Posterior distributions with implicit objective priors

Erlis Ruli
ruli@stat.unipd.it

(joint with L. Ventura, N. Sartori)
Philosophical matters aside, the prior distribution is what makes the Bayesian realm fundamentally different from the classical frequentist one.

For the Bayesian approach, the prior is both its strength (Beauty) and its weakness (Beast).

Beauty: because it permits to integrate, in a principled way, expert opinions or a priori information into the inferential conclusions, delivering thus more precise assessment of the unknowns.

Beast: use of the prior over the unknowns necessarily induces subjectivity into the inferential conclusions...and scientist don’t like subjectivity!
Jeffreys (1946) proposed the density (with respect to the $d$-dimensional Lebesgue measure) $\text{det}(I(\theta))^{1/2}$ as an "objective" or default prior for $\theta$...

though, Jeffreys (1961) recommended against its use in linear regression due to inconsistencies in the degrees of freedom in the error sum of squares.

Bernardo (1976) introduced the concept of reference priors (formalised latter on by Berger et al., 2009), a family of prior distributions obtained by minimising the K-L divergence between the prior and the posterior asymptotically.
Furthermore, in the scalar parameter case, Bernardo (1976) showed that, and under suitable regularity conditions, the Jeffreys prior is a reference prior.

In multidimensional problems the reference prior seems troublesome (depends on the order of the parameters);
- although there have been further developments on this issue (Berger et al. 2015), yet the multidimensional case remains controversial (Rousseau, 2015).

Reference priors tend to have good frequentist properties, although these are mostly ”side-effects” and are not always theoretically guaranteed.
Another remarkable family of default priors are the so-called Matching priors (Datta and Mukerjee, 2004), which aim is to achieve Bayes-frequentist agreement.

- There are many types of matching priors, e.g., quantiles, HPD, predictive distributions, etc. (see, Datta and Sweeting, 2005).

The penalised complexity or PC-prior (Simpson et al. 2017) is another useful idea for building model-based priors.

- It is a proper prior, defined with the aim of penalising the model on the basis of its complexity (as compared to a simpler model).

S. Walker’s yesterday talk ...
Motivation

In some practical applications, accurate estimation of unknown parameters is the main interest. For instance,

• operational risk of a bank institution is often the result of (possibly many) parametric estimation problems. Estimation here is a delicate issue, since if the quantified operational risk is high ⇒ capital risk must be high, thus leading to less bank profits (see, e.g, Danesi et al., 2016).

• Suitable European Commission regulations require that household appliances in the EU market must conform with certain ECO design requirements, such as electricity, water consumption, etc.. Essentially all UE manufacturers must measure, i.e. estimate, and declare certain performance measures of their household appliances. Estimates, often obtained via parametric inferential procedures, should be correct: under- and over-estimation ⇒ higher economic costs.
Motivation II

- These examples (and there are many other of the like) call for methods able to deliver accurate estimates, i.e.
  - for priors that match the true parameter value.
- Reference priors do not guarantee this.
- No matching priors exist which target the true parameter value, and however they can only be constructed for a single parameter at time.
- PC-priors fulfil a different purpose, e.g. model penalising complexity, and however they depend on a scaling parameter and therefore are not completely default.
- The ideal would be a default prior which, as the Jeffreys, is free of scaling constants and delivers accurate parameter estimates...
  - along with suitable measures of uncertainty, i.e. the whole posterior distribution is also of interest.
Bias reduction in a nutshell

- In regular models indexed by the parameter $\theta$, the asymptotic bias of the MLE (i.e. the MAP under the flat prior $\pi(\theta) \propto 1$) can be written as

$$b(\theta) = b_1(\theta)/n + b_2(\theta)/n^2 + ...$$

where $n$ is usually the sample size.

- Extensive frequentist literature is devoted to the bias-reduction problem by removing the first-order term $b_1(\theta)/n$.

Approaches followed can be classified in two groups:

- corrective: get the MLE $\hat{\theta}$ and correct afterwards (analytically, bootstrap, Jackknife, etc.);
- preventive: penalised MLE, i.e. maximise something like $L(\theta)\pi(\theta)$. 
• The “preventive” approach was first proposed by Firth (1993), whereas the “corrective” approach has a much longer history.

• In a nutshell, Firth showed that the solution a suitably modified score equation – in place of the classical score equation – delivers accurate estimates, in the sense that the $b_1(\theta)$ term of these newly-defined estimates is zero.
Notation and Firth (1993)’s rationale

To fix notation (following McCullagh, 1987), let $\theta = (\theta^1, \ldots, \theta^d)$ and let:

- $\ell(\theta) = \log\{L(\theta)\}$ be the likelihood function;
- $\ell_r(\theta) = \partial \ell(\theta)/\partial \theta^r$ be the $r$th component of the score function;
- $\ell_{rs}(\theta) = \partial^2 \ell(\theta)/(\partial \theta^r \partial \theta^s)$;
- $I(\theta)$ is the exp. Fisher information, where the $(r,s)$-cell is $k_{r,s} = n^{-1}E_{\theta}[\ell_r(\theta)\ell_s(\theta)]$, $k^{r,s}$ is the $(r,s)$-cell of its inverse, $k_{r,s,t} = n^{-1}E_{\theta}[\ell_r(\theta)\ell_s(\theta)\ell_t(\theta)]$, $k_{r,st} = n^{-1}E_{\theta}[\ell_r(\theta)\ell_{st}(\theta)]$, be joint null cumulants.

To get an estimate of $\theta$ with reduced bias, Firth (1993) suggests to solve the modified score function

$$\tilde{\ell}_r(\theta) = \ell_r(\theta) + a_r(\theta), \quad r = 1, \ldots, d,$$

where $a_r(\theta)$ is a suitable $O_p(1)$ term, for $n \to \infty$. 
Firth (1993)’s idea and the Jeffreys prior

- For general models and in the summation convention,

\[ a_r = k^{u,v}(k_{r,u,v} + k_{r,uv})/2. \]

- Let \( \hat{\theta}^* \), be the solution of (1). Then Firth (1993) showed that the \( b_1(\theta) \) term of \( \hat{\theta}^* \) vanishes, i.e. \( E_{\theta}(\hat{\theta}^*) = \theta + O(n^{-2}) \).

- Interestingly, if the model belongs to the canonical exponential family, i.e. if the model can be written in the form

\[ \exp \left[ \sum_{i=1}^{d} \theta_i s_i(y) - \kappa(\theta) \right] h(y), \quad y \in \mathbb{R}^d \]

then

\[ a_r = (1/2) \partial \log \det(I(\theta))/\partial \theta^r \]

that is, \( \hat{\theta}^* \) is the MAP under the Jeffreys prior!
Towards Bias-Reduction priors

- These results suggest that $a_r, r = 1, \ldots, d$, could be a nice candidate as a default prior for $\theta$, because:
  - it is built from the model at hand;
  - it delivers unbiased estimates;
  - it is free of tuning or scaling parameters, just like the Jeffreys;
- $a_r$ could also be seen as a kind of matching prior, with the aim achieving Bayes-frequentist synthesis in terms of the true parameter value.
- On the other hand, under this prior, only the MAP is guaranteed to be unbiased.
- Although the MAP is not perfect, it is fast to compute!
The Bias-Reduction priors are implicit!

- We call this prior the Bias-Reduction prior or **BR-prior**, and define it implicitly as

\[ \pi_{BR}^m(\theta) \text{ such that } \frac{\partial \log \pi_{BR}^m(\theta)}{\partial \theta^r} = a_r(\theta), r = 1, \ldots, d. \]

- Note again that for canonical exponential models the prior is

\[ \pi_{BR}^m(\theta) = \det(I(\theta))^{1/2}, \]

whereas for general models no explicit forms are available for its density.
Dealing with the implicitity

- In general models, use of $\pi_{BR}^m(\theta)$ leads to an “implicit” posterior, that is, a posterior for which we can evaluate derivatives of their log-density but not the log-density itself.

- Unfortunately, this is a kind of “intractability” which cannot be dealt with by classical methods such as MCMC, importance sampling or Laplace approximation.

- ABC isn’t of use either ...
We explore two methods for approximating such "implicit" posteriors:

(a) a global approximation method based on the quadratic Rao-score function.

(b) a local approximation of the log-posterior ratio via Taylor expansions and to be used in MCMC in place of the true log-posterior ratio.

Langevin diffusion Monte Carlo has been deemed as a useful alternative to (a) and (b) but it has not been explored yet (work in progress with P. Jacob)
Classical Metropolis-Hastings

- To introduce methods (a) and (b), first let us recall the usual Metropolis-Hastings acceptance probability of a candidate value $\theta^{(t+1)}$, drawn from $q(\cdot|\theta^{(t)})$ given the chain at state $\theta^{(t)}$:

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

where $\pi(\theta|y)$ denotes the posterior density.

- The acceptance probability depends, among other things, on the posterior ratio:

$$
\frac{\pi(\theta^{(t+1)}|y)}{\pi(\theta^{(t)}|y)} = \exp \left[ \tilde{\ell}(\theta^{(t+1)}) - \tilde{\ell}(\theta^{(t)}) \right],
$$

where $\tilde{\ell}(\theta) = \ell(\theta) + \log \pi(\theta)$. 
Method (a): global approximation via the Rao-score

- Let $\hat{\theta}^*$ be the MAP, i.e. the solution of the equation
  $\ell_\theta(\theta) = \partial \tilde{\ell}(\theta)/\partial \theta = 0$, then

\[
\exp \left[ \tilde{\ell}(\theta^{(t+1)}) - \tilde{\ell}(\theta^{(t)}) \right] = \exp \left[ \tilde{w}(\theta^{(t)})/2 - \tilde{w}(\theta^{(t+1)})/2 \right],
\]

where $\tilde{w}(\theta) = 2(\tilde{\ell}(\hat{\theta}^*) - \tilde{\ell}(\theta))$, is the penalised log-likelihood ratio statistic.

- For a fixed $\theta$, assuming the prior is $O(1)$ and for large $n$

\[
\tilde{w}(\theta) \sim \tilde{s}(\theta) = n^{-1} \tilde{\ell}_\theta(\theta)^\top I(\theta)^{-1} \tilde{\ell}_\theta(\theta).
\]

- Thus, for each $\theta^{(t)}$, we can approximate $\tilde{w}(\theta^{(t)})$ by $\tilde{s}(\theta^{(t)})$. 
Method (b): local approximation (Taylor expansion)

• Consider a Taylor approximation of \( \tilde{\ell}(\theta(t)) \) and \( \tilde{\ell}(\theta(t+1)) \) (assuming \( d = 1 \) for notational convenience)

\[
\tilde{\ell}(\theta(t)) \approx \tilde{\ell}(\bar{\theta}) + (\theta(t) - \bar{\theta})\tilde{\ell}_{\theta}(\bar{\theta}) + (\theta(t) - \bar{\theta})^2\tilde{\ell}_{\theta\theta}(\bar{\theta})/2!,
\]

\[
\tilde{\ell}(\theta(t+1)) \approx \tilde{\ell}(\bar{\theta}) + (\theta(t+1) - \bar{\theta})\tilde{\ell}_{\theta}(\bar{\theta}) + (\theta(t+1) - \bar{\theta})^2\tilde{\ell}_{\theta\theta}(\bar{\theta})/2!.
\]

• Then replacing these approximations in the log-posterior ratio we get

\[
\tilde{\ell}(\theta(t+1)) - \tilde{\ell}(\theta(t)) \approx (\theta(t+1) - \theta(t))\tilde{\ell}_{\theta}(\bar{\theta}) + [((\theta(t+1) - \bar{\theta})^2 - (\theta(t) - \bar{\theta})^2]\tilde{\ell}_{\theta\theta}(\bar{\theta})/2!.
\]

• Possible choices for \( \bar{\theta} \) are \( a\theta(t+1) + (1 - a)\theta(t) \), \( a \in [0, 1] \).
Method (b) pictorially

\[
\log\text{-posterior}\left(\theta(t)\right) = \log\text{-posterior}\left(\theta(t+1)\right)
\]

O'Bayes 2019, Univ. Warwick, 28 June - 02 July 2019
Some comments on (a) and (b)

- Method (a) is global in the sense that it can be seen as an approximation which targets the posterior distribution, replacing it with the quadratic Rao score function.

- Method (b) targets the log-posterior ratio in the M-H ratio, and offers a local approximation through Taylor expansion. It turns out that expanding in the middle of $\theta(t+1)$ and $\theta(t)$, i.e. $a = 1/2$, gives better approximations. Furthermore, $\theta(t+1)$ shouldn’t be too far from $\theta(t)$...
  - but this might lead to larger autocorrelation (slower convergence).
log-posterior ratio: (a) vs (b)

For the posterior distribution in the figure:

- we take a regular grid 
  \[ \{\theta_1, \theta_2, \ldots, \theta_{100}\} \] in [0.1, 7]
  and
- evaluate the log-posterior ratio 
  \[ \tilde{\ell}(\theta_i) - \tilde{\ell}(\theta_i + k \cdot se) \],

where 
\[ se = 1/\sqrt{I(\hat{\theta}^*)}. \]
log-posterior ratio: (a) vs (b)
Example 1:
The model is Poisson($\lambda$),
the prior is Gamma(4/a, a), $a = 2.5$,
the sample of size $n = 5$ is generated with $\lambda = a = 2.5$. 
Poisson($\lambda$): method (b)

- **4 x sd.prop** (acc.rate 30%)
- **3 x sd.prop** (acc.rate 38%)
- **2 x sd.prop** (acc.rate 50%)
- **0.5 x sd.prop** (acc.rate 85%)
Poisson($\lambda$): method (b)

4 x sd.prop
(acc.rate 30%)

3 x sd.prop
(acc.rate 38%)

2 x sd.prop
(acc.rate 50%)

0.5 x sd.prop
(acc.rate 85%)
Poisson($\lambda$): (a) vs (b)

Distributions for $\lambda$

- Prior
- Target
- Rao (1)
- Rao (2)
Example 2
The endometrial data set:
was first analysed by Heinze and Schemper (2002), and was
originally provided by Dr E. Asseryanis from the Medical University
of Vienna.
The MLE is problematic!

For NV we notice some degree of separation (in terms of the response HG), which presumably leads to a highly flat likelihood function for the associated regression coefficient.
Posteriors with the BR-prior (i.e. Jeffreys')

Acc. rates: Classical 40%, Rao 33%, Taylor 61%
Autocorrelations of the chains

MCMC: beta0

MCMC: beta1

MCMC: beta2

MCMC: beta3

Rao: beta0

Rao: beta1

Rao: beta2

Rao: beta3

Taylor: beta0

Taylor: beta1

Taylor: beta2

Taylor: beta3
Comments on Example 2

- Approximation based on Taylor expansion seem to work better than quadratic Rao score function.
- Differences between the two methods seem particularly relevant in cases with “problematic” parameters such as $\beta_1$, the coefficient of NV.
- The presence of such problematic parameters however seems to lead to highly correlated chains (both for classical MCMC and Taylor)...
- To go deeper into the last two points, let’s exaggerate things a bit by considering the following extreme scenario.
Example 3 (a posterior with non-standard shape):
Logistic regression with complete separation
The MLE is infinite!

20 observations with complete separation

Contours of the log-likelihood (solid)
log-posterior with the Jeffreys prior (dashed)

> (glm(y~x,family=binomial))
Call: glm(formula = y ~ x, family = binomial)

Coefficients:
(Intercept)           x
        -225.3     1878.8

Degrees of Freedom: 19 Total (i.e. Null); 18 Residual
Null Deviance: 27.73
Residual Deviance: 1.035e-07  AIC: 4

Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
Standard Metropolis-Hasting leads to very autocorrelated chains!
Adaptive MH vs (a) vs (b)

Histogram of adaptive MCMC

Contours of the log-likelihood (solid)
log-posterior with the Jeffreys prior (dashed), Rao score posterior (dots)

Contours of the log-likelihood (solid)
log-posterior with the Jeffreys prior (dashed), Taylor posterior (dots)
Adaptive MH vs (a) vs (b): comments

- The Rao score function – method (a) – seems to give a bimodal posterior.
- The approximation based on Taylor expansion – method (b) – gets closer to the target.
- However, the posterior sample drawn with method (b), using standard M-H, is highly autocorrelated...
Prior elicitation is a difficult task when no a priori information is available.

Default priors such as the Jeffreys, the reference or matching priors could be of practical use.

However, in multidimensional cases, matching and reference priors are typically hard to derive.

In practical applications we may be looking for unbiased parameter estimates.

Our proposal is then to use a Bias-Reduction prior which:
  ▶ can be used as a default and scaling-free prior for the whole vector of parameters
  ▶ delivers MAP estimates that are second-order unbiased.
Wrap up with final remarks

- In canonical exponential families, use of the BR-prior amounts to using the Jeffreys prior...
- In other cases, the BR-prior is available only via the first derivative of its log-density which in general does not coincide with the Jeffreys.
- Unfortunately, use of BR-priors leads to a kind computational intractability that seem not solvable by classical MCMC, IS, ABC, or Laplace.
Wrap up with final remarks

- We explored two methods for approximating the posterior with such implicit priors.
- The method based on Taylor expansion seem to work better.
- However, for its success proposal jumps must be small.
- Unfortunately, small proposal jumps means slower posterior exploration...
- How to speed up posterior exploration using small jumps is an open problem...

suggestions?
Some selected references


Thank you for your attention!
• In multidimensional cases, matching and reference priors are typically hard to derive.
• In practical applications we may be looking for unbiased parameter estimates.
• Our proposal is then to use a Bias-Reduction prior which:
  ▶ can be used as a default and scaling-free prior for the whole vector of parameters
  ▶ delivers MAP estimates that are second-order unbiased.