Posterior distributions with non explicit objective priors

Distribuzioni a posteriori basate su distribuzioni a priori oggettive non esplicite

Erlis Ruli, Nicola Sartori and Laura Ventura

Abstract We introduce two methods useful to derive a posterior distribution for a parameter of interest, when only the first derivative of a log-prior is available. This is typically the situation when dealing with multidimensional parameters and objective priors. An example is illustrated using a predictive matching prior.

Abstract *In questo contributo vengono introdotti due metodi utili per derivare una distribuzione a posteriori per un parametro di interesse, quando è disponibile solo la derivata prima del logaritmo della distribuzioni a priori. Tale situazione si presenta tipicamente in presenza di parametri multidimensionali e distribuzioni a priori oggettive. Il metodo viene illustrato in un modello logistico con una matching prior predittiva.*

Key words: Firth's adjustment, Logistic regression, Matching prior, MCMC, Rao score test statistic, Score function, Taylor expansion.

1 Introduction

Let $y = (y_1, \ldots, y_n)$ be the available data, considered for simplicity as a random sample of size n, i.e. as a realization of a random variable $Y = (Y_1, \ldots, Y_n)$ 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 $\theta \in \Theta \subseteq \mathbb{R}^k$, $k \geq 1$. We are interested in objective Bayesian inference on the unknown parameter θ, us-

Erlis Ruli
Department of Statistical Sciences, Univeristy of Padova, e-mail: ruli@stat.unipd.it

Nicola Sartori
Department of Statistical Sciences, Univeristy of Padova, e-mail: sartori@stat.unipd.it

Laura Ventura
Department of Statistical Sciences, Univeristy of Padova, e-mail: ventura@stat.unipd.it

ing the posterior distribution

$$\pi(\theta|y) \propto \pi(\theta) L(\theta), \tag{1}$$

where $\pi(\theta)$ is a prior for θ and $L(\theta) \propto p(y;\theta)$ is the likelihood function.

We consider the situation in which the prior distribution $\pi(\theta)$ is known only through its first derivative $\partial \log \pi(\theta)/\partial\theta$. This is typically the situation with defalut priors, such as matching priors (see, e.g., Datta and Mukerjee, 2004). In these cases, the posterior distribution (1) is not directly available, and it is only possibile to evaluate the first derivative of the log-posterior $t(\theta) = t(\theta;y) = \log \pi(\theta|y)$ given by

$$t_\theta(\theta) = t_\theta(\theta;y) = \frac{\partial}{\partial\theta} \log \pi(\theta|y) = \ell_\theta(\theta;y) + m(\theta), \tag{2}$$

where $\ell_\theta(\theta;y) = \partial \log L(\theta;y)/\partial\theta$ is the score function and $m(\theta) = \partial \log \pi(\theta)/\partial\theta$ is the derivative of the logarithm of the prior.

In this contribution we are interested in deriving the posterior density $\pi(\theta|y)$ such that $\partial \log \pi(\theta|y)/\partial\theta = t_\theta(\theta)$. In particular, we explore two methods for approximating $\pi(\theta|y)$ using MCMC and only $t_\theta(\theta)$ and its first derivative.

In the classical MCMC setting, the usual Metropolis-Hastings (MH) probability of acceptance of a candidate value $\theta^{(t+1)}$, given a chain at stage $\theta^{(t)}$, $\theta^{(t+1)} \sim q(\theta^{(t+1)}|\theta^{(t)})$, is

$$\min\left\{1, \frac{q(\theta^{(t)}|\theta^{(t+1)})}{q(\theta^{(t+1)}|\theta^{(t)})} \frac{\pi(\theta^{(t+1)}|y)}{\pi(\theta^{(t)}|y)}\right\}. \tag{3}$$

To evaluate (3), we must be able to evaluate

$$\frac{\pi(\theta^{(t+1)}|y)}{\pi(\theta^{(t)}|y)} = \exp\{t(\theta^{(t+1)}) - t(\theta^{(t)})\}, \tag{4}$$

in which normalizing constants, as is well known, are not needed. Here we propose two strategies for MCMC sampling even if $t(\theta) = \log \pi(\theta|y)$ is unknown, but its first and second derivatives are available in closed form. The first method (Section 2) considers an approximation based on a Rao score-type statistic based on (2). The second method (Section 3) is based on a local approximation through a Taylor expansion. We present an application to logistic regression with predictive matching priors (Section 4).

2 Method I: Approximation based on the Rao score statistic

A simple analytical way of using (2) is Bayesian statistic is to resort to a posterior distribution derived from a quadratic form of t_θ. This enables us to accomodate two important advantages of the Bayesian approach: the expressiveness of the posterior distribution and the convenient computational method of MCMC.

In particular, let $j(\theta) = -\ell_{\theta\theta}(\theta) = -\partial^2\ell(\theta)/\partial\theta^2$ be the observed Fisher information. Then the approximate posterior density takes the form

$$\pi(\theta|y) \dot{\propto} \exp\left(-\frac{1}{2}t_\theta(\theta)^2 j(\theta)^{-1}\right) = \exp\left(-\frac{1}{2}\tilde{s}(\theta)\right), \tag{5}$$

where $\tilde{s}(\theta) = t_\theta(\theta)^2 j(\theta)^{-1}$ is a Rao score-type statistic based on (2) and the symbol "$\dot{\propto}$" means asymptotic proportionality to first order. In (4), (5) can be used for straightforward MCMC updating for the corresponding Bayesian posterior without any iterative optimization steps (Chernozbukov and Hong, 2003).

Here the idea is to recast (4) in terms of log-likelihood ratio type statistics and then replace the formers by Rao score tests. In particular,

$$\begin{aligned}\frac{\pi(\theta^{(t+1)}|y)}{\pi(\theta^{(t)}|y)} &= \exp\{(t(\theta^{(t+1)}) - t(\tilde{\theta})) - (t(\theta^{(t)} - t(\tilde{\theta}))\} \\ &\doteq \exp\{\tilde{s}(\theta^{(t)})/2 - \tilde{s}(\theta^{(t+1)})/2\}\end{aligned}$$

where $\tilde{\theta}$ is the posterior mode, such that $t_\theta(\tilde{\theta}) = 0$, and "$\doteq$" means asymptotic equality to first order.

3 Method II:I Local approximation through Taylor expansion

Assume, for notational simplicity, θ scalar. A Taylor expansion of $t(\theta)$ at θ_0 gives the approximation

$$t(\theta) \simeq t(\theta_0) + (\theta - \theta_0)t_\theta(\theta_0) + \frac{(\theta - \theta_0)^2}{2}t_{\theta\theta}(\theta_0), \tag{6}$$

where $t_{\theta\theta}(\theta) = (\partial t_\theta(\theta))/(\partial\theta)$. Using (6) we can approximate (4) using

$$\begin{aligned}t(\theta^{(t+1)}) - t(\theta^{(t)}) \approx\ & (\theta^{(t+1)} - \theta^{(t)})t_\theta(\theta_0) \\ & + \frac{[(\theta^{(t+1)} - \theta_0)^2 - (\theta^{(t)} - \theta_0)^2]}{2}t_{\theta\theta}(\theta_0).\end{aligned} \tag{7}$$

Possible choices for θ_0 are $\theta^{(t)}$ or $\bar{\theta} = (\theta^{(t+1)} + \theta^{(t)})/2$. Note that

$$t_{\theta\theta}(\theta) = \ell_{\theta\theta}(\theta) + m_\theta(\theta) = \frac{\partial^2}{\partial\theta^2}\ell(\theta) + \frac{\partial}{\partial\theta}m(\theta) = \frac{\partial^2}{\partial\theta^2}\ell(\theta)\left\{1 + O(n^{-1})\right\}.$$

Hence, in (7) the quantity $t_{\theta\theta}(\theta)$ can be substituted by $\partial^2\ell(\theta)/\partial\theta^2$ with approximately the same level of accuracy.

4 Example: Predictive matching prior for logistic regression

In this section we discuss an example based on the logistic regression model and a predictive matching prior, i.e. a prior ensuring asymptotic equivalence of higher-order frequentist and Bayesian predictive densities (see, e.g., Datta and Mukerjee, 2004).

To give the expression of the proposed predictive matching prior, index notation and Einstein summation convention are convenient. Generic components of θ will be denoted by $\theta_r, \theta_s, \ldots$, with $r, s, \ldots = 1, \ldots, k$. First and second likelihood derivatives are ℓ_r and ℓ_{rs}. By equating the second-order asymptotic expansion of Corcuera and Giummolè (1999) of Bayesian predictive distributions and the frequentist modified estimative density of Komaki (1996), we obtain the proposed predictive matching prior, which is such that

$$t_{\theta_r}(\theta) = \frac{\partial_r \log \pi(\theta)}{\partial \theta_r} = -\frac{1}{2} i^{su}(\theta) \left(E_\theta\{\ell_{rsu}\} - E_\theta\{\ell_{rs}\ell_u\} \right), \tag{8}$$

where $E_\theta\{\cdot\}$ denotes expectation with respect to Y under θ and i^{rs} is the generic element of the inverse of $i(\theta) = E_\theta\{j(\theta)\}$. Note that the term on the right hand side of (8) corresponds to the Firth's adjustment (Firth, 1993) to the score function. In other words, if the prior density is choosen according to (8), then the right hand side of (2) 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 (8). This prior thus validates the use of the method introduced by Firth (1993) for point estimation in the Bayesian framework.

Consider the binary logistic regression, where Y_i is Bernoulli with probability $\pi_i = P(Y_i = 1|x_i)$, where x_i is a known k-variate vector of regressors, $i = 1, \ldots, n$. The model can be expressed as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T \beta,$$

where β is an unknown vector of regression coefficients. The log-likelihood function for β is

$$\ell(\beta) = \sum_{j=1}^{k} \beta_j \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} \log\left(1 + \exp^{\Sigma_{j=1}^{k} \beta_j x_{ij}}\right).$$

In this example the posterior distribution is analytically available since the predictive matching prior coincides with Jeffreys' prior, up to the normalisation constant, and hence classical MCMC can be performed. Therefore, we use it as a benchmark in order to assess the accuracy of the proposed approximation methods. We stress, however, that in many other practical cases, e.g. with non canonical links, (8) does not lead to Jeffreys' prior and and with Firth's adjustment, the posterior is

not available, and thus classical MCMC is not possible. The Fisher information is $i(\beta) = j(\beta) = X^T W X$, where X is the design matrix, and W is a diagonal matrix with elements $(w_1, \ldots, w_n)$, with $w_i = \pi_i(1-\pi_i)$ $(1 \leq i \leq n)$. The modified likelihood equation (2) (Firth, 1993) and its first derivative is

$$t^*_{\beta_r}(\beta) = \sum_{i=1}^{n} (y_i - \pi_i) x_{ir} + t_{\theta_r}(\theta), \quad 1 \leq r \leq k.$$

For this model based on the canonical link, the posterior for β is available and is given by

$$\pi_J(\beta|y,X) \propto L(\beta;y)|i(\beta)|^{1/2}. \tag{9}$$

We compare the posterior (9) with its approximate versions obtained with Methods I and II, using the *endometrial* dataset. The latter has been first analysed by Heinze and Schemper (2002) and reports histology grade (`HG`, the binary response variable) and three risk factors (`NV` a binary indicator for the presence of neovasculation, `PI` the pulsality index of arteria uterina and `EH` the endometrium height) for 79 cases of endometrial cancer.

Consider the model with all the covariates included, i.e.,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \texttt{NV}_i + \beta_2 \texttt{PI}_i + \beta_3 \texttt{EH}_i, \quad 1 \leq i \leq n. \tag{10}$$

Figure 1 compares the marginal posteriors of β_j $(0 \leq j \leq 3)$ obtained from (9) by classical MCMC with the corresponding approximations obtained by Method I (Taylor) and Method II (Rao). We generate 10^6 samples from (9) and from the two approximate posteriors, using the MH algorithm with a multivariate normal proposal distribution. The latter has scale matrix given by the negative of the inverse of the second derivative of $t(\theta)$. For (9) and the approximate posterior obtained by Method II the proposal was tuned to give approximately 0.4 acceptance rate; see Section 5 for the computational details. From Figure 1 we can conclude that the approximation obtained with Method I is very similar to the target (9), whereas the approximation obtained with Method II is less accurate.

5 Concluding remarks

In general, the higher is the acceptance rate the lower is the approximation error of Method I. We recognise that high acceptance rates in MCMC are generally not recommended because the posterior exploration of the MCMC algorithm for a finite time period may be too local. The issue of choosing an optimal acceptance rate is under investigation. However, a practical guidance to circumvent this issue would be as follows. Set a sequence of shrunken proposal matrices obtained by shrinking the main diagonal elements of the starting proposal scaling matrix, and with each of them generate an MCMC sample. Then, for this sequence of MCMC samples, which has increasing acceptance rates, monitor the shape of the resulting marginal

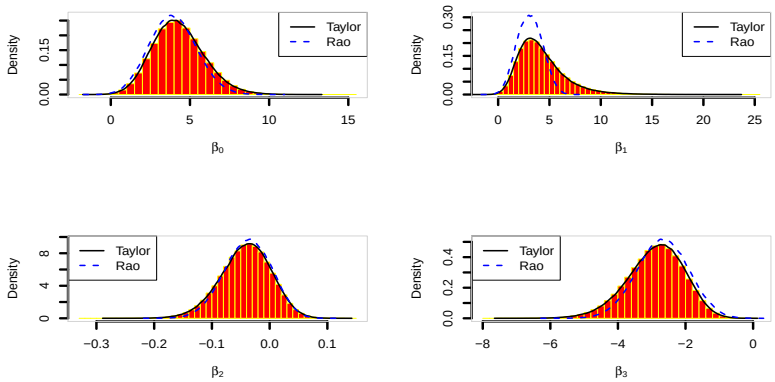

Fig. 1 Marginal posterior distributions for the logistic regression model with the *endometrial* data. The marginals of (10) are illustrated by histograms.

posteriors. If the shape of the latters is reasonably stable across two or three consequent posterior samples, then the MCMC sample with the lowest acceptance rate may by used for posterior inference. This is the strategy adopted in Section 4 which lead to an acceptance rate of 0.70.

Acknowledgement. This research work was partially supported by University of Padova (Progetti di Ricerca di Ateneo 2015, `CPDA153257`) and by PRIN 2015 (grant `2015EASZFS_003`).

References

1. Chernozbukov, V., Hong, H.: An MCMC approach to classical estimation, J. Econometrics **115**, 1234–1241 (2003)
2. Corcuera, J.M., Giummolè, F.: A generalized Bayes rule for prediction. Scand. J. Statist. **26**, 265–279 (1999)
3. Datta, G.S., Mukerjee, R.: Probability Matching Priors: Higher-Order Asymptotics. Lecture Notes in Statistics, Springer (2004)
4. Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika **80**, 27–38 (1993)
5. Heinze, G., Schemper, M: A Solution to the problem of separation in logistic regression. Statist. Med. **21**, 2409–2419 (2002)
6. Komaki, F.: On asymptotic properties of predictive distributions. Biometrika **83**, 299–313 (1996)