ISSN - 0103-2577

The Kullback - Leibler approximation of the marginal posterior density: an application to the linear functional model

JOSEMAR RODRIGUES

Nº 92

N O T A S D O I C M S C

São Carlos
Set./91

The Kullback - Leibler Approximation of the Marginal Posterior Density: An application to the linear functional model

by
Josemar Rodrigues
ICMSC - USP - 13.560 - C.P.-668 - São Carlos - SP - Brasil

Summary: Consider the problem of inference about a parameter θ_1 in the presence of a nuisance parameter θ_2. In a Bayesian framework, the inferences about θ_1 are based on the marginal posterior. In this paper, we consider the Kullback-Leibler measure of divergence for finding an approximation of the marginal posterior of θ_1. An application to the linear functional relationships is investigated in details given us a Bayesian justification for Fuller's results (1987). In the normal case and for small sample sizes, it is shown that the Kullback-Leibler approximation is more accurate than the modal approximation (Kass et al., 1990).

Key Words: Marginal posterior, Kullback-Leibler measure, linear functional models.

1. Introduction:

Let $p(y,\theta)$ the density function of the random vector y, where $\theta = (\theta_1, \theta_2)$. Our purpose is to make inferences about the parameter θ_1, while the nuisance parameter is θ_2. In the Bayesian approach, "overall"inferences about θ_1 are completely determined by the marginal posterior of θ_1 obtained by "integrating out" the nuisance parameter θ_2. Thus, the marginal posterior of θ_1 given y can be written as

$$p(\theta_1, /y) = \int p(\theta_1 /_{\theta_2,y})\, p(\theta_2/y)\, d\theta_2 , \qquad (1)$$

where $p(\theta_1/\theta_2, y)$ is the conditional density of θ_1, given θ_2 and $p(\theta_2/y)$ is the marginal posterior of θ_2. As has been emphasized by Box and Tiao (1973), caution should be exercised in integrating out nuisance parameters. If we wished to make inferences about θ_1 and it was found that $p(\theta_1/\theta_2, y)$ was very sensitive to changes in θ_2, it would be important to examine carefully the marginal posterior of θ_2. Albert (1989), to avoid this difficult, suggested a transformation $(\theta_1, h(\theta_1, \theta_2))$ such that θ_1 and $h(\theta_1, \theta_2)$ are approximately independent. In this paper, we consider the problem of finding $\hat{\theta}_2$ that minimizes the Kullback-Leibler measure (1951) of divergence between $p(\theta_1/y)$ and $p(\theta_1/\hat{\theta}_2, y)$. Aitchison (1980), in a recent paper, showed that this measure of divergence is coherent with the Bayesian approach for parametric

density estimation problems. We investigate the performance of the K.L.- approximation of the marginal posterior in the normal case and an interesting application of it to the linear functional error models (Fuller, 1987) is considered in details. Under special case, the K.L.- approximation is the modal approximation (Kass et al. 1990).

2. The Kullback-Leibler Approximation of the Marginal Posterior Density

Aitchison (1975),(1990) used Kullback-Leibler (1951) measure of divergence for the parametric density estimation problems. The divergence between $p(\theta_1/y)$ and $p(\theta_1/\hat{\theta}_2, y)$ is given by

$$d[\, p(\theta_1/y), p(\theta_1/\hat{\theta}_2, y\,] = \int p(\theta_1/y)\ log\ \left\{\frac{p(\theta_1/y)}{p(\theta_1/\hat{\theta}_2, y)}\right\}\ d\theta_1\ , \qquad (2)$$

which is positive unless $p(\theta_1/y)$ coincides with $p(\theta_1/\hat{\theta}_2, y)$. We could interpret (2) as the measure of the influence of $\hat{\theta}_2$ on the estimative of θ_1, or, in analogy with Lindley's measure, how much information $\hat{\theta}_2$ provides about θ_1.If $\hat{\theta}_2$ minimizes (2), we call $p(\theta_1/\hat{\theta}_2, y)$ the K.L.-approximation of $p(\theta_1/y)$. In the parametric density estimation, where $p(\theta_1/\theta_2, y)$ is the target, $p(\theta_1/\hat{\theta}_2, y)$ is the "best" shot, that is, $p(\theta_1/\hat{\theta}_2, y)$ looks like the predictive density function, (Aitchison, 1975). The next result gives the way of obtaining $\hat{\theta}_2$ in order to have the K.L.- approximation of $p(\theta_1/y)$.

Theorem 2.1:

Let $p(\theta_1/\theta_2, y) = exp[\theta_2 T_y(\theta_1) + d_y(\theta_2) + h_y(\theta_1)]$. Then $p(\theta_1/\hat{\theta}_2, y)$ is the K.L.-approximation of the marginal posterior density $p(\theta_1/y)$ if and only if $\hat{\theta}_2$ satisfies the following equation:

$$E_{\hat{\theta}_2}[T_y(\theta_1)] = E_P\{E_{\theta_2}[T_y(\theta_1)]\}\ , \quad \text{where} \qquad (3)$$

E_{θ_2} and E_p denote the expectation w.r.t. $p(\theta_1/\theta_2, y)$ and $p(\theta_2/y)$, respectively.

Proof:

$$d[p(\theta_1/y), p(\theta_1/\hat{\theta}_2, y)] = \int E_p[p(\theta_1/\theta_2, y]\ log\ \left\{\frac{p(\theta_1/y)}{p(\theta_1/\hat{\theta}_2, y)}\right\} d\theta_1 =$$

$$= \int E_p[p(\theta_1/\theta_2, y)]\ log\ p(\theta_1/y) d\theta_1 - L(\hat{\theta}_2)\ , \quad \text{where}$$

$$L[\hat{\theta_2}] = \int E_p[p(\theta_1/\theta_2, y)] \, . \, log \; p(\theta_1/\hat{\theta}_2, y) d\theta_1 =$$
$$= \hat{\theta_2} \; E_p \; E_{\theta_2} \; [T_y(\theta_1)] + d_y(\hat{\theta_2}) + E_p \; E_{\theta_2} \; [h_y(\theta_1)].$$

Now, $\frac{\partial L(\hat{\theta_2})}{\partial \hat{\theta_2}} = 0$ if and only if $E_p \; E_{\theta_2} \; [T_y(\theta_1)] = - d_y'(\hat{\theta_2}) = E_{\hat{\theta_2}}[T_y(\theta_1)]$, (4)

then the result follows from (4).

Taking $E_{\theta_2}[T_y(\theta_1)] = \theta_2$, we have from (3) that $p(\theta_1/\hat{\theta}_2, y)$ is the K.L.- approximation of the marginal posterior $p(\theta_1/y)$, where $\hat{\theta}_2$ is the posterior mean of θ_2. A similar result was obtained by DeGroot in his discussion of Geisser's paper (1982) for the parametric density estimation situation under the exponential family. If $p(\theta_1/\theta_2, y)$ is symmetric, the K.L.- approximation is the modal approximation (Kass et al., 1990). The modal approximations are commonly used in Bayesian data analysis because they are generally easy to compute and the interpretation of inferences is simplified. Next, we shown how Theorem 2.1 works in the normal model and provides a more accurate approximation than the modal approximation.

Normal Case: Assume that we have n independent observations $y' = (y_1, \ldots, y_n)$, from a normal population with unknown mean θ and unknown standard deviation σ. If our prior information is vague, we can represent this state of information about (θ, σ) by

$$p(\theta, \sigma) \propto \frac{1}{\sigma}, \; - \infty < \theta < \infty \; \sigma > 0.$$

We are interested in θ and σ is a nuisance parameter. It can be found in Zellner (1971) that:

Posterior of σ: $p(\sigma/y) \propto \sigma^{-(n+1)} \; exp\left\{ - \frac{(n-1)S^2}{2\sigma^2} \right\}$ (the inverted gamma),

Marginal posterior of t: $p(t/y) \propto \left(1 + \frac{t^2}{n-1} \right)^{-\frac{n}{2}}$ ("t" - distribution),

Conditional posterior of θ: $p(\theta/\sigma, y) \propto exp \left\{ - \frac{n(\theta - \bar{y})^2}{2\sigma^2} \right\}$ (the normal distribution),

where $t = \frac{\sqrt{n}(\theta - \bar{y})}{S}, S^2 = \frac{\sum (y - \bar{y})^2}{n-1}$ and $\bar{y} = \frac{\sum y_i}{n}$.

It is not difficult to obtain, from Theorem 2.1, that the solution of (3) is

$$\hat{\sigma} = \sqrt{\frac{n-1}{n}} .S \tag{5}$$

and the K.L.- approximation of the marginal posterior of t, is

$$\hat{p}_{KL}\,(t/y) \;\propto\; exp\left\{-\frac{nt^2}{2(n-1)}\right\}. \tag{6}$$

The modal approximation of the marginal posterior (Kass et al., 1990) is

$$\hat{p}_m\,(t/y) \propto exp\{-\frac{(n+1)t^2}{2(n-1)}\}. \tag{7}$$

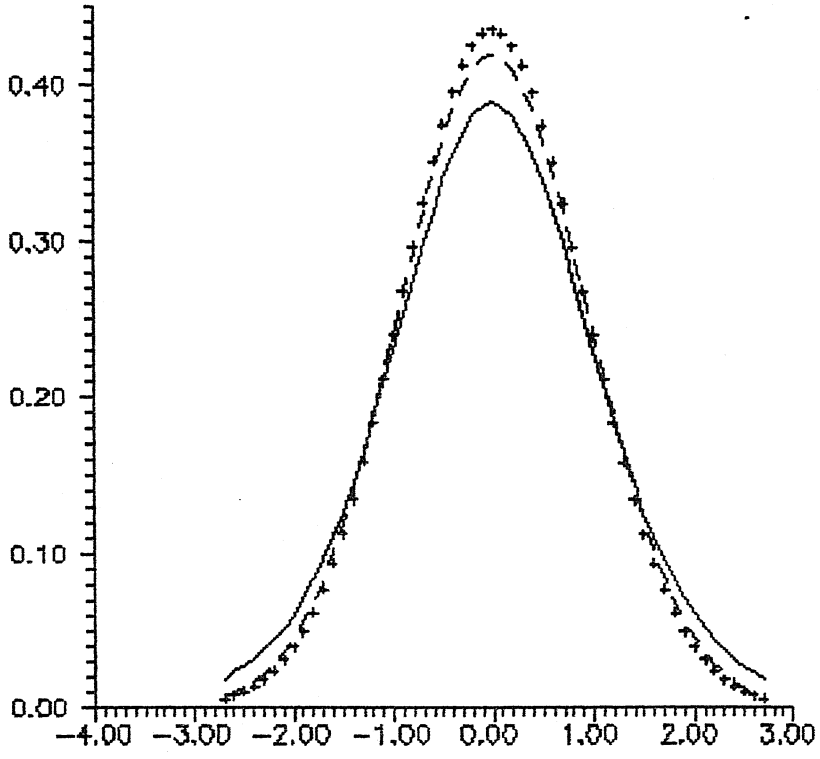

Figure 1: dashed line:K.L.-approximation; solid line:Exact marginal; +++ :Modal approximation.

Figure 1 plots the K.L.-approximation,the exact marginal posterior of t given y and the modal approximation. The K.L-approximation appears to be reasonably more closed to the exact marginal posterior, for the sample size $n = 11$, than the modal approximation. In the next section we give an interesting application of Theorem 1 to the linear functional error models described by Fuller (1987) and Sprent (1966).

3. The Linear Functional Models

Let us assume the following model:

$$Y_t = \beta x_t + e_t \ , \ X_t = x_t + u_t \ , \quad t = 1, \ldots, n \ , \ \text{where}$$

$$(e_t, u_t)' \sim N\left[(0,0)', \sigma^2 I\right] \ , \ \text{I: the identity matrix.} \tag{8}$$

In this type of model one is unable to observe the fixed x_t, directly. Instead of observing x_t, one observes X_t. The value X_t may be obtained by asking people questions or by reading an imperfect instrument. In this model we suppose β and x_t unknown and σ^2 known. Sprent (1966), without the normality assumption, applied the "Generalized Least-Squares Principle" to estimate the parameter β given the data $(Y_1, X_1), \ldots, (Y_n, X_n)$. The Bayesian justification of the generalized least-squares procedure were given by Lindley in his discussion of Sprent's paper (1966) and an exact Bayesian justification using Jeffreys prior by Rodrigues (1990). In these papers were found that

$$p(\beta/Y, X) \propto exp\left\{-\frac{1}{2}\left[(S_{XX} - \lambda\sigma^2)\frac{(\beta - \hat{\beta})^2}{\sigma^2(\beta^2+1)}\right]\right\} \ ; \ \text{where}$$

$$\hat{\beta} = \frac{S_{XY}}{S_{XX} - \lambda\sigma^2} \ , \ S_{XX} = \frac{\sum X_t^2}{n} \ , \ S_{XY} = \frac{\sum X_t Y_t}{n} \ , \ S_{YY} = \frac{\sum Y_t^2}{n} \ , \tag{9}$$

and λ is the smallest root of the determinantal equation $|B - \lambda \sum| = 0$,

$$B = \begin{bmatrix} S_{YY} & S_{XY} \\ S_{YX} & S_{XX} \end{bmatrix} \quad \text{and} \quad \sum = \sigma^2 I.$$

In this section, we consider the problem of inference about the x_t value that generated the data (X_t, Y_t) in the presence of the nuisance parameter $\beta(\sigma^2 : Known)$. Consider this problem from a Bayesian viewpoint. Let the assumptions of model (8) hold given $(X, Y) = (X_1 \ldots, X_n, Y_1, \ldots Y_n)$, β and $x_{-t} = (x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_n)$. After some algebraic manipulations, we can show that the likelihood function of x_t given (X, Y) , x_{-t} and β is

$$L\left[x_t/(X,Y), \beta, x_{-t}\right] \propto exp\left\{-\frac{1}{2\sigma_0^2}(x_t - \mu_t(\beta))^2\right\} \ , \tag{10}$$

where

$$\mu_t(\beta) = \frac{X_t + \beta Y_t}{1 + \beta^2} \quad \text{and} \quad \sigma_0^2 = \frac{\sigma^2}{1 + \beta^2} .$$

Since σ_0^2 is fixed and $L\ [\ x_t/(X,Y)\ ,\ \beta, x_{-t}\ ]$ is data translated in x_t given β(Box and Tiao, 1972), a noninformative prior for x_t given β is locally uniform, that is ,

$$p(x_t/\beta) \propto \text{const.} \tag{11}$$

From (10) and (11), the conditional posterior of x_t given (X,Y) , β and x_{-t} is

$$p(x_t/(X,Y),\beta) \propto \exp\left\{-\frac{1}{2\sigma_0^2}\ (x_t - \mu_t(\beta))^2\right\} . \tag{12}$$

In a Bayesian framework, the inference about x_t given (X,Y) is based on the marginal posterior

$$p(x_t/X,Y) = \int p\,(x_t/(X,Y)\ ,\ \beta)\ .\ p(\beta/X,Y)\ d\beta \tag{13}$$

Since (13) is very hard to find exactly, we look for the K.L.-approximation of it by applying Theorem 2.1 of Section 2. From (12) we have that

$$T_{(X,Y)}\ (x_t) = -\frac{x_t}{2\sigma_0^2} \quad \text{and} \quad E_\beta\ \left[T_{(X,Y)}\ (x_t)\right] = -\frac{\mu_t(\beta)}{2\sigma_0^2} = -\frac{X_t - \beta\ Y_t}{2\sigma^2} .$$

We have from (3) that $\hat{\beta}$ must satisfy the equation $-\frac{X_t + \hat{\beta}\ Y_t}{2\sigma^2} = E_p\left[-\frac{X_t + \beta\ Y_t}{2\sigma^2}\right]$, which implies that $\hat{\beta} = E_p\ [\beta] = \hat{\hat{\beta}}$.

The K.L.-approximation of the marginal posterior of x_t given (X,Y) is given by

$$p(x_t/(X,Y)\ ,\hat{\hat{\beta}}) \sim N\left[\frac{X_t + \hat{\hat{\beta}}\ Y_t}{1 + \hat{\hat{\beta}}^2}\ ,\ \frac{\sigma^2}{1 + \hat{\hat{\beta}}^2}\right] \tag{14}$$

Considering the null intercept, expression (14) is a Bayesian justification of the results obtained by Fuller (1987), p. 21. All inference about x_t can be based on the approximated marginal posterior $p(x_t/(X,Y)\ ,\ \hat{\beta})$. For example, an approximated 95% - high posterior density interval for x_t based on X, Y could be

$$\frac{X_t + \hat{\hat{\beta}}\ Y_t}{1 + \hat{\hat{\beta}}^2} \pm 1.96\ \sqrt{\frac{\sigma^2}{1 + \hat{\hat{\beta}}^2}} .$$

It is important to note that (14) is the modal approximation (Kass et al., 1990) of the exact marginal posterior of x_t given the data (X,Y).

References

[1] Aitchison, J. (1990) - On the coherence in Parametric Density Estimation, Biometrica, 77, 4, 905-8.

[2] Aitchison, J. (1975) - Goodness of prediction fit. Biometrika, 62, 547-54.

[3] Albert, H.J. (1989) - Nuisance Parameters and the use of exploratory Graphical Methods Bayesian Analysis, the American Statistican, V. 43, nº 4.

[4] Box, G.E.P. and Tião, G.C. (1973) - Bayesian Inference in Statistical Analysis, Wesley.

[5] Fuller, W.A. (1987) - Measurement Error Models, Wiley.

[6] Geisser, S. (1982) - Aspects of the predictive and estimative approaches in the determination of Probabilities, Current Topics in Biostatistics and Epidemiology, 75-85.

[7] Kass, R.E., Tierney, L. and Kadane, J.B.(1990) - Laplace Method in Bayesian Analysis, unpublished report.

[8] Kullback, S. & Leibler, R.A. (1951) - On Information and Suficiency, Annals Mathematical Statistics, 22, 79-86.

[9] Rodrigues, J. (1990) - A Bayesian Analysis of the Generalized Least-Squares Procedure for the Functional Relationships, Technical Report nº , ICMSC-USP, submitted.

[10] Sprent, S. (1986) - A Generalized Least-squares approach to linear functional relationships, J.R. Statist. Soc. B, 28, 278-297.

[11] Zellner, A. (1971) - An Introduction to Bayesian Inference in Econometrics, Wiley.

NOTAS DO ICMSC

Nº 91/91 - SCOTT, D.R.;-NUNES, M.G.V. - Focus-driven search for decinding what to say in reply to questions

Nº 90/91 - SOUZA, C.S.; NUNES, M.G.V. - On the role of text generation in knowledge based systems interfaces

Nº 89/91 - MORABITO, R.N.; ARENALES, M.N. - On solving large two-dimensional guillotine cutting problems

Nº 88/91 - MICALI, A.; VILLAMAYOR, O.E. - Algèbres de Clifford sur un corps de caractéristique 2

Nº 87/91 - RAPOPORT-CAMPODÔNICO, D.L. - On the construction of Lie-isotopic relativistic sto - chastic mechanics and Lie-isotopic potential theory from the Lie-isotopic geometry associated to a torsion potential

Nº 86/91 - ACHCAR, J.A. - Inferences for the Birnbaun-Saunders fadigue life model using Bayesian methods

Nº 85/91 - ACHCAR, J.A.; LOUZADA NETO, F. - Accelaerated life tests with one stress variable : a Bayesian analysis of the eyring model

Nº 84/90 - RODRIGUES, J. - Bayes predective likelihood function for the accelerated life tests via the orthogonal parameters

Nº 83/90 - RODRIGUES, J. - A Bayesian analysis of the generalized least-square procedure to functional relationship

Nº 82/90 - LIZANA PEÑA, M. - Exponential dichotomy for singularity perturbed linear functional differential equations with small delays