Ann Inst Stat Math (2011) 63:923–937
DOI 10.1007/s10463-010-0274-9

Adjustments of profile likelihood through predictive densities

Luigi Pace · Alessandra Salvan · Laura Ventura

Received: 30 November 2007 / Revised: 18 February 2009 / Published online: 24 February 2010

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 interpretation for modified profile likelihoods that complements results in the literature. Moreover, it suggests how to construct adjusted profile likelihoods using accurate predictive densities.

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

1 Introduction

Likelihood and prediction are related from the very definition of likelihood. Consider a parametric statistical model for data y, realization of a random variable Y, with parameter θ. Then the likelihood function is proportional to the probability of observing a future Y in a neighbourhood of the actually obtained data y, for each given θ. When $\theta = (\psi, \lambda)$ and the primary interest of inference is about ψ, the most widely used pseudo-likelihood for ψ is the profile likelihood. It can be interpreted, to first order, in terms of a predictive density. Indeed it represents, up to a multiplicative constant,

L. Pace
Department of Statistics, University of Udine, Udine, Italy
e-mail: luigi.pace@dss.uniud.it

A. Salvan (✉) · L. Ventura
Department of Statistics, University of Padova, Padova, Italy
e-mail: salvan@stat.unipd.it

L. Ventura
e-mail: ventura@stat.unipd.it

for each given ψ, the plug-in estimate of the probability of observing a future Y in a neighbourhood of the actually obtained data y. Plug-in refers here to maximum likelihood estimation of λ for the given value of ψ. Hence, the profile likelihood is proportional to the estimative predictive density, evaluated at y, of a future Y when ψ is treated as known.

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 accommodate to second order 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 refinements of the estimative predictive density suggested by Aitchison (1975) and Harris (1989) in prediction. See Severini (2000, Chapter 9) and Young and Smith (2005, Chapter 10) for recent accounts.

Here, we show that, in frequentist inference, there exists a form of consistency between likelihood theory and prediction also to second order. In particular, there are asymptotic connections between adjustments of the profile likelihood and refinements of the estimative predictive density. These connections provide a new rationale for modifications of the profile likelihood, that complements the results in Severini (1998a, 2007) and in Pace and Salvan (2006). Moreover, the result suggests how to construct adjusted profile likelihoods using accurate predictive densities.

Notation and background material are given in Sect. 2. For clarity and motivation, connections between adjustments of the profile likelihood and refinements of the estimative predictive density are explored first through two introductory examples in Sect. 3. The main result is discussed in Sect. 4, together with a procedure to construct adjusted profile likelihoods using accurate predictive densities. The proof of the result in Sect. 4 is given in the Appendix.

2 Notation and background

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 $d = p + q$.

Let us consider first the typical setting of likelihood theory for inference about ψ in the presence of the nuisance parameter λ. We assume throughout a regular parametric model satisfying the major conditions as detailed e.g. in Severini (2000, Sect. 3.4). 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 $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$. Whenever possible, inference about ψ is based on exact reduction by marginalisation or by conditioning, leading to marginal or conditional likelihoods. When no exact reduction is available, likelihood inference is usually based on the profile likelihood. The corresponding loglikelihood is

$$\ell_P(\psi) = \ell_P(\psi; y) = \ell(\hat{\theta}_\psi) = \ell(\psi, \hat{\lambda}_\psi). \tag{1}$$

As is well known, (1) shares most first-order properties of a genuine loglikelihood for ψ (see Barndorff-Nielsen and Cox 1994, Sect. 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

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

where

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

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 $\log|\partial\hat{\lambda}_\psi/\partial\hat{\lambda}| = O_p(n^{-1})$ when $\psi - \hat{\psi} = O_p(n^{-1/2})$. This leads to the approximate conditional loglikelihood of Cox and Reid (1987)

$$\ell_A(\psi) = \ell_A(\psi; y) = \ell_P(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)|,$$

which approximates $\ell_M(\psi)$ with error of order $O_p(n^{-1})$, i.e. to second order. See Severini (2000, Sect. 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

$$\bar{\ell}_M(\psi) = \bar{\ell}_M(\psi; y) = \ell_P(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)| - \log|\nu_{\lambda,\lambda}(\hat{\theta}_\psi,\hat{\theta};\hat{\theta})|, \qquad (4)$$

where

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

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

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

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

The modified profile loglikelihood and the related approximations presented above were all developed as aimed at approximating some target loglikelihood for ψ, corresponding to a suitable model reduction via conditioning or marginalisation. When such exact reduction is not available, approximations of form (4) are still applicable and often provide notable improvements over the profile loglikelihood. See e.g. simulation results in DiCiccio and Martin (1993), DiCiccio and Stern (1994), Sartori et al. (1999), Bellio and Sartori (2006). Therefore, broader perspectives for theoretical justification of adjustments of profile loglikelihood are called for. One view is related to bias adjustment of the profile score function as in McCullagh and Tibshirani (1990), DiCiccio et al. (1996), Stern (1997). A similar view leads to direct adjustment of the profile loglikelihood by removing the leading term of its bias with respect to a suitable target loglikelihood for ψ as in Pace and Salvan (2006). See also Arellano and Hahn (2007). This correction implies second-order unbiasedness of the corresponding estimating equation for ψ. All adjustments that remove the leading term of the profile loglikelihood bias are on average second-order equivalent to the target loglikelihood and to $\ell_M(\psi)$. For a generic adjusted profile loglikelihood $\ell_{AP}(\psi)$, second-order equivalence on average to $\ell_M(\psi)$ means that, for $\psi - \psi_0 = O(n^{-1/2})$ and neglecting additive constants, we have $E_{\theta_0}(\ell_{AP}(\psi) - \ell_M(\psi)) = O(n^{-1})$. The adjusted profile score based on such an $\ell_{AP}(\psi)$ gives an estimating equation for ψ which is second-order unbiased.

In Pace and Salvan (2006), various on average second-order equivalent versions of $\ell_M(\psi)$, denoted by $\ell_{AP}(\psi)$, are discussed. For the purposes of this paper, the relevant version is

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

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 Sect. 3.3 of the quoted paper.

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 Sect. 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 $\lambda_r, \lambda_s, \ldots,$ 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,

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

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 for general regular models.

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

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

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 discussed in Sect. 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. As is well known, 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

$$\ell_P(\sigma^2; y) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\frac{\hat{\sigma}_y^2}{\sigma^2}, \tag{10}$$

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

$$\ell_M(\sigma^2; Y) = \ell_P(\sigma^2; Y) - \frac{1}{2}\frac{\hat{\sigma}_y^2}{\sigma^2} + O_p(n^{-1/2}). \tag{11}$$

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,

$$\begin{aligned}\log\hat{p}(x; y, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2(1+n^{-1})}\sum_{i=1}^n (x_i - \bar{x}_n)^2 \\ -\frac{1}{2\sigma^2(1+n^{-1})}n(\bar{y}_n - \bar{x}_n)^2.\end{aligned} \tag{12}$$

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

$$\begin{aligned}
\hat{p}(x_1; y, \psi) &= p(x_1; \psi, \hat{\lambda}_\psi) \\
&\quad \times \left[1 + \frac{1}{2n\psi}\left\{\hat{\lambda}_\psi^2 x_1^2 - 2\hat{\lambda}_\psi(\psi+1)x_1 + \psi(\psi+1)\right\} + O(n^{-2})\right],
\end{aligned} \tag{13}$$

see Example 4.2 in Vidoni (1995).

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 (13), we get

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

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,

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

showing that $\log \hat{p}(x; y, \psi)$ agrees to second order on average with $\ell_M(\psi; y)$.

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 in the Appendix that, if the optimal predictive density of Komaki (1996), see expression (8), is used for the generic factor $\hat{p}(x_i; y, \psi)$ in $\hat{p}(x; y, \psi)$ given by (9), then, for $\psi - \psi_0 = O(n^{-1/2})$ and neglecting additive constants,

$$E_{\theta_0}\left\{\log \hat{p}(X; Y, \psi)\right\} = E_{\theta_0}\left\{\ell_M(\psi; Y)\right\} + O(n^{-1}). \tag{14}$$

Relation (14) 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 $E_{\theta_0}(\log \hat{p}(X; Y, \psi)) = E_{\theta_0}(\ell_P(\psi; Y)) + O(1)$.

Relation (14) can be employed to construct adjusted profile loglikelihoods using accurate predictive densities. An adjusted profile loglikelihood has the form $\ell_P(\psi) + M(\psi)$, where $M(\psi)$ is an adjustment of order $O_p(1)$. Suppose, without loss of generality for asymptotic arguments, that n is even and equal to $2m$. Let us write the data as $y = (y_A, y_B)$, with both $y_A = (y_{A1}, \ldots, y_{Am})$ and $y_B = (y_{B1}, \ldots, y_{Bm})$ random samples of size m. Let us write (9) as $\hat{p}(y_B; y_A, \psi) = \prod_{i=1}^m \hat{p}(y_{Bi}; y_A, \psi)$, where $\hat{p}(y_{Bi}; y_A, \psi)$ is either an exact predictive density or the refined estimative predictive density (8). Similarly, we can compute the predictive density, $\hat{p}(y_A; y_B, \psi)$, of Y_A based on y_B. Then the adjusted profile loglikelihood defined as

$$\bar{\ell}_{AP}(\psi) = \ell_P(\psi; y) + \bar{M}(\psi), \tag{15}$$

with

$$\bar{M}(\psi) = \frac{1}{2}\left\{\log \hat{p}(y_B; y_A, \psi) + \log \hat{p}(y_A; y_B, \psi) - \ell_P(\psi; y_A) - \ell_P(\psi; y_B)\right\},$$

is, by (14), second order equivalent on average to $\ell_M(\psi)$.

Example 1 (*cont.*): *Random sampling from a normal distribution* To compute the adjustment term in (15), let us consider (12) with x and y replaced by y_A and y_B. Then, using (10), $\bar{M}(\sigma^2)$ takes the form

$$\bar{M}(\sigma^2) = \frac{m}{4\sigma^2(m+1)}\left(\hat{\sigma}^2_{y_A} + \hat{\sigma}^2_{y_B}\right) - \frac{m^2}{2(m+1)}(\bar{y}_A - \bar{y}_B)^2, \tag{16}$$

with $\bar{y}_A$, $\bar{y}_B$, $\hat{\sigma}^2_{y_A}$ and $\hat{\sigma}^2_{y_B}$ denoting sample means and maximum likelihood estimates of variance for the two subsamples y_A and y_B each of size m.

It is straightforward to see that

$$E_{\theta_0}(\bar{M}(\sigma^2)) = -\frac{1}{2}\frac{\sigma_0^2}{\sigma^2},$$

so that, comparison with (11) shows that $\bar{\ell}_{AP}(\sigma^2) = \ell_P(\sigma^2) + \bar{M}(\sigma^2)$ is second order equivalent on average to $\bar{\ell}_M(\sigma^2)$.

The example above could be generalized in a straightforward way to situations where exact reductions are available (see the discussion in Sect. 2.2). Similarly, in exponential and curved exponential models, calculation of (8) is fairly simple and, according to the results in this paper, the adjusted profile loglikelihood (15) is second order equivalent on average to other available modifications such as (4) or the proposal in Barndorff-Nielsen (1995). Outside these models, the adjusted profile loglikelihood (15) seems to offer the advantage of not requiring calculation of sample space derivatives or calculation of non-null expectations as in (5). Comparing the small sample effects of (15) and (4) is outside the scope of this paper. Indeed, the behaviour of different adjustments of profile loglikelihod is expected to depend on the particular model under consideration. Further developments suggested by (15) concern: (i) the use of Bayesian predictive densities with suitable matching priors (Datta and Mukerjee 2004, Chapter 6); (ii) construction of robust adjusted profile loglikelihoods using the generalization of (8) by Vidoni (2008) which accounts for possible model misspecification.

Appendix

Relation (14) can be proved as follows. Using expression (8) for $\hat{p}(x_i; y, \psi)$, we get

$$\begin{aligned}\log \hat{p}(X; Y, \psi) &= \sum_{i=1}^{n} \log p_e(X_i; \psi) + \sum_{i=1}^{n} \log \Big[1 + \frac{1}{2} \Big\{ h_{rs}(\psi, \hat{\lambda}_\psi(Y); X_i) \\ &\qquad - \Gamma^t_{rs}(\psi, \hat{\lambda}_\psi(Y)) \ell_t(\psi, \hat{\lambda}_\psi(Y); X_i) \Big\} i^{rs}(\psi, \hat{\lambda}_\psi(Y)) \Big] \\ &= \sum_{i=1}^{n} \log p(X_i; \psi, \hat{\lambda}_\psi(Y)) \\ &\qquad + \frac{1}{2} i^{rs}(\psi, \hat{\lambda}_\psi(Y)) \sum_{i=1}^{n} \Big\{ h_{rs}(\psi, \hat{\lambda}_\psi(Y); X_i) \\ &\qquad - \Gamma^t_{rs}(\psi, \hat{\lambda}_\psi(Y)) \ell_t(\psi, \hat{\lambda}_\psi(Y); X_i) \Big\} + O_p(n^{-1}). \end{aligned} \tag{17}$$

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

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

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

Let us denote by λ_ψ the maximiser with respect to λ of $E_{\theta_0}(\ell(\psi, \lambda))$, with ψ fixed. We also let $\theta_\psi = (\psi, \lambda_\psi)$. We recall that $\hat{\lambda}_\psi$ is a consistent estimator of λ_ψ (cf. Severini 2000, Sect. 4.6). 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

$$\begin{aligned}\left(\hat{\lambda}_\psi(Y) - \hat{\lambda}_\psi(X)\right)_r &= \left(\hat{\lambda}_\psi(Y) - \lambda_\psi - \hat{\lambda}_\psi(X) + \lambda_\psi\right)_r \\ &= i^{rt}(\theta_\psi; \theta_0)\ell_t(\theta_\psi; Y) - i^{rt}(\theta_\psi; \theta_0)\ell_t(\theta_\psi; X) + O_p(n^{-1}) \\ &= i^{rt}(\theta_\psi; \theta_0)\left\{\ell_t(\theta_\psi; Y) - \ell_t(\theta_\psi; X)\right\} + O_p(n^{-1}), \qquad (18)\end{aligned}$$

while

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

Therefore,

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

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 (17) has the expansion

$$\begin{aligned} E_{\theta_0}\left\{\sum_{i=1}^{n} \log p(X_i; \psi, \hat{\lambda}_\psi(Y))\right\} &= E_{\theta_0}(\ell_P(\psi; X)) \\ &\quad - i_{rs}(\theta_\psi; \theta_0) i^{rt}(\theta_\psi; \theta_0) i^{su}(\theta_\psi; \theta_0) \nu_{t,u}(\theta_\psi, \theta_\psi; \theta_0) + O(n^{-1}) \\ &= E_{\theta_0}(\ell_P(\psi; X)) - i^{rs}(\theta_\psi; \theta_0)\nu_{r,s}(\theta_\psi, \theta_\psi; \theta_0) + O(n^{-1}) \\ &= E_{\theta_0}(\ell_P(\psi; X)) - \mathrm{tr}\left[\nu_{\lambda,\lambda}(\theta_\psi, \theta_\psi; \theta_0) i_{\lambda\lambda}(\theta_\psi; \theta_0)^{-1}\right] + O(n^{-1}). \end{aligned} \quad (19)$$

Let us now denote by A twice the adjustment term of order $O_p(1)$ on the right-hand side of (17), 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 (18),

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

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

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

so that, neglecting additive constants,

$$E_{\theta_0}(A) = \text{tr}\left[\nu_{\lambda,\lambda}(\theta_\psi, \theta_\psi; \theta_0) i_{\lambda\lambda}(\theta_\psi; \theta_0)^{-1}\right] + O(n^{-1}), \tag{20}$$

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, Sect. 9.2.2).

From (19) and (20), using results in Sect. 2.1 and neglecting additive constants, we obtain, for $\psi - \psi_0 = O(n^{-1/2})$,

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

This shows relation (14).

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika, 62*, 547–554.
Arellano, M., Hahn, J. (2007). Understanding bias in nonlinear panel models: some recent developments. In: R. Blundell, W. Newey, T. Persson (Eds) *Advances in economics and econometrics, ninth world congress*, Vol.3, Ch.12. Cambridge: Cambridge University Press.
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. (1995). Stable and invariant adjusted profile likelihood and directed likelihood for curved exponential models. *Biometrika, 82*, 489–499.
Barndorff-Nielsen, O. E., Cox, D. R. (1994). *Inference and asymptotics*. London: Chapman and Hall.
Barndorff-Nielsen, O. E., Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli, 2*, 319–340.
Bellio, R., Sartori, N. (2006). Practical use of modified maximum likelihoods for stratified data. *Biometrical Journal, 48*, 876–886.
Corcuera, J. M., Giummolè, F. (2000). First order optimal predictive densities. In: P. Marriott, M. Salmon, (Eds), *Applications of differential geometry to econometrics* (pp. 214–229). Cambridge: Cambridge University Press.
Cox, D. R., Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, B, 49*, 1–39.
Datta, G. S., Mukerjee, R. (2004). *Probability matching priors: higher order asymptotics*. New York: Springer.
DiCiccio, T. J., Martin, M. A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *Journal of the Royal Statistical Society, B, 55*, 305–316.
DiCiccio, T. J., Stern, S. E. (1994). Constructing approximately standard normal pivots from signed roots of adjusted likelihood ratio statistics. *Scandinavian Journal of Statistics, 21*, 447–460.
DiCiccio, T. J., Martin, M. A., Stern, S. E., Young, G. A. (1996). Information bias and adjusted profile likelihoods. *Journal of the Royal Statistical Society, B, 58*, 189–203.
Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika, 76*, 675–684.
Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika, 83*, 299–313.

McCullagh, P., Tibshirani, R. (1990). A simple method for the adjustement of profile likelihoods. *Journal of the Royal Statistical Society, B, 52*, 325–344.

Pace, L., Salvan, A. (1997). *Principles of statistical inference*. Singapore: World Scientific.

Pace, L., Salvan, A. (2006). Adjustments of the profile likelihood from a new perspective. *Journal of Statistical Planning and Inference, 136*, 3554–3564.

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

Sartori, N., Bellio, R., Salvan, A., Pace, L. (1999). The directed modified profile likelihood in models with many nuisance parameters. *Biometrika, 86*, 735–742.

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

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

Severini, T. (2000). *Likelihood methods in statistics*. Oxford: Oxford University Press.

Severini, T. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika, 94*, 529–542.

Stern, S. E. (1997). A second-order adjustment to the profile likelihood in the case of a multidimensional parameter of interest. *Journal of the Royal Statistical Society, B, 59*, 653–665.

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

Vidoni, P. (2008). Improved predictive model selection. *Journal of Statistical Planning and Inference, 138*, 3713–3721.

Young, G. A., Smith, R. L. (2005). *Essentials of statistical inference*. Cambridge: Cambridge University Press.