P(\psi) = \sup_{\lambda_{(\psi)} \in \Lambda_\psi} \ell(\psi, \lambda_{(\psi)})$. Note that only loglikelihood summands independent of ψ, i.e. that are the same across all k models, are negligible in $\ell(\psi, \lambda_{(\psi)})$. Selection based on maximisation of $\ell_P(\psi)$ amounts to choosing the model that best fits the data. Such a procedure will always select the less parsimonious model when the competing models are nested and adjustments are called for.

The estimative, or plug-in, methods considered in likelihood theory, prediction and model selection, neglect sampling variability of the estimated nuisance components. Particularly serious inaccuracies may occur when the dimension of the nuisance component, λ or $\lambda_{(\psi)}$, is large relative to n. See e.g. Sartori (2003) and Vidoni (1995).

3 Background: adjusted methods

In likelihood theory, various modifications of the profile loglikelihood (1) have been proposed, starting from Barndorff-Nielsen (1980, 1983); see Severini (2000, Chapter 9) for a recent account. All the available adjustments to the profile likelihood are equivalent to second order and share the common feature of reducing the score bias to the order $O(n^{-1})$. Reduction of the score bias is one of the motivations for adjusting the profile likelihood. Other proposals have the aim of approximating some target likelihood, defined by a suitable marginality or conditionality reduction. Pace and Salvan (2005) propose a unifying perspective that does not depend on the continuous or discrete nature of the parameter of interest and looks suitable for the purposes of the present paper. In short, available modifications of the profile loglikelihood are shown to be equivalent to first-order bias adjustments of the profile loglikelihood considered as an estimate of the expectation of the least favourable target loglikelihood. The latter is defined as the loglikelihood corresponding, for each ψ, to the model element having minimum Kullback–Leibler divergence with respect to the true data generating distribution. Although the least favourable target loglikelihood is a genuine loglikelihood, it is not available in practice because it depends on the true parameter value.

In prediction, a number of improvements of the estimative predictive density have been considered. In particular, the proposals in Harris (1989) and Vidoni (1995) aim to reduce the Kullback-Leibler divergence between the unknown density of the future observation and the predictive density. Komaki (1996) obtains the improvement over the estimative predictive density which is asymptotically optimal in terms of average Kullback–Leibler divergence. In the same vein, Corcuera and Giummolè (2000) derive improvements asymptotically optimal with reference to the class of α-divergences, that includes the Kullback–Leibler divergence. All the resulting predictive densities appear in the form of (2) plus an appropriate correction term, which involves estimated likelihood quantities. For a broad discussion and a different approach, see Barndorff-Nielsen and Cox (1996).

Even in the context of model selection, several adjustments to the criterion based on the profile likelihood have been introduced, starting from Akaike's information criterion (AIC, Akaike, 1973) and Takeuchi's information criterion (TIC, Takeuchi, 1976). See Burnham and Anderson (2002, Chapters 2 and 7) for a recent account; see also Claeskens and Hjort (2003). Information criteria, such as AIC and TIC, are based on bias corrections of the profile loglikelihood as an estimate of an expected Kullback-Leibler divergence, see e.g. Shibata (1997). Using a principle of parsimony, AIC will select the model with the fewest parameters that fits the data well. TIC is a generalisation of AIC having a more refined penalisation term, depending on information matrices for $\lambda_{(\psi)}$. Also other related model selection criteria, such as the Bayesian information criterion (BIC) of Schwarz (1978), amount to a penalisation of the profile loglikelihood.

We briefly recall below the expression of notable instances of penalised loglikelihoods or modified estimative densities. Hereafter, $E_\theta(\cdot)$ and $V_\theta(\cdot)$ will denote expectation and variance under θ. In possibly misspecified models, symbols like $E_0^Y(\cdot)$, $E_0^X(\cdot)$, and so on, will be used to denote expectation of a function of Y or X under the true distribution.

Consider first the setting of likelihood theory. Let us denote by $\ell_\psi = \ell_\psi(\theta)$ and $\ell_\lambda = \ell_\lambda(\theta)$ blocks of the score vector $\ell_\theta = \partial\ell(\theta)/\partial\theta$. Moreover, let $j_{\psi\psi} = j_{\psi\psi}(\theta)$, $j_{\psi\lambda} = j_{\psi\lambda}(\theta)$ and $j_{\lambda\lambda} = j_{\lambda\lambda}(\theta)$ be blocks of the observed information $j = j(\theta) = -\partial^2\ell(\theta)/(\partial\theta\partial\theta^\top)$. Similarly, we will denote by $i_{\psi\psi} = i_{\psi\psi}(\theta)$, $i_{\psi\lambda} = i_{\psi\lambda}(\theta)$ and $i_{\lambda\lambda} = i_{\lambda\lambda}(\theta)$ blocks of the expected information $i = i(\theta) = E_\theta(j(\theta))$. Assume that the minimal sufficient statistic for the model is a one-to-one function of $(\hat\psi, \hat\lambda, a)$, where a is an ancillary statistic, either exactly or approximately, so that $\ell(\psi, \lambda; y) = \ell(\psi, \lambda; \hat\psi, \hat\lambda, a)$. Then, the modified profile loglikelihood of Barndorff-Nielsen (1980, 1983) is

$$\ell_{_M}(\psi) = \ell_{_M}(\psi; y) = \ell_{_P}(\psi) - \frac{1}{2}\log \mid j_{\lambda\lambda}(\hat\theta_\psi) \mid - \log\left|\frac{\partial\hat\lambda_\psi}{\partial\hat\lambda}\right| ,$$

where

$$\left|\frac{\partial\hat\lambda_\psi}{\partial\hat\lambda}\right| = \frac{\mid \ell_{_{\lambda;\hat\lambda}}(\hat\theta_\psi) \mid}{\mid j_{\lambda\lambda}(\hat\theta_\psi) \mid},$$

involving the sample space derivatives $\ell_{_{\lambda;\hat\lambda}}(\psi, \lambda) = \partial^2\ell(\psi, \lambda; \hat\psi, \hat\lambda, a)/(\partial\lambda\, \partial\hat\lambda^\top)$. Calculation of sample space derivatives is straightforward only in special classes of models, notably exponential and group families. When ψ and λ are orthogonal, i.e. when $i_{\psi\lambda} = 0$, such a calculation can be avoided because $\log\left|\partial\hat\lambda_\psi/\partial\hat\lambda\right| = O_p(n^{-1})$ when $\psi - \hat\psi = O(n^{-1/2})$. This gives the approximate conditional likelihood of Cox and Reid (1987)

$$\ell_{_A}(\psi) = \ell_{_A}(\psi; y) = \ell_{_P}(\psi) - \frac{1}{2}\log \mid j_{\lambda\lambda}(\hat\theta_\psi) \mid ,$$

which approximates $\ell_{_M}(\psi)$ with error of order $O_p(n^{-1})$. Recently, attention has been devoted to approximate calculation of sample space derivatives. For a review, see Severini (2000, Section 9.5). In particular, the approximation of $\ell_{_M}(\psi)$ developed in Severini (1998) is

$$\bar\ell_{_M}(\psi) = \bar\ell_{_M}(\psi; y) = \ell_{_P}(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat\theta_\psi)| - \log|\nu_{\lambda,\lambda}(\hat\theta_\psi, \hat\theta; \hat\theta)| , \qquad (3)$$

where

$$\nu_{\lambda,\lambda}(\theta_1, \theta_2; \theta_0) = E_{\theta_0}(\ell_\lambda(\theta_1)\ell_\lambda(\theta_2)^\top) .$$

An asymptotically equivalent version of (3) is obtained by replacing $\nu_{\lambda,\lambda}(\hat\theta_\psi, \hat\theta; \hat\theta)$ with its empirical analogue $\hat\nu_{\lambda,\lambda}(\hat\theta_\psi, \hat\theta)$, where

$$\hat\nu_{\lambda,\lambda}(\theta_1, \theta_2) = \sum_{i=1}^{n} \ell_\lambda^{(i)}(\theta_1)\ell_\lambda^{(i)}(\theta_2)^\top , \qquad (4)$$

with $\ell_\lambda^{(i)}(\theta) = \partial \log p_{_{Y_i}}(y_i; \psi, \lambda)/\partial\lambda$ (cf. Severini, 2000, Section 9.5.5). Note that $\hat\nu_{\lambda,\lambda}(\theta_1, \theta_2)$ is the empirical analogue of

$$\nu^0_{\lambda,\lambda}(\theta_1, \theta_2) = E_0(\ell_\lambda(\theta_1)\ell_\lambda(\theta_2)^\top) , \qquad (5)$$

useful to cope with possibly misspecified models.

In the setting of prediction, let us consider adjustments of the estimative predictive density $p_e(x;\psi) = p_X(x;\hat{\theta}_\psi)$. For curved exponential families and $h = 1$, Komaki (1996) obtains the optimal improvement over $p_e(x;\psi)$ in terms of average Kullback–Leibler divergence, up to and including terms of order $O(n^{-1})$. To give the expression of the resulting modified estimative density $p_K(x;\psi)$, index notation and Einstein summation convention are convenient. Generic components of λ will be denoted by λ_r, λ_s, …, with $r, s, \ldots = 1, \ldots, m$. Let $\ell(\psi,\lambda;x) = \log p_X(x;\psi,\lambda)$, $\ell_r(\psi,\lambda;x) = \partial \log p_X(x;\psi,\lambda)/\partial\lambda_r$ and $\ell_{rs}(\psi,\lambda;x) = \partial^2 \log p_X(x;\psi,\lambda)/(\partial\lambda_r\partial\lambda_s)$. Hence,

$$p_K(x;\psi) = p_e(x;\psi)\left[1 + \frac{1}{2}\left\{h_{rs}(\psi,\hat{\lambda}_\psi;x) - \Gamma^t_{rs}(\psi,\hat{\lambda}_\psi)\ell_t(\psi,\hat{\lambda}_\psi;x)\right\} i^{rs}(\psi,\hat{\lambda}_\psi)\right], \tag{6}$$

where

$$h_{rs}(\psi,\lambda;x) = \ell_{rs}(\psi,\lambda;x) + \ell_r(\psi,\lambda;x)\ell_s(\psi,\lambda;x)\,,$$
$$\Gamma^t_{rs}(\psi,\lambda) = i^{tu}(\psi,\lambda)\,E_{\psi,\lambda}\{h_{rs}(\psi,\lambda;X)\ell_u(\psi,\lambda;X)\}$$

and $i^{rs}(\psi,\lambda)$ denotes the generic element of the inverse matrix of $i_{\lambda\lambda}(\psi,\lambda)$. Corcuera and Giummolè (2000) show that (6) holds also for general regular models and extend Komaki's result to α-divergences. They consider $h \geq 1$ fixed and possibly dependent observations.

Let us consider a model selection problem where the generic model M_ψ has a parameter $\lambda_{(\psi)}$ of dimension m_ψ, $\psi \in \Psi = \{1, \ldots, k\}$. Here, $j_{\lambda\lambda}(\theta)$ denotes the observed information for $\lambda_{(\psi)}$ in model M_ψ. A general version of Takeuchi's information criterion (see Burnham and Anderson, 2002, formula (7.38)) is

$$TIC = 2\left[-\ell_P(\psi) + \mathrm{tr}\left\{j_{\lambda\lambda}(\hat{\theta}_\psi)^{-1}\hat{\nu}_{\lambda,\lambda}(\hat{\theta}_\psi,\hat{\theta}_\psi)\right\}\right],$$

with $\hat{\nu}_{\lambda,\lambda}(\hat{\theta}_\psi,\hat{\theta}_\psi)$ defined according to (4). The above criterion corresponds to the penalised profile loglikelihood

$$\ell_{TIC}(\psi) = \ell_P(\psi) - \mathrm{tr}\left\{j_{\lambda\lambda}(\hat{\theta}_\psi)^{-1}\hat{\nu}_{\lambda,\lambda}(\hat{\theta}_\psi,\hat{\theta}_\psi)\right\}. \tag{7}$$

In practice, the summand corresponding to trace term in (7) is often conveniently evaluated as m_ψ, and this gives rise to the well-known Akaike's information criterion

$$AIC = 2(-\ell_P(\psi) + m_\psi)\,,$$

corresponding to the penalised profile loglikelihood

$$\ell_{AIC}(\psi) = \ell_P(\psi) - m_\psi\,.$$

Note that there seem to be some relations between the ingredients in (7) and those in (3) when the quantity $\nu_{\lambda,\lambda}(\hat{\theta}_\psi,\hat{\theta};\hat{\theta})$ appearing in (3) is replaced by its empirical analogue $\hat{\nu}_{\lambda,\lambda}(\hat{\theta}_\psi,\hat{\theta})$. Moreover, the same ingredients appearing in the penalisation term of $\ell_{TIC}(\psi)$ appear in the well-known sandwich estimate of the covariance matrix of $\hat{\lambda}_\psi$ (Huber, 1967; White, 1982)

$$\hat{V}(\hat{\lambda}_\psi) = j_{\lambda\lambda}(\hat{\theta}_\psi)^{-1}\hat{\nu}_{\lambda,\lambda}(\hat{\theta}_\psi,\hat{\theta}_\psi)j_{\lambda\lambda}(\hat{\theta}_\psi)^{-1}\,. \tag{8}$$

4 Theoretical target criteria

In this section we present theoretical target criteria for modifications of the profile loglikelihood or of the estimative predictive density in the three different settings we are considering. As emerges from the review in Section 3, all such target criteria are related with Kullback–Leibler divergence. We recall that the Kullback–Leibler divergence between two possible densities, $p^0_{Z}(z)$ and $p^1_{Z}(z)$, for a random vector Z is given by

$$D\{p^0_{Z}(z); p^1_{Z}(z)\} = \int \log\left\{\frac{p^0_{Z}(z)}{p^1_{Z}(z)}\right\} p^0_{Z}(z)\, dz\,. \tag{9}$$

4.1 Likelihood theory

In Pace and Salvan (2005) a theoretical target criterion is introduced for adjustments of the profile loglikelihood. It is based on model restriction, amounting to calculating the original likelihood along a curve in the parameter space, parameterised by ψ. One must ensure that no unrealistic information regarding ψ is introduced by model restriction. This leads to the least favourable curve (Stein, 1956), along which the expected information for ψ is equal to the adjusted (or partial) information for ψ in the original model; see also Severini and Wong (1992). The least favourable curve minimises a Kullback–Leibler divergence. Let $\theta_0 = (\psi_0, \lambda_0)$ denote the true parameter value and let $\tilde{\lambda}_\psi$ be a function of ψ such that $\tilde{\lambda}_{\psi_0} = \lambda_0$. Then, the least favourable curve is $\theta_\psi = (\psi, \lambda_\psi)$, where λ_ψ minimises $D\{p(y;\theta_0); p(y;\psi,\tilde{\lambda}_\psi)\}$ among all curves $\tilde{\lambda}_\psi$. The function λ_ψ is also the maximiser with respect to λ for fixed ψ of

$$E_{\theta_0}(\ell(\psi,\lambda)) = -D\{p(y;\theta_0); p(y;\psi,\lambda)\} + c_0\,,$$

where $c_0 = E_{\theta_0}(\log p(Y;\theta_0))$, cf. Severini (2000, Section 4.8). Under regularity conditions, the constrained MLE of λ for a given value of ψ, i.e. $\hat{\lambda}_\psi$, is a consistent estimator of λ_ψ (Huber, 1967).

The least favourable target loglikelihood is defined as

$$\ell_{T}(\psi) = \ell(\theta_\psi)\,. \tag{10}$$

Although $\ell_{T}(\psi)$ is a genuine loglikelihood, it is not available in practice, because λ_ψ depends on θ_0. The profile loglikelihood $\ell_{P}(\psi)$ is an estimative, or plug-in, counterpart of $\ell_{T}(\psi)$ and has a bias of order $O(1)$ as an estimate of the theoretical target criterion

$$T^{lt}(\psi;\theta_0) = E_{\theta_0}(\ell_{T}(\psi)) = -D\{p(y;\theta_0); p(y;\theta_\psi)\} + c_0\,. \tag{11}$$

Note that ψ in $T^{lt}(\psi;\theta_0)$ is given while θ_0 is unknown.

4.2 Prediction

Consider the problem of prediction from the theoretical point of view of selecting the member of a given class of densities having minimum Kullback–Leibler divergence with respect to $p^0_{X}(x)$, the true unknown density of $X = (X_1, \ldots, X_h)$, where

$X_1, \ldots, X_h$ are independent and identically distributed. Such an optimal element will be called a target predictive density. It will depend on the true distribution of X and, therefore, it will not be directly available for prediction. An estimation step is thus required and subsequent bias adjustment is in order. When the given class of densities is correctly specified, that is when $p^0_X(x)$ belongs to it, this preliminary definition of a target predictive density is redundant because it gives $p^0_X(x)$ itself.

Here we consider in particular the situation where the class has elements $p_X(x; \psi, \lambda)$, where $\lambda \in \Lambda$ and ψ is fixed. We suppose in addition that $p^0_X(x) = p_X(x; \psi_0, \lambda_0)$, where usually $\psi_0 \neq \psi$. Then the target predictive density minimises

$$D(p_X(x; \psi_0, \lambda_0); p_X(x; \psi, \lambda))$$

with respect to λ with ψ fixed. The minimum is attained at λ_ψ as defined in Subsection 4.1 and the target predictive density is

$$p_X(x; \psi, \lambda_\psi) = p_X(x; \theta_\psi) \,.$$

Equivalently, $p_X(x; \theta_\psi)$ maximises, with respect to λ, $E^X_{\theta_0}\{\log p_X(X; \psi, \lambda)\}$ and the maximum is the theoretical target criterion

$$T^{pr}(\psi; \theta_0) = E^X_{\theta_0}\{\log p_X(X; \theta_\psi)\} \,. \tag{12}$$

Note that $T^{pr}(\psi; \theta_0)$ is proportional to $T^{lt}(\psi; \theta_0)$ given by (11) with proportionality constant h/n.

The estimative predictive density $p_e(x; \psi)$ is the plug-in counterpart of $p_X(x; \theta_\psi)$ and its logarithm has a bias of order $O(1)$ as an estimate of (12).

4.3 Model selection

Using the notation introduced in Section 2, candidate models are indexed by $\psi \in \Psi = \{1, \ldots, k\}$. A candidate model is fitted by maximizing $\ell(\psi, \lambda_{(\psi)})$ with ψ fixed, giving $\hat\lambda_\psi = \hat\lambda_\psi(y)$ as the MLE of $\lambda_{(\psi)}$. Let $\hat\theta_\psi = \hat\theta_\psi(y) = (\psi, \hat\lambda_\psi)$. The estimative predictive density function (2) for a future observation x of X, an independent copy of Y, based on the model with densities $p_X(x; \theta)$, is $p_X(x; \hat\theta_\psi)$.

Let $p^0_X(x)$ and $p^0_Y(y)$ denote the true unknown densities of X and Y. Akaike's and Takeuchi's information criteria aim at minimizing with respect to ψ the expected Kullback–Leibler divergence between $p^0_X(x)$ and $p_X(x; \hat\theta_\psi(Y))$, i.e.

$$\begin{aligned} ED_0(\psi) &= E^Y_0\left[D\left\{p^0_X(x); p_X(x; \hat\theta_\psi(Y))\right\}\right] \\ &= \int\int \log\left(\frac{p^0_X(x)}{p_X(x; \hat\theta_\psi(y))}\right) p^0_X(x) p^0_Y(y)\, dx\, dy \;, \end{aligned}$$

see e.g. Burnham and Anderson (2002, formula (7.19)). The rationale underlying the definition of $ED_0(\psi)$ is the same as in cross-validation. The estimate $\hat\theta_\psi = \hat\theta_\psi(y)$ is used to predict x, through $p_X(x; \hat\theta_\psi)$. Then, an independent replication is used to evaluate predictive fit. The asymptotic relationship between model selection by cross-validation and by Akaike's criterion is discussed in Stone (1977).

Minimising $ED_0(\psi)$ is equivalent to maximising the theoretical quantity

$$\begin{aligned} T_0^{ms}(\psi) &= E_0^Y \left\{ \int \log p_X(x; \hat{\theta}_\psi(Y)) p_X^0(x)\, dx \right\} \\ &= E_0^Y \left[E_0^X \left\{ \log p_X(X; \hat{\theta}_\psi(Y)) \right\} \right] = E_0^Y \left[E_0^X \left\{ \log p(X; \hat{\theta}_\psi(Y)) \right\} \right] . \end{aligned}$$

Indeed,

$$T_0^{ms}(\psi) = -ED_0(\psi) + c_0' ,$$

where $c_0' = E_0^Y \left(E_0^X (\log p_X^0(X)) \right) = E_0^X (\log p_X^0(X))$.

Assuming that the overall model M is correctly specified, $p_X^0(x)$ and $p_Y^0(y)$ are given by $p_X(x; \theta_0)$ and $p(y; \theta_0)$, respectively, where $\theta_0 = (\psi_0, \lambda_0)$, with $\lambda_0 = \lambda_{(\psi_0)}$. The expected Kullback–Leibler divergence between $p_X(x; \theta_0)$ and $p_X(x; \hat{\theta}_\psi(Y))$ becomes

$$\begin{aligned} ED_{\theta_0}(\psi) &= E_{\theta_0}^Y \left[D \left\{ p_X(x; \theta_0); p_X(x; \hat{\theta}_\psi(Y)) \right\} \right] \\ &= \int \int \log \left(\frac{p_X(x; \theta_0)}{p_X(x; \hat{\theta}_\psi(y))} \right) p_X(x; \theta_0) p(y; \theta_0)\, dx\, dy . \end{aligned}$$

Again, minimising $ED_{\theta_0}(\psi)$ is equivalent to maximising the theoretical target criterion

$$\begin{aligned} T^{ms}(\psi; \theta_0) &= E_{\theta_0}^Y \left\{ \int \log p_X(x; \hat{\theta}_\psi(Y)) p_X(x; \theta_0)\, dx \right\} \\ &= E_{\theta_0}^Y \left[E_{\theta_0}^X \left\{ \log p_X(X; \hat{\theta}_\psi(Y)) \right\} \right] = E_{\theta_0}^Y \left[E_{\theta_0}^X \left\{ \log p(X; \hat{\theta}_\psi(Y)) \right\} \right] \end{aligned} \quad (13)$$

The profile loglikelihood $\ell_P(\psi) = \log p(y; \psi, \hat{\lambda}_\psi)$ is a biased estimator of $T_0^{ms}(\psi)$ with bias of order $O(1)$. Analogously, if the overall model M is correctly specified, the profile loglikelihood has bias of order $O(1)$ as an estimator of $T^{ms}(\psi; \theta_0)$.

5 Bias adjustment of plug-in estimates of theoretical target criteria

In this section we will show that adjusted procedures recalled in Section 3 can be viewed as bias corrections of plug-in estimates of the corresponding theoretical target criteria. Moreover, we will discuss some asymptotic relations among adjusted estimative criteria arising in the three different settings.

5.1 Biases of the profile loglikelihood when estimating theoretical target criteria

The bias under θ_0 of $\ell_P(\psi)$ as an estimate of $T^{lt}(\psi; \theta_0)$ is

$$b^{lt}(\psi; \theta_0) = E_{\theta_0}(\ell_P(\psi)) - T^{lt}(\psi; \theta_0) . \quad (14)$$

Under the assumptions of standard likelihood asymptotics, in Pace and Salvan (2005) various asymptotically equivalent approximations are obtained for $b^{lt}(\psi;\theta_0)$. In particular,

$$\begin{aligned} b^{lt}(\psi;\theta_0) &= b^{lt}_I(\psi;\theta_0) + O(n^{-1}) \\ &= b^{lt}_{II}(\psi;\theta_0) + O(n^{-1}) \ , \end{aligned}$$

where

$$b^{lt}_I(\psi;\theta_0) = \frac{1}{2}\mathrm{tr}\{i_{\lambda\lambda}(\theta_\psi;\theta_0)V_{\theta_0}(\hat\lambda_\psi)\}\,, \tag{15}$$

$$b^{lt}_{II}(\psi;\theta_0) = \frac{1}{2}\mathrm{tr}\{i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}\nu_{\lambda,\lambda}(\theta_\psi,\theta_\psi;\theta_0)\}\,, \tag{16}$$

with $i_{\lambda\lambda}(\theta_\psi;\theta_0) = E_{\theta_0}(j_{\lambda\lambda}(\theta_\psi))$. Moreover, for $\psi-\psi_0 = O(n^{-1/2})$,

$$b^{lt}(\psi;\theta_0) = b^{lt}_{III}(\psi;\theta_0) + O(n^{-1})\,,$$

with

$$b^{lt}_{III}(\psi;\theta_0) = \frac{1}{2}\mathrm{tr}\{i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}\nu_{\lambda,\lambda}(\theta_\psi,\theta_0;\theta_0)i_{\lambda\lambda}(\theta_0)^{-1}\nu_{\lambda,\lambda}(\theta_\psi,\theta_0;\theta_0)\}\,. \tag{17}$$

In prediction, the bias of $\log p_e(x;\psi)$ as an estimate of $T^{pr}(\psi;\theta_0)$,

$$b^{pr}(\psi;\theta_0) = E_{\theta_0}(\log p_e(x;\psi)) - T^{pr}(\psi;\theta_0)\,,$$

is

$$b^{pr}(\psi;\theta_0) = \frac{h}{n}b^{lt}(\psi;\theta_0)\,.$$

Finally, in model selection, the bias of $\ell_P(\psi)$ as an estimate of $T^{ms}_0(\psi)$ is

$$b^{ms}_0(\psi) = E_0(\ell_P(\psi)) - T^{ms}_0(\psi)\ . \tag{18}$$

Two key relations in the model selection literature (see e.g. Burnham and Anderson, 2002, pages 369-370) are

$$T^{ms}_0(\psi) = E_0(\ell(\psi,\lambda^*_\psi)) - \frac{1}{2}\mathrm{tr}\left[\nu^0_{\lambda,\lambda}(\theta^*_\psi,\theta^*_\psi)\{E_0(j_{\lambda\lambda}(\theta^*_\psi))\}^{-1}\right] + O(n^{-1}) \tag{19}$$

and

$$E_0(\ell_P(\psi)) = E_0(\ell(\psi,\lambda^*_\psi)) + \frac{1}{2}\mathrm{tr}\left[\nu^0_{\lambda,\lambda}(\theta^*_\psi,\theta^*_\psi)\{E_0(j_{\lambda\lambda}(\theta^*_\psi))\}^{-1}\right] + O(n^{-1})\,, \tag{20}$$

with $\nu^0_{\lambda,\lambda}(\theta_1,\theta_2)$ defined as in (5) and θ^*_ψ given by

$$\theta^*_\psi = (\psi,\lambda^*_\psi) \tag{21}$$

where λ^*_ψ is the value of λ that minimises the Kullback-Leibler divergence between the true density of Y and $p_Y(y;\psi,\lambda)$.

Using (19) and (20) we get

$$b^{ms}_0(\psi) = \mathrm{tr}\left[\nu^0_{\lambda,\lambda}(\theta^*_\psi,\theta^*_\psi)\{E_0(j_{\lambda\lambda}(\theta^*_\psi))\}^{-1}\right] + O(n^{-1})\,. \tag{22}$$

When the overall model M is correctly specified, the bias under θ_0 of $\ell_P(\psi)$ as an estimate of $T^{ms}(\psi;\theta_0)$ is

$$b^{ms}(\psi;\theta_0) = E_{\theta_0}(\ell_P(\psi)) - T^{ms}(\psi;\theta_0)\ . \tag{23}$$

Relation (19) becomes

$$\begin{aligned} T^{ms}(\psi;\theta_0) &= E_{\theta_0}(\ell(\psi,\lambda_\psi)) - \frac{1}{2}\mathrm{tr}\left[\nu_{\lambda,\lambda}(\theta_\psi,\theta_\psi;\theta_0) i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}\right] + O(n^{-1}) \\ &= T^{lt}(\psi;\theta_0) - b^{lt}_{II}(\psi;\theta_0) + O(n^{-1}) \qquad (24) \end{aligned}$$

so that

$$\begin{aligned} -T^{ms}(\psi;\theta_0) &= -T^{lt}(\psi;\theta_0) + b^{lt}_{II}(\psi;\theta_0) + O(n^{-1}) \\ E_{\theta_0}(\ell_P(\psi)) - T^{ms}(\psi;\theta_0) &= E_{\theta_0}(\ell_P(\psi)) - T^{lt}(\psi;\theta_0) + b^{lt}_{II}(\psi;\theta_0) + O(n^{-1}) \\ b^{ms}(\psi;\theta_0) &= 2b^{lt}_{II}(\psi;\theta_0) + O(n^{-1})\,. \end{aligned}$$

Therefore,

$$b^{ms}(\psi;\theta_0) = 2b^{lt}(\psi;\theta_0) + O(n^{-1})\,. \tag{25}$$

5.2 Bias adjustments

Adjusted procedures of Section 3 can now be easily recognised as obtained by correcting crude plug-in estimates using a plug-in estimate of the relevant bias term.

In likelihood theory, consistent estimates of $b^I(\psi;\theta_0)$, $b^{II}(\psi;\theta_0)$ and $b^{III}(\psi;\theta_0)$ are defined by replacing $i_{\lambda\lambda}(\theta_\psi;\theta_0)$ with $j_{\lambda\lambda}(\theta_\psi)$, θ_ψ with $\hat\theta_\psi$ and θ_0 with $\hat\theta$. The following adjustments of the profile loglikelihood are then obtained:

$$\begin{aligned} \ell^I_{AP}(\psi) &= \ell_P(\psi) - \frac{1}{2}\mathrm{tr}\{j_{\lambda\lambda}(\hat\theta_\psi)V_{\hat\theta}(\hat\lambda_\psi)\} \qquad (26) \\ \ell^{II}_{AP}(\psi) &= \ell_P(\psi) - \frac{1}{2}\mathrm{tr}\{j_{\lambda\lambda}(\hat\theta_\psi)^{-1}\nu_{\lambda,\lambda}(\hat\theta_\psi,\hat\theta_\psi;\hat\theta)\}\ , \qquad (27) \\ \ell^{III}_{AP}(\psi) &= \ell_P(\psi) - \frac{1}{2}\mathrm{tr}\{j_{\lambda\lambda}(\hat\theta_\psi)^{-1}\nu_{\lambda,\lambda}(\hat\theta_\psi,\hat\theta_\psi;\hat\theta)j_{\lambda\lambda}(\hat\theta)^{-1}\nu_{\lambda,\lambda}(\hat\theta_\psi,\hat\theta_\psi;\hat\theta)\}\ . \qquad (28) \end{aligned}$$

Relations of the above versions of $\ell_{AE}(\psi)$ to other adjustments of the profile loglikelihood are discussed in Pace and Salvan (2005). In particular, under the assumptions of standard likelihood theory, for $\psi - \hat\psi = 0(n^{-1/2})$, locally equivalent versions of $\ell^I_{AE}(\psi)$, $\ell^{II}_{AE}(\psi)$ and $\ell^{III}_{AE}(\psi)$ are, respectively,

$$\begin{aligned} \ell^I_M(\psi) &= \ell_P(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\hat\theta_\psi)| - \frac{1}{2}\log|V_{\hat\theta}(\hat\lambda_\psi)|\,, \\ \ell^{II}_M(\psi) &= \ell_P(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat\theta_\psi)| - \frac{1}{2}\log|\nu_{\lambda,\lambda}(\hat\theta_\psi,\hat\theta_\psi;\hat\theta)|\,, \\ \ell^{III}_M(\psi) &= \ell_P(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat\theta_\psi)| - \log|\nu_{\lambda,\lambda}(\hat\theta_\psi,\hat\theta;\hat\theta)|\,, \end{aligned}$$

where $\ell^{III}_M(\psi)$ coincides with (3).

In prediction, a bias corrected version of $\log p_e(x;\psi)$ as an estimate of $T^{pr}(\psi;\theta_0)$ is obtained by Komaki (1996) and by Corcuera and Giummolé (2000), see formula (6).

Finally, in model selection, (7) is obtained by subtracting from $\ell_{P}(\psi)$ the estimated bias given by (22) where θ^*_ψ is replaced with $\hat\theta_\psi$, $\nu^0_{\lambda,\lambda}(\theta^*_\psi,\theta^*_\psi)$ with $\hat\nu_{\lambda,\lambda}(\hat\theta_\psi,\hat\theta_\psi)$ and $E_0(j_{\lambda\lambda}(\theta^*_\psi))$ with $j_{\lambda\lambda}(\hat\theta_\psi)$.

When the overall model M is correctly specified, in view of (25), we obtain the following version of $\ell_{TIC}(\psi)$

$$\ell^{cs}_{TIC}(\psi) = \ell_{P}(\psi) - \mathrm{tr}\left\{ j_{\lambda\lambda}(\hat\theta_\psi)^{-1}\nu_{\lambda,\lambda}(\hat\theta_\psi,\hat\theta_\psi;\hat\theta)\right\}\,. \tag{29}$$

We note that (29) differs from (27) by a factor 1/2 affecting the correction term.

6 Adjusted profile loglikelihood from an optimal predictive density

In a model with parameter θ, a predictive density calculated at x is proportional to the estimated probability of observing the future data in a neighbourhood of x in the light of data y observed from a model with the same unknown θ. On the other hand, a likelihood function calculated at θ for data y is proportional to the probability under θ of observing future data, generated as y, in a neighbourhood of the observed y. Similarly, when a nuisance component is present in θ, we conjecture that a (pseudo-)likelihood function for an interest parameter ψ should be related to a predictive density calculated at the observed y in a submodel having parameter λ with densities $p(y;\psi,\lambda)$.

Example 1: Random sampling from a normal distribution.
Let us consider $y=(y_1,\ldots,y_n)$ as a random sample from a normal distribution with unknown mean μ and fixed variance σ^2. Let $\bar Y_n$ be the sample mean. Let X_1 be an independent future observation from the same distribution. Based on the exact pivot $X_1-\bar Y_n$, the predictive density of X_1 is

$$\hat p(x_1;y,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma\sqrt{1+n^{-1}}}\exp\left\{-\frac{1}{2}\frac{(x_1-\bar y_n)^2}{\sigma^2(1+n^{-1})}\right\}\,,$$

i.e. normal with mean $\bar y_n$ and variance $\sigma^2(1+n^{-1})$.

We argue that this predictive density can be exploited to obtain an agreed pseudo-loglikelihood for σ^2, such as the modified profile loglikelihood. Note that, both in prediction and in the likelihood context, elimination of the nuisance parameter μ is required. We recall that the modified profile loglikelihood for σ^2 is

$$\ell_{M}(\sigma^2;y) = \ell_{P}(\sigma^2;y) + \frac{1}{2}\log\sigma^2\,,$$

with

$$\ell_{P}(\sigma^2;y) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\frac{\hat\sigma^2_y}{\sigma^2}\,,$$

where $\hat{\sigma}^2_y$ is the MLE of σ^2 based on y. Let $\theta_0 = (\mu_0, \sigma^2_0)$. Then for $\sigma^2 - \sigma^2_0 = O(n^{-1/2})$,

$$\ell_{\scriptscriptstyle M}(\sigma^2; y) = \ell_{\scriptscriptstyle P}(\sigma^2; y) - \frac{1}{2}\frac{\hat{\sigma}^2_y}{\sigma^2} + O_p(n^{-1/2})\,.$$

The predictive density of n independent copies of X_1, i.e. of the random sample $X = (X_1, \ldots, X_n)$, is

$$\begin{aligned} \hat{p}(x; y, \sigma^2) &= \prod_{i=1}^n \hat{p}(x_i; y, \sigma^2) \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2}(1+n^{-1})^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2(1+n^{-1})} \sum_{i=1}^n (x_i - \bar{y}_n)^2 \right\}. \end{aligned}$$

Hence,

$$\log \hat{p}(x; y, \sigma^2) = c - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2(1+n^{-1})}\sum_{i=1}^n (x_i - \bar{x}_n)^2 - \frac{1}{2\sigma^2(1+n^{-1})} n(\bar{y}_n - \bar{x}_n)^2\,.$$

Let $\hat{\sigma}^2_x = n^{-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. Then,

$$\log \hat{p}(X; Y, \sigma^2) = c - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\frac{\hat{\sigma}^2_x}{\sigma^2} + \frac{1}{2}\frac{\hat{\sigma}^2_x}{\sigma^2} - \frac{1}{2\sigma^2} n(\bar{Y}_n - \bar{X}_n)^2 + O_p(n^{-1})\,.$$

Under θ_0 the quantity $n(\bar{Y}_n - \bar{X}_n)^2$ has the same distribution as $2\sigma^2_0 W$, where W is a chi-square on one degree of freedom. Therefore,

$$\log \hat{p}(X; Y, \sigma^2) = c - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\frac{\hat{\sigma}^2_x}{\sigma^2} + \frac{1}{2}\frac{\hat{\sigma}^2_x}{\sigma^2} - \frac{\sigma^2_0}{\sigma^2} W + O_p(n^{-1})\,.$$

Moreover, using $E_{\theta_0}(\sigma^2_0 W) = E_{\theta_0}(\hat{\sigma}^2_x) + O(n^{-1})$,

$$\begin{aligned} E_{\theta_0}(\log \hat{p}(X; Y, \sigma^2)) &= E_{\theta_0}\left(c - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\frac{\hat{\sigma}^2_x}{\sigma^2} + \frac{1}{2}\frac{\hat{\sigma}^2_x}{\sigma^2} \right) - \frac{E_{\theta_0}(\hat{\sigma}^2_x)}{\sigma^2} + O(n^{-1}) \\ &= E_{\theta_0}\left(c - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\frac{\hat{\sigma}^2_x}{\sigma^2} - \frac{1}{2}\frac{\hat{\sigma}^2_x}{\sigma^2} \right) + O(n^{-1}) \\ &= E_{\theta_0}(\ell_{\scriptscriptstyle M}(\sigma^2; X)) + O(n^{-1}) \\ &= E_{\theta_0}(\ell_{\scriptscriptstyle M}(\sigma^2; Y)) + O(n^{-1})\,. \end{aligned}$$

The last identity follows from the fact that Y is a copy of X. This shows that the pseudo-loglikelihood $\ell_{\scriptscriptstyle M}(\sigma^2) = \ell_{\scriptscriptstyle M}(\sigma^2; y)$ for σ^2 has the desired likelihood interpretation. In more detail, $\ell_{\scriptscriptstyle M}(\sigma^2)$ represents, up to constants, the log-probability, for the tentative value σ^2 of the parameter of interest, of observing future data, generated as y, in a neighbourhood of the observed y.

Example 2: Random sampling from a gamma distribution.
Let us consider $y = (y_1, \ldots, y_n)$ as a random sample from a gamma distribution with

unknown scale parameter λ and fixed shape parameter ψ. Let X_1 be an independent future observation from the same distribution, i.e. with density

$$p(x_1;\psi,\lambda)=\frac{1}{\Gamma(\psi)}\lambda^{\psi}x_1^{\psi-1}\exp\{-\lambda x_1\}\,,\quad x_1>0\,.$$

The MLE of λ with ψ fixed is $\hat\lambda_\psi=\hat\lambda_\psi(y)=\psi/\bar y_n$. Based on the exact pivot $\hat\lambda_\psi X_1$, the predictive density of X_1 may be expanded as (Vidoni, 1995, Example 4.2)

$$\hat p(x_1;y,\psi)=p(x_1;\psi,\hat\lambda_\psi)\left[1+\frac{1}{2n\psi}\left\{\hat\lambda_\psi^2x_1^2-2\hat\lambda_\psi(\psi+1)x_1+\psi(\psi+1)\right\}\right]\,.$$

The modified profile loglikelihood for ψ is

$$\ell_M(\psi;y)=\ell_P(\psi;y)-\frac{1}{2}\log\psi\,,$$

where

$$\ell_P(\psi;y)=(\psi-1)\sum\log y_i-n\psi+n\psi\log\psi-n\psi\log\bar y_n-n\log\Gamma(\psi)\,.$$

Let $\theta_0=(\psi_0,\lambda_0)$. Then for $\psi-\psi_0=O(n^{-1/2})$, and neglecting constants,

$$\ell_M(\psi;y)=\ell_P(\psi;y)-\frac{1}{2}\frac{\psi}{\hat\psi}+O(n^{-1/2})\,,$$

where $\hat\psi=\hat\psi(y)$. The predictive density of n independent copies of X_1, i.e. of the random sample $X=(X_1,\ldots,X_n)$, is

$$\hat p(x;y,\psi)=\prod_{i=1}^n\hat p(x_i;y,\psi)\,.$$

Consequently,

$$\begin{aligned}\log\hat p(X;Y,\psi)&=(\psi-1)\sum\log X_i-n\log\Gamma(\psi)-\hat\lambda_\psi(Y)\sum X_i+n\psi\log\hat\lambda_\psi(Y)\\&\quad+\sum\frac{1}{2n\psi}\left\{\hat\lambda_\psi^2(Y)X_i^2-2\hat\lambda_\psi(Y)(\psi+1)X_i+\psi(\psi+1)\right\}+O_p(n^{-1})\\&=\ell_P(\psi;X)-n\psi(F-1)+n\psi\log F\\&\quad+\sum\frac{1}{2n\psi}\left\{\hat\lambda_\psi^2(Y)X_i^2-2\hat\lambda_\psi(Y)(\psi+1)X_i+\psi(\psi+1)\right\}+O_p(n^{-1})\,,\end{aligned}$$

where $F=\hat\lambda_\psi(Y)/\hat\lambda_\psi(X)$ under $\theta_0=(\psi_0,\lambda_0)$ is distributed as the ratio of two independent gamma variates with common shape parameter $n\psi_0$ and common unit scale. As a consequence, $E_{\theta_0}(F)=1+(n\psi_0)^{-1}+O(n^{-2})$ and $E_{\theta_0}(\log F)=O(n^{-2})$. Moreover, under θ_0, we have $\hat\lambda_\psi(Y)=\lambda_\psi+O_p(n^{-1/2})=\psi\lambda_0/\psi_0+O_p(n^{-1/2})$, $\sum X_i/n=\psi_0/\lambda_0+O_p(n^{-1/2})$ and $\sum X_i^2/n=\psi_0(\psi_0+1)/\lambda_0^2+O_p(n^{-1/2})$.

Hence, under θ_0,

$$\log\hat p(X;Y,\psi)=\ell_P(\psi;X)-n\psi(F-1)+n\psi\log F+\frac{\psi}{2\psi_0}-\frac{1}{2}+O_p(n^{-1/2})$$

so that, neglecting additive constants,

$$\begin{aligned}
E_{\theta_0}(\log \hat{p}(X;Y,\psi)) &= E_{\theta_0}(\ell_P(\psi;X)) - \frac{\psi}{2\psi_0} + O(n^{-1}) \\
&= E_{\theta_0}\left(\ell_P(\psi;X) - \frac{\psi}{2\hat{\psi}}\right) + O(n^{-1}) \\
&= E_{\theta_0}(\ell_M(\psi;X)) + O(n^{-1}) \\
&= E_{\theta_0}(\ell_M(\psi;Y)) + O(n^{-1})\,.
\end{aligned}$$

The result suggested by the two examples above carries over in wide generality. Let us consider prediction of X based on a random sample $y = (y_1, \ldots, y_n)$ from $Y = (Y_1, \ldots, Y_n)$. We suppose that X is independent of Y and has n independent and identically distributed components with X_1 having the same distribution as Y_1, with density $p(x_1;\psi,\lambda)$. As before, we treat ψ as known. In the following the MLE of λ with ψ fixed is $\hat{\lambda}_\psi = \hat{\lambda}_\psi(y)$.

A predictive density of n independent copies of X_1, i.e. a joint predictive density of X, factorises as

$$\hat{p}(x;y,\psi) = \prod_{i=1}^{n} \hat{p}(x_i;y,\psi)\,,$$

where $\hat{p}(x_i;y,\psi)$ is a predictive density for X_i.

If expression (6) is used for $\hat{p}(x_i;y,\psi)$, we get

$$\begin{aligned}
\log \hat{p}(X;Y,\psi) &= \sum_{i=1}^{n} \log p_e(X_i;\psi) \\
&\quad + \sum_{i=1}^{n} \log \left[1 + \frac{1}{2}\left\{h_{rs}(\psi,\hat{\lambda}_\psi(Y);X_i) - \Gamma_{rs}^{t}(\psi,\hat{\lambda}_\psi(Y))\ell_t(\psi,\hat{\lambda}_\psi(Y);X_i)\right\} i^{rs}(\psi,\hat{\lambda}_\psi(Y \right. \\
&= \sum_{i=1}^{n} \log p(X_i;\psi,\hat{\lambda}_\psi(Y)) \\
&\quad + \frac{1}{2} i^{rs}(\psi,\hat{\lambda}_\psi(Y)) \sum_{i=1}^{n} \left\{h_{rs}(\psi,\hat{\lambda}_\psi(Y);X_i) - \Gamma_{rs}^{t}(\psi,\hat{\lambda}_\psi(Y))\ell_t(\psi,\hat{\lambda}_\psi(Y);X_i)\right\} + O_p(n
\end{aligned}$$

Let us consider first the expansion of $\sum_{i=1}^{n} \log p(X_i;\psi,\hat{\lambda}_\psi(Y))$ as a function of $\hat{\lambda}_\psi(Y)$ around $\hat{\lambda}_\psi(X)$. We obtain

$$\begin{aligned}
\sum_{i=1}^{n} \log p(X_i;\psi,\hat{\lambda}_\psi(Y)) &= \sum_{i=1}^{n} \log p(X_i;\psi,\hat{\lambda}_\psi(X)) \\
&\quad + \left(\hat{\lambda}_\psi(Y) - \hat{\lambda}_\psi(X)\right)_r \sum_{i=1}^{n} \ell_r(\psi,\hat{\lambda}_\psi(X);X_i) \\
&\quad + \frac{1}{2}\left(\hat{\lambda}_\psi(Y) - \hat{\lambda}_\psi(X)\right)_{rs} \sum_{i=1}^{n} \ell_{rs}(\psi,\hat{\lambda}_\psi(X);X_i) + O_p(n^{-1/2})\,,
\end{aligned}$$

where $\left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)_{rs} = \left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)_r \left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)_s$. Above, the first summand on the right-hand side is the profile loglikelihood for ψ based on X. The second summand vanishes because it involves the likelihood equation for λ with ψ fixed. Hence,

$$\sum_{i=1}^n \log p(X_i;\psi,\hat\lambda_\psi(Y)) = \ell_{P}(\psi;X) - \frac{1}{2}\left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)_{rs} j_{rs}(\psi,\hat\lambda_\psi(X);X) + O_p(n^{-1/2}),$$

where $j_{rs}(\psi,\lambda;x) = -\sum_{i=1}^n \ell_{rs}(\psi,\lambda;x_i)$. Consider now that, using results in the Appendix of Pace and Salvan (2005),

$$\begin{aligned}\left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)_r &= \left(\hat\lambda_\psi(Y)-\lambda_\psi-\hat\lambda_\psi(X)+\lambda_\psi\right)_r\\ &= i^{rt}(\theta_\psi;\theta_0)\ell_t(\theta_\psi;Y) - i^{rt}(\theta_\psi;\theta_0)\ell_t(\theta_\psi;X) + O_p(n^{-1})\\ &= i^{rt}(\theta_\psi;\theta_0)\left\{\ell_t(\theta_\psi;Y)-\ell_t(\theta_\psi;X)\right\} + O_p(n^{-1}),\end{aligned}$$

while

$$\begin{aligned}\left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)_{rs} &= i^{rt}(\theta_\psi;\theta_0)\left\{\ell_t(\theta_\psi;Y)-\ell_t(\theta_\psi;X)\right\} i^{su}(\theta_\psi;\theta_0)\left\{\ell_u(\theta_\psi;Y)-\ell_u(\theta_\psi;X)\right\}\\ &\quad +O_p(n^{-3/2})\\ &= i^{rt}(\theta_\psi;\theta_0)i^{su}(\theta_\psi;\theta_0)\\ &\quad \{\ell_t(\theta_\psi;Y)\ell_u(\theta_\psi;Y)-\ell_t(\theta_\psi;Y)\ell_u(\theta_\psi;X)\\ &\quad -\ell_t(\theta_\psi;X)\ell_u(\theta_\psi;Y)+\ell_t(\theta_\psi;X)\ell_u(\theta_\psi;X)\} + O_p(n^{-3/2}).\end{aligned}$$

Therefore, with $\theta_0=(\psi_0,\lambda_0)$,

$$E_{\theta_0}\left[\left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)_{rs}\right] = 2i^{rt}(\theta_\psi;\theta_0)i^{su}(\theta_\psi;\theta_0)\nu_{t,u}(\theta_\psi,\theta_\psi;\theta_0) + O(n^{-2}),$$

so that

$$E_{\theta_0}\left[\left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)\left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)^\top\right] = 2i^{\lambda\lambda}(\theta_\psi;\theta_0)\nu_{\lambda,\lambda}(\theta_\psi,\theta_\psi;\theta_0)i^{\lambda\lambda}(\theta_\psi;\theta_0) + O(n^{-2}),$$

where the leading term is estimated by twice the sandwich estimate.

Moreover,

$$j_{rs}(\psi,\hat\lambda_\psi(X);X) = i_{rs}(\theta_\psi;\theta_0) + O_p(n^{1/2}).$$

Hence,

$$\begin{aligned}E_{\theta_0}\left\{\sum_{i=1}^n \log p(X_i;\psi,\hat\lambda_\psi(Y))\right\} &= E_{\theta_0}(\ell_{P}(\psi;X))\\ &\quad -i_{rs}(\theta_\psi;\theta_0)i^{rt}(\theta_\psi;\theta_0)i^{su}(\theta_\psi;\theta_0)\nu_{t,u}(\theta_\psi,\theta_\psi;\theta_0) + O(n^{-1})\\ &= E_{\theta_0}(\ell_{P}(\psi;X)) - i^{rs}(\theta_\psi;\theta_0)\nu_{r,s}(\theta_\psi,\theta_\psi;\theta_0) + O(n^{-1})\\ &= E_{\theta_0}(\ell_{P}(\psi;X)) - \mathrm{tr}\left[\nu_{\lambda,\lambda}(\theta_\psi,\theta_\psi;\theta_0)i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}\right] + O(n^{-1}).\end{aligned}$$

Note that

$$E_{\theta_0}\left\{\sum_{i=1}^n \log p(X_i;\psi,\hat\lambda_\psi(Y))\right\} = E_{\theta_0}(\ell^{cs}_{TIC}(\psi;X)) + O(n^{-1}).$$

On the other hand,

$$\begin{aligned}
& i^{rs}(\psi,\hat{\lambda}_\psi(Y))\sum_{i=1}^{n}\Big\{h_{rs}(\psi,\hat{\lambda}_\psi(Y);X_i)-\Gamma^t_{rs}(\psi,\hat{\lambda}_\psi(Y))\ell_t(\psi,\hat{\lambda}_\psi(Y);X_i)\Big\}\\
= \; & i^{rs}(\psi,\hat{\lambda}_\psi(X))\sum_{i=1}^{n}\Big\{h_{rs}(\psi,\hat{\lambda}_\psi(X);X_i)-\Gamma^t_{rs}(\psi,\hat{\lambda}_\psi(X))\ell_t(\psi,\hat{\lambda}_\psi(X);X_i)\Big\}+O_p(n^{-1/2})\\
= \; & i^{rs}(\psi,\hat{\lambda}_\psi(X))\left\{\sum_{i=1}^{n}l_{rs}(\psi,\hat{\lambda}_\psi(X);X_i)+\sum_{i=1}^{n}l_r(\psi,\hat{\lambda}_\psi(X);X_i)l_s(\psi,\hat{\lambda}_\psi(X);X_i)\right.\\
& \left.-\Gamma^t_{rs}(\psi,\hat{\lambda}_\psi(X))\sum_{i=1}^{n}\ell_t(\psi,\hat{\lambda}_\psi(X);X_i)\right\}+O_p(n^{-1/2})\\
= \; & i^{rs}(\psi,\hat{\lambda}_\psi(X))\Big\{-j_{rs}(\psi,\hat{\lambda}_\psi(X);X)+\hat{\nu}_{r,s}(\hat{\theta}_\psi(X),\hat{\theta}_\psi(X))\Big\}+O_p(n^{-1/2})\\
= \; & \{i^{rs}(\theta_\psi;\theta_0)+O(1)\}\Big\{-i_{rs}(\theta_\psi;\theta_0)+O(n^{1/2})+\nu_{r,s}(\theta_\psi,\theta_\psi;\theta_0)+O(n^{1/2})\Big\}
\end{aligned}$$

so that, neglecting additive constants,

$$\begin{aligned}
E_{\theta_0}\{\log\hat{p}(X;Y,\psi)\} &= E_{\theta_0}(\ell_P(\psi;X))-\mathrm{tr}\left[\nu_{\lambda,\lambda}(\theta_\psi,\theta_\psi;\theta_0)i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}\right]\\
&\quad+\frac{1}{2}\mathrm{tr}\left[\nu_{\lambda,\lambda}(\theta_\psi,\theta_\psi;\theta_0)i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}\right]+O(n^{-1}\\
&= E_{\theta_0}\{\ell_M(\psi;X)\}+O(n^{-1})\\
&= E_{\theta_0}\{\ell_M(\psi;Y)\}+O(n^{-1})\,.
\end{aligned}$$

7 Simulation results

7.1 Adjstments of profile likelihood based on model selection criteria

For model with nuisance parameters, inference about a scalar parameter of interest ψ is often based on a pseudolikelihood function. In particular, the profile and the modified profile loglikelihood can be used. There seems to be another route for obtaining likelihood inference about an interest parameter, motivated in the model selection approach. In particular, as pointed out in the previuos sections, a modified profile likelihood can be defined through Akaike's information criterion. Akaike's information criterion states that we adopt as final estimate the one which will give the maximum expected loglikelihood, or the maximum of a suitable modification of the profile loglikelihood function. In a similar manner, a modified profile likelihood can be defined starting from model selection procedures based on simple cross-validation techniques and on suitable predictive densities.

In this Section we discuss two examples. In particular, we perform Monte Carlo experiments (based on 10,000 trials) with the aim of evaluating the finite-sample properties of the estimators based on $\ell^{cs}_{TIC}(\psi)$, in comparison with MLE and the estimators based on the modified profile likelihood. The estimators are compared in terms of the usual centering and dispersion measures, i.e. bias (BI) and mean

		$\hat\psi_P$	$\hat\psi_{AP}$	$\hat\psi^*$	$\hat\psi^*_A$	$\hat\psi_{TIC}$
$n = 5$	bias	1.340	0.875	0.617	0.563	0.407
	sd	3.407	2.652	2.246	2.161	1.921
	PU	0.292	0.405	0.467	0.494	0.559
$n = 10$	bias	0.428	0.285	0.198	0.189	0.142
	sd	1.003	0.865	0.786	0.779	0.739
	PU	0.368	0.452	0.502	0.507	0.534
$n = 20$	bias	0.180	0.121	0.083	0.081	0.062
	sd	0.447	0.407	0.384	0.383	0.373
	PU	0.376	0.440	0.482	0.483	0.509
$n = 50$	bias	0.069	0.048	0.034	0.034	0.027
	sd	0.023	0.221	0.215	0.215	0.213
	PU	0.426	0.462	0.493	0.493	0.504
$n = 100$	bias	0.029	0.019	0.012	0.012	0.010
	sd	0.154	0.151	0.149	0.149	0.148
	PU	0.472	0.501	0.501	0.501	0.502

Table 1: Comparison for the parameter ψ of the inverse Gaussian distribution.

square error (MSE), which depend on the parameterisation of the model, and also in terms of the probability of underestimation (PU), that gives median bias. The first example refers to a simple situation, where only one nuisance parameter is present. The second example refers to situations where the dimension of the nuisance parameter is large.

Example 1. Consider independent observations y_i, $i = 1, \ldots, n$, having an inverse Gaussian distribution. The loglikelihood function is

$$\ell(\psi, \lambda) = \frac{n}{2} \log \psi - \frac{\psi}{2\lambda^2} \sum_{i=1}^{n} \frac{(y_i - \lambda)^2}{y_i} \ ,$$

where $\lambda > 0$, $\psi > 0$. Let ψ be the parameter of interest. In this case, all the quantities involved in the modified profile likelihoods for ψ (see e.g. Severini, 2000, p. 335) are easy to compute.

Table 1 presents the results of a simulation study including in the comparison the MLE, the estimator based on the modified profile likelihood, i.e. $\hat\psi_{AP}$, the estimator based on the modified directed likelihood $r^*(\psi) = 0$ and its approximate closed form expression (see Giummolè and Ventura, 2002), and the estimator based on $\ell^{cs}_{TIC}(\psi)$, i.e. $\hat\psi_{TIC}$. It can be noted that, in this situation with a scalar nuisance parameter, the estimator $\hat\psi_{TIC}$ behaves surprisingly well, compared to the estimator based on the modified profile likelihood and to $\hat\psi^*$ and $\hat\psi^*_A$. Also different reparameterisations where considered in the simulation studies, and similar results to those given in Table 1 where found.

Referring to the construction of confidence regions, we perform a Monte Carlo experiment whose objective is to evaluate the accuracy of the confidence regions.

	$1-\alpha$	W_P	W_{AP}	r^*	W_{TIC}
$n=5$	0.90	0.819	0.886	0.897	0.907
	0.95	0.891	0.938	0.949	0.959
	0.99	0.968	0.989	0.991	0.991
$n=10$	0.90	0.869	0.895	0.903	0.904
	0.95	0.926	0.949	0.952	0.955
	0.99	0.981	0.989	0.990	0.991
$n=20$	0.90	0.875	0.896	0.899	0.903
	0.95	0.941	0.948	0.949	0.948
	0.99	0.986	0.988	0.990	0.991
$n=50$	0.90	0.892	0.900	0.900	0.900
	0.95	0.947	0.951	0.949	0.948
	0.99	0.987	0.989	0.990	0.989

Table 2: Comparison for the parameter ψ of the inverse Gaussian distribution.

Table 2 shows the results of a Monte Carlo experiment (based on 10,000 trials) that compares confidence intervals for ψ based on the loglikelihood ratio statistics derived from the profile loglikelihood, $W_P(\psi)$, from the modified profile loglikelihood, $W_{AP}(\psi)$, from the Akaike modified loglikelihood, $W_{TIC}(\psi)$, and on the modified directed likelihood, $r^*(\psi)$. ¿From Table 2 we can see that inference based on $r^*(\psi)$ is accurate for very small sample sizes, and that $W_{AP}(\psi)$ and $W_{TIC}(\psi)$ give quite similar results.

Example 2. Let y_{ij}, for $i=1,\ldots,q$ and $j=1,\ldots,m$, be independent independent gamma random variables with Y_{ij} having density

$$p(y;\psi,\lambda_i)=\frac{\lambda_i^{\psi}}{\Gamma(\psi)}\exp\{-\lambda_i y\}y^{\psi-1}\ .$$

Let ψ be the parameter of interest and $\lambda_1,\ldots,\lambda_q$ be nuisance parameters. Also in this case, all the quantities involved in the modified profile likelihoods for ψ are easy to compute.

Table 3 summarize the results of a simulation study based on 10,000 replications. The estimator $\hat{\psi}$, the estimator based on the modified profile likelihood $\hat{\psi}_{AM}$ and the estimator based on $\ell^{cs}_{TIC}(\psi)$, i.e. $\hat{\psi}_{TIC}$ are compared both in terms of bias and MSE. Various values of m and q are considered.

It can be noted that, in this situation with many nuisance parameters, the estimator $\hat{\psi}_{TIC}$ behaves well when q is small. However, when q increases, the estimator based on the modified profile likelihood becomes preferable according to the MSE.

7.2 Model selection based on adjusted profile likelihoods

Paralleling the previous section, the aim of this section is to assess the behaviour of the modified profile loglikelihood in order to the select the best model. To this end, in this Section we discuss an example concerning two simple exponential models (as

		$\hat{\psi}_P$	$\hat{\psi}_{AP}$	$\hat{\psi}_{TIC}$
$m = 5$	bias	1.22	0.83	0.45
$q = 1$	MSE	37.58	23.82	13.26
$m = 10$	bias	0.37	0.25	0.14
$q = 1$	MSE	0.77	0.58	0.43
$m = 3$	bias	1.985	1.421	0.895
$q = 3$	MSE	3.987	1.545	0.400
$m = 3$	bias	1.514	1.104	0.727
$q = 10$	MSE	0.503	0.123	0.111
$m = 3$	bias	1.467	1.072	0.710
$q = 15$	MSE	0.343	0.064	0.103
$m = 5$	bias	1.220	1.027	0.841
$q = 20$	MSE	0.079	0.021	0.037
$m = 5$	bias	1.203	1.013	0.831
$q = 50$	MSE	0.053	0.0081	0.0333

Table 3: Comparison for the parameter ψ of the gamma distribution.

in Clayton, *et al.*, 1986). A Monte Carlo experiment was performed to compare model selection rates of the following procedures: the modified profile loglikelihood, the AIC and the optimal predictive density of Vidoni (2000).

Let us suppose that there are two plausible statistical models, based on the simple exponential distribution, for describing a dichotomously labelled set of data $y = (y_1, y_2) = (y_{1,1}, \ldots, y_{1,n_1}, y_{2,1}, \ldots, y_{2,n_2})$, where $n_1 + n_2 = n$. Under the first model (M_1) the sample y is distributed with density $p(y; \lambda) = \lambda \exp(-\lambda y)$, $y \geq 0$, $\lambda > 0$. Thus, the sampling distribution does not depend on the label. Under the second model (M_2) the label is assumed to be relevant so that y_i, $i = 1, 2$, is a random sample from an exponential distribution with density $p(y; \lambda_i) = \lambda_i \exp(-\lambda_i y)$, with $\lambda_1 \neq \lambda_2$. Moreover, y_1 and y_2 are independent.

The estimates of the probabilities of correct selection are obtained by considering 10,000 samples for each sample size $n = 4, 8, 12, 20$, with $n_1 = n_2$. Different parameter configurations are considered, by fixing $\lambda_1 = 1$ and setting $\lambda_2 = 1, 1.5, 2, 2.5$. The configuration $\lambda_1 = \lambda_2 = 1$, clearly means that the model M_1 is true, while the situation $\lambda_1 = 1$ and $\lambda_2 \neq 1$ refers to the model M_2. The estimates are given in Table 4. Our choices of λ_2 and n were suggested by the simulation studies of Clayton *et al.* (1986) and Vidoni (2000).

Inspection of the table gives an idea on the behaviour of the three alternative criteria in this particular selection problem. The classical TIC procedure for model selection performs better than the two other proposals under M_1, i.e. when $\lambda_1 = \lambda_2 = 1$. The TIC makes more correct selections under M_1, but this is not true under M_2. In fact, under model M_2 the criterion based on the modified profile loglikelihood has a better behaviour.

Although a number of other simulation studies are needed for a deeper analysis, these preliminary results amphasise that the modified profile loglikelihood gives rise to a model selection procedure which is competitive with the existing criteria.

	n	ℓ_{AP}	ℓ_{VID}	ℓ_{TIC}
$\lambda_2 = 1$	4	0.602	0.743	0.832
	8	0.635	0.798	0.835
	12	0.657	0.816	0.845
	20	0.677	0.834	0.886
	n	ℓ_{AE}	ℓ_{VID}	ℓ_{AIC}
$\lambda_2 = 1.5$	4	0.550	0.400	0.231
	8	0.559	0.401	0.296
	12	0.594	0.435	0.361
	20	0.681	0.530	0.464
	n	ℓ_{AE}	ℓ_{VID}	ℓ_{AIC}
$\lambda_2 = 2$	4	0.717	0.590	0.339
	8	0.786	0.672	0.502
	12	0.857	0.762	0.624
	20	0.935	0.874	0.800
	n	ℓ_{AE}	ℓ_{VID}	ℓ_{AIC}
$\lambda_2 = 2.5$	4	0.834	0.734	0.441
	8	0.916	0.852	0.668
	12	0.961	0.925	0.802
	20	0.989	0.977	0.934

Table 4: Estimated probabilities of correct selection between populations specified by $p_1 = e^{-x}$ and $p_2 = \lambda_2 e^{-\lambda_2 x}$ for three selection criteria.

ACKNOWLEDGMENTS. This work was partially supported by grants from Ministero dell'Università e della Ricerca Scientifica e Tecnologica, Italy.

References

Akaike, H. (1973). Information theory and the maximum likelihood principle. *2nd International Symposium on Information Theory*, Akademiai Kiado, Budapest.

Barndorff-Nielsen, O.E. (1980). Conditionality resolutions. *Biometrika*, **67**, 293–310.

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343-365.

Barndorff-Nielsen, O.E., Cox, D.R. (1994). *Inference and Asymptotics.* Chapman and Hall, London.

Barndorff-Nielsen, O.E., Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli*, **2**, 319-340.

Burnham, K.P., Anderson, D.R. (2002). *Model Selection and Multimodel Inference. A practical Information–Theoretic Approach.* Second edition, Springer, New York.

Clayton, M.K., Geisser, S., Jennings, D.E. (1986). A comparison of several model selection procedures'. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 425–439

Claeskens, G., Hjort, N.L. (2003). The focused information criterion (with discussion). *J. Amer. Statist. Assoc.*, **98**, 900–945.

Corcuera, J.M., Giummolè, F. (2000). First order optimal predictive densities. In *Applications of Differential Geometry to Econometrics* (Eds. P. Marriott and M. Salmon), 214–229, Cambridge University Press, Cambridge.

Cox, D.R., Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B, **49**, 1–39.

Giummolè, F., Ventura, L. (2002). Practical point estimation from higher-order pivot. *J. Statist. Comput. Simul.*, **72**, 419–430.

Harris, I.R. (1989). Predictive fit for natural exponential families. *Biometrika*, **76**, 675-684.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, **1**, 221–234, Univ. California Press, Berkeley.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, **83**, 299–314.

Pace, L., Salvan, A. (2005). A new motivation for adjustments of the profile likelihood. *Journal of Statistical Planning and Inference*, to appear.

Pace, L., Salvan, A., Ventura, L. (2005). Likelihood based discrimination between separate scale and regression models. *Journal of Statistical Planning and Inference*, to appear.

Quesenberry, C. P. (1985). Model construction: Selection of distributions. In *Encyclopedia of Statistical Sciences*, Vol. 5, 583–589.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, **90**, 533–549.

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

Severini, T., Wong, W.H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.*, **20**, 1768–1802.

Severini, T. (1998). An approximation to the modified profile likelihood function. *Biometrika*, **85**, 403–411.

Severini, T. (2000). *Likelihood Methods in Statistics.* Oxford University Press, Oxford.

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica*, **7**, 375–394.

Stein, C. (1956). Efficient nonparametric testing and estimation. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, **1**, 187–195.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc.* B, **39**, 44–47.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences*, **153**, 12–18. (in Japanese)

Vidoni, P. (1995). A simple predictive density based on the p^*- formula. *Biometrika*, **82**, 855–863.

Vidoni, P. (2000). Model selection using the estimative and the approximate p^* predictive densities. *Ann. Inst. Statist. Math.*, **52**, 57–70.

White, H. (1982). Maximum likelihood estimation in misspecified models. *Econometrica*, **50**, 1-25.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it

Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITA
DEGLI STUDI
DI PADOVA
DIPARTIMENTO
DI SCIENZE
STATISTICHE

Department of Statistical Sciences
University of Padua
Italy