Ann. Inst. Statist. Math.
Vol. 52, No. 1, 57–70 (2000)

MODEL SELECTION USING THE ESTIMATIVE AND THE APPROXIMATE p^* PREDICTIVE DENSITIES†

PAOLO VIDONI

Department of Statistics, University of Udine, via Treppo 18, I-33100 Udine, Italy

(Received December 16, 1996; revised August 20, 1998)

Abstract. Model selection procedures, based on a simple cross-validation technique and on suitable predictive densities, are taken into account. In particular, the selection criterion involving the estimative predictive density is recalled and a procedure based on the approximate p^* predictive density is defined. This new model selection procedure, compared with some other well-known techniques on the basis of the squared prediction error, gives satisfactory results. Moreover, higher-order asymptotic expansions for the selection statistics based on the estimative and the approximate p^* predictive densities are derived, whenever a natural exponential model is assumed. These approximations correspond to meaningful modifications of the Akaike's model selection statistic.

Key words and phrases: Akaike's criterion, cross-validation procedure, misspecification statistic, natural exponential model, predictive sample reuse method, squared prediction error.

1. Introduction

This paper concerns the problem of selecting a suitable model from a class of plausible statistical models and, for this purpose, selection criteria based on a simple cross-validation technique and involving some well-known predictive densities are primarily considered. Whenever a particular model is selected, it has to be viewed as an adequate description, which may be fruitfully employed for some further inferential or predictive analysis, of the underlying phenomenon, rather than a true representation of the process which has generated a given set of data.

Let $X_1, \ldots, X_n$ be n independent and identically distributed random variables and let $\boldsymbol{M}_j$, $j = 1, \ldots, k$, be k plausible statistical models. The probability density functions, specified within the model $\boldsymbol{M}_j$, $j = 1, \ldots, k$, constitute the family $F_j = \{p_j(x; \omega_j), \omega_j \in \Omega_j \subseteq \boldsymbol{R}^{d_j}\}$, $j = 1, \ldots, k$, where ω_j is an unknown d_j-dimensional parameter. In the following, natural exponential families are mainly considered as statistical models.

Suppose that the set of data $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a realisation of the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$. The aim here is not to choose the model which has generated the observation but to choose, within $\boldsymbol{M}_j$, $j = 1, \ldots, k$, the model which offers the most satisfactory explanation to the data. Thus, the key question is not which model is correct, but, rather, which model would serve best the purpose of the analysis. Here, the selected model is supposed to be used as a basis for making predictive inference and, for this reason, the alternative selection procedure are compared in term of their predictive capability.

† This research was partially supported by the Italian National Research Council grant n.96.01542.CT10.

In this context, model selection criteria based on a predictive approach are taken into account and attention is devoted to procedures which may be applied to non-nested models as well; that is, to the case where an arbitrary member or subset of F_j, $j = 1, \ldots k$, can not be obtained by imposing constraints on the parameters of any F_i, $i \neq j$ or as a limit in distribution of members of any F_i, $i \neq j$. Geisser ((1993), Chapter 4) reviews the main techniques for selecting statistical models adopting a predictive viewpoint and Clayton *et al.* (1986) make an interesting comparison of several predictive and non-predictive model selection criteria.

This paper mainly concerns cross-validation or, adopting the terminology proposed by Geisser and Eddy (1979), predictive sample reuse procedures, based on suitable predictive densities. In particular, the selection criterion involving the estimative predictive density is recalled and a new one, based on the approximate p^* predictive density (Vidoni (1995)), is introduced. By means of a simple preliminary simulation study, this new selection procedure is compared with other four alternative criteria on the basis of their predictive capability. The results, concerning the estimation of the squared prediction error, show that the approximate p^* predictive density gives rise to a useful predictive selection procedure.

Finally, when the assumed statistical model is a natural exponential family, meaningful higher-order asymptotic expansions for the selection statistics based on the estimative and the approximate p^* predictive densities are derived. These results enable an interesting characterisation of these two model selection techniques and allow a substantial simplification in the selection statistics involved, which can be useful for computations.

2. Predictive density functions

2.1 *Preliminaries*

This section provides a brief review on predictive densities and their major properties. Only non-Bayesian predictive densities are taken into account.

Let us consider the simple situation where the observable random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ consists of independent identically distributed observations on a random variable X, having probability density function $p(x;\omega)$, $\omega \in \Omega \subseteq \boldsymbol{R}^d$. The future or as yet unobserved random variable Z is independent of $\boldsymbol{X}$ and has the same distribution as X. Any estimator of the true probability density function $p(z;\omega)$ of the future random variable Z, based on the sample $\boldsymbol{X}$, is called a predictive density function.

The simplest approach to prediction consists in using the estimative probability density function $p_e(z) = p(z;\hat{\omega})$ obtained by substituting $\hat{\omega} = \hat{\omega}(\boldsymbol{X})$ for ω; $\hat{\omega}$ is an appropriate estimator of ω, usually the maximum likelihood estimator. In spite of its intuitiveness, $p_e(z)$ may not be entirely adequate for prediction, especially when the dimension of ω is large in comparison with n.

A number of recent papers aim to improve the estimative density. The contribution of Harris (1989), Vidoni (1995) and Komaki (1996), where the goodness of the approximation is measured by the Kullback-Liebler divergence, are in this direction. In particular, Harris (1989) proposed the parametric bootstrap predictive density, given by

$$p_{pb}(z;\omega) = \int p(z;t)p_{\hat{\omega}}(t;\omega)dt, \tag{2.1}$$

computed at $\omega = \hat{\omega}$, where $p_{\hat{\omega}}(\cdot;\omega)$ is the probability density function of the maximum likelihood estimator. Density (2.1) has some desirable properties; namely, within natural

exponential models, it is asymptotically superior to $p_e(z)$ in terms of average Kullback-Leibler divergence. Unfortunately, it is usually not in a reasonable closed form and it needs to be computed numerically even for simple models.

Vidoni (1995) pointed out that, although Harris's proposal is often unsuitable for exact calculations, it allows fairly simple approximations through straightforward asymptotic arguments. In particular, when $\hat{\omega}(\boldsymbol{X})$ is a sufficient statistic, such as within natural exponential models, it is possible to derive an high-order, closed-form approximation to (2.1), which consists of approximating $p_{\hat{\omega}}(\cdot;\omega)$ by Barndorff-Nielsen's (1983) p^*-formula and then using a Laplace approximation with $O(n^{-1})$ correction terms for integrating out the parameter.

When the maximum likelihood estimator is not itself a sufficient statistic, an appropriate sampling distribution, according to the conditionality principle, to be used as a weighting function in (2.1) is the conditional density of $\hat{\omega}(\boldsymbol{X})$, given an ancillary statistic a. This predictive density is called conditional parametric bootstrap predictive density and it may be approximated using the same asymptotic arguments of the previous case, with the Barndorff-Nielsen's p^*-formula considered in the conditional form. These approximations define the approximate p^* predictive density, which appears in the form $\tilde{p}_{p^*}(z;\hat{\omega}) = p(z;\hat{\omega})\{1 + \frac{1}{2}H(z;\hat{\omega},a)\}$, where the term $H(z;\hat{\omega},a)$, given explicitly by Vidoni (1995), is of order $O(n^{-1})$ and involves, as variable terms, the first two derivatives of $\ell(\omega;z) = \log p(z;\omega)$ with respect to ω, evaluated at $\omega = \hat{\omega}$, and, as coefficients, likelihood quantities based on the observable random sample $\boldsymbol{X}$.

2.2 *The approximate p^* predictive density for natural exponential models*

Since the selection procedures considered in the next sections are mainly applied for discriminating between natural exponential families, a brief idea of the derivation of $\tilde{p}_{p^*}(z;\hat{\omega})$ for these models is given below. For the general case, see Vidoni (1995). Hereafter it is convenient to use index notation and the Einstein summation convention, according to which if an index occurs more than once in a single term then summation over that index is understood.

Let us assume that the statistical model for each observation $X_1,\ldots,X_n$, Z is a natural exponential family of order d, with density function given by

$$p(x;\omega) = h(x)\exp\{\omega^r x^r - K(\omega)\} \tag{2.2}$$

where $x = (x^1,\ldots,x^d)$ and $\omega = (\omega^1,\ldots,\omega^d) \in \Omega \subseteq \boldsymbol{R}^d$. In this case, the p^*-formula (see Barndorff-Nielsen (1983)), which gives an approximation to the density of the maximum likelihood estimator $\hat{\omega}(\boldsymbol{X})$, with relative error of order $O(n^{-3/2})$, is

$$p^*(\hat{\omega};\omega) = c(\omega)|j(\hat{\omega})|^{1/2}\exp\{\ell_x(\omega;\hat{\omega}) - \ell_x(\hat{\omega};\hat{\omega})\}. \tag{2.3}$$

Here, $\ell_x(\omega;\hat{\omega}) = \ell_x(\omega;\boldsymbol{x}) = n\{K_r(\hat{\omega})\omega^r - K(\omega)\}$ is the log-likelihood function, $|j(\hat{\omega})| = |[nK_{rs}(\hat{\omega})]|$ is the determinant of the observed information matrix evaluated at $\omega = \hat{\omega}$ and $c(\omega)$ is a normalising constant; $K_r(\hat{\omega})$ and $K_{rs}(\hat{\omega})$ are the partial derivatives of $K(\omega)$ with respect to the corresponding components of ω, computed at $\omega = \hat{\omega}$.

By using (2.3) as a weighing function in (2.1), a predictive density can be defined as

$$p_{p^*}(z;w) = \int p(z;t)p^*(t;\omega)dt = \int p(z;t)c(\omega)|j(t)|^{1/2}\exp\{\ell_x(\omega;t) - \ell_x(t;t)\}dt, \tag{2.4}$$

computed at $\omega = \hat{\omega}$, which is an approximation to the parametric bootstrap predictive density, with relative error of order $O(n^{-3/2})$. The predictive density (2.4) can be further

approximated, retaining the same order of error, in the following two steps. First, by writing explicitly the normalising constant $c(\omega)$, the function $p_{p^*}(z;\hat{\omega})$ may be express as a ratio of integrals in the standard form (Tierney *et al.* (1989)); that is,

$$p_{p^*}(z;\omega) = \frac{\int p(z;t)|j(t)|^{1/2}\exp\{-r(t;\omega)\}dt}{\int |j(t)|^{1/2}\exp\{-r(t;\omega)\}dt}, \tag{2.5}$$

computed at $\omega = \hat{\omega}$, with $r(t;\omega) = \ell_x(t;t) - \ell_x(\omega;t) = n[K_r(t)(t^r - \omega^r) - \{K(t) - K(\omega)\}]$. Secondly, since $r(t;\omega)$ is usually a smooth function with unique minimum at $t = \omega$, by applying to the numerator and to the denominator of (2.5) an higher-order Laplace approximation for the integral (Vidoni (1995), Equation A1), we obtain a relatively simple close form expression. This approximation defines the approximate p^* predictive density given, in this particular instance, by

$$\tilde{p}_{p^*}(z;\hat{\omega}) = p_e(z)\left\{1 + \frac{1}{2}H(z;\hat{\omega})\right\},$$

with

$$H(z;\hat{\omega}) = \{(z^r - \hat{K}_r)(z^s - \hat{K}_s)\hat{K}^{rs} - (z^r - \hat{K}_r)\hat{K}^{sr}\hat{K}^{tu}\hat{K}_{tus} - d\}n^{-1}, \tag{2.6}$$

where $\hat{K}_r = K_r(\hat{\omega})$, $\hat{K}^{rs}$ is the (r,s) element of the inverse of the matrix $[K_{rs}(\hat{\omega})]$ and $\hat{K}_{rst}$ is the third partial derivative of $K(\omega)$ with respect to the corresponding components of ω, evaluated at $\omega = \hat{\omega}$. This predictive density is an approximation, with relative error of order $O(n^{-3/2})$, to the parametric bootstrap predictive density, which maintains the superiority over the estimative distribution pointed out by Harris (1989), while being much simpler to compute.

Example 1. *The gamma distribution.* Suppose that $X_1, \ldots, X_n$, Z are mutually independent and identically distributed with a common gamma density

$$p(x;\lambda,\beta) = \{\Gamma(\beta)\}^{-1}x^{\beta-1}e^{-\lambda x}\lambda^{\beta} \qquad (x \geq 0, \lambda > 0, \beta > 0).$$

Let us consider the shape parameter β known. This is an exponential model, which can be expressed in the natural form (2.2), with $\omega = -\lambda$ and $K(\omega) = -\beta\log(-\omega)$. Moreover, the maximum likelihood estimator is $\hat{\omega} = -\beta\bar{X}^{-1} = -\beta n(\sum X_i)^{-1}$. Thus, it is not difficult to calculate the approximate p^* predictive density, which is

$$\tilde{p}_{p^*}(z;\hat{\omega}) = p(z;\beta/\bar{X},\beta)\left[1 + \frac{1}{2}\{z^2\beta/(\bar{X})^2 - 2z(\beta+1)/(\bar{X}) + \beta + 1\}n^{-1}\right].$$

For the more interesting situation with both the scale and the shape parameter unknown, see Vidoni (1995). With $\beta = 1$, we have the approximate p^* predictive density associated to an exponential distribution

$$\tilde{p}_{p^*}(z;\hat{\omega}) = p(z;1/\bar{X},1)\left[1 + \frac{1}{2}\{(z/\bar{X})^2 - 4(z/\bar{X}) + 2\}n^{-1}\right]. \tag{2.7}$$

Example 2. *The normal distribution.* Suppose that $X_1, \ldots, X_n, Z$ are mutually independent and normally distributed with both the mean μ and the variance σ^2 unknown. The normal distribution is a two-dimensional natural exponential family, with natural

observation $(x_1, x_2) = (x, x^2)$, natural parameter $\omega = (\omega_1, \omega_2) = (\mu\sigma^{-2}, -\frac{1}{2}\sigma^{-2})$ and $K(\omega) = -\frac{1}{2}\log(-2\omega_2) - \frac{1}{4}(\omega_1^2/\omega_2)$; moreover, $\hat{\mu} = \bar{X} = \frac{1}{n}\sum X_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$. By differentiating the function $K(\omega)$, it is not difficult to compute the correction term $H(z; \hat{\omega})$, associated to the normal model, and to determine the approximate p^* predictive density, which is

$$\tilde{p}_{p^*}(z; \hat{\omega}) = p(z; \hat{\mu}; \hat{\sigma}^2)\left[1 + \frac{1}{4}\{(z - \hat{\mu})^4/\hat{\sigma}^4 - 6(z - \hat{\mu})^2/\hat{\sigma}^2 + 3\}n^{-1}\right].$$

When σ^2 is known, the approximate p^* predictive density has the simple form

$$\tilde{p}_{p^*}(z; \hat{\omega}) = p(z; \hat{\mu}; \sigma^2)\left[1 + \frac{1}{2}\{(z - \hat{\mu})^2/\sigma^2 - 1\}n^{-1}\right].$$

3. A predictive approach to model selection

3.1 *Selection criteria based on predictive densities*

We consider now the use of predictive densities in model selection problems. In this framework, a possibility is to use a cross-validation, or predictive sample reuse, procedure and to select, with regard to the observation $\boldsymbol{x}$, the model which has, in some sense, the best predictive ability, according to the predictive density which is considered. In this section, the selection criterion based on the estimative predictive density is recalled and a new one, involving the approximate p^* predictive density, is defined.

The idea behind cross-validation techniques is to split the data into two parts and to use the first part to fit a model and the second to judge the goodness of the prediction based on the model which is considered. The simplest version of cross-validation consists of leaving out one observation at a time. Thus, the predictive density taken into account has to be computed using $\boldsymbol{x}_{(i)}$, the data with x_i omitted. Its value at $z = x_i$ shows how well the fitted model predicts the excluded data point; all the n subdivisions of $\boldsymbol{x}$ have to be considered.

Geisser and Eddy (1979) use the estimative predictive density $p_e(\cdot)$ and define a selection procedure, termed predictive sample reuse quasi-likelihood (PSRQL), which points to the model maximising the selection statistic

$$S_{QL} = \sum_{i=1}^{n} \log p(x_i; \hat{\omega}_{(i)}) = \sum_{i=1}^{n} \ell(\hat{\omega}_{(i)}; x_i), \tag{3.1}$$

where $\ell(\hat{\omega}_{(i)}; x_i) = \log p(x_i; \hat{\omega}_{(i)})$ and $\hat{\omega}_{(i)} = \hat{\omega}(\boldsymbol{x}_{(i)})$ is the maximum likelihood estimator based on $\boldsymbol{x}_{(i)}$. Hereafter, the index j, which labels the models, is omitted to simplify the notation. This criterion does not need a remarkable computational effort but it is based on the estimative predictive density, which, as pointed out in Section 2, could give inaccurate results, whenever the dimension of ω is large in comparison with n.

An alternative criterion, called predictive sample reuse quasi-likelihood p^* (PSRQLP*), may be defined by considering the approximate p^* predictive density $\tilde{p}_{p^*}(\cdot)$ instead of $p_e(\cdot)$. This criterion selects the model which maximises

$$S_{QLP^*} = \sum_{i=1}^{n} \log \tilde{p}_{p^*}(x_i; \hat{\omega}_{(i)}) = S_{QL} + \sum_{i=1}^{n} \log\left\{1 + \frac{1}{2}H(x_i; \hat{\omega}_{(i)}, a)\right\}. \tag{3.2}$$

For natural exponential families, the $O(n^{-1})$ modifying term $H(x_i; \hat{\omega}_{(i)}, a)$ is given by (2.6), with $z = x_i$ and $\hat{\omega} = \hat{\omega}_{(i)}$. This new model selection criterion is not as immediate to compute as the previous one but, at least within natural exponential models, it is based on a better estimator for the probability density function of the future observation. In particular, since for small values of n the improvement in terms of average Kullback-Liebler divergence is substantial, this discriminating procedure is expected to provide a better selection, according to the predictive capability of the alternative models.

3.2 *A simple simulation study*

The criterion based on the approximate p^* predictive density is compared with other four well-known selection procedures. Since the selected model is supposed to be considered for a subsequent prediction analysis, we would like to choose the model which best serves this purpose, among a set of competitors. In this respect, it seems natural to compare the alternative model selection procedures by assessing their predictive capability, that is the capability of choosing, in some sense, the best model for making predictive inference. Here, a simulation is performed in order to estimate the expected square error of prediction of the alternative procedures, with regard to a simple selection problem already considered by Geisser and Eddy (1979) and Clayton *et al.* (1986). This is only a preliminary study to have a first idea on the potential usefulness of the new criterion. In order to perform a deeper comparative analysis, further simulations are needed but this is beyond the scope of the present paper.

Let us suppose that there are two plausible statistical models, based on the simple exponential distribution, for describing a dichotomously labelled set of data $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2) = (x_{1,1}, \ldots, x_{1,n_1}, x_{2,1}, \ldots, x_{2,n_2})$, where $n_1 + n_2 = n$. Under $\boldsymbol{M}_1$, the associated random vector $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) = (X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2})$ is a set of independent random variables with density $p(x; \lambda) = \lambda \exp(-\lambda x)$, $x \geq 0$, $\lambda > 0$. Thus, the sampling distribution does not depend on the label. Under $\boldsymbol{M}_2$, the label is assumed to be relevant so that the independent random variables $X_{h,i}$, $h = 1, 2$, $i = 1, \ldots, n_h$, follow an exponential distribution with density $p(x; \lambda_h) = \lambda_h \exp(-\lambda_h x)$, $x \geq 0, \lambda_h > 0$, with $\lambda_1 \neq \lambda_2$.

The selection criterion based on the approximate p^* predictive density chooses the model with the largest $S_{QLP^*}(\boldsymbol{M}_j)$, $j = 1, 2$. According to (2.7) and (3.2), the selection statistics are

$$S_{QLP^*}(\boldsymbol{M}_1) = \sum_{h=1}^{2} \sum_{i=1}^{n_h} \Big[-\log(\bar{x}_{(h,i)}) - g_{h,i}(\boldsymbol{x}) + \log\Big\{1 + \frac{1}{2}(g_{h,i}(\boldsymbol{x})^2 - 4g_{h,i}(\boldsymbol{x}) + 2)(n-1)^{-1}\Big\}\Big],$$

where $g_{h,i}(\boldsymbol{x}) = x_{h,i}/\bar{x}_{(h,i)}$ and $\bar{x}_{(h,i)}$ is the sample mean based on $\boldsymbol{x}$ with $x_{h,i}$ omitted;

$$S_{QLP^*}(\boldsymbol{M}_2) = \sum_{h=1}^{2} \sum_{i=1}^{n_h} \Big[-\log(\bar{x}_{h(h,i)}) - q_{h,i}(\boldsymbol{x}_h) + \log\Big\{1 + \frac{1}{2}(q_{h,i}(\boldsymbol{x}_h)^2 - 4q_{h,i}(\boldsymbol{x}_h) + 2)(n_h - 1)^{-1}\Big\}\Big],$$

where $q_{h,i}(\boldsymbol{x}_h) = x_{h,i}/\bar{x}_{h(h,i)}$ and $\bar{x}_{h(h,i)}$ is the sample mean based on $\boldsymbol{x}_h$ with $x_{h,i}$ omitted. This procedure is compared with the Akaike's (1973) information criterion (AIC), the large sample Bayes criterion (LSB) proposed by Schwarz (1978), the PSRQL

criterion, previously mentioned, and a Bayesian predictive sample reuse criterion. This last procedure considers a Bayesian predictive density with a diffuse prior on the unknown parameter and it is termed predictive sample reuse quasi-Bayes (PSRQB) criterion. For these criteria the selection statistics can be found in Geisser and Eddy (1979) and Clayton *et al.* (1986).

The aim of the simulation is to estimate the expected square error of prediction associated to the five alternative selection procedures, for different parameter configurations and sample sizes. More precisely, we need to predict future observations from each population, namely Z_1 and Z_2, where Z_h, $h = 1, 2$, is distributed as an observation from the h-th population. Under $\boldsymbol{M}_1$ the two populations are the same. Given the predictors $\hat{Z}_1$ and $\hat{Z}_2$, the expected square error of prediction is

$$E\{(Z_1 - \hat{Z}_1)^2 + (Z_2 - \hat{Z}_2)^2\}$$
$$= \mathrm{Var}(Z_1) + \mathrm{Var}(Z_2) + E\{(\hat{Z}_1 - E(Z_1))^2\} + E\{(\hat{Z}_2 - E(Z_2))^2\};$$

thus, in the simulation, only the last two terms have to be estimated. Indeed, if $\boldsymbol{M}_1$ is selected, we consider $\hat{Z}_1 = \hat{Z}_2 = \bar{X}$, the sample mean based on $\boldsymbol{X}$, while, if $\boldsymbol{M}_2$ is selected, then $\hat{Z}_1 = \bar{X}_1$ and $\hat{Z}_2 = \bar{X}_2$, where $\bar{X}_h$, $h = 1, 2$, is the sample mean based on $\boldsymbol{X}_h$.

The estimates of the expected squared error of prediction are obtained by considering 10,000 samples for each sample size n, ranging from 8 to 40, with $n_1 = n_2 = n/2$. Different parameter configurations are considered, by fixing $\lambda_1 = 1$ and setting $\lambda_2 = 1(0.5)2.5$; the configuration $\lambda_1 = 1$, $\lambda_2 = 1$, clearly means that the model $\boldsymbol{M}_1$ is true, while $\lambda_1 = 1$, $\lambda_2 \neq 1$ refers to the model $\boldsymbol{M}_2$. The estimates are given in Table 1 and the corresponding standard errors are estimated to be between 0.005 and 0.001, as n increases from 8 to 40. The last column of Table 1 gives the expected squared error of prediction when λ_1 and λ_2 are known, namely when the true model is known. It represents the unavoidable part of the error occurring in the prediction, which is due to the variance of the future observations Z_1 and Z_2.

Inspection of the table gives a preliminary idea on the behaviour of the alternative criteria in this particular selection problem. As noted by Clayton *et al.* (1986), the LSB procedure performs better under $\boldsymbol{M}_1$ and usually poorer under $\boldsymbol{M}_2$. This is clearly related to the fact that the LSB criterion is consistent; that is, when the restricted model is true, the probability of correct selection tends to 1, as n increases. Furthermore, the other four methods do not present significantly different estimates and, at least in this simple example, they are indistinguishable with respect to the prediction error. A further additional comment is that the PSRQL method seems to perform a little better under $\boldsymbol{M}_2$, for small values of n, while the PSRQLP* criterion, under $\boldsymbol{M}_1$, it is usually closer to the LSB procedure than the other three criteria and it performs better than the Schwarz's (1978) criterion for $n_1 = n_2 = 4$.

Although a number of additional simulation are needed for a deeper analysis, these preliminary results emphasise that the approximate p^* predictive density, which is theoretically superior to the estimative one, gives rise to a predictive model selection procedure which is competitive with the existing criteria.

4. Asymptotic expansions for the selection statistics

4.1 *The procedure based on the estimative predictive density*

There are two main motivations behind the asymptotic expansions considered in the present section; namely, to obtain suitable higher-order approximations to S_{QL} and

Table 1. Estimates of the expected squared error of prediction.*

λ_2	Size of each sample	Selection criterion					
		AIC	BIC	PSRQB	PSRQL	PSRQLP*	λ_1, λ_2 known
1	4	2.399	2.395	2.396	2.395	2.383	2.00
	8	2.194	2.175	2.193	2.192	2.189	
	12	2.130	2.113	2.131	2.130	2.129	
	20	2.078	2.064	2.078	2.077	2.077	
1.5	4	1.776	1.775	1.775	1.770	1.768	1.44
	8	1.629	1.626	1.628	1.626	1.627	
	12	1.577	1.579	1.577	1.576	1.577	
	20	1.531	1.539	1.531	1.531	1.532	
2	4	1.587	1.587	1.585	1.578	1.582	1.25
	8	1.436	1.446	1.435	1.435	1.437	
	12	1.381	1.398	1.382	1.383	1.383	
	20	1.328	1.348	1.328	1.329	1.329	
2.5	4	1.500	1.502	1.500	1.494	1.502	1.16
	8	1.338	1.354	1.340	1.342	1.343	
	12	1.276	1.296	1.277	1.281	1.279	
	20	1.224	1.238	1.225	1.226	1.225	

*Based on 10,000 samples for each sample size with standard error estimated to be between 0.005 and 0.001, as n increases from 8 to 40.

S_{QLP^*}, which turns out to be easier to compute, and to find out the features of the alternative models, besides goodness-of-fit, which are involved in these model selection procedures. Here, index notation and Einstein summation convention are maintained; however, with a slight abuse of notation, summation over the index i, which refers to the predictive sample reuse technique, is expressed explicitly. Moreover, when the sign "$\bullet$" appears in a formula, it means that the terms following are asymptotically smaller of order at least $n^{-1/2}$ than the preceding ones; the sign "$\bullet\bullet$" means a drop of order n^{-1} and so on. This notation may be useful for an immediate determination of the order of the terms involved in an asymptotic expansion.

Let us consider S_{QL}, given by (3.1), and expand $\ell(\hat{\omega}_{(i)}; x_i)$ around $\hat{\omega}_{(i)} = \hat{\omega}$, in such a way that

$$S_{QL} = \sum_{i=1}^{n} \Big\{ \ell(\hat{\omega}; x_i) + (\hat{\omega}_{(i)} - \hat{\omega})^r \ell_r(\hat{\omega}; x_i) + \frac{1}{2}(\hat{\omega}_{(i)} - \hat{\omega})^{rs} \ell_{rs}(\hat{\omega}; x_i) + O_p(n^{-3}) \Big\}, \tag{4.1}$$

where $(\hat{\omega}_{(i)} - \hat{\omega})^{rs} = (\hat{\omega}_{(i)} - \hat{\omega})^r (\hat{\omega}_{(i)} - \hat{\omega})^s$ and $\ell_r(\hat{\omega}; x_i)$, $\ell_{rs}(\hat{\omega}; x_i)$ are the partial derivatives of $\ell(\omega, x_i) = \log p(x; \omega)$ with respect to the corresponding components of ω, evaluated at $\omega = \hat{\omega}$. In order to obtain an alternative expression for S_{QL}, the following Taylor expansion, around $\hat{\omega}_{(i)} = \hat{\omega}$, is useful

$$\ell_r(\hat{\omega}_{(i)}; \boldsymbol{x}) = \ell_r(\hat{\omega}; \boldsymbol{x}) + (\hat{\omega}_{(i)} - \hat{\omega})^s \ell_{rs}(\hat{\omega}; \boldsymbol{x}) \tag{4.2}$$

$$+\frac{1}{2}(\hat{\omega}_{(i)}-\hat{\omega})^{st}\ell_{rst}(\hat{\omega};\boldsymbol{x})+O_p(n^{-2}).$$

Here $\ell_{R_m}(\hat{\omega}_{(i)};\boldsymbol{x})$ and $\ell_{R_m}(\hat{\omega};\boldsymbol{x})$, with $R_m=(r_1,\ldots,r_m)$, $m\in\mathbf{N}^+$, are the m-th partial derivatives of the log-likelihood function $\ell(\omega;\boldsymbol{x})=\sum\ell(\omega;x_i)$, with respect to the components of ω with indices in R_m, computed at $\omega=\hat{\omega}_{(i)}$ and $\omega=\hat{\omega}$, respectively. Since $\hat{\omega}$ and $\hat{\omega}_{(i)}$ are such that $\ell_r(\hat{\omega};\boldsymbol{x})=0$ and $\ell_r(\hat{\omega}_{(i)};\boldsymbol{x})=\ell_r(\hat{\omega}_{(i)};x_i)$, multiplication of (4.2) by $\ell^{ru}(\hat{\omega};\boldsymbol{x})$, namely the (r,u) element of the inverse of the matrix $[\ell_{ru}(\hat{\omega};\boldsymbol{x})]$, gives

$$(4.3)\quad (\hat{\omega}_{(i)}-\hat{\omega})^u=\ell_r(\hat{\omega}_{(i)};x_i)\hat{\ell}^{ru}\overset{\bullet\bullet}{-}\frac{1}{2}\ell_v(\hat{\omega}_{(i)};x_i)\ell_w(\hat{\omega}_{(i)};x_i)\hat{\ell}^{vs}\hat{\ell}^{wt}\hat{\ell}^{ru}\hat{\ell}_{rst}\overset{\bullet\bullet}{+}O_p(n^{-3}).$$

Here $\hat{\ell}_{R_m}=\ell_{R_m}(\hat{\omega};\boldsymbol{x})$, with $R_m=(r_1,\ldots,r_m)$, $m\in\mathbf{N}^+$; moreover, (4.3) assures that $(\hat{\omega}_{(i)}-\hat{\omega})^u=O_p(n^{-1})$. Replacing (4.3) in (4.1), the selection statistic S_{QL} can be rewritten as a modification of the profile log-likelihood function $\ell(\hat{\omega};\boldsymbol{x})$ such as

$$\begin{aligned}(4.4)\qquad S_{QL}=\ell(\hat{\omega};\boldsymbol{x})&\overset{\bullet\bullet}{+}\sum_{i=1}^{n}\{\ell_r(\hat{\omega};x_i)\ell_s(\hat{\omega};x_i)\}\hat{\ell}^{rs}\\&\overset{\bullet\bullet}{+}\frac{3}{2}\sum_{i=1}^{n}\{\ell_r(\hat{\omega};x_i)\ell_s(\hat{\omega};x_i)\ell_{tu}(\hat{\omega};x_i)\}\hat{\ell}^{rt}\hat{\ell}^{su}\\&-\frac{1}{2}\sum_{i=1}^{n}\{\ell_r(\hat{\omega};x_i)\ell_s(\hat{\omega};x_i)\ell_t(\hat{\omega};x_i)\}\hat{\ell}^{ru}\hat{\ell}^{sv}\hat{\ell}^{tw}\hat{\ell}_{vuw}\overset{\bullet\bullet}{+}O_p(n^{-2}).\end{aligned}$$

Whenever the assumed statistical model coincides with the true one, the following asymptotic relations hold (see, for example, Barndorff-Nielsen and Cox (1994), Chapter 5)

$$\frac{1}{n}\sum_{i=1}^{n}\ell_r(\hat{\omega};x_i)\ell_s(\hat{\omega};x_i)=\upsilon_{r,s}+o_p(1),\qquad \frac{1}{n}\ell_{rs}(\hat{\omega};\boldsymbol{x})=\upsilon_{rs}+o_p(1),$$

with $\upsilon_{rs}=E\{\ell_{rs}(\omega;X);\omega\}$ and $\upsilon_{r,s}=E\{\ell_r(\omega;X)\ell_s(\omega;X);\omega\}$. As a consequence of the well-known identity $\upsilon_{rs}=-\upsilon_{r,s}$, it is almost immediate to obtain

$$S_{QL}=\ell(\hat{\omega};\boldsymbol{x})+\upsilon_{r,s}\upsilon^{rs}+o_p(1)=\ell(\hat{\omega};\boldsymbol{x})-d+o_p(1),$$

where υ^{rs} is the (r,s) element of the inverse of the matrix $[\upsilon_{rs}]$. Note that $\ell(\hat{\omega};\boldsymbol{x})-d$ defines the Akaike's (1973) model selection criterion; thus, as shown in Stone (1977), the cross-validation procedure based on the estimative predictive distribution turns out to be first-order equivalent to the Akaike's selection statistic. However, this equivalence is only a preliminary feature of S_{QL} and, as shown below for natural exponential families, more information on S_{QL} is gained by considering the higher-order terms in the asymptotic expansion (4.4).

For natural exponential models, with probability density function given by (2.2), we have

$$\begin{aligned}&\ell_r(\omega;x_i)=x_i^r-K_r(\omega),\qquad \ell_{R_m}(\omega;x_i)=-K_{R_m}(\omega),\\(4.5)\qquad&\ell_r(\omega;\boldsymbol{x})=\sum_{i=1}^{n}x_i^r-nK_r(\omega),\qquad \ell_{R_m}(\omega;\boldsymbol{x})=-nK_{R_m}(\omega),\end{aligned}$$

with $R_m = (r_1, \ldots, r_m)$, $m \in \mathbf{N}^+/\{1\}$. Let us define

$$\frac{1}{n}\hat{M}_{rs} = \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \hat{K})^{rs} - \hat{K}_{rs}\}, \qquad \frac{1}{n}\hat{M}_{rst} = \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \hat{K})^{rs} - \hat{K}_{rst}\},$$

$$\frac{1}{n}\hat{M}_{rstu} = \frac{1}{n}\Big\{\sum_{i=1}^{n}(x_i - \hat{K})^{rstu} - n\hat{K}_{rstu} + n[3]\hat{K}_{rs}\hat{K}_{tu} - [6]\sum_{i=1}^{n}(x_i - \hat{K})^{rs}\hat{K}_{tu}\Big\}, \tag{4.6}$$

where $[h]$, $h \in \mathbf{N}^+$, indicates a sum of h terms obtained by permutation of the subscripts and $(x_i - \hat{K})^{rs} = (x_i^r - \hat{K}_r)(x_i^s - \hat{K}_s)$, $(x_i - \hat{K})^{rst} = (x_i - \hat{K})^{rs}(x_i^t - \hat{K}_t)$ and $(x_i - \hat{K})^{rstu} = (x_i - \hat{K})^{rst}(x_i^u - \hat{K}_u)$. Relations (4.6) correspond to the sample version of the quantities constituting, respectively, the second, third and fourth expected balance relations, computed at $\omega = \hat{\omega}$. The expected balance relations, also called Bartlett identities, are obtained by differentiating $\upsilon_r = E\{\ell_r(\omega; X); \omega\} = 0$, with respect to the components of ω (see, for example, Barndorff-Nielsen and Cox (1994), Section 5.2). In particular, the second expected balance relation corresponds to the above mentioned identity $\upsilon_{rs} = -\upsilon_{r,s}$, while the third and the fourth expected balance relations are

$$\upsilon_{rst} + \upsilon_{r,st}[3] + \upsilon_{r,s,t} = 0,$$
$$\upsilon_{rstu} + \upsilon_{r,stu}[4] + \upsilon_{rs,tu}[3] + \upsilon_{r,s,tu}[6] + \upsilon_{r,s,t,u} = 0,$$

with $\upsilon_{R_m,\ldots,S_h} = E\{\ell_{R_m}(\omega; X)\cdots\ell_{S_h}(\omega; X); \omega\}$, $R_m = (r_1, \ldots, r_m)$, $S_h = (s_1, \ldots, s_h)$, $m, h \in \mathbf{N}^+$. If the assumed statistical model is correctly specified, the sample statistics (4.6) are of order $O_p(n^{-1/2})$, otherwise they present values which differ systematically from zero. In the first case, by considering (4.5) and (4.6), formula (4.4) may be rewritten as

$$S_{QL} = \ell(\hat{\omega}; \boldsymbol{x}) \overset{\bullet\bullet}{-} d \overset{\bullet}{-} \hat{M}_{rs}\hat{K}^{rs}n^{-1} \overset{\bullet}{-} \frac{3}{2}dn^{-1} - \frac{1}{2}\hat{R}_{23}n^{-1} \overset{\bullet}{-} \frac{3}{2}\hat{M}_{rs}\hat{K}^{rs}n^{-2} - \frac{1}{2}\hat{M}_{rst}\hat{K}^{ru}\hat{K}^{sv}\hat{K}^{tw}\hat{K}_{uvw}n^{-2} \overset{\bullet}{+} O_p(n^{-2}), \tag{4.7}$$

where $\hat{R}_{23} = \hat{K}_{rst}\hat{K}_{uvw}\hat{K}^{ru}\hat{K}^{sv}\hat{K}^{tw}$ is a multivariate generalisation of the square of the third standardized cumulant (McCullagh (1987), Section 2.8), evaluated at $\omega = \hat{\omega}$. Thus, S_{QL} can be viewed as a modification of the Akaike's model selection statistic involving the dimension of the parameter space, a sample index of skewness and the sample version of the second and the third expected balance relations. Moreover, formula (4.7) provides, for natural exponential models, a simple expression for the higher-order terms not considered in the approximation given by Stone (1977).

4.2 *The procedure based on the approximate p^* predictive density*

Assuming an underlying natural exponential model, a higher-order asymptotic expansion may be derived for S_{QLP^*} as well. By expanding the logarithmic function in (3.2), we have

$$S_{QLP^*} = \sum_{i=1}^{n}\ell(\hat{\omega}_{(i)}; x_i) \overset{\bullet\bullet}{+} A \overset{\bullet\bullet}{+} B \overset{\bullet\bullet}{+} O_p(n^{-2}), \tag{4.8}$$

where $A = \frac{1}{2}\sum H(x_i;\hat{\omega}_{(i)})$ and $B = -\frac{1}{8}\sum\{H(x_i;\hat{\omega}_{(i)})\}^2$, with the $O(n^{-1})$ term $H(x_i;\hat{\omega}_{(i)})$ given by (2.6), with x_i and $\hat{\omega}_{(i)}$ substituted for z and $\hat{\omega}$.

The first term in (4.8) is equal to S_{QL}, which may be approximated by (4.7). In order to obtain an asymptotic approximation for the second and the third terms in (4.8), it is necessary to consider the following stochastic Taylor expansions, around $\hat{\omega}_{(i)} = \hat{\omega}$

$$
\begin{aligned}
K_r(\hat{\omega}_{(i)}) &= \hat{K}_r + (\hat{\omega}_{(i)} - \hat{\omega})^s \hat{K}_{rs} + O_p(n^{-2}),\\
K^{rs}(\hat{\omega}_{(i)}) &= \hat{K}^{rs} + (\hat{\omega}_{(i)} - \hat{\omega})^t \hat{K}^{ru}\hat{K}^{sv}\hat{K}_{tuv} + O_p(n^{-2}),\\
K_{rst}(\hat{\omega}_{(i)}) &= \hat{K}_{rst} + (\hat{\omega}_{(i)} - \hat{\omega})^u \hat{K}_{rstu} + O_p(n^{-2})
\end{aligned}
$$

and the relation (4.3) computed for natural exponential families. By means of an algebraic procedure similar to that considered for S_{QL}, since $\hat{K}^{rs} = \hat{K}_{tu}\hat{K}^{rt}\hat{K}^{su}$, $\hat{\omega}_{(i)} = \hat{\omega} + O_p(n^{-1})$ and $\hat{K}_{tu}\hat{K}^{tu} = d$, it follows that

$$
\begin{aligned}
(4.9)\qquad S_{QLP^*} = \ell(\hat{\omega};\boldsymbol{x}) &\overset{\bullet\bullet}{-} \frac{1}{2}d - \frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{K})^{rs}\hat{K}^{rs}n^{-1} \overset{\bullet\bullet}{-} \frac{1}{2}dn^{-1} - \frac{1}{8}d^2n^{-1}\\
&+ \frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{K})^{rs}\\
&\cdot\left\{\frac{1}{2}d\hat{K}^{rs} + \hat{K}^{rt}\hat{K}^{su}\hat{K}^{vw}\hat{K}_{tuvw}\right.\\
&\quad\left. - \hat{K}^{rt}\hat{K}^{su}\hat{K}^{vw}\hat{K}^{pq}\left(\hat{K}_{twp}\hat{K}_{uvq} + \frac{5}{4}\hat{K}_{tuv}\hat{K}_{wpq}\right)\right\}n^{-2}\\
&+ \frac{1}{4}\sum_{i=1}^{n}(x_i - \hat{K})^{rst}\hat{K}^{rs}\hat{K}^{tu}\hat{K}^{vw}\hat{K}_{uvw}n^{-2}\\
&- \frac{1}{8}\sum_{i=1}^{n}(x_i - \hat{K})^{rstu}\hat{K}^{rs}\hat{K}^{tu}n^{-2} \overset{\bullet\bullet}{+} O_p(n^{-2}).
\end{aligned}
$$

Finally, by adding and subtracting conveniently $\hat{K}_{rs}$, $\hat{K}_{rst}$ and $-\hat{K}_{rstu} + [3]\hat{K}_{rs}\hat{K}_{tu} - [6](x_i - \hat{K})^{rs}\hat{K}^{rs}$, formula (4.9) may be rewritten as

$$
\begin{aligned}
(4.10)\qquad S_{QLP^*} = \ell(\hat{\omega};\boldsymbol{x}) &\overset{\bullet\bullet}{-} d \overset{\bullet}{-} \frac{1}{2}\hat{M}_{rs}\hat{K}^{rs}n^{-1} \overset{\bullet}{-} \frac{1}{2}dn^{-1} - \frac{1}{4}d^2n^{-1} - \frac{1}{2}\hat{R}_{23}n^{-1}\\
&- \frac{3}{8}\hat{R}_{13}n^{-1} + \frac{3}{8}\hat{R}_4n^{-1} \overset{\bullet}{-} \frac{1}{2}d\hat{M}_{rs}\hat{K}^{rs}n^{-2} + \frac{1}{2}\hat{M}_{rs}\hat{K}^{rt}\hat{K}^{su}\hat{K}^{vw}\\
&\cdot\left(\hat{K}_{tuvw} - \hat{K}^{pq}\hat{K}_{twp}\hat{K}_{uvq} - \frac{5}{4}\hat{K}^{pq}\hat{K}_{tuv}\hat{K}_{wpq}\right)n^{-2}\\
&+ \frac{1}{4}\hat{M}_{rst}\hat{K}^{rs}\hat{K}^{tu}\hat{K}^{vw}\hat{K}_{uvw}n^{-2} - \frac{1}{8}\hat{M}_{rstu}\hat{K}^{rs}\hat{K}^{tu}n^{-2}\\
&\overset{\bullet}{+} O_p(n^{-2}),
\end{aligned}
$$

where $\hat{R}_{13} = \hat{K}_{rst}\hat{K}_{uvw}\hat{K}^{rs}\hat{K}^{tu}\hat{K}^{vw}$ and $\hat{R}_{23} = \hat{K}_{rst}\hat{K}_{uvw}\hat{K}^{ru}\hat{K}^{sv}\hat{K}^{tw}$ are the multivariate generalisations of the square of the third standardized cumulant, computed at $\omega = \hat{\omega}$, and $\hat{R}_4 = \hat{K}_{rstu}\hat{K}^{rs}\hat{K}^{tu}$ is the multivariate generalisation of the fourth standardized cumulant, computed at $\omega = \hat{\omega}$ (McCullagh (1987), Section 2.8). The sample statistics $\hat{M}_{rs}$, $\hat{M}_{rst}$ and $\hat{M}_{rstu}$ are defined by (4.6).

The final approximation (4.10) holds whenever the assumed statistical model is correctly specified and it is a suitable modification of the profile log-likelihood function $\ell(\hat{\omega}; \boldsymbol{x})$ which, up to terms of order $O_p(1)$, corresponds to the Akaike's (1973) criterion. Here, the modifying term involves further quantities beyond those given in the expansion for S_{QL}. In particular, we can find in (4.10) the sample version of the fourth expected balance relation, computed at $\omega = \hat{\omega}$, and the standardised cumulants $\hat{R}_{13}$ and $\hat{R}_4$, which do not appear in the asymptotic expansion for S_{QL}.

5. Comments and conclusions

This paper provides a new model selection criterion involving the approximate p^* predictive density, by means of a simple cross-validation technique. A preliminary simulation study shows that this new model selection procedure, compared with some other well-known techniques on the basis of the squared prediction error, gives satisfactory results. Moreover, higher-order asymptotic expansions for the selection statistics based on the estimative and the approximate p^* predictive distribution are derived. These results are given, in particular, for natural exponential models and may be fruitfully considered, up to terms of order $O_p(n^{-1})$, as a simple alternative to the selection statistics (3.1) and (3.2), when the dimension of the observed sample $\boldsymbol{x}$ is large. These approximations are obtained with the key assumption that the model is correctly specified. Although, within a model selection procedure, such an assumption is not completely plausible, one can still use these approximate selection statistics when the assumed statistical models are not remarkably different from the true one. Otherwise, it may be convenient to consider the original selection statistics or the expansions (4.4) and (4.9), which do not require any underlying hypothesis on the true model.

Example 1. (*continued*) *The gamma distribution.* The approximations (4.7) and (4.10), up to terms of order $O_p(n^{-1})$, are almost immediate to compute for the gamma distribution. In particular, if the shape parameter β is known, $\hat{R}_{13} = \hat{R}_{23} = 4/\beta$, $\hat{R}_4 = 6/\beta$ and then

$$\begin{aligned} S_{QL} &= \sum_{i=1}^{n} \log p(x_i; \beta/\bar{x}, \beta) - 1 - \sum_{i=1}^{n} \{\beta(x_i - \bar{x})^2/(\bar{x})^2 - 1\} n^{-1} \\ &\quad - \frac{1}{2}(3 + 4\beta^{-1}) n^{-1} + O_p(n^{-3/2}), \\ S_{QLP^*} &= \sum_{i=1}^{n} \log p(x_i; \beta/\bar{x}, \beta) - 1 - \frac{1}{2} \sum_{i=1}^{n} \{\beta(x_i - \bar{x})^2/(\bar{x})^2 - 1\} n^{-1} \\ &\quad - \frac{1}{4}(3 + 5\beta^{-1}) n^{-1} + O_p(n^{-3/2}), \end{aligned}$$

where $\bar{x}$ is the sample mean based on $\boldsymbol{x}$. With $\beta = 1$, we obtain the approximations associated to an exponential distribution.

Example 2. (*continued*) *The normal distribution.* Whenever a normal distribution with both the mean μ and the variance σ^2 unknown is assumed, the computations require an additional effort. Since $(x_1, x_2) = (x, x^2)$, $\omega = (\omega_1, \omega_2) = (\mu\sigma^{-2}, -\frac{1}{2}\sigma^{-2})$, and $K(\omega) = -\frac{1}{2}\log(-2\omega_2) - \frac{1}{4}(\omega_1^2/\omega_2)$ we obtain the partial derivatives $K_1(\omega) = \mu$, $K_2(\omega) = \sigma^2 + \mu^2$, $K_{11}(\omega) = \sigma^2$, $K_{12}(\omega) = K_{21}(\omega) = 2\mu\sigma^2$ and $K_{22}(\omega) = 2\sigma^4 + 4\mu^2\sigma^2$;

moreover $R_{13} = R_{23} = R_4 = 0$. Thus, the approximations (4.7) and (4.10), up to terms of order $O_p(n^{-1})$, are

$$S_{QL} = \sum_{i=1}^{n} \log p(x_i; \hat{\mu}, \hat{\sigma}^2) - 2 - \hat{M}_{rs}\hat{K}^{rs}n^{-1} - 3n^{-1} + O_p(n^{-3/2}),$$

$$S_{QLP^*} = \sum_{i=1}^{n} \log p(x_i; \hat{\mu}, \hat{\sigma}^2) - 2 - \frac{1}{2}\hat{M}_{rs}\hat{K}^{rs}n^{-1} - 2n^{-1} + O_p(n^{-3/2}).$$

Here

$$\hat{M}_{rs}\hat{K}^{rs}n^{-1} = \sum_{i=1}^{n} \left\{ \hat{f}_1(x_i)^2(\hat{\sigma}^2 + 2\hat{\mu}^2)/\hat{\sigma}^2 - (2\hat{\mu}/\hat{\sigma})\hat{f}_1(x_i)\hat{f}_2(x_i) + \frac{1}{2}\hat{f}_2(x_i)^2 - 2 \right\} n^{-1},$$

with $\hat{f}_1(x_i) = (x_i - \hat{\mu})/\hat{\sigma}$, $\hat{f}_2(x_i) = (x_i^2 - \hat{\sigma}^2 - \hat{\mu}^2)/\hat{\sigma}^2$ and $\hat{\mu}$, $\hat{\sigma}^2$ the maximum likelihood estimates based on $\boldsymbol{x}$. With $\hat{\sigma}^2$ known, we obtain the simple formulae

$$S_{QL} = \sum_{i=1}^{n} \log p(x_i; \hat{\mu}, \hat{\sigma}^2) - 1 - \sum_{i=1}^{n} [\{(x_i - \hat{\mu})^2/\sigma^2\} - 1]n^{-1} - \frac{3}{2}n^{-1} + O_p(n^{-3/2}),$$

$$S_{QLP^*} = \sum_{i=1}^{n} \log p(x_i; \hat{\mu}, \hat{\sigma}^2) - 1 - \frac{1}{2}\sum_{i=1}^{n} [\{(x_i - \hat{\mu})^2/\sigma^2\} - 1]n^{-1} - \frac{3}{4}n^{-1} + O_p(n^{-3/2}).$$

The interpretation of the approximations (4.7) and (4.10) and, consequently, of the selection statistics S_{QL} and S_{QLP^*} demands further attention; here, some preliminary considerations are presented. As mentioned previously, when the assumed parametric statistical model is correctly specified, S_{QL} and S_{QLP^*} correspond, up to terms of order $O_p(1)$, to the Akaike's (1973) model selection criterion. Furthermore, the higher-order modifying terms involve, besides the dimension d of the parameter space, the fourth and the square of the third standardized cumulants evaluated at $\omega = \hat{\omega}$, namely $\hat{R}_4$, $\hat{R}_{13}$ and $\hat{R}_{23}$ and the quantities $\hat{M}_{rs}$, $\hat{M}_{rst}$ and $\hat{M}_{rstu}$, defined by (4.6). Since these sample statistics refer, respectively, to the second, third and forth expected balance relations, it is reasonable to argue that, in absence of model misspecification, they are nearly negligible. In particular, $\hat{M}_{rs}$, which corresponds to the information identity, is used by White (1982), Royal (1986) and Orme (1990) in order to define tests of model misspecification. Analogously, $\hat{M}_{rst}$ and $\hat{M}_{rstu}$ may be viewed as statistics which measure the misspecification of the model, with particular reference to skewness and kurtosis. Thus, these further terms seems to allow for a more accurate evaluation of the predictive ability of the model which is considered.

In particular, expansion (4.10) points out that, up to terms of order $O_p(n^{-3/2})$, the selection statistic S_{QLP^*} corresponds to a modification of the profile log-likelihood function. This modification apparently penalises models with many parameters and models with a remarkable skewness and a low kurtosis. However, these penalisations are not substantial since they are adjusted by the corresponding sample balance relations and, together, provide a measure of model misspecification based on violations of the expected balance relations, with reference to the data $\boldsymbol{x}$. With regard to the selection statistic S_{QL}, the corresponding asymptotic expansion (4.7) presents less terms than those obtained for S_{QLP^*} and, for this reason, it is supposed to have a lower discriminating ability, as far as a further prediction analysis is concerned.

Acknowledgements

The author would like to thank the referees whose comments led to improvements in the presentation.

References

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle, *Second Symposium on Information Theory* (eds. N. B. Petron and F. Caski), 267–281, Akademiai Kiado, Budapest.

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

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

Clayton, M. K., Geisser, S. and Jennings, D. E. (1986). A comparison of several model selection procedures, *Bayesian Inference and Decision Techniques* (eds. P. Goel and A. Zellner), 425–439, Elsevier Science Publishers, North Holland, Amsterdam.

Geisser, S. (1993). *Predictive Inference: An Introduction*, Chapman and Hall, New York.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection, *J. Amer. Statist. Assoc.*, **74**, 153–160.

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**, 299–314.

McCullagh, P. (1987). *Tensor Methods in Statistics*, Chapman and Hall, London.

Orme, C. (1990). The small-sample performance of the information-matrix test, *J. Econometrics*, **46**, 309–331.

Royall, R. M. (1986). Model robust confidence interval using maximum likelihood estimator, *International Statistical Review*, **54**, 221–226.

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

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

Tierney, L., Kass, R. E. and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions, *J. Amer. Statist. Assoc.*, **84**, 710–716.

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

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