

# Bayesian model selection and averaging with mombf

David Rossell

The `mombf` package implements Bayesian model selection (BMS) and model averaging (BMA) for regression (linear, asymmetric linear, median and quantile regression, accelerated failure times) and mixture models. This is the main package implementing *non-local priors* (NLP) but some other priors are also implemented, e.g. Zellner's prior in regression, Normal-IWishart priors for mixtures. NLPs are briefly reviewed here, see Johnson and Rossell (2010, 2012) for their model selection properties and Rossell and Telesca (2017) for parameter estimation and posterior sampling. The main `mombf` features are:

- Density, cumulative density, quantiles and random numbers for NLPs
- BMS (Section 2, Johnson and Rossell (2010, 2012)) and BMA (Section 5, Rossell and Telesca (2017)) in linear regression.
- Exact BMS and BMA under orthogonal and block-diagonal linear regression (Section 6, Papaspiliopoulos and Rossell (2016)).
- BMS and BMA for certain generalized linear models (Section 7, Johnson and Rossell (2012); Rossell et al. (2013))
- BMS in linear regression with non-normal residuals (Rossell and Rubio, 2018). Particular cases are Bayesian versions of asymmetric least squares, median and quantile regression.
- BMS for Accelerated Failure Time models.
- BMS for mixture models (Section 8, currently only Normal mixtures) (Fúquene et al., 2018).

This manual introduces basic notions underlying NLPs and the main R functions implementing model selection and averaging. Most of these are

internally implemented in C++ so, while they are not optimal in any sense they are designed to be minimally scalable to large sample sizes  $n$  and high dimensions (large  $p$ ).

## 1 Quick start

The main BMS function is `modelSelection`. BMA is also available for some models, mainly linear regression and Normal mixtures. Details are in subsequent sections, here we illustrate quickly how to get posterior model probabilities, marginal posterior inclusion probabilities, BMA point estimates and posterior intervals for the regression coefficients and predicted outcomes.

```
> library(mombf)
> set.seed(1234)
> x <- matrix(rnorm(100*3),nrow=100,ncol=3)
> theta <- matrix(c(1,1,0),ncol=1)
> y <- x %*% theta + rnorm(100)
> #Default MOM prior on parameters
> priorCoef <- momprior(tau=0.348)
> #Beta-Binomial prior for model space
> priorDelta <- modelbbprior(1,1)
> #Model selection
> fit1 <- modelSelection(y ~ x[,1]+x[,2]+x[,3], priorCoef=priorCoef, priorDelta=priorDelta)
```

Enumerating models...

Computing posterior probabilities..... Done.

```
> #Posterior model probabilities
> postProb(fit1)
```

	modelid	family	pp
7	2,3	normal	9.854873e-01
8	2,3,4	normal	7.597369e-03
15	1,2,3	normal	6.771575e-03
16	1,2,3,4	normal	1.437990e-04
3	3	normal	3.240602e-17
5	2	normal	7.292230e-18
4	3,4	normal	2.150174e-19
11	1,3	normal	9.892869e-20
6	2,4	normal	5.615517e-20
13	1,2	normal	2.226164e-20
12	1,3,4	normal	1.477780e-21
14	1,2,4	normal	3.859388e-22
1		normal	2.409908e-25
2	4	normal	1.300748e-27
9	1	normal	2.757778e-28
10	1,4	normal	3.971521e-30

```

> #BMA estimates, 95% intervals, marginal post prob
> coef(fit1)

              estimate      2.5%      97.5%      margpp
(Intercept) 0.007230966 -0.02624289 0.04085951 0.006915374
x[, 1]      1.134700387  0.93487948 1.33599873 1.000000000
x[, 2]      1.135810652  0.94075622 1.33621298 1.000000000
x[, 3]      0.000263446  0.00000000 0.00000000 0.007741168
phi         1.100749637  0.83969879 1.44198567 1.000000000

> #BMA predictions for y, 95% intervals
> ypred= predict(fit1)
> head(ypred)

      mean      2.5%      97.5%
1 -0.8936883 -1.1165154 -0.67003262
2 -0.2162846 -0.3509188 -0.08331286
3  1.3152329  1.0673711  1.56348261
4 -3.2299241 -3.6826696 -2.77728625
5 -0.4431820 -0.6501280 -0.23919345
6  0.7727824  0.6348189  0.90977798

> cor(y, ypred[,1])

      [,1]
[1,] 0.8468436

```

## 2 Basics on non-local priors

The basic motivation for NLPs is what one may denominate the *model separation principle*. The idea is quite simple, suppose we are considering a (possibly infinite) set of probability models  $M_1, M_2, \dots$  for an observed dataset  $y$ , if these models overlap then it becomes hard to tell which of them generated  $y$ . The notion is important because the vast majority of applications consider nested models: if say  $M_1$  is nested within  $M_2$  then these two models are not well-separated. Intuitively, if  $y$  are truly generated from  $M_1$  then  $M_1$  will receive high integrated likelihood however that for  $M_2$  will also be relatively large given that  $M_1$  is contained in  $M_2$ . We remark that the notion remains valid when none of the posed models are true, in that case  $M_1$  is the model of smallest dimension minimizing Kullback-Leibler divergence to the data-generating distribution of  $y$ . A usual mantra is that performing Bayesian model selection via posterior model probabilities (equivalently, Bayes factors) automatically incorporates Occam's razor, e.g.  $M_1$  will eventually be favoured over  $M_2$  as the sample size  $n \rightarrow \infty$ . This statement is correct but can be misleading: there is no guarantee that parsimony is enforced to an adequate extent, indeed it turns out to be insufficient in many

practical situations even for small  $p$ . This issue is exacerbated for large  $p$  to the extent that one may even lose consistency of posterior model probabilities (Johnson and Rossell, 2012) unless sufficiently strong sparsity penalties are introduced into the model space prior. See (Rossell, 2018) for theoretical results on what priors achieve model selection consistency and a discussion on how to set priors that balance sparsity versus power to detect truly active coefficients.

Intuitively, NLPs induce a probabilistic separation between the considered models which, aside from being philosophically appealing (to us), one can show mathematically leads to stronger parsimony. When we compare two nested models and the smaller one is true the resulting BFs converge faster to 0 than when using conventional priors and, when the larger model is true, they present the usual exponential convergence rates in standard procedures. That is, the extra parsimony induced by NLPs is data-dependent, as opposed to inducing sparsity by formulating sparse model prior probabilities or increasingly vague prior distributions on model-specific parameters.

To fix ideas we first give the general definition of NLPs and then proceed to show some examples. Let  $y \in \mathcal{Y}$  be the observed data with density  $p(y | \theta)$  where  $\theta \in \Theta$  is the (possibly infinite-dimensional) parameter. Suppose we are interested in comparing a series of models  $M_1, M_2, \dots$  with corresponding parameter spaces  $\Theta_k \subseteq \Theta$  such that  $\Theta_j \cap \Theta_k$  have zero Lebesgue measure for  $j \neq k$  and, for precision, there exists an  $l$  such that  $\Theta_l = \Theta_j \cap \Theta_k$  so that the whole parameter space is covered.

**Definition 1** *A prior density  $p(\theta | M_k)$  is a non-local prior under  $M_k$  iff  $\lim_{\theta \rightarrow \theta_0} p(\theta | M_k) = 0$  as  $\theta \rightarrow \theta_0$  for any  $\theta_0 \in \Theta_j \subset \Theta_k$ .*

In words,  $p(\theta | M_k)$  vanishes as  $\theta$  approaches any value that would be consistent with a submodel  $M_j$ . Any prior not satisfying Definition 1 is a *local prior* (LP). As a canonical example, suppose that  $y = (y_1, \dots, y_n)$  with independent  $y_i \sim N(\theta, \phi)$  and known  $\phi$ , and that we entertain the two following models:

$$\begin{aligned} M_1 &: \theta = 0 \\ M_2 &: \theta \neq 0 \end{aligned}$$

Under  $M_1$  all parameter values are fully specified, the question is thus reduced to setting  $p(\theta | M_2)$ . Ideally this prior should reflect one's knowledge or beliefs about likely values of  $\theta$ , conditionally on the fact that  $\theta \neq 0$ . The left panel in Figure 1 shows two LPs, specifically the unit information prior  $\theta \sim N(0, 1)$  and a heavy-tailed alternative  $\theta \sim \text{Cachy}(0, 1)$  as recommended by Jeffreys. These assign  $\theta = 0$  as their most likely value a priori, even

though  $\theta = 0$  is not even a possible value under  $M_2$ , which we view as philosophically unappealing. The right panel shows three NLPs (called MOM, eMOM and iMOM, introduced below). Their common defining feature is their vanishing as  $\theta \rightarrow 0$ , thus probabilistically separating  $M_2$  from  $M_1$  or, to borrow terminology from the stochastic processes literature, inducing a repulsive force between  $M_1$  and  $M_2$ . As illustrated in the figure beyond this defining feature the user is free to choose any other desired property, e.g. the speed at which  $p(\theta | M_2)$  vanishes at the origin, prior dispersion, tail thickness or in multivariate cases the prior dependence structure.

Once the NLP has been specified inference proceeds as usual, e.g. posterior model probabilities are

$$p(M_k | y) = \frac{p(y | M_k)p(M_k)}{\sum_j p(y | M_j)p(M_j)} \quad (1)$$

where  $p(y | M_k) = \int p(y | \theta)dP(\theta | M_k)$  is the integrated likelihood under  $M_k$  and  $p(M_k)$  the prior model probability. Similarly, inference on parameters can be carried out conditional on any chosen model via  $p(\theta | M_k, y) \propto p(y | \theta)p(\theta | M_k)$  or via Bayesian model averaging  $p(\theta | y) = \sum_k p(\theta | M_k, y)p(M_k | y)$ . A useful construction (Rossell and Telesca, 2017) is that any NLP can be expressed as

$$p(\theta | M_k) = \frac{p(\theta | M_k)}{p^L(\theta | M_k)}p^L(\theta | M_k) = d_k(\theta)p^L(\theta | M_k), \quad (2)$$

where  $p^L(\theta | M_k)$  is a LP and  $d_k(\theta) = p(\theta | M_k)/p^L(\theta | M_k)$  a penalty term. Simple algebra shows that

$$p(y | M_k) = p^L(y | M_k)E^L(d_k(\theta) | M_k, y), \quad (3)$$

where  $E^L(d_k(\theta) | M_k, y) = \int d_k(\theta)dP^L(\theta | M_k, y)$  is the posterior mean of the penalty term under the underlying LP. That is, the integrated likelihood under a NLP is equal to that under a LP times the posterior expected penalty under that LP. The construction allows to use NLPs in any situation where LPs are already implemented, all one needs is  $p^L(y | M_k)$  or an estimate thereof and posterior samples under  $p^L(\theta | y, M_k)$ . We remark that most functions in `mombf` do not rely on construction (2) but instead work directly with NLPs, as this is typically more efficient computationally. For instance, there are closed-form expressions and Laplace approximations to  $p(y | M_k)$  (Johnson and Rossell, 2012), and one may sample from  $p(\theta | M_k, y)$  via simple latent truncation representations (Rossell and Telesca, 2017).

Up to this point we kept the discussion as generic as possible, Section 3 illustrates NLPs for variable selection and mixture models. For extensions to

other settings see Consonni and La Rocca (2010) for directed acyclic graphs under an objective Bayes framework, Chekouo et al. (2015) for gene regulatory networks, Collazo et al. (2016) for chain event graphs, or ? for finite mixture models. We also remark that this manual focuses mainly on practical aspects. Readers interested in theoretical NLP properties should see Johnson and Rossell (2010) and Rossell and Telesca (2017) for an asymptotic characterization under asymptotically Normal models with fixed  $\dim(\Theta)$ , essentially showing that  $E^L(d_k(\theta) | M_k, y)$  leads to stronger parsimony, ? for similar results in mixture models, and Rossell and Rubio (work in progress) for robust linear regression with non-normal residuals where data may be generated by a model other than those under consideration (M-complete). Regarding high-dimensional results Johnson and Rossell (2012) prove that under certain linear regression models  $p(M_t | y) \xrightarrow{P} 1$  as  $n \rightarrow \infty$  where  $M_t$  is the data-generating truth when using NLPs and  $p = O(n^\alpha)$  with  $\alpha < 1$ . The authors also proved the conceptually stronger result that  $p(M_t | y) \xrightarrow{P} 0$  under LPs, which implies that NLPs are a necessary condition for strong consistency in high dimensions (unless extra parsimony is induced via  $p(M_k)$  or increasingly diffuse  $p(\theta | M_k)$  as  $n$  grows, but this may come at a loss of signal detection power). Shin et al. (2015) extend the consistency result to ultra-high linear regression with  $p = O(e^{n^\alpha})$  with  $\alpha < 1$  under certain specific NLPs.

## 3 Some default non-local priors

### 3.1 Regression models

Definition 1 in principle allows one to define NLPs in any manner that is convenient, as long as  $p(\theta | M_k)$  vanishes for any value  $\theta_0$  that would be consistent with a submodel of  $M_k$ . `mombf` implements some simple priors that lead to convenient implementation and interpretation while offering a reasonable modelling flexibility, but naturally we encourage everyone to come up with more sophisticated alternatives as warranted by their specific problem at hand. It is important to distinguish between two main strategies to define NLPs, namely imposing additive vs. product penalties. Additive NLPs were historically the first to be introduced (Johnson and Rossell, 2010) and primarily aimed to compare only two models, whereas product NLPs were introduced later on (Johnson and Rossell, 2012) for the more general setting where considers multiple models. Throughout let  $\theta = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$  be a vector of regression coefficients and  $\phi$  a dispersion parameter such as the residual variance in linear regression.

Suppose first that we wish to test  $M_1 : \theta = (0, \dots, 0)$  versus  $M_2 : \theta \neq (0, \dots, 0)$ . An additive NLP takes the form  $p(\theta | M_k) = d(q(\theta))p^L(\theta | M_k)$ , where  $q(\theta) = \theta'V\theta$  for some positive-definite  $p \times p$  matrix  $V$ , the penalty  $d(q(\theta)) = 0$  if and only if  $q(\theta) = 0$  and  $p^L(\theta | M_k)$  is an arbitrary LP with the only restriction that  $p(\theta | M_k)$  is proper. For instance,

$$\begin{aligned} p_M(\theta | \phi, M_k) &= \frac{\theta'V\theta}{p\tau\phi} N(\theta; 0, \tau\phi V^{-1}) \\ p_E(\theta | \phi, M_k) &= c_E e^{-\frac{\tau\phi}{\theta'V\theta}} N(\theta; 0, \tau\phi V^{-1}) \\ p_I(\theta | \phi, M_k) &= \frac{\Gamma(p/2)}{|V|^{\frac{1}{2}} (\tau\phi)^{\frac{p}{2}} \Gamma(\nu/2) \pi^{p/2}} (\theta'V\theta)^{-\frac{\nu+p}{2}} e^{-\frac{\tau\phi}{\theta'V\theta}} \end{aligned} \quad (4)$$

are the so-called moment (MOM), exponential moment (eMOM) and inverse moment (iMOM) priors, respectively, and  $c_E$  is the moment generating function of an inverse chi-square random variable evaluated at -1. By default  $V = I$ , but naturally other choices are possible.

Suppose now that we wish to consider all  $2^p$  models arising from setting individual elements in  $\theta$  to 0. Product NLPs are akin to (4) but now the penalty term  $d(\theta)$  is a product of univariate penalties.

$$\begin{aligned} p_M(\theta | \phi, M_k) &= \prod_{i \in M_k} \frac{\theta_i^2}{\tau\phi_k} N(\theta_i; 0, \tau\phi_k) \\ p_E(\theta | \phi, M_k) &= \prod_{i \in M_k} \exp \left\{ \sqrt{2} - \frac{\tau\phi_k}{\theta_i^2} \right\} N(\theta_i; 0, \tau\phi_k), \\ p_I(\theta | \phi, M_k) &= \prod_{i \in M_k} \frac{(\tau\phi_k)^{\frac{1}{2}}}{\sqrt{\pi}\theta_i^2} \exp \left\{ -\frac{\tau\phi_k}{\theta_i^2} \right\}. \end{aligned} \quad (5)$$

This implies that  $d(\theta) \rightarrow 0$  whenever any individual  $\theta_i \rightarrow 0$ , in contrast with (4) which requires the whole vector  $\theta = 0$ . More generally, one can envision settings requiring a combination of additive and product penalties. For instance in regression models for continuous predictors product penalties are generally appropriate, but for categorical predictors one would like to either include or exclude all the corresponding coefficients simultaneously, in this sense additive NLPs resemble group-lasso type penalties and have the nice property of being invariant to the chosen reference category. At this moment `mombf` primarily implements product NLPs and in some cases additive NLPs, we plan to incorporate combined product and additive penalties in the future.

Figure 1 displays the prior densities in the univariate case, where (4) and (5) are equivalent.  $p_M$  induces a quadratic penalty as  $\theta \rightarrow 0$ , and has the

computational advantage that for posterior inference the penalty can often be integrated in closed-form, as it simply requires second order moments.  $p_E$  and  $p_I$  vanish exponentially fast as  $\theta \rightarrow 0$ , inducing stronger parsimony in the Bayes factors than  $p_M$ . This exponential term converges to 1 as  $q(\theta)$  increases, thus the eMOM has Normal tails and the iMOM can be easily checked to have tails proportional to those of a Cauchy. Thick tails can be interesting to address the so-called *information paradox*, namely that the posterior probability of the alternative model converges to 1 as the residual sum of squares from regressing  $y$  on a series of covariates converges to 0 (Liang et al., 2008; Johnson and Rossell, 2010), although in our experience this is often not an issue unless  $n$  is extremely small. The priors above can be extended to include nuisance regression parameters that are common to all models, and also to consider higher powers  $q(\theta)^r$  for some  $r > 1$ , but for simplicity we restrict attention to (4).

`modelSelection` automatically sets default prior distributions that from our experience are sensible in most applications. If you are happy with these defaults are not interested in how they were obtained you can skip to the next section. Of course, we encourage you to think what prior settings are appropriate for your problem at hand. In linear regression by default we set  $\tau$  so that  $P(|\theta/\sqrt{\phi}| > 0.2) = 0.99$ , that is the signal-to-noise ratio  $|\theta_j|/\sqrt{\phi}$  is a priori expected to be  $> 0.2$ . The reason for choosing the 0.2 threshold is that in many applications smaller signals are not practically relevant (e.g. the implied contribution to the  $R^2$  coefficient is small). Function `priorp2g` finds  $\tau$  for any given threshold. Other useful functions are `pmom`, `pemom` and `pimom` for distribution functions, and `qmom`, `qemom` and `qimom` for quantiles. For instance, under a MOM prior  $\tau = 0.348$  gives  $P(|\theta/\sqrt{\phi}| > 0.2) = 0.99$ ,

```
> priorp2g(.01, q=.2, prior='normalMom')
```

```
[1] 0.3483356
```

```
> 2*pmom(-.2, tau=0.3483356)
```

```
[1] 0.99
```

For Accelerated Failure Time models  $e^{\theta_j}$  is the increase in median survival time for a unit standard deviation increase in  $x_j$  (for continuous variables) or between two categories (for discrete variables). A small change, say  $< 15\%$  (i.e.  $\exp(|\theta_j|) < 1.15$ ), is often viewed as practically irrelevant. By default we set  $\tau$  such that

$$P(|\beta_j| > \log(1.15)) = 0.99. \quad (6)$$



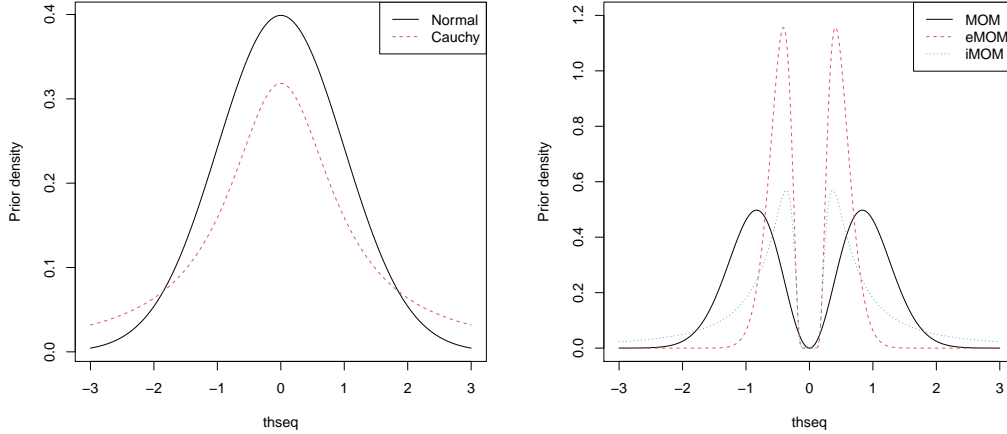


Figure 1: Priors for  $\theta$  under a model  $M_2 : \theta \neq 0$ . Left: local priors. Right: non-local priors

The probability in (6) is under the marginal priors  $\pi_M(\theta_j)$ ,  $\pi_E(\theta_j)$  and  $\pi_I(\theta_j)$  which depend on  $\tau$  and on  $(a_\phi, b_\phi)$ . By default we set  $a_\phi = b_\phi = 3$ , as then the tails of  $\pi_M(\theta_j)$  and  $\pi_E(\theta_j)$  are proportional to a t distribution with 3 degrees of freedom and the marginal prior variance is finite. This gives  $\tau = 0.192$  for  $\pi_M$  and  $g_E = 0.091$  for  $\pi_E$ .

Prior densities can be evaluated as shown below.

```
> thseq <- seq(-3,3,length=1000)
> plot(thseq,dnorm(thseq),type='l',ylab='Prior density')
> lines(thseq,dt(thseq,df=1),lty=2,col=2)
> legend('topright',c('Normal','Cauchy'),lty=1:2,col=1:2)

> thseq <- seq(-3,3,length=1000)
> plot(thseq,dmom(thseq,tau=.348),type='l',ylab='Prior density',ylim=c(0,1.2))
> lines(thseq,demom(thseq,tau=.119),lty=2,col=2)
> lines(thseq,dimom(thseq,tau=.133),lty=3,col=4)
> legend('topright',c('MOM','eMOM','iMOM'),lty=1:3,col=c(1,2,4))
```

Another way to elicit  $\tau$  is to mimic the Unit Information Prior and set the prior variance to 1, for the MOM prior this leads to the same  $\tau$  as the earlier rule  $P(|\theta/\sqrt{\phi}| > 0.2) = 0.99$ .

### 3.2 Mixture models

Let  $y = (y_1, \dots, y_n)$  be the observed data, where  $n$  is the sample size. Denote by  $M_k$  a mixture model with  $k$  components and density

$$p(y_i | \vartheta_k, M_k) = \sum_{j=1}^k \eta_j p(y_i | \theta_j), \quad (7)$$

where  $\eta = (\eta_1, \dots, \eta_k)$  are the mixture weights,  $\theta = (\theta_1, \dots, \theta_k)$  component-specific parameters, and  $\vartheta_k = (\eta, \theta)$  denotes the whole parameter vector. For instance, for Normal mixtures  $\theta_j = (\mu_j, \Sigma_j)$ , where  $\mu_j$  is the mean and  $\Sigma_j$  the covariance matrix. A standard prior choice for such location-scale mixtures is the Normal-IW-Dir prior

$$\tilde{p}(\vartheta_k | M_k) = \left[ \prod_{j=1}^k N(\mu_j; \mu_0, g\Sigma_j) \text{IW}(\Sigma_j; \nu_0, S_0) \right] \text{Dir}(\eta; \tilde{q}) \quad (8)$$

for some given prior parameters  $(\mu_0, g, \nu_0, S_0, \tilde{q})$ . This choice defines a local prior, which we view as unappealing in this setting: it assigns positive prior density to two components having the same parameters and, if  $\tilde{q} \leq 1$ , also to having some mixture weights  $\eta_j = 0$ .

A non-local prior on  $\vartheta_k$  must penalize parameter values that would make the  $k$ -component mixture equivalent to a mixture with  $< k$  components. As formalized in Fúquene et al. (2018) for any generically identifiable mixture this is achieved by penalizing zero weights ( $\eta_j = 0$ ) and any two components having the same parameter values (i.e.  $\theta_j = \theta_l$  for  $j \neq l$ ). Specifically, for location scale mixtures where  $\theta_j = (\mu_j, \Sigma_j)$  we consider the MOM-IW-Dir prior  $p(\vartheta_k | M_k) =$

$$C_k^{-1} \left[ \prod_{j < l} (\mu_j - \mu_l)' A^{-1} (\mu_j - \mu_l) \right] \left[ \prod_{j=1}^k N(\mu_j; \mu_0, gA) \text{IW}(\Sigma_j; \nu_0, S_0) \right] \text{Dir}(\eta; q) \quad (9)$$

where  $C_k$  is the prior normalization constant (implemented in `mombf`),  $A^{-1} = \frac{1}{k} \sum_{j=1}^k \Sigma_j^{-1}$  the average precision matrix,  $(\mu_0, g, \nu_0, S_0)$  are given prior parameters (`mombf` implements defaults for all of them) and one must have  $q > 1$  for (9) to define a non-local prior.

Computing Bayes factors and posterior model probabilities turns out to be surprisingly easy due to two results by Fúquene et al. (2018). First, the integrated likelihood under a MOM-IW-Dir prior

$$p(y | M_k) = \tilde{p}(y | M_k) \int d_k(\vartheta_k) \tilde{p}(\vartheta_k | y, M_k),$$

where  $\tilde{p}(y | M_k)$  is the integrated likelihood associated to the Normal-IW-Dir prior,  $\tilde{p}(\vartheta_k | y, M_k) \propto p(y | \vartheta_k, M_k)\tilde{p}(\vartheta_k | M_k)$  the corresponding posterior and  $d_k(\vartheta_k) =$

$$C_k^{-1} \left[ \prod_{j < l} (\mu_j - \mu_l)' A^{-1} (\mu_j - \mu_l) \right] \left[ \prod_{j=1}^k N(\mu_j; \mu_0, gA) \text{IW}(\Sigma_j; \nu_0, S_0) \right] \frac{\text{Dir}(\eta; q)}{\text{Dir}(\eta; \tilde{q})} \quad (10)$$

is a penalty term straightforward to estimate provided one has access to posterior samples from  $\tilde{p}(\vartheta_k | y, M_k)$ .

The second result is the following simple connection between Bayes factors and empty-component posterior probabilities:

$$\frac{\tilde{p}(y | M_{k-1})}{\tilde{p}(y | M_k)} = \frac{1}{ka_k} \sum_{j=1}^k P(n_j = 0 | y, M_k)$$

where  $a_k = \Gamma(kq)\Gamma(n+kq-q)/(\Gamma(kq-q)\Gamma(n+kq))$  and  $n_j$  is the number of individuals allocated to cluster  $j$ . The right-hand side requires the average marginal posterior probability of one cluster being empty  $P(n_j = 0 | y, M_k)$ . Given  $T$  posterior samples  $\vartheta_k^{(1)}, \dots, \vartheta_k^{(T)}$  this can be estimated in a Rao-Blackwellized fashion as  $\hat{P}(n_j = 0 | y, M_k) =$

$$\frac{1}{T} \sum_{t=1}^T P(n_j = 0 | y, \vartheta_k^{(t)}, M_k) = \frac{1}{T} \sum_{t=1}^T \prod_{i=1}^n P(z_i \neq j | y, \vartheta_k^{(t)}, M_k),$$

where  $z_i \in \{1, \dots, k\}$  for  $i = 1, \dots, n$  are latent cluster indicators. `mombf` implements this computational strategy, see Section 8 for examples.

We remark that  $\hat{P}(n_j = 0 | y, M_k)$  only requires cluster allocation probabilities  $P(z_i \neq j | y, \vartheta_k^{(t)}, M_k)$ . The latter are readily available as a by-product of any standard MCMC algorithm and for any parametric mixture model. That is relative to running a standard MCMC one can obtain  $\hat{P}(n_j = 0 | y, M_k)$  at very little cost, further the estimator is readily applicable to non-conjugate models (in contrast to other estimators, see Marin (2008)).

## 4 Variable selection for linear models

The main function for model selection is `modelSelection`, which returns model posterior probabilities under linear regression models allowing for Normal, asymmetric Normal, Laplace and asymmetric Laplace residuals. A second interesting function is `nlpMarginal`, which computes the integrated likelihood for a given model under the same settings. We illustrate their use with

a simple simulated dataset. Let us generate 100 observations for the response variable and 3 covariates, where the true regression coefficient for the third covariate is 0.

```
> set.seed(2011*01*18)
> x <- matrix(rnorm(100*3),nrow=100,ncol=3)
> theta <- matrix(c(1,1,0),ncol=1)
> y <- x %*% theta + rnorm(100)
```

To start with we assume Normal residuals (the default). We need to specify the prior distribution for the regression coefficients, the model space and the residual variance. We specify a product iMOM prior on the coefficients with prior dispersion  $\tau = .133$ , which targets the detection of standardized effect sizes above 0.2. Regarding the model space we use a Beta-binomial(1,1) prior (Scott and Berger, 2010). Finally, for the residual variance we set a minimally informative inverse gamma prior. For defining other prior distributions see the help for `msPriorSpec` *e.g.* `momprior`, `emomprior` and `zellnerprior` define MOM, eMOM and Zellner priors, and `modelunifprior`, `modelcomplexprior` set uniform and complexity priors on the model space (the latter as defined in Castillo et al. (2015)).

```
> priorCoef <- imomprior(tau=.133)
> priorDelta <- modelbbprior(alpha.p=1,beta.p=1)
> priorVar <- igprior(.01,.01)
```

`modelSelection` enumerates all models when its argument `enumerate` is set to TRUE, otherwise it uses a Gibbs sampling scheme to explore the model space (saved in the slot `postSample`). It returns the visited model with highest posterior probability and the marginal posterior inclusion probabilities for each covariate (when using Gibbs sampling these are estimated via Rao-Blackwellization to improve accuracy).

```
> fit1 <- modelSelection(y=y, x=x, center=FALSE, scale=FALSE,
+ priorCoef=priorCoef, priorDelta=priorDelta, priorVar=priorVar)
```

```
Enumerating models...
```

```
Computing posterior probabilities..... Done.
```

```
> fit1$postMode
```

```
x1 x2 x3
1 1 0
```

```
> fit1$margpp
```

```
          x1          x2          x3
1.00000000 1.00000000 0.04011662
```

```
> postProb(fit1)
```

```

  modelid family      pp
7     1,2 normal 9.598834e-01
8     1,2,3 normal 4.011662e-02
5         1 normal 2.748191e-13
3         2 normal 9.543343e-14
6     1,3 normal 1.172358e-15
4     2,3 normal 6.549246e-16
1         normal 8.609828e-20
2         3 normal 2.748829e-22

> fit2 <- modelSelection(y=y, x=x, center=FALSE, scale=FALSE,
+ priorCoef=priorCoef, priorDelta=priorDelta, priorVar=priorVar,
+ enumerate=FALSE, niter=1000)

Greedy searching posterior mode... Done.
Running Gibbs sampler..... Done.

> fit2$postMode

x1 x2 x3
 1  1  0

> fit2$margpp

      x1      x2      x3
1.00000000 1.00000000 0.04011662

> postProb(fit2,method='norm')

  modelid family      pp
1     1,2 normal 0.95988338
2     1,2,3 normal 0.04011662

> postProb(fit2,method='exact')

  modelid family      pp
1     1,2 normal 0.96111111
2     1,2,3 normal 0.03888889

```

The highest posterior probability model is the simulation truth, indicating that covariates 1 and 2 should be included and covariate 3 should be excluded. `fit1` was obtained by enumerating the  $2^3 = 8$  possible models, whereas `fit2` ran 1,000 Gibbs iterations, delivering very similar results. `postProb` estimates posterior probabilities by renormalizing the probability of each model conditional to the set of visited models when `method='norm'` (the default), otherwise it uses the proportion of Gibbs iterations spent on each model.

Below we run `modelSelection` again but now using Zellner's prior, with prior dispersion set to obtain the so-called Unit Information Prior. The posterior mode is still the data-generating truth, albeit its posterior probability

has decreased substantially. This illustrates the core issue with NLPs: they tend to concentrate more posterior probability around the true model (or that closest in the Kullback-Leibler sense). This difference in behaviour relative to LPs becomes amplified as the number of considered models becomes larger, which may result in the latter giving a posterior probability that converges to 0 for the true model (Johnson and Rossell, 2012).

```
> priorCoef <- zellnerprior(tau=nrow(x))
> fit3 <- modelSelection(y=y, x=x, center=FALSE, scale=FALSE, niter=10^2,
+ priorCoef=priorCoef, priorDelta=priorDelta, priorVar=priorVar,
+ method='Laplace')
```

```
Enumerating models...
Computing posterior probabilities..... Done.
```

```
> postProb(fit3)

  modelid family      pp
7     1,2 normal 7.214937e-01
8     1,2,3 normal 2.785063e-01
5         1 normal 1.079508e-13
3         2 normal 3.565310e-14
6     1,3 normal 1.096444e-14
4     2,3 normal 3.827255e-15
1         normal 3.640151e-20
2         3 normal 1.394484e-21
```

Finally, we illustrate how to relax the assumption that residuals are Normally distributed. We may set the argument `family` to `'twopiecenormal'`, `'laplace'` or `'twopiecelaplace'` to allow for asymmetry (for two-piece Normal and two-piece Laplace) or thicker-than-normal tails (for Laplace and asymmetric Laplace). For instance, the maximum likelihood estimator under Laplace residuals is equivalent to median regression and under asymmetric Laplace residuals to quantile regression, thus these options can be interpreted as robust alternatives to Normal residuals. A nice feature is that regression coefficients can still be interpreted in the usual manner. These families add flexibility while maintaining analytical and computational tractability, e.g. they lead to convex optimization and efficient approximations to marginal likelihoods, and additionally to robustness we have found they can also lead to increased sensitivity to detect non-zero coefficients. Alas, computations under Normal residuals are inevitably faster, hence whenever this extra flexibility is not needed it is nice to be able to fall back onto the Normal family, particularly when  $p$  is large. `modelSelection` and `nlpMarginal` incorporate this option by setting `family=='auto'`, which indicates that the residual distribution should be inferred from the data. When  $p$  is small a full model enumeration is conducted, but when  $p$  is large the Gibbs scheme spends most

time on models with high posterior probability, thus automatically focusing on the Normal family when it provides a good enough approximation and resorting to one of the alternatives when warranted by the data.

For instance, in the example below there's roughly 0.95 posterior probability that residuals are Normal, hence the Gibbs algorithm would spend most time on the (faster) Normal model. The two-piece Normal and two-piece Laplace (also known as asymmetric Laplace) incorporate an asymmetry parameter  $\alpha \in [-1, 1]$ , where  $\alpha = 0$  corresponds to the symmetric case (i.e. Normal and Laplace residuals). We set a NLP on  $\text{atanh}(\alpha) \in (-\infty, \infty)$  so that under the asymmetric model we push prior mass away from  $\alpha = 0$ , which intuitively means we are interested in finding significant departures from asymmetry and otherwise we fall back onto the simpler symmetric case.

```
> priorCoef <- imomprior(tau=.133)
> fit4 <- modelSelection(y=y, x=x, center=FALSE, scale=FALSE,
+ priorCoef=priorCoef, priorDelta=priorDelta, priorVar=priorVar,
+ priorSkew=imomprior(tau=.133), family='auto')
```

```
Enumerating models...
Computing posterior probabilities..... Done.
```

```
> head(postProb(fit4))
```

	modelid	family	pp
7	1,2	normal	9.486877e-01
8	1,2,3	normal	3.964871e-02
15	1,2	laplace	8.899186e-03
23	1,2	twopiecenormal	2.561040e-03
16	1,2,3	laplace	9.678812e-05
31	1,2	twopiecelaplace	5.825143e-05

All examples above use `modelSelection`, which is based on product NLPs (5). `mombf` also provides some (limited) functionality for additive NLPs (4). The code below contains an example based on the Hald data, which has  $n = 13$  observations, a continuous response variable and  $p = 4$  predictors. We load the data and fit a linear regression model.

```
> data(hald)
> dim(hald)

[1] 13 5

> lm1 <- lm(hald[,1] ~ hald[,2] + hald[,3] + hald[,4] + hald[,5])
> summary(lm1)

Call:
lm(formula = hald[, 1] ~ hald[, 2] + hald[, 3] + hald[, 4] +
    hald[, 5])
```

Residuals:

Min	1Q	Median	3Q	Max
-3.1750	-1.6709	0.2508	1.3783	3.9254

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	62.4054	70.0710	0.891	0.3991
hald[, 2]	1.5511	0.7448	2.083	0.0708 .
hald[, 3]	0.5102	0.7238	0.705	0.5009
hald[, 4]	0.1019	0.7547	0.135	0.8959
hald[, 5]	-0.1441	0.7091	-0.203	0.8441

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of freedom

Multiple R-squared: 0.9824, Adjusted R-squared: 0.9736

F-statistic: 111.5 on 4 and 8 DF, p-value: 4.756e-07

```
> V <- summary(lm1)$cov.unscaled
```

```
> diag(V)
```

(Intercept)	hald[, 2]	hald[, 3]	hald[, 4]	hald[, 5]
820.65457471	0.09271040	0.08756026	0.09520141	0.08403119

Bayes factors between a fitted model and a submodel where some of the variables are dropped can be easily computed from the `lm` output using functions `mombf` and `imombf`. As an example here we drop the second coefficient from the model. Parameter  $g$  corresponds to the prior dispersion  $\tau$  in our notation. There are several options to estimate numerically iMOM Bayes factors (for MOM they have closed form), here we compare adaptive quadrature with a Monte Carlo estimate.

```
> mombf(lm1,coef=2,g=0.348)
```

```
      [,1]  
[1,] 1.646985
```

```
> imombf(lm1,coef=2,g=.133,method='adapt')
```

```
      [,1]  
[1,] 1.596056
```

```
> imombf(lm1,coef=2,g=.133,method='MC',B=10^5)
```

```
      [,1]  
[1,] 1.593805
```



## 5 Parameter estimation

A natural question after performing model selection is obtaining estimates for the parameters. Rossell and Telesca (2017) developed a general posterior sampling framework for NLPs based on Gibbs sampling an augmented probability model that expresses NLPs as mixtures of truncated distributions. The methodology is implemented in function `rnlp`. Its basic use is simple, by setting parameter `msfit` to the output of `modelSelection` the function produces posterior samples under each model  $\gamma$  visited by `modelSelection`. The number of samples is proportional to its posterior probability  $p(\gamma | y)$ , thus averaging the output gives Bayesian model averaging estimates  $E(\theta | y) = \sum_{\gamma} E(\theta | \gamma, y)p(\gamma | y)$  and likewise for the residual variance  $E(\phi | y)$ . For convenience the method `coef` extracts such BMA estimates, along with BMA 95% posterior intervals and marginal posterior inclusion probabilities (i.e. when such inclusion probability is large there's Bayesian evidence that the variable has an effect).

```
> priorCoef <- momprior(tau=.348)
> priorDelta <- modelbbprior(alpha.p=1,beta.p=1)
> fit1 <- modelSelection(y=y, x=x, center=FALSE, scale=FALSE,
+ priorCoef=priorCoef, priorDelta=priorDelta)

Enumerating models...
Computing posterior probabilities..... Done.

> b <- coef(fit1)
> head(b)

      estimate      2.5%      97.5%      margpp
intercept 0.00000000 0.0000000 0.0000000 0.00000000
x1        1.010510796 0.8057897 1.209052 1.00000000
x2        0.942050432 0.7422590 1.136970 1.00000000
x3       -0.003898274 0.0000000 0.0000000 0.02579503
phi       1.000346475 0.7605886 1.311363 1.00000000

> th <- rnlp(msfit=fit1,priorCoef=priorCoef,niter=10000)
> colMeans(th)

      intercept      beta1      beta2      beta3      phi
0.000000000 1.011679161 0.943859780 -0.003293146 1.000865372

> head(th)

      intercept      beta1      beta2 beta3      phi
[1,]      0 0.9509956 0.9063265      0 0.9400200
[2,]      0 0.9044596 0.8576924      0 0.7979446
[3,]      0 0.8005331 1.1655173      0 0.8821355
[4,]      0 0.9191956 0.9034925      0 1.0869058
[5,]      0 1.0286702 0.7846458      0 1.1089983
[6,]      0 1.0614485 1.0511627      0 0.9881454
```

Another interesting use of `rnlp` is to obtain posterior samples under a generic non-local posterior

$$p(\theta | y) \propto d(\theta)N(\theta; m, V),$$

where  $d(\theta)$  is a non-local prior penalty and  $N(\theta; m, V)$  is the normal posterior that one would obtain under the underlying local prior. For instance, suppose our prior is proportional to Zellner's prior times a product MOM penalty  $p(\theta) \propto \prod_j \theta_j^2 N(\theta; 0, n\tau\phi(X'X)^{-1})$  where  $\phi$  is the residual variance, then the posterior is proportional to  $p(\theta | y) \propto \prod_j \theta_j^2 N(\theta; m, V)$  where  $m = s_\tau(X'X)^{-1}X'y$ ,  $V = \phi(X'X)^{-1}s_\tau^2$  where  $s_\tau = n\tau/(1 + n\tau)$  is the usual ridge regression shrinkage factor. We may obtain posterior samples as follows. Note that the posterior mean is close to that obtained above.

```
> tau= 0.348
> shrinkage= nrow(x)*tau/(1+nrow(x)*tau)
> V= shrinkage * solve(t(x) %*% x)
> m= as.vector(shrinkage * V %*% t(x) %*% y)
> phi= mean((y - x%*%m)^2)
> th= rnlp(m=m,V=phi * V,priorCoef=momprior(tau=tau))
> colMeans(th)

      beta1      beta2      beta3
1.0110759  0.9308370 -0.1790356
```

## 6 Exact inference for block-diagonal regression

Papaspiliopoulos and Rossell (2016) proposed a fast computational framework to compute exact posterior model probabilities, variable inclusion probabilities and parameter estimates for Normal linear regression when the  $X'X$  matrix is block-diagonal. Naturally this includes the important particular case of orthogonal regression where  $X'X$  is diagonal. The framework performs a fast model search that finds the best model of each size (i.e. with  $1, 2, \dots, p$  variables) and a fast deterministic integration to account for the fact that the residual variance is unknown (the residual variance acts as a "cooling" parameter that affects how many variables are included, hence the associated uncertainty must be dealt with appropriately). The function `postModeOrtho` tackles the diagonal  $X'X$  case and `postModeBlockDiag` the block-diagonal case.

The example below simulates  $n = 210$  observations with  $p = 200$  variables where all regression coefficients are 0 except for the last three (0.5, 0.75, 1) and the residual variance is one. We then perform variable selection under Zellner's and the MOM prior.

```

> set.seed(1)
> p <- 200; n <- 210
> x <- scale(matrix(rnorm(n*p),nrow=n,ncol=p),center=TRUE,scale=TRUE)
> S <- cov(x)
> e <- eigen(cov(x))
> x <- t(t(x %*% e$vectors)/sqrt(e$values))
> th <- c(rep(0,p-3),c(.5,.75,1)); phi <- 1
> y <- x %*% matrix(th,ncol=1) + rnorm(n,sd=sqrt(phi))
> priorDelta=modelbinomprior(p=1/p)
> priorVar=igprior(0.01,0.01)
> priorCoef=zellnerprior(tau=n)
> pm.zell <-
+ postModeOrtho(y,x=x,priorCoef=priorCoef,priorDelta=priorDelta,
+ priorVar=priorVar,bma=TRUE)
> head(pm.zell$models)

      modelid      pp
4      198,199,200 0.8257052262
5       54,198,199,200 0.0390061004
107    11,198,199,200 0.0062427546
108   186,198,199,200 0.0042350105
110    36,198,199,200 0.0037652564
6     11,54,198,199,200 0.0003357223

> priorCoef=momprior(tau=0.348)
> pm.mom <- postModeOrtho(y,x=x,priorCoef=priorCoef,priorDelta=priorDelta,
+ priorVar=priorVar,bma=TRUE)
> head(pm.mom$models)

      modelid      pp
4      198,199,200 9.779392e-01
5       54,198,199,200 1.144910e-02
107    11,198,199,200 1.209685e-03
108   186,198,199,200 7.314291e-04
110    36,198,199,200 6.262828e-04
6     11,54,198,199,200 1.717659e-05

```

postModelBlockDiag returns a list with the best model of each size and its corresponding (exact) posterior probability, displayed in Figure 2 (left panel). It also returns marginal inclusion probabilities and BMA estimates, shown in the right panel. The code required to produce these figures is below.

```

> par(mar=c(5,5,1,1))
> nvars <- sapply(strsplit(as.character(pm.zell$models$modelid),split=','),length)
> plot(nvars,pm.zell$models$pp,ylab=expression(paste("p(",gamma,"|y)")),
+ xlab=expression(paste("|",gamma,"|")),cex.lab=1.5,ylim=0:1,xlim=c(0,50))
> sel <- pm.zell$models$pp>.05
> text(nvars[sel],pm.zell$models$pp[sel],pm.zell$models$modelid[sel],pos=4)
> nvars <- sapply(strsplit(as.character(pm.mom$models$modelid),split=','),length)
> points(nvars,pm.mom$models$pp,col='gray',pch=17)
> sel <- pm.mom$models$pp>.05

```

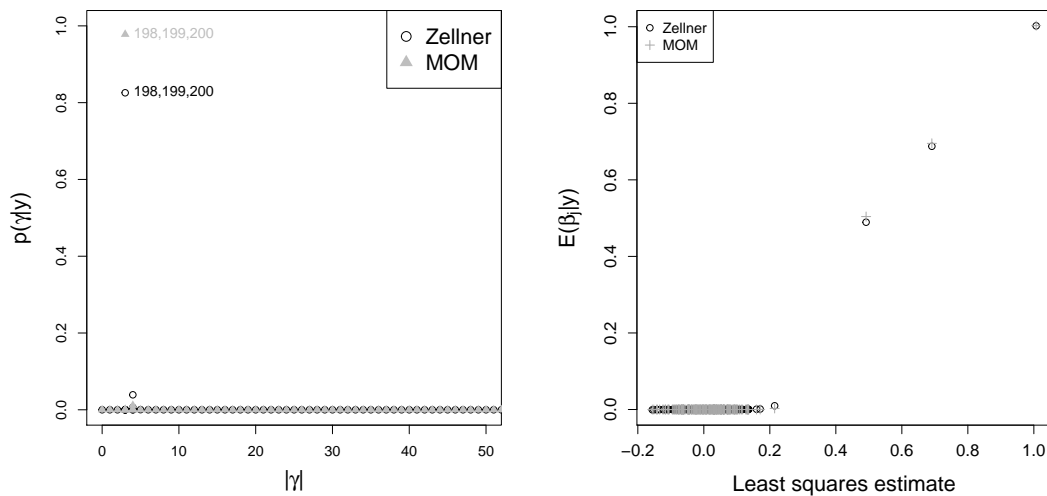


Figure 2: Posterior probability under simulated orthogonal data

```
> text(nvars[sel], pm.mom$models$pp[sel], pm.mom$models$modelid[sel], pos=4, col='gray')
> legend('topright', c('Zellner', 'MOM'), pch=c(1,17), col=c('black', 'gray'), cex=1.5)

> par(mar=c(5,5,1,1))
> ols <- (t(x) %*% y) / colSums(x^2)
> plot(ols, pm.zell$bma$coef, xlab='Least squares estimate',
+ ylab=expression(paste('E(', beta[j], '|y)')), cex.lab=1.5, cex.axis=1.2, col=1)
> points(ols, pm.mom$bma$coef, pch=3, col='darkgray')
> legend('topleft', c('Zellner', 'MOM'), pch=c(1,3), col=c('black', 'darkgray'))
```

We now illustrate similar functionality under block-diagonal  $X'X$ . To this end we consider a total  $p = 100$  variables split into 10 blocks of 10 variables each, generated in such a way that they all have unit variance and within-blocks pairwise correlation of 0.5. The first block has three non-zero coefficients, the second block two and the remaining blocks contain no active variables.

```
> set.seed(1)
> p <- 100; n <- 110
> blocksize <- 10
> blocks <- rep(1:(p/blocksize), each=blocksize)
> x <- scale(matrix(rnorm(n*p), nrow=n, ncol=p), center=TRUE, scale=TRUE)
> S <- cov(x)
> e <- eigen(cov(x))
> x <- t(t(x) %*% e$vectors) / sqrt(e$values)
> Sblock <- diag(blocksize)
> Sblock[upper.tri(Sblock)] <- Sblock[lower.tri(Sblock)] <- 0.5
> vv <- eigen(Sblock)$vectors
```



5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	X49	X50	X51	X52	X53	X54	X55	X56	X57	X58	X59	X60	X61	X62	X63	X64	X65	X66	X67
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	X68	X69	X70	X71	X72	X73	X74	X75	X76	X77	X78	X79	X80	X81	X82	X83	X84	X85	X86
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	X87	X88	X89	X90	X91	X92	X93	X94	X95	X96	X97	X98	X99	X100					
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0					

Figure 3 shows a LASSO-type plot with the posterior means under the best model of each size visited by Coolblock. We appreciate how the truly active variables 8, 9, 10, 19 and 20 are picked up first.

```

> maxvars=50
> ylim=range(pm$postmean.model[,-1])
> plot(NA,NA,xlab='Model size',
+      ylab='Posterior mean given model',
+      xlim=c(0,maxvars),ylim=ylim,cex.lab=1.5)
> visited <- which(!is.na(pm$models$pp))
> for (i in 2:ncol(pm$postmean.model)) {
+   lines(pm$models$nvars[visited],pm$postmean.model[visited,i])
+ }
> text(maxvars, pm$postmean.model[maxvars,which(th!=0)+1],
+      paste('X',which(th!=0),sep=' '), pos=3)

```

## 7 Bayes factors for generalized linear models

At the moment `mombf` provides some limited functionality for generalized linear models. `pmomLM` implements probit models following the framework in Rossell et al. (2013), see also Nikooienejad et al. (2016) for an alternative

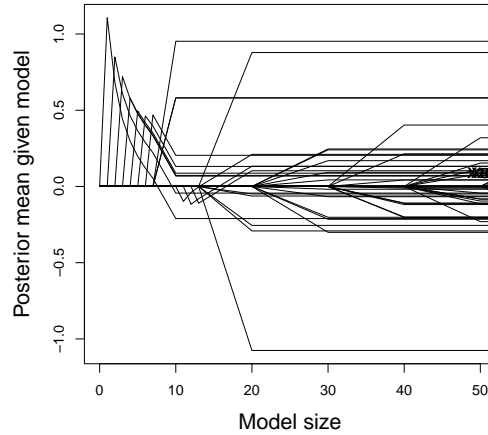


Figure 3: Coolblock algorithm: posterior mean of regression coefficients under best model of each size

framework that bypasses the need to sample  $\theta$  to improve the mixing of the Markov Chain.

As an illustration we simulate  $n = 50$  observations from a probit regression model with  $p = 2$  correlated predictors and  $\theta = (\log(2), 0)$ . The predictors are stored in the matrix  $x$ , the success probabilities in the vector  $p$  and the observed responses in the vector  $y$ . For reproducibility purposes we set the random number generator seed.

```
> set.seed(4*2*2008)
> n <- 50; theta <- c(log(2),0)
> x <- matrix(NA,nrow=n,ncol=2)
> x[,1] <- rnorm(n,0,1); x[,2] <- rnorm(n,.5*x[,1],1)
> p <- pnorm(x %%% matrix(theta,ncol=1))
> y <- rbinom(n,1,p)
```

```
> th <- pmomLM(y=y,x=x,xadj=rep(1,n),niter=10000)
```

Running MCMC.....Done.

```
> head(th$postModel)
```

	[,1]	[,2]
[1,]	1	1
[2,]	1	0
[3,]	1	0
[4,]	1	0
[5,]	1	0
[6,]	1	0

```
> table(apply(th$postModel,1,paste,collapse=', '))
0,0 0,1 1,0 1,1
240 24 7477 1259
```

An alternative to `pmomLM`, which considers all possible models under product NLPs, is to compute Bayes factors between pairs of models under additive NLPs using functions `momknown` and `imomknown`.

Before computing Bayes factors, we fit a probit regression model with the function `glm`. The maximum likelihood estimates are stored in `thetahat` and the asymptotic covariance matrix in `V`. These functions take as primary arguments a vector of regression coefficients and their covariance matrix, and hence they can be used in any setting where one has a statistic that is asymptotically sufficient and normally distributed. The resulting Bayes factors are approximate. The functions also allow for the presence of a dispersion parameter `sigma`, *i.e.* the covariance of the regression coefficients is  $\sigma^2 V$ , but they assume that `sigma` is known. The probit regression model that we simulated has no over-dispersion and hence it corresponds to `sigma=1`. We first compare the full model with the model resulting from excluding the second covariate (the first term corresponds to the intercept), setting  $g = 0.5$  for illustration.

```
> glm1 <- glm(y~x[,1]+x[,2],family=binomial(link = "probit"))
> thetahat <- coef(glm1)
> V <- summary(glm1)$cov.scaled
> g <- .5
> bfmom.1 <- momknown(thetahat[2],V[2,2],n=n,g=g,sigma=1)
> bfimom.1 <- imomknown(thetahat[2],V[2,2],n=n,nuisance.theta=2,g=g,sigma=1)
> bfmom.1
```

```
      [,1]
[1,] 4.262401
```

```
> bfimom.1
```

```
      [,1]
[1,] 3.336888
```

Both priors result in evidence for including the first covariate. We now check whether the second covariate can be dropped.

```
> bfmom.2 <- momknown(thetahat[3],V[3,3],n=n,g=g,sigma=1)
> bfimom.2 <- imomknown(thetahat[3],V[3,3],n=n,nuisance.theta=2,g=g,sigma=1)
> bfmom.2
```

```
      [,1]
[1,] 0.02784354
```



```
> bfmom.2
```

```
      [,1]  
[1,] 0.008250121
```

Both Mom and iMom BF provide strong evidence in favor of the simpler model, *i.e.* excluding  $x[,2]$ . To compare the full model with the model that has no covariates (*i.e.* only the constant term remains) we use the same routines, passing a vector as the first argument and a matrix as the second argument.

```
> bfmom.0 <- momknown(thetahat[2:3],V[2:3,2:3],n=n,g=g,sigma=1)  
> bfmom.0 <- imomknown(thetahat[2:3],V[2:3,2:3],n=n,nuisance.theta=2,g=g,sigma=1)  
> bfmom.0
```

```
      [,1]  
[1,] 0.5272556
```

```
> bfmom.0
```

```
      [,1]  
[1,] 0.953978
```

Based on the resulting BF being close to 1, it is not clear whether the full model is preferable to the model with no covariates.

The BF can be used to easily compute posterior probabilities for each of the four considered models: no covariates, only  $x[,1]$ , only  $x[,2]$  and both  $x[,1]$  and  $x[,2]$ . We assume equal probabilities *a priori*.

```
> prior.prob <- rep(1/4,4)  
> bf <- c(bfmom.0,bfmom.1,bfmom.2,1)  
> pos.prob <- prior.prob*bf/sum(prior.prob*bf)  
> pos.prob  
[1] 0.090632677 0.732686026 0.004786169 0.171895128
```

The model with the highest posterior probability is the one including only  $x[,1]$ , *i.e.* the correct model, and the model with the lowest posterior probability is that including only  $x[,2]$ .

## 8 Model selection for mixtures

`bfnormmix` is the main function to obtain posterior probabilities on the number of mixture components, as well as posterior samples for any given number of components. We start by simulating univariate Normal data truly arising from a single component.

```

> set.seed(1)
> n=200; k=1:3
> x= rnorm(n)

```

We run `bfnormmix` based on  $B=10000$  MCMC iterations and default prior parameters. The prior dispersion  $g$  in (9) is an important parameter that determines how separated one expects components to be a priori. Roughly speaking its default value assigns 0.95 prior probability to any two components giving a multimodal density, this default helps focus the analysis on detecting well-separated components and enforcing parsimony whenever components are poorly separated. The Dirichlet prior parameter  $q$  is also important, `bfnormmix` allows one to specify different values for the MOM-IW-Dir prior in (9) (argument `q`) and the Normal-IW-Dir prior in (8) (argument `q.niw`). The default `q=3` helps discard components with small weights, since these could be due to over-fitting or outliers.

```

> fit= bfnormmix(x=x,k=k,q=3,q.niw=1,B=10^4)
> postProb(fit)

```

	k	pp.momiw	pp.niw	logprobempty	logbf.momiw	logpen	logbf.niw
1	1	0.94708961	0.73065710	-Inf	0.000000	0.000000	0.000000
2	2	0.03806607	0.18764538	-3.943915	-3.214070	-1.854680	-1.359390
3	3	0.01484432	0.08169752	-3.783590	-4.155777	-1.964856	-2.190921

```

      model
1 Normal, VVV
2 Normal, VVV
3 Normal, VVV

```

Under a MOM-Dir-IW prior one assigns high posterior probability  $P(M_1 | y) \approx 0.96$  to the data-generating truth, for the Normal-Dir-IW this probability is  $\approx 0.753$ .

We can use `postSamples` to retrieve samples from the Normal-Dir-IW posterior, and `coef` to obtain the corresponding posterior means.

```

> mcmcout= postSamples(fit)
> names(mcmcout)

[1] "k=1" "k=2" "k=3"

> names(mcmcout[[2]])

[1] "eta"          "mu"           "cholSigmaInv" "momiw.weight"

> colMeans(mcmcout[[2]]$eta)

[1] 0.5800639 0.4199361

> parest= coef(fit)
> names(parest)

```

```

[1] "k=1" "k=2" "k=3"

> parest[[2]]

$eta
[1] 0.5800639 0.4199361

$mu
[1] 0.078319526 0.001958708

$SigmaInv
$SigmaInv[[1]]
      [,1]
[1,] 8.085496

$SigmaInv[[2]]
      [,1]
[1,] 12.22981

```

We remark that one can reweight these samples to obtain posterior samples from the non-local MOM-Dir-IW posterior, these weights are given by the term  $d_k(\vartheta_k)$  in (10) and are stored in `momiw.weight`. As illustrated below in our example under the 2-component model the location parameters under the MOM-Dir-IW posterior are more separated than under the Normal-Dir-IW, as one would expect from the repulsive force between components.

```

> w= postSamples(fit)[[2]]$momiw.weight
> apply(mcmcOut[[2]]$eta, 2, weighted.mean, w=w)

[1] 0.5360198 0.4639802

> apply(mcmcOut[[2]]$mu, 2, weighted.mean, w=w)

[1] 0.2621437 0.1387601

```

To further illustrate we simulate another dataset, this time where the data-generating truth are  $k = 2$  components. Interestingly the MOM-IW-Dir favours  $k = 2$  more than the Normal-IW-Dir and both clearly discard  $k = 1$ , as expected given that the two components are well-separated (the means are 3 standard deviations away from each other).

```

> set.seed(1)
> n=200; k=1:3; probs= c(1/2,1/2)
> mu1= -1.5; mu2= 1.5
> x= rnorm(n) + ifelse(runif(n)<probs[1],mu1,mu2)
> fit= bfnormmix(x=x,k=k,q=3,q.niw=1,B=10^4)
> postProb(fit)

```

	k	pp.momiw	pp.niw	logprobempty	logbf.momiw	logpen	logbf.niw
1	1	1.045463e-48	1.749420e-48	-Inf	0.0000	0.0000000	0.0000
2	2	8.000360e-01	6.057393e-01	-114.766800	110.2565	0.7930316	109.4635
3	3	1.999640e-01	3.942607e-01	-4.185683	108.8700	-0.1640506	109.0341

model

1 Normal, VVV  
2 Normal, VVV  
3 Normal, VVV

## References

- I. Castillo, J. Schmidt-Hieber, and A.W. van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.
- T. Chekouo, F.C. Stingo, J.D. Doecke, and K.-A. Do. mirna–target gene regulatory networks: A bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics*, 71(2):428–438, 2015.
- Rodrigo A Collazo, Jim Q Smith, et al. A new family of non-local priors for chain event graph model selection. *Bayesian Analysis*, (in press), 2016.
- G. Consonni and L. La Rocca. On moment priors for Bayesian model choice with applications to directed acyclic graphs. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M Smith, and M. West, editors, *Bayesian Statistics 9 - Proceedings of the ninth Valencia international meeting*, pages 119–144. Oxford University Press, 2010.
- J. Fúquene, M.F.J. Steel, and D. Rossell. On choosing mixture components via non-local priors. *arXiv 1604.00314v3*, pages 1–43, 2018.
- V.E. Johnson and D. Rossell. Prior densities for default bayesian hypothesis tests. *Journal of the Royal Statistical Society B*, 72:143–170, 2010.
- V.E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 24(498):649–660, 2012.
- F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008.
- C.P. Marin, J. M. Robert. Approximating the marginal likelihood in mixture models. *Bulleting of the Indian Chapter of ISBA*, 1:2–7, 2008.

- Amir Nikooienejad, Wenyi Wang, and Valen E Johnson. Bayesian variable selection for binary outcomes in high dimensional genomic studies using non-local priors. *Bioinformatics*, btv764:1–8, 2016.
- O. Papaspiliopoulos and D. Rossell. Scalable variable selection and model averaging under block-orthogonal design. *arXiv*, pages 1–19, 2016.
- D. Rossell. A framework for posterior consistency in model selection. *arXiv*, 1806.04071:1–58, 2018.
- D. Rossell and F.J. Rubio. Tractable bayesian variable selection: beyond normality. *Journal of the American Statistical Association*, 113(524):1742–1758, 2018.
- D. Rossell and D. Telesca. Non-local priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112:254–265, 2017.
- D. Rossell, D. Telesca, and V.E. Johnson. High-dimensional Bayesian classifiers using non-local priors. In *Statistical Models for Data Analysis XV*, pages 305–314. Springer, 2013.
- J.G. Scott and J.O Berger. Bayes and empirical Bayes multiplicity adjustment in the variable selection problem. *The Annals of Statistics*, 38(5): 2587–2619, 2010.
- M. Shin, A. Bhattacharya, and V.E. Johnson. Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Texas A&M University (technical report)*, pages 1–33, 2015.