

A note on the relationships between Bayesian and non-Bayesian predictive inference

Ventura Laura and Racugno Walter

Abstract Let us consider a random vector Y distributed according to a statistical model depending on an unknown parameter vector. Suppose we are interested in prediction, where the object of inference is a future or a yet unobserved random variable X , not depending on Y . The aim of this contribution is to provide some theoretical insight into the relationships between Bayesian and non-Bayesian higher-order asymptotic expansions for predictive densities. More precisely, we characterize prior probability distributions which guarantee the equivalence between Bayesian asymptotic expansions of the predictive distribution and frequentist accurate refinements of the estimative predictive density. An illustration in the context of the scalar skew-normal model is discussed. As a further result, invoking asymptotic connections between adjustments of the profile likelihood and asymptotic expansions of predictive densities, we illustrate how Bayesian predictive densities allows for the construction of adjusted profile loglikelihoods.

Key words: Bias reduction, default prior, higher-order asymptotics, maximum likelihood estimator, prediction, profile likelihood, skew-normal model.

1 Introduction

Let $y = (y_1, \dots, y_n)$ be the available data, realisation of a random variable $Y = (Y_1, \dots, Y_n)$ having independent and identically distributed components. Moreover, let $p(y; \theta) = \prod_{i=1}^n p(y_i; \theta)$ denote the probability density function of Y , with $\theta \in \Theta \subseteq \mathbb{R}^p$, $p \geq 1$. We are interested in prediction, where the object of inference is a future or a yet unobserved random variable X , not depending on Y . Under the

Laura Ventura

Department of Statistics, University of Padova, Italy, e-mail: ventura@stat.unipd.it

Walter Racugno

Department of Mathematics, University of Cagliari, Italy, e-mail: racugno@unica.it

frequentist approach accurate refinements of the estimative predictive density can be used; we refer here to the results in Komaki (1996) and Corcura and Giummolè (2000). Another approach is to use asymptotic expansions of the Bayesian predictive distribution, based on a suitable prior on θ ; for these expansions see Komaki (1996), Corcura and Giummolè (1999) and Datta and Mukerjee (2004).

The goal of this contribution is to discuss the relationships between Bayesian and non-Bayesian higher-order asymptotic expansions for predictive densities. In particular, we characterize prior functions under which the two approaches lead to approximately identical predictive densities in certain sense. These default priors offer a useful bridge between frequentist and Bayesian approaches. Questions of this type have been considered in the statistical literature in different inferential settings and with different requirements. For instance, the relationships between Bayesian and non-Bayesian interval estimates have been studied since Welch and Peers (1963). This approach is based on a probability matching criterion, which aims to match asymptotically the coverage of a Bayesian credible interval, or of a Bayesian credible set of a future observation, with the corresponding frequentist probability. See, e.g., Datta and Mukerjee (2004), and references therein, for an overview on these and other methods.

As a further result, invoking a second-order link between adjustments of the profile likelihood and asymptotic expansions of predictive densities (Pace *et al.*, 2010), we illustrate how accurate expansions of Bayesian predictive densities can be employed to construct adjusted profile loglikelihoods.

Notation and background material are given in Section 2. In Sections 3 and 4 we discuss priors for which the asymptotic expansions of Bayesian and frequentist predictive densities agree to a certain order. An illustration in the context of the scalar skew-normal model is given. Section 5 discusses some final remarks and the procedure to construct adjusted profile likelihoods using accurate Bayesian predictive densities.

2 Background on higher-order predictive densities

Let us denote by $\ell(\theta) = \log L(\theta) = \log p(y; \theta)$ the loglikelihood function based on y and by $\pi(\theta)$ a prior density on θ . The corresponding posterior distribution is $\pi(\theta|y) \propto \pi(\theta)L(\theta)$. Moreover, let $\hat{\theta}$ be the maximum likelihood estimate (MLE) of θ , $\ell_\theta(\theta) = \partial \ell(\theta) / \partial \theta$ the score function and $j(\theta) = -\ell_{\theta\theta} = -\partial^2 \ell(\theta) / (\partial \theta \partial \theta^\top)$ the observed information. Let $i(\theta) = E_\theta\{j(\theta)\} = n\kappa(\theta)$ be the expected information, where $E_\theta\{\cdot\}$ denotes expectation under θ and $\kappa(\theta)$ is the expected information per unit observation. In the simplest instance of prediction, the object of inference is a future or a yet unobserved random variable X , not depending on $Y = (Y_1, \dots, Y_n)$. Let $p_X(x; \theta)$ be the density of X .

The basic frequentist approach to prediction of X , on the basis of the observed y from Y , consists in using the estimative predictive density function

$$p_e(x) = p_x(x; \hat{\theta}) , \quad (1)$$

obtained by substituting the unknown θ with its MLE or, more generally, with an asymptotically efficient estimator of θ . The first-order solution (1) neglects sampling variability of the estimated parameter, and serious inaccuracies may occur when n is small (see e.g. Vidoni, 1995). For curved exponential families, Komaki (1996) obtains the optimal improvement over $p_e(x)$ in terms of average Kullback–Leibler divergence, up to and including terms of order $O(n^{-1})$, i.e. to second order. Corcuera and Giummolè (2000) show that Komaki’s approximation is, to second order, the optimal improvement over (1) 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. To give the expression of the resulting modified estimative density $p_{me}(x)$ index notation and Einstein summation convention are convenient. Generic components of θ will be denoted by θ_r , θ_s , ..., with $r, s, \dots = 1, \dots, p$. Let $\ell(\theta; x) = \log p_x(x; \theta)$, $\ell_r = \partial \ell(\theta; x) / \partial \theta_r$ and $\ell_{rs} = \partial^2 \ell(\theta; x) / (\partial \theta_r \partial \theta_s)$. Moreover, a hat over a likelihood quantity indicates evaluation at $\hat{\theta}$. The modified estimative density can be expressed as

$$p_{me}(x) = p_e(x) (1 + M_1(x)) , \quad (2)$$

with

$$M_1(x) = \frac{1}{2n} \left\{ \hat{\ell}_{rs} + \frac{1-\alpha}{2} \hat{\ell}_r \hat{\ell}_s + \frac{1+\alpha}{2} \hat{\kappa}_{rs} - \Gamma_{rs}^t(\hat{\theta}; \alpha) \hat{\ell}_t \right\} \hat{\kappa}^{rs} ,$$

where

$$\Gamma_{rs}^t(\theta; \alpha) = \kappa^{ut} \left(E_\theta \{ \ell_{rs} \ell_u \} + \frac{1-\alpha}{2} E_\theta \{ \ell_r \ell_s \ell_u \} \right)$$

and κ^{rs} is the generic element of $\kappa(\theta)^{-1}$. For $\alpha = -1$, the Kullback–Leibler divergence is obtained, and (2) is equivalent to the expression given in Komaki (1996).

In the Bayesian setting, Corcuera and Giummolè (1999) obtain the asymptotic expansion of Bayesian predictive distributions for curved exponential densities, up to and including terms of order $O(n^{-1})$, i.e. to second order. In particular, the Bayesian predictive distribution $\pi_p(x|y) = \int p_x(x; \theta) \pi(\theta|y) d\theta$ can be expanded as

$$\pi_p(x|y) = p_e(x) (1 + M_1(x) + M_2(x)) , \quad (3)$$

where

$$M_2(x) = \frac{1}{n} \left\{ \partial_r \log \pi(\hat{\theta}) - \Gamma_{rs}^s(\hat{\theta}; (1-\alpha)/2) \right\} \hat{\kappa}^{rt} \hat{\ell}_t , \quad (4)$$

with $\partial_r \log \pi(\theta) = \partial \log \pi(\theta) / \partial \theta_r$. Thus, the approximate Bayesian predictive distribution can be obtained by adding the quantity $M_2(x)$ to the correction term of the modified estimative density in (2). Note that for $\alpha = -1$, the asymptotic expansion of Komaki (1996) is obtained, which holds for general regular models. In this case, the quantity $\Gamma_{rs}^s(\theta; 1)$ in (4) simply reduces to

$$\Gamma_{rs}^s(\theta; 1) = \kappa^{su} E_\theta \{ \ell_{rs} \ell_u \} .$$

3 Asymptotic connections: Part I

In this section we discuss conditions on $\pi(\theta)$ ensuring asymptotic equivalence of higher-order frequentist and Bayesian predictive densities, given by (2) and (3), respectively.

For curved exponential models, a sufficient condition for $\pi(\theta)$ to give the equivalence of the two asymptotic expansions is $\partial_r \log \pi(\theta) - \Gamma_{rs}^s(\theta; (1-\alpha)/2) = 0$. Equivalently, we are looking for a prior such that

$$\partial_r \log \pi(\theta) = \kappa^{su} (E_\theta \{\ell_{rs} \ell_u\} + \beta E_\theta \{\ell_r \ell_s \ell_u\}) , \quad (5)$$

with $\beta = (1+\alpha)/4$. Note that the condition (5) is equivalent to the family of invariant priors discussed by Hartigan (1964, 1965), which includes the Jeffreys' prior for $\beta = 1/2$ (and it corresponds to $\alpha = 1$). When the Kullback-Leibler divergence is used, i.e. when $\alpha = -1$ (or equivalently $\beta = 0$), then, for general regular models, condition (5) reduces to

$$\partial_r \log \pi(\theta) = \kappa^{su} E_\theta \{\ell_{rs} \ell_u\} . \quad (6)$$

In the scalar parameter case, it is straightforward to show that the solution to (5) takes the form

$$\pi(\theta) \propto \exp \left(\int i(\theta)^{-1} [E_\theta \{\ell_{\theta\theta} \ell_\theta\} + \beta E_\theta \{\ell_\theta^3\}] d\theta \right) , \quad (7)$$

which, using the third Bartlett relation, can be rewritten as

$$\pi(\theta) \propto \exp \left(\int i(\theta)^{-1} [(1-3\beta)E_\theta \{\ell_{\theta\theta} \ell_\theta\} - \beta E_\theta \{\ell_{\theta\theta\theta}\}] d\theta \right) , \quad (8)$$

with $\ell_{\theta\theta\theta} = \partial^3 \ell(\theta) / \partial \theta^3$. For $\alpha = -1$ (or $\beta = 0$), equations (7) and (8) reduce to

$$\pi(\theta) \propto \exp \left(\int i(\theta)^{-1} E_\theta \{\ell_{\theta\theta} \ell_\theta\} d\theta \right) .$$

For the exponential family in the canonical form $p(y; \theta) = \exp(\theta y - K(\theta)) p_0(y)$, where $K(\cdot)$ and $p_0(\cdot)$ are known functions, the prior (8) is $\pi(\theta) \propto i(\theta)^\beta$. For $\beta = 1/2$ the usual Jeffreys' prior is obtained. If interest is on the Kullback-Leibler divergence, then $\pi(\theta) \propto 1$ (see Hartigan, 1964, 1965).

As a second example, consider the scale model $p(y; \theta) = (1/\theta) p_0(y/\theta)$, where $p_0(\cdot)$ is a given function. Let $g(\cdot) = -\log p_0(\cdot)$. We have $E_\theta(\ell_{\theta\theta\theta}) = c_1/\theta^3$, $E_\theta(\ell_{\theta\theta} \ell_\theta) = c_2/\theta^3$ and $i(\theta) = c_3/\theta^2$, with $c_1 = \int (y^3 g'''(y) + 6y^2 g''(y) + 6y g'(y) - 2) p_0(y) dy$, $c_2 = \int (3y g'(y) + y^2 g''(y) - 2y^2 g'(y)^2 - y^3 g'(y) g''(y) - 1) p_0(y) dy$ and $c_3 = \int (2y g'(y) + y^2 g''(y) - 1) p_0(y) dy$. The prior distribution (8) is $\pi(\theta) \propto \theta^k$, with $k = ((1-3\beta)c_2 - \beta c_1)/c_3$. This prior coincides with the Jeffreys' prior for a one-parameter scale model when $k = -1$. In this case, $\pi(\theta)$ has matching properties in

the sense of ensuring frequentist validity, up to $o(n^{-1})$, of the posterior quantiles of X .

As a third example, consider the location model $p(y; \theta) = p_0(y - \theta)$, where $p_0(\cdot)$ is a given function. Let us assume $\beta = 0$. Since all the likelihood quantities involved in $\pi(\theta)$ do not depend on θ , then $\pi(\theta) \propto \exp(k\theta)$, with $k = -(\int g''(y)g'(y)p_0(y)dy)/(\int g''(y)p_0(y)dy)$ and $g(y) = -\log p_0(y)$. For the particular case of the $N(\theta, 1)$ distribution, we have $k = 0$ and thus $\pi(\theta)$ agrees with the Jeffreys' prior distribution $\pi(\theta) \propto 1$.

4 Asymptotic connections: Part II

Choosing a prior of the form (5) or (6) under the Bayesian approach corresponds to using the MLE in the estimative density in the frequentist setting. Indeed, following Corcuera and Giummolè (1999), the factor (4) can be interpreted as a modification in the estimative density (1) of the estimator used. In particular, the second-order Bayesian predictive density (3) can be rewritten as

$$\pi_p(x|y) = p_X(x; \hat{\theta}^*) (1 + M_1(x)) , \quad (9)$$

where $\hat{\theta}^*$ denotes a modified MLE of the form $\hat{\theta}^* = \hat{\theta} + b(\hat{\theta})$, with

$$b^r(\theta) = \kappa^{rs} [\partial_s \log \pi(\theta) - \Gamma_{st}^t(\theta; (1 - \alpha)/2)] . \quad (10)$$

The correction term (10) has the effect of modifying the bias of the MLE, since

$$E_\theta\{\hat{\theta}_r^* - \theta_r\} = E_\theta\{\hat{\theta}_r - \theta_r\} + b^r(\theta) = d^r(\theta) + b^r(\theta) + o(n^{-1}) ,$$

where

$$d^r(\theta) = \frac{1}{2n} \kappa^{rs} \kappa^{tu} (E_\theta\{\ell_{stu}\} + 2E_\theta\{\ell_{st}\ell_u\})$$

is the expression for the asymptotic bias of the MLE (see e.g. Pace and Salvan, 1997, Chapter 9). Thus, the asymptotic predictive densities (2) and (3) agree if the estimator $\hat{\theta}^*$ is used in the estimative density, with $b(\theta)$ satisfying (10).

Equation (10) points out a correspondence between the prior used in the Bayesian analysis and the modification required for the MLE. For instance, if the Jeffreys' prior is used in (10), then the modified MLE is $\hat{\theta}_r^* = \hat{\theta}_r + b^r(\hat{\theta})$, with

$$b^r(\theta) = \frac{(1 - \alpha)}{4} \kappa^{rs} \kappa^{tu} E_\theta\{\ell_r \ell_t \ell_u\} .$$

We now turn to the problem of characterizing priors on the basis of a chosen estimator $\hat{\theta}^*$, using relation (10). If it is of interest to consider a suitable modified MLE of the form $\hat{\theta}^* = \hat{\theta} + b(\hat{\theta})$, then the two asymptotic expansions (2) and (3) agree if the prior $\pi(\theta)$ satisfies

$$\partial_r \log \pi(\theta) = \kappa_{sr} b^s(\theta) + \Gamma_{rs}^s(\theta; (1 - \alpha)/2) . \quad (11)$$

For $\alpha = -1$, condition (11) simplifies to

$$\partial_r \log \pi(\theta) = \kappa_{sr} b^s(\theta) + \kappa^{su}(\theta) E_\theta \{ \ell_{rs} \ell_u \} . \quad (12)$$

In particular, let us consider the first-order bias corrected MLE given by

$$\hat{\theta}_r^* = \hat{\theta}_r - d^r(\hat{\theta}) .$$

This is equivalent to require that $b^r(\theta) = -d^r(\theta)$. For curved exponential families, we are looking for a prior such that

$$\partial_r \log \pi(\theta) = \frac{1}{2} \kappa^{su}(\theta) ((\beta - 1) E_\theta \{ \ell_{rsu} \} - E_\theta \{ \ell_{rs} \ell_u \}) . \quad (13)$$

For general regular models and for $\alpha = -1$, equation (13) reduces to

$$\partial_r \log \pi(\theta) = -\frac{1}{2} \kappa^{su}(\theta) (E_\theta \{ \ell_{rsu} \} + E_\theta \{ \ell_{rs} \ell_u \}) . \quad (14)$$

Note that the term at the right hand side of (14) corresponds to the Firth's adjustment (Firth, 1993) to the score function; see also Datta and DiCiccio (2001, Section 5.3). Taking the derivative with respect to θ of $\log \pi(\theta|y) = \ell(\theta) + \log \pi(\theta) + \text{constant}$, we have

$$\partial_r \log \pi(\theta|y) = \ell_r(\theta) + \partial_r \log \pi(\theta) . \quad (15)$$

If the prior density is chosen according to (14), then the right hand side of (15) is exactly the modified likelihood equation discussed by Firth (1993). In view of this, for general regular models, Firth's estimate coincides with the mode of the posterior distribution obtained using the default prior defined by (14). This result validates the use of the method introduced by Firth (1993) for point estimation in the Bayesian framework.

An example. Let us consider the scalar skew normal model (Azzalini, 1985) with density function $p(y; \lambda) = 2\phi(y)\Phi(\lambda y)$, where ϕ and Φ denote the $N(0, 1)$ density and distribution functions, respectively, and $y, \lambda \in \mathbb{R}$. Point estimation of the shape parameter λ is a quite challenging problem since the likelihood function can be monotone. Some recent contributions we focus on, both in the frequentist and Bayesian settings, are Sartori (2006), which concentrates on Firth's method, and Liseo and Loperfido (2006), which show that the Jeffreys' prior for λ is proper.

The condition (14) associated with this model is (see also Sartori, 2006)

$$\frac{\partial}{\partial \lambda} \log \pi(\lambda) = -\frac{\lambda}{2} \frac{a_{42}(\lambda)}{a_{22}(\lambda)} ,$$

where $a_{ij}(\lambda) = E_\lambda \{ Y^i \phi(\lambda Y)^j / \Phi(\lambda Y)^j \}$. These quantities can be computed numerically or approximated using results in Bayes and Branco (2007). In this latter case,

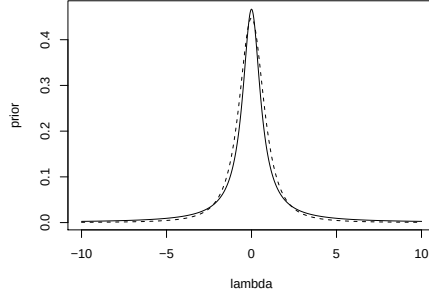


Fig. 1 Skew normal model: Prior $\pi(\lambda)$ (solid line) and Jeffreys' prior $\pi_J(\lambda)$ (dotted line).

$\partial \log \pi(\lambda) / \partial \lambda = -(3\lambda/2)(1 + 2\lambda^2/(\pi^2/4))^{-1}$, and thus

$$\pi(\lambda) \propto \left(1 + \frac{2\lambda^2}{(\pi^2/4)}\right)^{-\frac{3\pi^2/4}{8}}. \quad (16)$$

The right hand side of (16) is proportional to the density function of a t distribution with scale parameter $(2\pi^2/(3\pi^2 + 16))^{1/2} \cong 0.66$ and $(3\pi^2 - 16)/16 \cong 0.85$ degrees of freedom. This prior is not equal to the Jeffreys' prior $\pi_J(\lambda)$ (see Liseo and Loperfido, 2006), which is still a t distribution but with scale parameter $\pi/4 \cong 0.79$ and 2 degrees of freedom (Bayes and Branco, 2007). The plots of the two priors $\pi(\lambda)$ and $\pi_J(\lambda)$ is illustrated in Figure 1.

5 Final remarks

In this contribution the relationships between Bayesian and non-Bayesian higher-order asymptotic expansions for predictive densities are discussed. Prior functions under which the two approaches lead to identical predictive densities have been characterized. In particular, a class of priors related to bias corrected maximum likelihood is illustrated. This result validates in the Bayesian framework the use of the method introduced by Firth (1993).

A further application of the results discussed in this paper is related to the construction of adjusted profile loglikelihoods using accurate Bayesian predictive densities. Suppose that $\theta = (\psi, \lambda)$, where the primary interest of inference is about ψ . Let us denote by $p_X(x; \theta) = p_X(x; \psi, \lambda)$ the density of X , where the parameter ψ is treated as fixed, while λ is unknown. Moreover, let us denote by $\hat{\lambda}_\psi$ the constrained MLE of λ for a given value of ψ and by $\ell_p(\psi) = \ell_p(\psi; y) = \ell(\psi, \hat{\lambda}_\psi)$ the profile loglikelihood. Several modifications of $\ell_p(\psi)$ have been proposed in the literature, of the form $\ell_{mp}(\psi) = \ell_p(\psi) + a(\psi)$, where $a(\psi)$ is a smooth function of order $O_p(1)$; see, e.g., Pace and Salvan (2006), and references therein,

and see Pace *et al.* (2010) for the second-order link between modifications of the profile likelihood and refinements of the estimative predictive density. 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 $y_A = (y_{A1}, \dots, y_{Am})$ and $y_B = (y_{B1}, \dots, y_{Bm})$ random samples of size m . Let us write the predictive density of y_B as $\hat{p}(y_B|y_A, \psi) = \prod_{i=1}^m \pi_p(y_{Bi}|y_A, \psi)$, where $\pi(y_{Bi}|y_A, \psi)$ is the approximate predictive density in (3). Similarly, we can compute the predictive density, $\hat{p}(y_A|y_B, \psi)$, of y_A based on y_B . Then an adjusted profile loglikelihood can be defined as (see Pace *et al.*, 2010)

$$\ell_{mp}(\psi) = \ell_p(\psi; y) + \bar{a}(\psi), \quad (17)$$

with $\bar{a}(\psi) = \frac{1}{2} \{ \log \hat{p}(y_B|y_A, \psi) + \log \hat{p}(y_A|y_B, \psi) - \ell_p(\psi; y_A) - \ell_p(\psi; y_B) \}$. The behaviour of the adjusted profile loglikelihood (17) is currently under study.

References

1. Azzalini, A.: A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171–178 (1985)
2. Bayes, C.L., Branco, M.D.: Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Braz. J. Probab. Statist.* **21**, 141–163 (2007)
3. Corcuera, J.M., Giummolè, F.: A generalized Bayes rule for prediction. *Scand. J. Statist.* **26**, 265–279 (1999)
4. Corcuera, J.M., Giummolè, F.: First order predictive densities. In: Marriott, P., Salmon, M. (eds.) *Applications of Differential Geometry to Econometrics*, pp. 214–229. Cambridge University Press, Cambridge (2000)
5. Datta, G.S., DiCiccio, T.J.: On expected volumes of multidimensional confidence sets associated with the usual and adjusted likelihoods. *J. R. Statist. Soc. B* **63**, 691–703 (2001)
6. Datta, G.S., Mukerjee, R.: *Probability Matching Priors: Higher Order Asymptotics*. Springer-Verlag, New York (2004)
7. Firth, D.: Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993)
8. Hartigan, J.A.: Invariant prior densities. *Ann. Math. Statist.* **35**, 836–845 (1964)
9. Hartigan, J.A.: The asymptotically unbiased density. *Ann. Math. Statist.* **36**, 1137–1152 (1965)
10. Komaki, F.: On asymptotic properties of predictive distributions. *Biometrika* **83**, 299–313 (1996)
11. Liseo, B., Loperfido, N.: A note on reference priors for the scalar skew-normal distribution. *J. Statist. Plann. Inf.* **136**, 373–389 (2006)
12. Pace, L., Salvan, A.: *Principles of Statistical Inference*. World Scientific, Singapore (1997)
13. Pace, L., Salvan, A.: A new motivation for adjustments of the profile likelihood. *J. Statist. Plann. Inf.* **136**, 3554–3564 (2006)
14. Pace, L., Salvan, A., Ventura, L.: On a likelihood interpretation of adjusted profile likelihoods through refined predictive densities. *Ann. Inst. Statist. Math.* (2010) doi: 10.1007/s10463-010-0274-9
15. Sartori, N.: Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J. Statist. Plann. Inf.* **136**, 4259–4275 (2006)
16. Vidoni, P.: A simple predictive density based on the p^* formula. *Biometrika* **82**, 855–863 (1995)
17. Welch, B.L., Peers, H.W.: On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. B* **25**, 318–329 (1963)