Working Paper Series, N. 14, November 2007

Department of Statistical Sciences
University of Padua
Italy

On a likelihood interpretation of adjusted profile likelihoods through refined predictive densities

L. Pace
Department of Statistics
University of Udine
Italy

A. Salvan, L. Ventura
Department of Statistical Sciences
University of Padua
Italy

Abstract: In this paper a second-order link between adjusted profile likelihoods and refinements of the estimative predictive density is shown. The result provides a new straightforward interpretation for modified profile likelihoods, that complements results in Severini (1998a) and in Pace and Salvan (2006). Moreover, it outlines a form of consistency to second order between likelihood theory and prediction in frequentist inference.

Keywords: Estimative predictive density, Modified profile likelihood, Nuisance parameter, Predictive pivot, Profile likelihood, Second-order asymptotics.

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

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

On a likelihood interpretation of adjusted profile likelihoods through refined predictive densities

L. Pace
Department of Statistics
University of Udine
Italy

A. Salvan, L. Ventura
Department of Statistical Sciences
University of Padua
Italy

Abstract: In this paper a second-order link between adjusted profile likelihoods and refinements of the estimative predictive density is shown. The result provides a new straightforward interpretation for modified profile likelihoods, that complements results in Severini (1998a) and in Pace and Salvan (2006). Moreover, it outlines a form of consistency to second order between likelihood theory and prediction in frequentist inference.

Keywords: Estimative predictive density, Modified profile likelihood, Nuisance parameter, Predictive pivot, Profile likelihood, Second-order asymptotics.

1 Introduction

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. Moreover, let $p(y;\theta) = p(y;\psi,\lambda) = \prod_{i=1}^{n} p_{Y_i}(y_i;\psi,\lambda)$ denote the density of Y, with $\theta = (\psi,\lambda) \in \Theta \subseteq \mathbb{R}^d$, where ψ is a p-dimensional parameter of interest and λ is a q-dimensional nuisance parameter, with $d = p + q$.

When the whole parameter θ is of interest, the likelihood function represents, up to a multiplicative constant, the probability of observing Y in a neighbourhood of the actually obtained data y, for each given $\theta \in \Theta$. Even when the primary interest of inference is about the component ψ, a similar interpretation holds for the profile likelihood function. It represents, up to a multiplicative constant, the plug-in estimate of the probability of observing Y in a neighbourhood of the actually obtained data y, for each given ψ. Plug-in refers here to maximum likelihood estimation of λ for each given value of ψ. On the other hand, the profile likelihood is proportional to the estimative predictive density, evaluated at y, of a future Y when ψ is treated as known. A first-order link between likelihood theory and prediction in frequentist

inference is thus depicted. This paper aims at enquiring second-order links.

The profile likelihood, as well as the estimative predictive density, does not take into account uncertainty introduced by sampling variability of the maximum likelihood estimator of the nuisance component λ with ψ treated as known. In order to accomodate for the bias ensuing from mere plug-in estimation, various proposals have been put forward, starting from the modified profile likelihood of Barndorff-Nielsen (1980, 1983) in likelihood theory and from the suggestions of Aichinson (1975) and Harris (1989) in prediction. See Severini (2000, Chapter 9) and Young and Smith (2005, Chapter 10) for recent accounts.

We study here the asymptotic connection between adjustments of the profile likelihood in likelihood theory and refinements of the estimative predictive density in prediction. We show that there exists a direct, second-order, likelihood interpretation for modified profile likelihoods. This new rationale for modifications of the profile likelihood complements the results in Severini (1998a) and in Pace and Salvan (2006).

In more detail, the main result we show can be summarized as follows. Let (y, x) be a $2n$-dimensional data vector from (Y, X), where X is an independent copy of Y. Consider first the refined estimative predictive density of Komaki (1996) for X based on y for a given ψ. Sub-sample y is thus used to eliminate the nuisance component λ in order to obtain an inferentially accurate predictive density for X depending only on ψ. Consider next that such a predictive density, when evaluated at the observed x, defines a pseudo-likelihood for ψ. It turns out that, on average, inference about ψ based on the resulting pseudo-likelihood, i.e. on the $2n$-dimensional sample (y, x), is, to second order, equivalent to inference about ψ based on the modified profile likelihood using data y alone. This result outlines a form of consistency to second order between likelihood theory and prediction in frequentist inference.

Notation and background material are given in Section 2. For clarity and motivation, connections between adjustments of the profile likelihood and refinements of the estimative predictive density are explored through two introductory examples in Section 3. The main result is proved in Section 4.

2 Notation and background

Let us consider first the typical setting of likelihood theory for inference about ψ in the presence of the nuisance parameter λ. 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 of $\theta = (\psi, \lambda)$. Moreover, let $\hat{\lambda}_\psi$ be the constrained maximum likelihood estimate of λ for a given value of ψ and let $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$. In the presence of a nuisance parameter, inference on the interest parameter is, whenever possible, based on exact reduction by marginalisation or by conditioning, leading to marginal or conditional likelihoods. When no exact reduction by marginalisation or conditioning is available, likelihood inference is often based on a pseudo-likelihood, i.e. on a function of ψ and y having properties similar to those of a genuine likelihood for ψ. The most commonly used pseudo-loglikelihood 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) 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 $X = (X_1, \ldots, X_h)$, $h \geq 1$, independent of $Y = (Y_1, \ldots, Y_n)$, and having independent and identically distributed components, where X_1 has the same distribution as Y_1. Let us denote by $p_{_X}(x;\theta) = p_{_X}(x;\psi,\lambda)$ the density of X. For notational consistency with the likelihood setting in the presence of nuisance parameters, ψ has to be treated as known, while λ is unknown. 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 the unknown λ with its maximum likelihood estimate for the given ψ, based on y, denoted by $\hat{\lambda}_\psi = \hat{\lambda}_\psi(y)$.

The estimative, or plug-in, device considered in first-order likelihood theory and prediction neglects sampling variability of the estimated nuisance parameter. Particularly serious inaccuracies may occur when the dimension of λ is large relative to n. See e.g. Sartori (2003) and Vidoni (1995) for likelihood theory and prediction, respectively.

We briefly recall below the expression of notable instances of adjustments of profile loglikelihood and of refinements of the estimative predictive density.

2.1 Modifications of profile likelihood

Let us denote by $\ell_\psi(\theta)$ and $\ell_\lambda(\theta)$ blocks of the score (column) vector $\partial\ell(\theta)/\partial\theta$. Moreover, let $j_{\psi\psi}(\theta)$, $j_{\psi\lambda}(\theta)$ and $j_{\lambda\lambda}(\theta)$ be blocks of the observed information $j(\theta) = -\partial^2\ell(\theta)/(\partial\theta\partial\theta^\top)$. Similarly, we will denote by $i_{\psi\psi}(\theta)$, $i_{\psi\lambda}(\theta)$ and $i_{\lambda\lambda}(\theta)$ blocks of the expected information $i(\theta) = E_\theta(j(\theta))$, where $E_\theta(\cdot)$ denotes expectation under θ. 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

$$\ell_{_M}(\psi) = \ell_{_M}(\psi;y) = \ell_{_P}(\psi) - \frac{1}{2}\log \mid j_{\lambda\lambda}(\hat{\theta}_\psi) \mid - \log \left| \frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}} \right| , \tag{3}$$

where

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

involves 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 $i_{\psi\lambda}(\theta) = 0$, such a calculation can be avoided because $\log \left|\partial\hat{\lambda}_\psi/\partial\hat{\lambda}\right| = O_p(n^{-1})$ when

$\psi - \hat{\psi} = O_p(n^{-1/2})$. This leads to the approximate conditional loglikelihood 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 $O_p(n^{-1})$, i.e. to second order. See Severini (2000, Section 9.5) for a review of approximate calculation of sample space derivatives. In particular, the approximation to $\ell_{M}(\psi)$ developed in Severini (1998b) is

$$\bar{\ell}_{M}(\psi) = \bar{\ell}_{M}(\psi; y) = \ell_{P}(\psi) + \frac{1}{2}\log |j_{\lambda\lambda}(\hat{\theta}_{\psi})| - \log |\nu_{\lambda,\lambda}(\hat{\theta}_{\psi}, \hat{\theta}; \hat{\theta})| , \tag{4}$$

where

$$\nu_{\lambda,\lambda}(\theta_1, \theta_2; \theta_0) = E_{\theta_0}(\ell_{\lambda}(\theta_1)\ell_{\lambda}(\theta_2)^{\top}) \tag{5}$$

and $\theta_0 = (\psi_0, \lambda_0)$ denotes the true parameter value. An asymptotically equivalent version of (4) 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

$$\hat{\nu}_{\lambda,\lambda}(\theta_1, \theta_2) = \sum_{i=1}^{n} \ell_{\lambda}^{(i)}(\theta_1)\ell_{\lambda}^{(i)}(\theta_2)^{\top} , \tag{6}$$

with $\ell_{\lambda}^{(i)}(\theta) = \partial \log p_{Y_1}(y_i; \psi, \lambda)/\partial\lambda$ (cf. Severini, 2000, Section 9.5.5).

In Pace and Salvan (2006) various on average second-order equivalent versions of $\ell_{M}(\psi)$, denoted by $\ell_{AP}(\psi)$, are discussed. Second-order equivalence on average means that $E_{\theta_0}(\ell_{AP}(\psi) - \ell_{M}(\psi)) = c + O(n^{-1})$, where c is a constant. In particular, for the purposes of this paper, the following version is relevant

$$\ell_{AP}(\psi) = \ell_{P}(\psi) - \frac{1}{2}\mathrm{tr}\{j_{\lambda\lambda}(\hat{\theta}_{\psi})^{-1}\nu_{\lambda,\lambda}(\hat{\theta}_{\psi}, \hat{\theta}_{\psi}; \hat{\theta})\} . \tag{7}$$

This is the straightforward generalization to $q > 1$ of the adjusted profile loglikelihood $\ell_{P}(\psi) - a^{II}(\psi)$, with $a^{II}(\psi)$ as given in Pace and Salvan (2006, Section 3.3).

2.2 Refinements of the estimative predictive density

Even in the setting of prediction, exact reductions are sometimes possible, in particular when an exact pivot for λ of the form $T(Y, X, \psi)$, for short an exact predictive pivot, is available (see Barndorff-Nielsen and Cox, 1996, and the examples in Section 3). This reduction parallels the construction of marginal likelihoods in likelihood theory. When no exact predictive pivot exists, asymptotic methods may be considered, with the estimative predictive density (2) playing the same role as the profile likelihood in likelihood theory.

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 $O(n^{-1})$, i.e. to second order. 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 $r, s, \ldots = 1, \ldots, q$. Let $\ell(\theta; x) = \log p_{_X}(x; \psi, \lambda)$, $\ell_r(\theta; x) = \partial \log p_{_X}(x; \psi, \lambda)/\partial\lambda_r$ and $\ell_{rs}(\theta; x) = \partial^2 \log p_{_X}(x; \psi, \lambda)/(\partial\lambda_r \partial\lambda_s)$. Then,

$$p_{_K}(x; \psi) = p_e(x; \psi) \left[1 + \frac{1}{2} \left\{ h_{rs}(\hat{\theta}_\psi; x) - \Gamma^t_{rs}(\hat{\theta}_\psi) \ell_t(\hat{\theta}_\psi; x) \right\} i^{rs}(\hat{\theta}_\psi) \right] , \qquad (8)$$

where

$$h_{rs}(\theta; x) = \ell_{rs}(\theta; x) + \ell_r(\theta; x)\ell_s(\theta; x) ,$$

$$\Gamma^t_{rs}(\theta) = i^{tu}(\theta)\, E_\theta\{h_{rs}(\theta; X)\ell_u(\theta; X)\}$$

and $i^{rs}(\theta)$ denotes the generic element of the inverse matrix of $i_{\lambda\lambda}(\theta)$.

Corcuera and Giummolè (2000) show that (8) is, to second order, the optimal improvement over (2) in terms of average Kullback-Leibler divergence also for general regular models. Moreover, they give the form of the optimal correction under a general α-divergence.

3 From prediction to likelihood: two examples

"Exact" or refined estimative predictive densities treating ψ as known can be exploited to define a pseudo-loglikelihood for ψ. Suppose in particular that $T = T(Y, X_1, \psi)$ is an exact pivot for λ, based on X_1, Y and, possibly, ψ. Let R_α be a set such that $\Pr_\psi\{T(Y, X_1, \psi) \in R_\alpha\} = 1 - \alpha$. Then, a prediction set based on T with exact level $1 - \alpha$ is

$$S_\alpha(y) = \{x_1 : T(y, x_1, \psi) \in R_\alpha\} .$$

A formal predictive density $\hat{p}(x_1; y, \psi)$ such that

$$\int_{S_\alpha(y)} \hat{p}(x_1; y, \psi) dx_1 = 1 - \alpha$$

for every $\alpha \in (0, 1)$ will be called an exact predictive density for X_1 based on y for a given ψ.

Given an exact predictive density for X_1, the corresponding predictive density of n independent copies of X_1, i.e. of the random sample $X = (X_1, \ldots, X_n)$, is

$$\hat{p}(x; y, \psi) = \prod_{i=1}^{n} \hat{p}(x_i; y, \psi) . \qquad (9)$$

A natural pseudo-loglikelihood for ψ may be defined by treating $\log \hat{p}(x; y, \psi)$ as a function of ψ for the observed (y, x). We argue in the two examples below that such a pseudo-loglikelihood for ψ agrees on average with $\ell_{_M}(\psi)$, to second order. When no exact pivot is available second-order agreement is maintained if $\hat{p}(x_i; y, \psi)$ in (9) is replaced by the refined estimative predictive density of the form (8), as will be shown in Section 4.

Example 1: Random sampling from a normal distribution

Let (y, x), with $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$, be a random sample of size $2n$ from a normal distribution with mean μ and variance σ^2. In order to obtain a pseudo-loglikelihood for σ^2 based on (9), let us consider y 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 and X_1 be an independent future observation from the same distribution. Based on the exact pivot $X_1 - \bar{Y}_n$, the exact 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 $\sigma^2(1+n^{-1})$.

The modified profile loglikelihood (3) for σ^2 takes the form

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

with

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

where $\hat{\sigma}_y^2 = n^{-1}\sum_{i=1}^n (y_i - \bar{y}_n)^2$ is the maximum likelihood estimate of σ^2 based on y. Let $\theta_0 = (\mu_0, \sigma_0^2)$. Then for $\sigma^2 - \sigma_0^2 = O(n^{-1/2})$, considering $\hat{\sigma}_y^2$ as a random variable, under θ_0 we have the stochastic expansion

$$\log\frac{\sigma^2}{\hat{\sigma}_y^2} = -\log\left(1 + \frac{\hat{\sigma}_y^2}{\sigma^2} - 1\right) = -\left(\frac{\hat{\sigma}_y^2}{\sigma^2} - 1\right) + O_p(n^{-1/2})\,,$$

giving

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

The predictive density of the random sample $X = (X_1, \ldots, X_n)$ is, in view of (9),

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

Hence, neglecting constants,

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

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

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

Under θ_0 the quantity $n(\bar{y}_n - \bar{x}_n)^2$ is a realization of $2\sigma_0^2 W$, where W is a chi-square on one degree of freedom. Therefore,

$$\log\hat{p}(X; Y, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\frac{\hat{\sigma}_x^2}{\sigma^2} + \frac{1}{2}\frac{\hat{\sigma}_x^2}{\sigma^2} - \frac{\sigma_0^2}{\sigma^2} W + O_p(n^{-1})\,.$$

Moreover, using $E_{\theta_0}(\sigma_0^2 W) = E_{\theta_0}(\hat{\sigma}_x^2) + O(n^{-1})$,

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

The last identity follows from the fact that Y is a copy of X. This shows that $\ell_{M}(\sigma^2)$ agrees on average to second-order with the pseudo-loglikelihood for σ^2 derived from $\log \hat{p}(x; y, \sigma^2)$.

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 maximum likelihood estimate of λ with ψ fixed is $\hat{\lambda}_\psi = \hat{\lambda}_\psi(y) = \psi/\bar{y}_n$. Based on the pivot $\hat{\lambda}_\psi X_1$, the exact predictive density of X_1 may be expanded as (Vidoni, 1995, Example 4.2)

$$\begin{aligned} \hat{p}(x_1; y, \psi) &= p(x_1; \psi, \hat{\lambda}_\psi) \\ &\quad \left[1 + \frac{1}{2n\psi}\left\{\hat{\lambda}_\psi^2 x_1^2 - 2\hat{\lambda}_\psi(\psi+1)x_1 + \psi(\psi+1)\right\} + O(n^{-2})\right]\,. \end{aligned} \tag{10}$$

The modified profile loglikelihood (3) for ψ is

$$\ell_{M}(\psi) = \ell_{M}(\psi; y) = \ell_{P}(\psi; y) - \frac{1}{2}\log\psi\,,$$

where

$$\ell_{P}(\psi; y) = (\psi - 1)\sum_{i=1}^{n}\log y_i - n\psi + n\psi\log\psi - n\psi\log\bar{y}_n - n\log\Gamma(\psi)\,.$$

Let $\theta_0 = (\psi_0, \lambda_0)$. Then for $\psi - \psi_0 = O(n^{-1/2})$, and neglecting constants, under θ_0,

$$\ell_{M}(\psi; Y) = \ell_{P}(\psi; Y) - \frac{1}{2}\frac{\psi}{\hat{\psi}} + O_p(n^{-1/2})\,,$$

where $\hat{\psi} = \hat{\psi}(Y)$. Using in (9) the predictive density (10), we get

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

where $F = \hat{\lambda}_\psi(Y)/\hat{\lambda}_\psi(X)$ is distributed, under $\theta_0 = (\psi_0, \lambda_0)$, 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 $\hat{\lambda}_\psi(Y) = \psi\lambda_0/\psi_0 + O_p(n^{-1/2})$, $\sum X_i/n = \psi_0/\lambda_0 + O_p(n^{-1/2})$ and $\sum X_i^2/n = \psi_0(\psi_0 + 1)/\lambda_0^2 + O_p(n^{-1/2})$.

Hence, under θ_0,

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

Note that, when taking expectations, the above terms of order $O_p(n^{-1/2})$ vanish. Hence, neglecting additive constants,

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

4 Adjusted profile loglikelihood from an optimal predictive density

In this section we show that the result of Examples 1 and 2 of the previous section 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.

We show below that, if expression (8) is used for $\hat{p}(x_i; y, \psi)$ in (9), $i = 1, \ldots, n$, then

$$E_{\theta_0}\left\{\log \hat{p}(X; Y, \psi)\right\} = c + E_{\theta_0}\left\{\ell_{M}(\psi; Y)\right\} + O(n^{-1}) \,, \tag{11}$$

where c is a constant. Relation (11) suggests the following direct likelihood interpretation for $\ell_{_M}(\psi)$. Based on a $2n$-dimensional data vector (y, x), the function $\log \hat{p}(x; y, \psi)$ constitutes a pseudo-loglikelihood for ψ. It amounts to use the sub-sample y to eliminate the nuisance component λ in order to obtain an inferentially accurate predictive density for X depending only on ψ. When evaluated at the observed sub-sample x, such a predictive density produces an inferentially accurate pseudo-loglikelihood for ψ, that, on average, agrees with $\ell_{_M}(\psi; y)$, up to terms of order $O(n^{-1})$. This contrasts with what happens for $\ell_{_P}(\psi; y)$ for which $E_{\theta_0}(\log \hat{p}(X; Y, \psi)) = c + E_{\theta_0}(\ell_{_P}(\psi; Y)) + O(1)$.

Relation (11) can be proved as follows. Using expression (8) for $\hat{p}(x_i; y, \psi)$, we get

$$\begin{aligned}
\log \hat{p}(X; Y, \psi) &= \sum_{i=1}^{n} \log p_e(X_i; \psi) + \sum_{i=1}^{n} \log \Big[1 + \frac{1}{2} \Big\{ h_{rs}(\psi, \hat{\lambda}_\psi(Y); X_i) \\
&\qquad - \Gamma^t_{rs}(\psi, \hat{\lambda}_\psi(Y)) \ell_t(\psi, \hat{\lambda}_\psi(Y); X_i) \Big\} i^{rs}(\psi, \hat{\lambda}_\psi(Y)) \Big] \\
&= \sum_{i=1}^{n} \log p(X_i; \psi, \hat{\lambda}_\psi(Y)) \\
&\qquad + \frac{1}{2} i^{rs}(\psi, \hat{\lambda}_\psi(Y)) \sum_{i=1}^{n} \Big\{ h_{rs}(\psi, \hat{\lambda}_\psi(Y); X_i) \\
&\qquad - \Gamma^t_{rs}(\psi, \hat{\lambda}_\psi(Y)) \ell_t(\psi, \hat{\lambda}_\psi(Y); X_i) \Big\} + O_p(n^{-1}) \,. \qquad (12)
\end{aligned}$$

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

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

where $\Big(\hat{\lambda}_\psi(Y) - \hat{\lambda}_\psi(X) \Big)_{rs} = \Big(\hat{\lambda}_\psi(Y) - \hat{\lambda}_\psi(X) \Big)_r \Big(\hat{\lambda}_\psi(Y) - \hat{\lambda}_\psi(X) \Big)_s$. 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,

$$\begin{aligned}
\sum_{i=1}^{n} \log p(X_i; \psi, \hat{\lambda}_\psi(Y)) &= \ell_{_P}(\psi; X) \\
&\qquad - \frac{1}{2} \Big(\hat{\lambda}_\psi(Y) - \hat{\lambda}_\psi(X) \Big)_{rs} j_{rs}(\psi, \hat{\lambda}_\psi(X); X) \\
&\qquad + O_p(n^{-1/2}) \,,
\end{aligned}$$

where $j_{rs}(\psi,\lambda;x) = -\sum_{i=1}^n \ell_{rs}(\psi,\lambda;x_i)$.

Let us denote by λ_ψ the maximiser with respect to λ of $E_{\theta_0}(\ell(\psi,\lambda))$, with ψ fixed, cf. Severini (2000, Section 4.8). We also let $\theta_\psi = (\psi,\lambda_\psi)$. Under regularity conditions, $\hat\lambda_\psi$ is a consistent estimator of λ_ψ (Cox, 1961; Huber, 1967). Moreover, we let $i_{rs}(\theta_\psi;\theta_0)$ denote a generic element of $i_{\lambda\lambda}(\theta_\psi;\theta_0) = E_{\theta_0}(j_{\lambda\lambda}(\theta_\psi))$. A generic element of $i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}$ is denoted by $i^{rs}(\theta_\psi;\theta_0)$.

Using results in the Appendix of Pace and Salvan (2006), we obtain

$$\begin{aligned}\left(\hat\lambda_\psi(Y)-\hat\lambda_\psi(X)\right)_r &= \left(\hat\lambda_\psi(Y)-\lambda_\psi-\hat\lambda_\psi(X)+\lambda_\psi\right)_r\\ &= i^{rt}(\theta_\psi;\theta_0)\ell_t(\theta_\psi;Y) - i^{rt}(\theta_\psi;\theta_0)\ell_t(\theta_\psi;X) + O_p(n^{-1})\\ &= i^{rt}(\theta_\psi;\theta_0)\left\{\ell_t(\theta_\psi;Y)-\ell_t(\theta_\psi;X)\right\} + O_p(n^{-1})\,, \qquad (13)\end{aligned}$$

while

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

Therefore,

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

where $\nu_{t,u}(\theta_1,\theta_2;\theta_0)$ is the (t,u) element of $\nu(\theta_1,\theta_2;\theta_0)$ defined by (5).

Moreover,

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

Hence, the leading term on the right-hand side of (12) has the expansion

$$\begin{aligned}E_{\theta_0}\left\{\sum_{i=1}^n \log p(X_i;\psi,\hat\lambda_\psi(Y))\right\} &= E_{\theta_0}(\ell_{P}(\psi;X))\\ &\quad -i_{rs}(\theta_\psi;\theta_0)i^{rt}(\theta_\psi;\theta_0)i^{su}(\theta_\psi;\theta_0)\nu_{t,u}(\theta_\psi,\theta_\psi;\theta_0)\\ &\quad +O(n^{-1})\\ &= E_{\theta_0}(\ell_{P}(\psi;X)) - i^{rs}(\theta_\psi;\theta_0)\nu_{r,s}(\theta_\psi,\theta_\psi;\theta_0)\\ &\quad +O(n^{-1})\\ &= E_{\theta_0}(\ell_{P}(\psi;X)) - \mathrm{tr}\left[\nu_{\lambda,\lambda}(\theta_\psi,\theta_\psi;\theta_0)i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}\right]\\ &\quad +O(n^{-1})\,. \qquad (14)\end{aligned}$$

Let us now denote by A twice the adjustment term of order $O_p(1)$ on the right-hand side of (12), i.e. we let

$$A = i^{rs}(\psi,\hat\lambda_\psi(Y))\sum_{i=1}^n\left\{h_{rs}(\psi,\hat\lambda_\psi(Y);X_i) - \Gamma^t_{rs}(\psi,\hat\lambda_\psi(Y))\ell_t(\psi,\hat\lambda_\psi(Y);X_i)\right\}\,.$$

Then, using (13),

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

where $\hat\nu_{r,s}(\theta_1,\theta_2)$ is the (r,s) element of $\hat\nu(\theta_1,\theta_2)$ defined by (6) and the likelihood equation $\sum_{i=1}^n\ell_t(\psi,\hat\lambda_\psi(X);X_i)=0$ has been used.

With a further expansion around λ_ψ of terms depending on $\hat\lambda_\psi(X)$ and using formula (9.17) in Pace and Salvan (1997), we get

$$\begin{aligned} A &= \Big\{i^{rs}(\theta_\psi;\theta_0)+O_p(n^{-3/2})\Big\}\\ &\quad \Big\{-i_{rs}(\theta_\psi;\theta_0)+O_p(n^{1/2})+\nu_{r,s}(\theta_\psi,\theta_\psi;\theta_0)+O_p(n^{1/2})\Big\}\,, \end{aligned}$$

so that

$$E_{\theta_0}(A)=c+\mathrm{tr}\left[\nu_{\lambda,\lambda}(\theta_\psi,\theta_\psi;\theta_0)i_{\lambda\lambda}(\theta_\psi;\theta_0)^{-1}\right]+O(n^{-1})\,, \tag{15}$$

where c is a constant and where the expectation of the terms of order $O_p(n^{r/2})$, with odd r, is easily seen to be of order $O(n^{(r-1)/2})$ (see e.g. Pace and Salvan 1997, Section 9.2.2).

¿From (14) and (15), using results in Section 2.1 and neglecting additive constants, we obtain

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

This shows relation (11).

References

[1] Aichinson, J. (1975). Goodness of prediction fit. *Biometrika*, 62, 547–554.

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

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

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

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

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

[7] Cox, D.R. (1961). Tests of separate families of hypotheses. *Proc. 4th Berkley Symp.*, 1, 105–123.

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

[9] Harris, I.R. (1989). Predictive fit for natural exponential families. *Biometrika*, 76, 675–684.

[10] 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.

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

[12] Pace, L., Salvan, A. (1997). *Principles of Statistical Inference*, World Scientific, Singapore.

[13] Pace, L., Salvan, A. (2006). A new motivation for adjustments of the profile likelihood. *J. Statist. Plan. Inf.*, 136, 3554–3564.

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

[15] Severini, T. (1998a). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika*, 85, 507–522.

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

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

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

[19] Young, G.A., Smith, R.L. (2005). *Essentials of Statistical Inference.* Cambridge University Press, Cambridge.

Acknowledgements

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

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

Department of Statistical Sciences
University of Padua
Italy