Default Bias-Reduced Bayesian Inference

Erlis Ruli

ruli@stat.unipd.it

(joint work with L. Ventura, N. Sartori)

Università degli Studi di Padova | Dipartimento di Scienze Statistiche

StaTalk 2019 @UniTs
22 November 2019

Why does it matter?

In some (many?) industrial and business decisions, statistical inferences play a crucial roles. For instance,

- quantification of a bank operational risk (Danesi et al., 2016) determines the bank's capital risk, i.e. the amount of money to be promptly available in order to deal with possible future losses. Inaccurate estimation of the capital risk leads to higher economic costs.

- household appliances in the EU market must conform with certain ECO design requirements, such as electricity (A+++, A++, etc.), water consumption, etc.. UE manufacturers must estimate and declare performance measures of their appliances... again, inaccurate estimation lead to higher economic costs.

- of course the list is much larger, e.g. think about medical instruments, diagnostic markers, etc.

Is Bayes accuracte?

Given our model P_θ (sufficiently regular), the data y, the likelihood function $L(\theta; y)$ and a prior $p(\theta)$, the posterior distribution is

$$p(\theta|y) \propto L(\theta; y)p(\theta).$$

Typically, a point estimate of θ is required and we could use the maximum a posteriori (MAP)

$$\tilde{\theta} = \arg\max_\theta p(\theta|y).$$

Question

How “accurate” is $\tilde{\theta}$?

Rules of the game

Classical full parametric inference problem in which, θ^0 is the true but unknown parameter value, P_{θ^0} is the true model and $\tilde{\theta}$ is our Bayes guess for θ^0.

We deal only with regular models P_θ, i.e. models for which the Fisher information $I(\theta) = E_\theta\left[(d\log L(\theta;y)/d\theta)^2\right]$ exists.

The bias $b(\theta^0) = E_{\theta^0}(\tilde{\theta}) - \theta^0$ is one popular way of measuring the accuracy of an estimator; $E_\theta(\cdot)$ is the expectation with respect to model P_{θ^0}.

Ideally we'd like zero bias, i.e. maximum accuracy, but in practice that's seldomly possible.

The typical behaviour of bias

If $p(\theta) \propto 1$, then $\tilde{\theta}$ is the maximum likelihood estimator (MLE). In this case, in independent samples of size n, we know that, typically

$$E_{\theta^0}(\tilde{\theta}) = \theta^0 + b_1(\theta^0)n^{-1} + b_2(\theta^0)n^{-2} + \cdots , \tag{1}$$

$b_k(\theta^0)$'s , $k = 1, 2, \dots,$ are higher-order bias terms that do not depend on n.

If we have a guess for $b_1(\theta^0)$ our estimator $\tilde{\theta}$ would be second-order unbiased. There are some non-Bayesian ways for getting rid of $b_1(\theta^0)$, when $\tilde{\theta}$ is the MLE (more on this latter).

Therefore, if the prior is flat, the MAP is as accurate as the MLE, i.e. is first-order unbiased.

What about the bias of $\tilde{\theta}$ in typical Bayesian analyses?

Is typical Bayes accurate?

In practice, the prior $p(\theta) \propto 1$ is seldomly used on the whole parameter vector; perhaps much typical choices are:

- subjective or proper priors
- default and often improper priors such as the Jeffreys'(Jeffreys, 1946), the reference (Bernardo, 1976), matching (Datta & Mukerjee, 2004) priors
- or the more recent Penalised Complexity (Simpson et al., 2017)

In some specific models, some of these priors could lead to accurate estimators, i.e. second-order unbiased (more on this latter) but none of them can guarantee this accuracy in general.

Roughly speaking, if the prior is not too data-dominated, the bias of $\tilde{\theta}$ will behave, at best, as in (**??**).

Even a small bias could be practically relevant

Typical Bayes does not guarantee – in full generality and even in the reasonable class of regular models – higher-accuracy in estimation.

You might think that

"the bias is an $O(n^{-1})$ term, so for large amounts of data it won't be a practical problem".

TRUE. But, there are at least two reasons as to why even the first-order term $b_1(\theta)$ could be relevant in practice:

- large samples could be economically impossible since measurement can be extremely costly, e.g. 3000$ per observation in the case of testing a washing machine for ECO design requirements
- even a tiny bias can have a large practical impact, especially when estimating tails of a distribution such as in operational risk.

Desiderata for accurate Bayes estimation

We desire therefore a prior that matches the true parameter value closer than the typical ones, and, possibly, free of hyper-parameters...just like the Jeffreys' or the reference.

We saw that such a “matching” is not always guaranteed by the aforementioned priors, including $p(\theta) \propto 1$.

Note: there is nothing wrong with those priors, they just don't fit our purpose of getting accurate estimates.

Obviously, with this desired prior, we want to get the whole posterior distribution, and not just $\tilde{\theta}$.

How to build such a desired prior ?

Bias reduction in a nutshell

Fortunately, there is an extensive frequentist literature devoted to the bias-reduction problem in which one tries to remove, i.e. estimate, the term $b_1(\theta)/n$.

Two approaches for doing this:

corrective: compute the MLE first, and correct afterwards (analytically, bootstrap, Jackknife, etc.);

preventive: penalised MLE, i.e. maximise something like $L(\theta)p(\theta)$, for a suitable $p(\theta)$.

Preventive bias-reduction

The “preventive” approach was first proposed by Firth (1993), whereas the “corrective” one is much older.

In a nutshell: Firth showed that, solving a suitably modified score equation – in place of the classical score equation – delivers more accurate estimates, in the sense that the $b_1(\theta)$ term of these newly-defined estimates turns out to be zero.

In order to be more detailed, we need further notation...

Notation and Firth (1993)'s rationale

Following McCullagh (1987), let $\theta = (\theta^1, \ldots, \theta^d)$ and set

- $\ell(\theta) = \log\{L(\theta; y)\}$ the likelihood function;
- $\ell_r(\theta) = \partial \ell(\theta)/\partial \theta^r$ the rth component of the score function;
- $\ell_{rs}(\theta) = \partial^2 \ell(\theta)/(\partial \theta^r \partial \theta^s)$;
- $I(\theta)$ the Fisher information, with (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.

Firth (1993) suggests to solve the modified score function

$$\tilde{\ell}_r(\theta) = \underbrace{\ell_r(\theta)}_{\text{score}} + \underbrace{a_r(\theta)}_{\text{modification factor}}, \quad r = 1, \ldots, d, \qquad (2)$$

where $a_r(\theta)$ is a suitable $O_p(1)$ term, for $n \to \infty$.

Firth (1993) meets Jeffreys' prior ?!

For general models (using summation convention)

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

If $\tilde{\theta}^*$, is the solution of (**??**), then Firth (1993) showed that the $b_1(\theta)$ term of $\tilde{\theta}^*$ vanishes, i.e. $E_{\theta^0}(\tilde{\theta}^*) = \theta^0 + O(n^{-2})$.

Interestingly enough, 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

$$a_r = (1/2)\partial[\log|I(\theta)|]/\partial\theta^r\,.$$

That is, $\tilde{\theta}^*$ is the MAP under the Jeffreys prior!

Towards priors with higher accuracy

Firth(1993)'s results suggest that a_r $(r \leq d)$, could be a suitable candidate as a default prior for the accurate estimation of θ, since:

- it is built from the model at hand;
- it delivers second-order unbiased estimates;
- it is free of tuning or scaling parameters, just like the Jeffreys;

From a Bayesian perspective, a_r is a kind of matching "prior", that tries to acheive Bayes-frequentist synthesis in terms of the true parameter value θ^0, when the estimator is the MAP.

Although the MAP is not the only Bayes estimator for θ^0, with respect to others, it is fast to compute.

The Bias-Reduction prior

Thus, a_r is the ingredient we are looking for in order to build our prior. We call this the Bias-Reduction prior or BR-prior, and we define it implicitly as

$$p^m_{BR}(\theta) = \{\theta :\, \partial \log p^m_{BR}(\theta)/\partial\theta^r = a_r(\theta)\,, r = 1, \ldots, d\,\}. \qquad (3)$$

Note that, for canonical exponential models, the BR-prior is explicit,

$$p^m_{BR}(\theta) = \mathrm{det}(I(\theta))^{1/2}\,,$$

but for general models is available only in the form of (**??**).

Dealing with the implicity

Use of ${\pi^m}_{BR}(\theta)$ in general models, leads to an “implicit” posterior, that is, a posterior for which derivatives of the log-density are available 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.

Approximate Bayesian Computation (ABC) isn't of use either ...

Dealing with the implicity (cont'ed)

For approximating such implicit posteriors, we explore two methods:

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

(b) a local approximation of the log-posterior ratio for MCMC algorithms.

Classical Metropolis-Hastings

To introduce methods (a) and (b), first, let's 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 $\theta^{(t)}$:

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

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

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

where $\tilde{\ell}(\theta) = \ell(\theta) + \log p(\theta)$.

Method (a): global approximation via the Rao-score

- Given $\tilde{\theta}$ the MAP of θ, i.e. the solution of the equation $\tilde{\ell}_\theta(\theta) = \partial\tilde{\ell}(\theta)/\partial\theta = 0$, 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 $\tilde{w}(\theta) = 2(\tilde{\ell}(\tilde{\theta}) - \tilde{\ell}(\theta))$, is the penalised log-likelihood ratio statistic.

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

  $$\tilde{w}(\theta) \;\dot{\sim}\; \tilde{s}(\theta) = n^{-1}\tilde{\ell}_\theta(\theta)^\mathsf{T} I(\theta)^{-1}\tilde{\ell}_\theta(\theta) \,.$$

- 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)

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

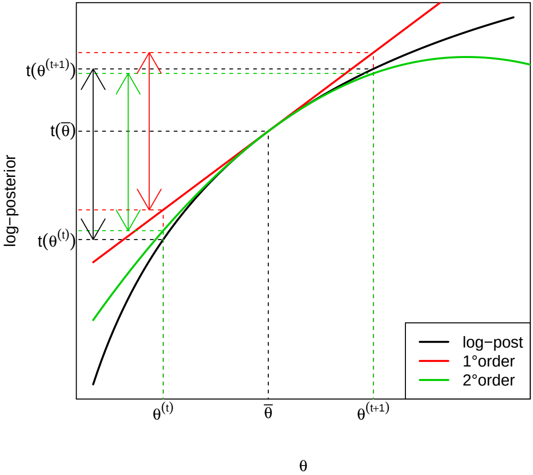
$$\tilde{\ell}(\theta^{(t+1)}) \approx \tilde{\ell}(\bar{\theta}) + (\theta^{(t+1)} - \bar{\theta})\tilde{\ell}_\theta(\bar{\theta}) + (\theta^{(t+1)} - \bar{\theta})^2\tilde{\ell}_{\theta\theta}(\bar{\theta})/2!\,.$$

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

$$\begin{aligned}\tilde{\ell}(\theta^{(t+1)}) - \tilde{\ell}(\theta^{(t)}) \quad &\approx \quad (\theta^{(t+1)} - \theta^{(t)})\,\tilde{\ell}_\theta(\bar{\theta})\; + \\ &\quad\; [(\theta^{(t+1)} - \bar{\theta})^2 - (\theta^{(t)} - \bar{\theta})^2]\tilde{\ell}_{\theta\theta}(\bar{\theta})/2!.\end{aligned}$$

- Possible choices for $\bar{\theta}$ are $a\theta^{(t+1)} + (1-a)\theta^{(t)}$, $a \in [0, 1]$.

Method (b) pictorially



Some comments on (a) and (b)

Method (a) is a global approximation in the sense that it approximates the whole posterior density, by (a certain function of) the quadratic Rao-score function.

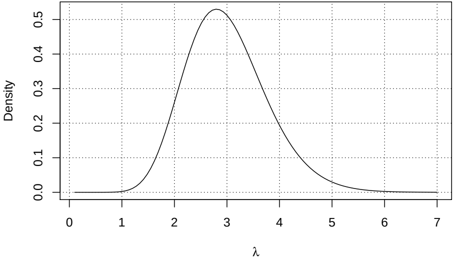
Method (b) targets the log-posterior ratio in the M-H ratio, and offers a local approximation through Taylor expansions...

Approximation of the log-posterior ratio: (a) vs (b)

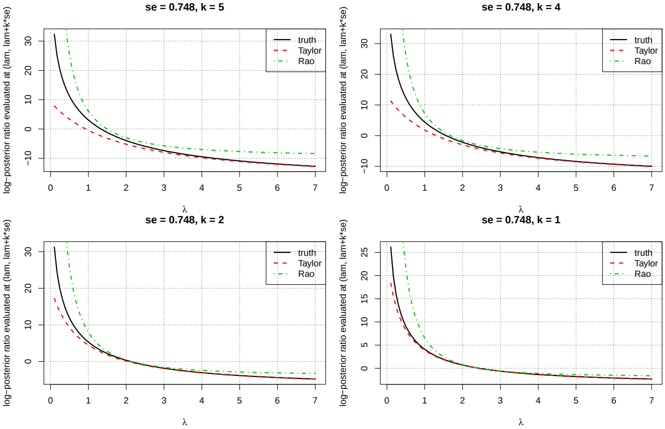
For the posterior distribution in the figure:

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

where $se = 1\Big/\sqrt{I(\tilde{\theta})}$.
$k > 0$ controls the degree of "locality" of the Taylor approximations; the lower k more local is the approximation.

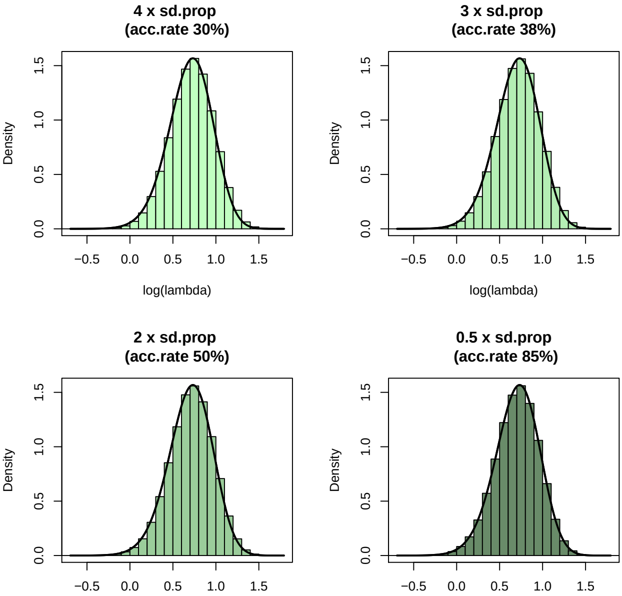


Approximation of the log-posterior ratio: (a) vs (b)



Example 1:
The model is Poisson(λ),
the prior is Gamma(4/a, a), a $= 2.5$,
the sample of size $n = 5$ is generated with $\lambda = a = 2.5$.

Poisson(λ): method (b)



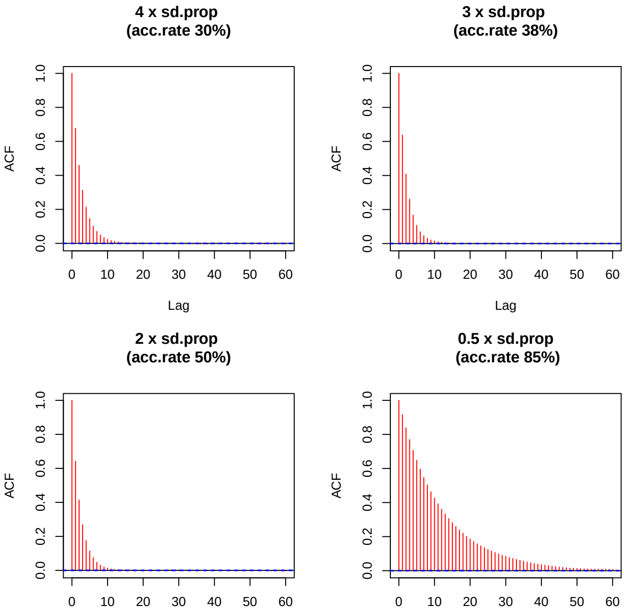
4 x sd.prop (acc.rate 30%)

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

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

0.5 x sd.prop (acc.rate 85%)

Poisson(λ): method (b)



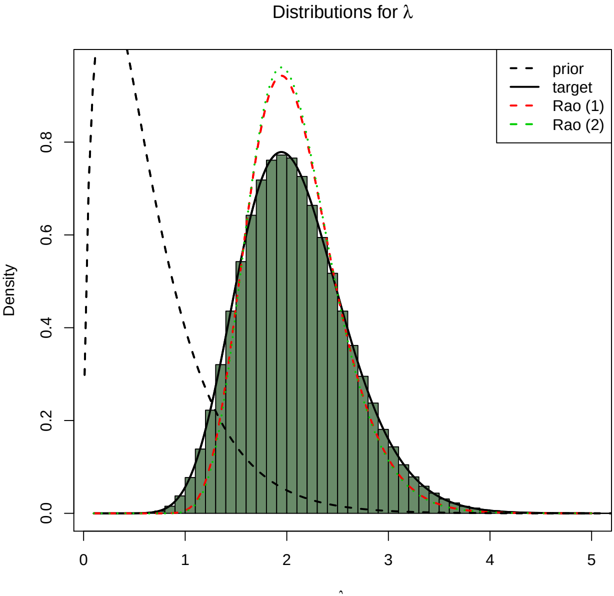
4 x sd.prop (acc.rate 30%)

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

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

0.5 x sd.prop (acc.rate 85%)

Poisson(λ): (a) vs (b)

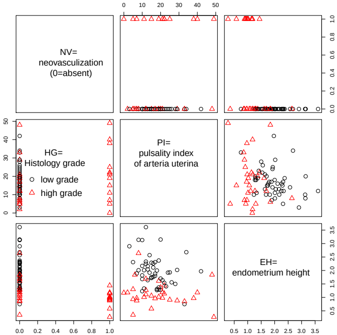


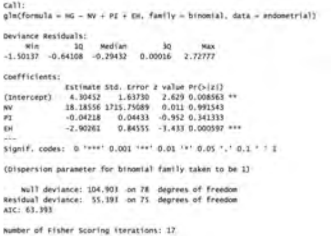
Example 2

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!



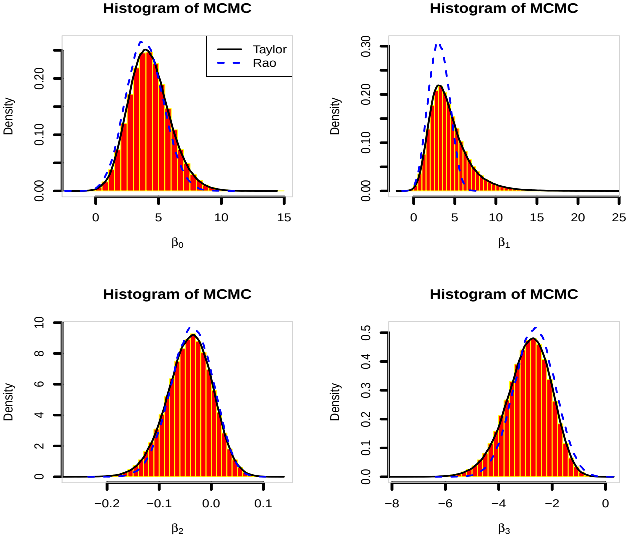


```
Call:
glm(formula = HG ~ NV + PI + EH, family = binomial, data = endometrial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50137  -0.64108  -0.29432   0.00016   2.72777

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.30452    1.63730   2.629 0.008563 **
NV           18.18556 1715.75089   0.011 0.991543
PI           -0.04218    0.04433  -0.952 0.341333
EH           -2.90261    0.84555  -3.433 0.000597 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  55.393  on 75  degrees of freedom
AIC: 63.393

Number of Fisher Scoring iterations: 17
```

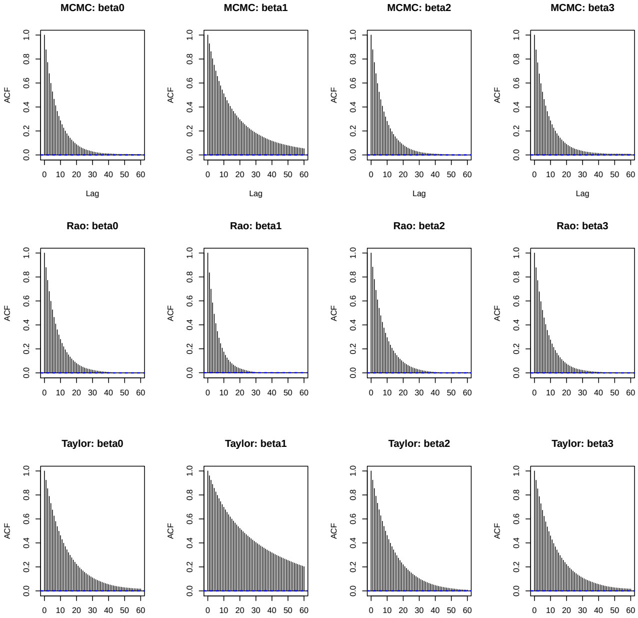
For NV we notice some degree of separation (in terms of the response HG), which presumably leads to a highly flat likelihood function for the associated regression coefficient.

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

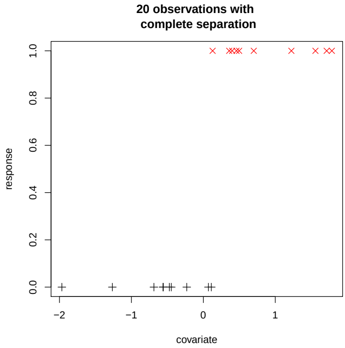
Taylor: beta3

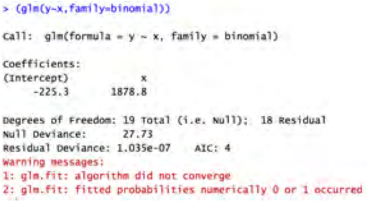
Comments on Example 2

- Approximation based on Taylor expansion seem to work better than the quadratic Rao-score function.
- Differences between the two methods seem particularly relevant in cases with "problematic" parameters such as β_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.

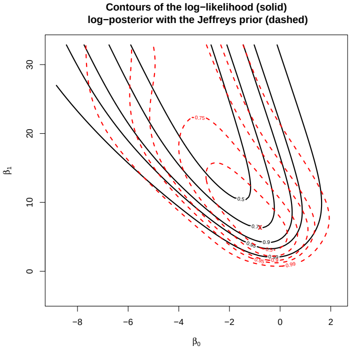
Example 3 (a posterior with non-standard shape): Logistic regression with complete separation

The MLE is infinite!

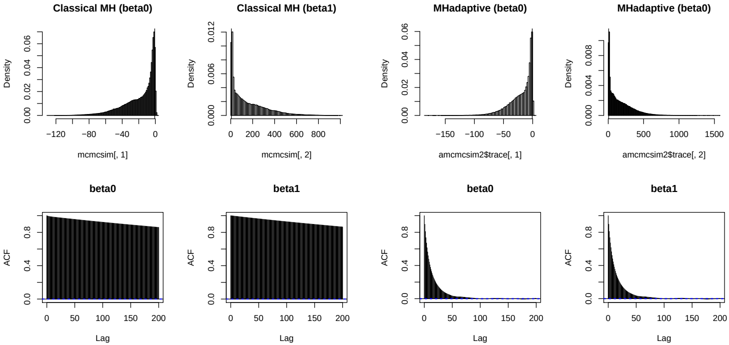

20 observations with complete separation



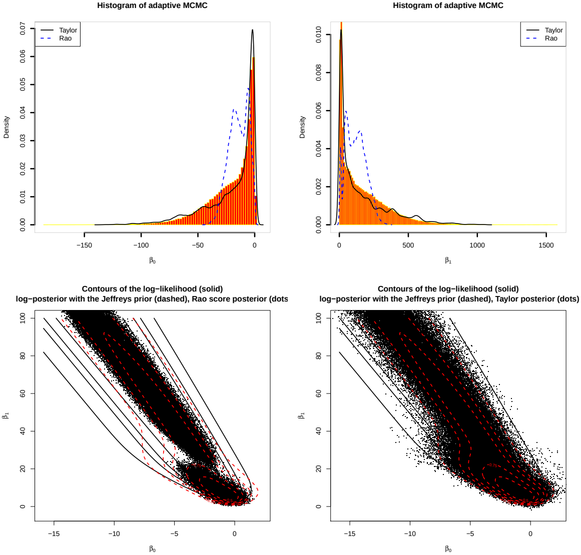
```
> (glm(y~x,family=binomial))

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
```


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

Standard Metropolis-Hasting leads to very autocorrelated chains!



Adaptive MH vs (a) vs (b)



Histogram of adaptive MCMC

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...

Wrap up with final remarks

- 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 accurate 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 expansions seems 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...

Some selected references

1. Berger, Bernardo & Sun (2009). The formal definition of reference priors. *Ann. Statist.* **37**, 905–938.
2. Berger, Bernardo & Sun (2015). Overall objective priors. *Bayesian Anal.* **10**, 189–221.
3. Danesi, Piacenza, Ruli & Ventura (2016). Optimal B-robust posterior distributions for operational risk. *J. Op. Risk* **11**, 35–54.
4. Datta & Sweeting (2005). Probability matching priors. In Handbook of Statistics 25 (D. K. Dey and C. R. Rao, eds.). North-Holland, Amsterdam.
5. Datta & Mukerjee (2004). Probability Matching Priors: Higher-Order Asymptotics. Lecture Notes in Statistics, Springer.
6. Simpson, Rue, Riebler, Martins & Sørbye (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32**, 1–28.
7. Jeffreys (1964). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. A* **186**, 453–461.
8. Firth (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993)

Wrap up with final remarks

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.

The Taylor method works better with small proposal jumps. But small proposal jumps means slower posterior exploration... How to speed up posterior exploration using small jumps is an open problem...

Suggestions?