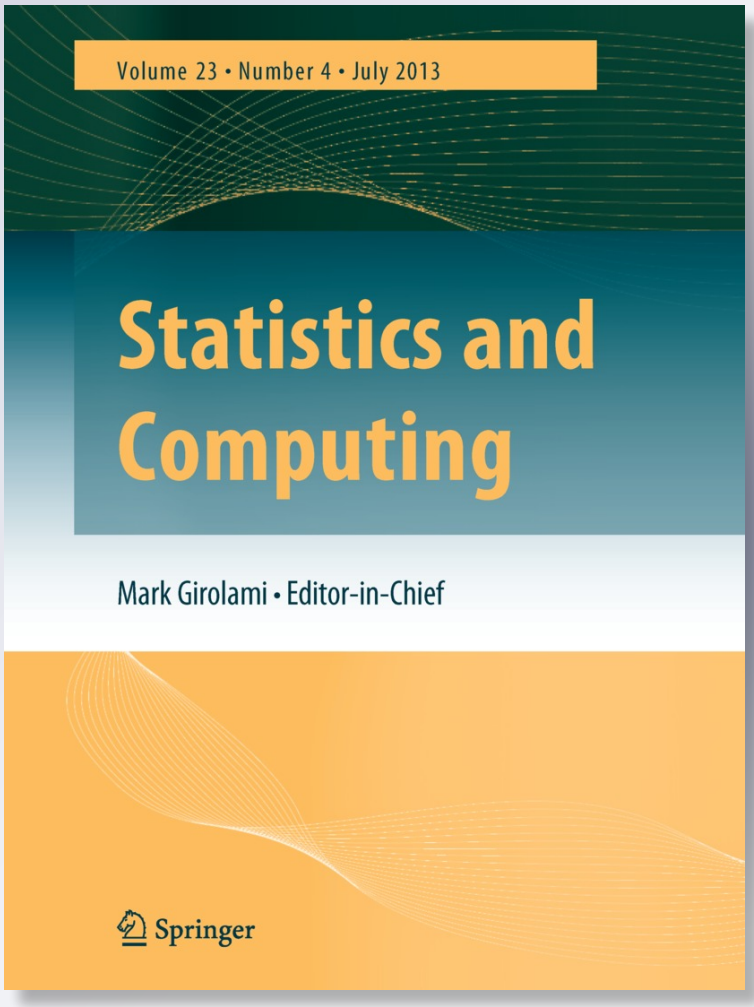
Computation for intrinsic variable selection in normal regression models via expected-posterior prior

D. Fouskakis & I. Ntzoufras

Statistics and Computing

ISSN 0960-3174
Volume 23
Number 4

Stat Comput (2013) 23:491-499
DOI 10.1007/s11222-012-9325-9



Springer

Stat Comput (2013) 23:491–499
DOI 10.1007/s11222-012-9325-9

Computation for intrinsic variable selection in normal regression models via expected-posterior prior

D. Fouskakis · I. Ntzoufras

Received: 19 July 2011 / Accepted: 8 March 2012 / Published online: 7 April 2012

Abstract In this paper, we focus on the variable selection problem in normal regression models using the expected-posterior prior methodology. We provide a straightforward MCMC scheme for the derivation of the posterior distribution, as well as Monte Carlo estimates for the computation of the marginal likelihood and posterior model probabilities. Additionally, for large spaces, a model search algorithm based on MC^3 is constructed. The proposed methodology is applied in two real life examples, already used in the relevant literature of objective variable selection. In both examples, uncertainty over different training samples is taken into consideration.

Keywords Bayesian variable selection · Expected-posterior priors · Imaginary data · Intrinsic priors · Jeffreys prior · Objective model selection methods · Normal regression models

1 Introduction

Let $\mathcal{M}$ be the model space, consisting of all combinations of the available covariates; for every $m_\ell \in \mathcal{M}$ with parameters $(\boldsymbol{\beta}_\ell, \sigma^2)$ the likelihood is specified by

$$\boldsymbol{Y}|\mathrm{X}_\ell, \boldsymbol{\beta}_\ell, \sigma^2, m_\ell \sim N_n\left(\mathrm{X}_\ell \boldsymbol{\beta}_\ell, \sigma^2 \mathrm{I}_n\right)$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is a multivariate random variable expressing the response for each subject, X_ℓ is a $n \times d_\ell$ design/data matrix containing the values of the explanatory variables in its columns, I_n is the $n \times n$ identity matrix, $\boldsymbol{\beta}_\ell$ is a vector of length d_ℓ with the effects of each covariate on the response data $\boldsymbol{Y}$ and σ^2 is the error variance of any model.

When using improper prior distributions to express prior ignorance for the model parameters, Bayes factors cannot be evaluated because of the presence of the unknown normalizing constants. This has urged the Bayesian community to develop various methodologies to overcome the problem of prior specification in variable selection problems including the approaches based on Zellner's (1986) g-priors amongst others; see in Fernandez et al. (2001), Liang et al. (2008), Celeux et al. (2012) and Dellaportas et al. (2012) for some recent advances and comparisons.

One of the proposed approaches is also the intrinsic Bayes factors (IBFs), introduced by Berger and Pericchi (1996). In order to provide a full Bayesian interpretation of IBFs, they have also defined intrinsic prior (IP) distributions. The intrinsic prior methodology has been applied for objective variable selection problems in normal regression models, by Casella and Moreno (2006), Moreno and Girón (2008), Girón et al. (2006) and Casella et al. (2009).

Intrinsic priors are closely related to the expected-posterior prior distributions of Pérez (1998) and Pérez and Berger (2002) which have nice interpretation based on imaginary training data coming from prior predictive distributions. The expected-posterior priors overcome some of the difficulties that appear in Bayesian model comparison and variable selection when using improper priors, like the indeterminacy of the Bayes factors, since the unknown normalizing constants cancel out in the marginal likelihood ratios. Moreover, all prior distributions are calculated automatically, having a notion of compatibility, since they are based

D. Fouskakis (✉)
Department of Mathematics, National Technical University of Athens, Zografou Campus, Athens 15780, Greece
e-mail: fouskakis@math.ntua.gr

I. Ntzoufras
Department of Statistics, Athens University of Economics and Business, 76 Patision Street, Athens 10434, Greece
e-mail: ntzoufras@aueb.gr

on averages of posterior distributions of similar imaginary-predictive data. Another advantage is that these priors take into account the different interpretation of the coefficients of each model. They are also connected not only to the intrinsic priors, but to the Zellner's (1986) g-priors that use a specific "imaginary" dataset instead of averaging across a predictive distribution. For a complete and more detailed list of the advantages of the expected-posterior priors see Pérez (1998) and Pérez and Berger (2002).

In this paper we implement the expected-posterior prior methodology on variable selection problems in normal regression models. We construct a straightforward MCMC scheme for the derivation of the posterior distribution, as well as a Monte Carlo estimate for the computation of the Bayes factors and posterior model probabilities under the intrinsic prior. The proposed methodology is applied to a variety of random training samples and therefore the uncertainty over different training samples is considered.

2 Expected-posterior priors

Pérez and Berger (2002) have defined the expected-posterior prior (EPP) as the posterior distribution of the parameter vector of the model under consideration averaged over all possible imaginary data $\boldsymbol{y}^*$ coming from the predictive distribution $f(\boldsymbol{y}^*|m_0)$ of a reference model m_0 (Pérez and Berger 2002, Def. 1, p. 493). Hence the EPP for the parameters of any model $m_\ell \in \mathcal{M}$ is given by

$$\pi_\ell^I\left(\boldsymbol{\beta}_\ell, \sigma^2|\mathrm{X}_\ell^*\right) = \int \pi_\ell^N\left(\boldsymbol{\beta}_\ell, \sigma^2|\boldsymbol{y}^*, \mathrm{X}_\ell^*\right) m_0^N\left(\boldsymbol{y}^*|\mathrm{X}_0^*\right) d\boldsymbol{y}^*, \tag{1}$$

where X_ℓ^* and X_0^* are the design matrices for the imaginary data under models m_ℓ and m_0 respectively, $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma^2|\boldsymbol{y}^*, \mathrm{X}_\ell^*)$ is the posterior of $(\boldsymbol{\beta}_\ell, \sigma^2)$ for model m_ℓ using an improper baseline prior $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma^2)$, given data $\boldsymbol{y}^*$, and $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)$ is the prior predictive distribution, evaluated at $\boldsymbol{y}^*$, for model m_0 under the prior $\pi_0^N(\boldsymbol{\beta}_0, \sigma^2)$. For the reference model ($m_\ell = m_0$) this prior degenerates to $\pi_0^N(\boldsymbol{\beta}_0, \sigma^2)$.

In the above equation, if we use the Bayes theorem to replace $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma^2|\boldsymbol{y}^*, \mathrm{X}_\ell^*)$ by the corresponding likelihood-prior product and write the marginal likelihood $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)$ as an integral of the likelihood over the prior of the parameters of the reference model, then we end up with the intrinsic prior as defined in Berger and Pericchi (1996).

A question that naturally arises is which model should be selected as a reference model. In order (1) to coincide with the intrinsic prior, m_0 must be nested to all models m_ℓ under consideration. Therefore, in variable selection problems, a natural choice for the reference model is the constant model.

3 Prior specification

We use the independence Jeffreys prior (or reference prior) as the baseline prior distribution. Hence for $m_\ell \in \mathcal{M}$, where $\mathcal{M}$ is the model space, we have

$$\pi_\ell^N\left(\boldsymbol{\beta}_\ell, \sigma^2\right) = \frac{c_\ell}{\sigma^2}, \tag{2}$$

where c_ℓ is an unknown normalizing constant. This prior is a special case of the prior used by Pérez (1998) with $q_i = 1$ in his notation.

Under the above setup, for every $m_\ell \in \mathcal{M}$, if we assume imaginary data $\boldsymbol{y}^*$ of size n^* and design matrix X_ℓ^*, the intrinsic prior $\pi_\ell^I(\boldsymbol{\beta}_\ell, \sigma^2|\mathrm{X}_\ell^*)$ has the following form:

$$\begin{aligned}\pi_\ell^I\left(\boldsymbol{\beta}_\ell, \sigma^2|\mathrm{X}_\ell^*\right) = \int \frac{m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*)} f\left(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \sigma^2, m_\ell, \mathrm{X}_\ell^*\right) \\ \times \pi_\ell^N\left(\boldsymbol{\beta}_\ell, \sigma^2\right) d\boldsymbol{y}^*,\end{aligned} \tag{3}$$

where $f(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \sigma^2, m_\ell, \mathrm{X}_\ell^*)$ is the likelihood of model m_ℓ with parameters $(\boldsymbol{\beta}_\ell, \sigma^2)$ evaluated at $\boldsymbol{y}^*$ and $m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*)$ is the prior predictive distribution (or the marginal likelihood), evaluated at $\boldsymbol{y}^*$, of model m_ℓ under the baseline prior $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma^2)$, i.e.

$$\begin{aligned}m_\ell^N\left(\boldsymbol{y}^*|\mathrm{X}_\ell^*\right) &= \int\int f\left(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \sigma^2, m_\ell, \mathrm{X}_\ell^*\right) \\ &\quad \times \pi_\ell^N\left(\boldsymbol{\beta}_\ell, \sigma^2\right) d\boldsymbol{\beta}_\ell\, d\sigma^2 \\ &= c_\ell\, (\pi)^{\frac{d_\ell - n^*}{2}} \left|\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*\right|^{-\frac{1}{2}} \frac{\Gamma(\frac{n^*-d_\ell}{2})}{RSS_\ell^{*\frac{n^*-d_\ell}{2}}}\end{aligned} \tag{4}$$

with

$$\begin{aligned}RSS_\ell^* &= \left(\boldsymbol{y}^* - \mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell^*\right)^T\left(\boldsymbol{y}^* - \mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell^*\right) \\ &= \boldsymbol{y}^{*T}\left(\mathrm{I}_{n^*} - \mathrm{X}_\ell^*\left(\mathrm{X}_\ell{}^{*T}\mathrm{X}_\ell^*\right)^{-1}\mathrm{X}_\ell{}^{*T}\right)\boldsymbol{y}^*\end{aligned} \tag{5}$$

being the residual sum of squares using $(\boldsymbol{y}^*, \mathrm{X}_\ell^*)$ as data and $\widehat{\boldsymbol{\beta}}_\ell^* = (\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*T}\boldsymbol{y}^*$; see also in Pérez (1998, Eq. 3.5) for an equivalent expression.

4 Computation of the posterior distribution

Under the intrinsic prior distribution described in Sect. 3, the posterior distribution of model parameters $(\boldsymbol{\beta}_\ell, \sigma^2)$ is now given by

$$\begin{aligned}&\pi_\ell^I\left(\boldsymbol{\beta}_\ell, \sigma^2|\boldsymbol{y}, \mathrm{X}_\ell, \mathrm{X}_\ell^*\right) \\ &\quad\propto f\left(\boldsymbol{y}|\boldsymbol{\beta}_\ell, \sigma^2, m_\ell, \mathrm{X}_\ell\right)\pi_\ell^I\left(\boldsymbol{\beta}_\ell, \sigma^2|\mathrm{X}_\ell^*\right) \\ &\quad\propto \int f\left(\boldsymbol{y}|\boldsymbol{\beta}_\ell, \sigma^2, m_\ell, \mathrm{X}_\ell\right)\pi_\ell^N\left(\boldsymbol{\beta}_\ell, \sigma^2|\boldsymbol{y}^*, \mathrm{X}_\ell^*\right) \\ &\qquad\times m_0^N\left(\boldsymbol{y}^*|\mathrm{X}_0^*\right) d\boldsymbol{y}^* \\ &\quad\propto \int \pi_\ell^N\left(\boldsymbol{\beta}_\ell, \sigma^2|\boldsymbol{y}, \boldsymbol{y}^*, \mathrm{X}_\ell, \mathrm{X}_\ell^*\right) m_\ell^N\left(\boldsymbol{y}|\boldsymbol{y}^*, \mathrm{X}_\ell, \mathrm{X}_\ell^*\right)\end{aligned}$$

$$\times m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)\,d\boldsymbol{y}^*$$
$$\propto \int f_{N_{d_\ell}}(\boldsymbol{\beta}_\ell;\widetilde{\boldsymbol{\beta}}^N,\widetilde{\Sigma}^N\sigma^2)\,f_{IG}(\sigma^2;\widetilde{a}_\ell^N,\widetilde{b}_\ell^N)$$
$$\times m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)\,d\boldsymbol{y}^*$$

where

$$\widetilde{\boldsymbol{\beta}}^N=\widetilde{\Sigma}^N(\mathrm{X}_\ell^T\boldsymbol{y}+\mathrm{X}_\ell{}^{*T}\boldsymbol{y}^*),$$
$$\widetilde{\Sigma}^N=\{\mathrm{X}_\ell{}^{*T}\mathrm{X}_\ell^*+\mathrm{X}_\ell^T\mathrm{X}_\ell\}^{-1},$$
$$\widetilde{a}_\ell^N=n/2+n^*/2-d_\ell/2,\quad\text{and}$$
$$\widetilde{b}_\ell^N=RSS_\ell^N/2+RSS_\ell^*/2$$

with

$$RSS_\ell^N=(\boldsymbol{y}-\mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell^*)^T(\mathrm{I}_n+\mathrm{X}_\ell(\mathrm{X}_\ell^T\mathrm{X}_\ell)^{-1}\mathrm{X}_\ell^T)(\boldsymbol{y}-\mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell^*).$$

Therefore we can construct an MCMC scheme to sample from the joint posterior

$$\pi_\ell^I(\boldsymbol{\beta}_\ell,\sigma^2,\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)$$
$$\propto f_{N_{d_\ell}}(\boldsymbol{\beta}_\ell;\widetilde{\boldsymbol{\beta}}^N,\widetilde{\Sigma}^N\sigma^2)\,f_{IG}(\sigma^2;\widetilde{a}_\ell^N,\widetilde{b}_\ell^N)$$
$$\times m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)$$

in the following way:

1. Generate $\boldsymbol{y}^*$ from
$$f(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)\propto\frac{m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell,\mathrm{X}_\ell^*)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell,\mathrm{X}_\ell^*)}\times m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*).$$
2. Generate σ^2 from $IG(\widetilde{a}_\ell^N,\widetilde{b}_\ell^N)$.
3. Generate $\boldsymbol{\beta}_\ell$ from $N_{d_\ell}(\widetilde{\boldsymbol{\beta}}^N,\widetilde{\Sigma}^N\sigma^2)$.

We can generate the imaginary data $\boldsymbol{y}^*$ using a Metropolis-Hastings algorithm proposing new candidate values $\boldsymbol{y}^{*'}$ from a proposal distribution with density

$$q(\boldsymbol{y}^{*'})=m_\ell^N(\boldsymbol{y}^{*'}|\,\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)$$
$$=f_{St_{n^*}}\Bigg(\boldsymbol{y}^{*'};n-d_\ell,\mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell,\frac{RSS_\ell}{n-d_\ell}(\mathrm{I}_{n^*}+\mathrm{X}_\ell^*(\mathrm{X}_\ell^T\mathrm{X}_\ell)^{-1}\mathrm{X}_\ell^{*T})\Bigg)\qquad(6)$$

and acceptance probability

$$\alpha=\min\left\{1,\frac{m_0^N(\boldsymbol{y}^{*'}|\mathrm{X}_0^*)m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*)}{m_\ell^N(\boldsymbol{y}^{*'}|\mathrm{X}_\ell^*)m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)}\right\}$$
$$=\min\left\{1,\left(\frac{RSS_\ell^{*'}}{RSS_\ell^*}\right)^{(n^*-d_\ell)/2}\times\left(\frac{RSS_0^{*'}}{RSS_0^*}\right)^{-(n^*-d_0)/2}\right\}\qquad(7)$$

where $RSS_\ell=(\boldsymbol{y}-\mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell)^T(\boldsymbol{y}-\mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell)$ is the residual sum of squares using $(\boldsymbol{y},\mathrm{X}_\ell)$ as data and RSS_ℓ^*, $RSS_\ell^{*'}$ are given in (5) using $(\boldsymbol{y}^*,\mathrm{X}_\ell^*)$, $(\boldsymbol{y}^{*'},\mathrm{X}_\ell^*)$ as data respectively.

5 Variable selection computation

In this section we provide two alternative approaches for the evaluation of the models under consideration. In Sect. 5.1 we construct an efficient Monte Carlo scheme for the estimation of the marginal likelihood for any given training sample X^*, while in Sect. 5.2 we introduce an MCMC algorithm, more appropriate for large model spaces, which directly estimates the posterior model probabilities over all possible training subsamples.

5.1 Monte Carlo estimation of the marginal likelihood

The marginal likelihood of any model $m_\ell\in\mathcal{M}$ is given by

$$m_\ell^I(\boldsymbol{y}|\mathrm{X}_\ell,\mathrm{X}_\ell^*)$$
$$=\int\!\!\int f(\boldsymbol{y}|\boldsymbol{\beta}_\ell,\sigma^2,m_\ell,\mathrm{X}_\ell)\pi_\ell^I(\boldsymbol{\beta}_\ell,\sigma^2|\mathrm{X}_\ell^*)\,d\boldsymbol{\beta}_\ell\,d\sigma^2$$
$$=\int\!\!\int\!\!\int f(\boldsymbol{y}|\boldsymbol{\beta}_\ell,\sigma^2,m_\ell,\mathrm{X}_\ell)\pi_\ell^N(\boldsymbol{\beta}_\ell,\sigma^2|\boldsymbol{y}^*,\mathrm{X}_\ell^*)\times m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)\,d\boldsymbol{\beta}_\ell\,d\sigma^2\,d\boldsymbol{y}^*$$
$$=\int\!\!\int\!\!\int f(\boldsymbol{y}|\boldsymbol{\beta}_\ell,\sigma^2,m_\ell,\mathrm{X}_\ell)f(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell,\sigma^2,m_\ell,\mathrm{X}_\ell^*)\times\pi_\ell^N(\boldsymbol{\beta}_\ell,\sigma^2)\frac{m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*)}\,d\boldsymbol{\beta}_\ell\,d\sigma^2\,d\boldsymbol{y}^*$$
$$=\int\left\{\int\!\!\int f(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell,\sigma^2,m_\ell,\mathrm{X}_\ell^*)\times\pi_\ell^N(\boldsymbol{\beta}_\ell,\sigma^2|\boldsymbol{y},\mathrm{X}_\ell)\,d\boldsymbol{\beta}_\ell\,d\sigma^2\right\}\times m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell)\frac{m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*)}\,d\boldsymbol{y}^*$$
$$=\int m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell)\frac{m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*)}\,d\boldsymbol{y}^*$$
$$=\int m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell)\times\frac{m_0^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_0,\mathrm{X}_0^*)m_0^N(\boldsymbol{y}|\mathrm{X}_0)/m_0^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_0,\mathrm{X}_0^*)}{m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell)/m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_\ell,\mathrm{X}_\ell^*)}\,d\boldsymbol{y}^*$$
$$=m_0^N(\boldsymbol{y}|\mathrm{X}_0)\int\frac{m_0^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_0,\mathrm{X}_0^*)/m_0^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_0,\mathrm{X}_0^*)}{m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)/m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_\ell,\mathrm{X}_\ell^*)}\times m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)\,d\boldsymbol{y}^*.\qquad(8)$$

In the above expression the predictive densities $m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)$ and $m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_\ell,\mathrm{X}_\ell^*)$ for any model ℓ, under the baseline prior $\pi_\ell^N(\boldsymbol{\beta}_\ell,\sigma^2)$, are given by

$$m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_\ell,\mathrm{X}_\ell^*)$$
$$=f_{St_n}(\boldsymbol{y};n^*-d_\ell,\mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell^*,(\mathrm{I}_n+\mathrm{X}_\ell(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^T)\widehat{\sigma}_U^{*2})\qquad(9)$$

and

$$m_\ell^N\left(\boldsymbol{y}^*|\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*\right) = f_{St_{n^*}}\left(\boldsymbol{y}^*;n-d_\ell,\mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell,\left(\mathrm{I}_{n^*}+\mathrm{X}_\ell^*\left(\mathrm{X}_\ell^T\mathrm{X}_\ell\right)^{-1}\mathrm{X}_\ell^{*T}\right)\widehat{\sigma}_U^2\right), \tag{10}$$

where $\widehat{\boldsymbol{\beta}}_\ell^*$ is the maximum likelihood estimate of $\boldsymbol{\beta}_\ell$, $\widehat{\sigma}_U^{*2} = RSS_\ell^*/(n^*-d_\ell)$ is the unbiased residual variance for m_ℓ with RSS_ℓ^* being the corresponding residual sum of squares using $(\boldsymbol{y}^*,\mathrm{X}_\ell^*)$ as data; $\widehat{\boldsymbol{\beta}}_\ell$, $\widehat{\sigma}_U^2$ and RSS_ℓ are the corresponding measures using $(\boldsymbol{y},\mathrm{X}_\ell)$ as data. The quantity $m_0^N(\boldsymbol{y}|\mathrm{X}_0)$ denotes the marginal likelihood of the reference model m_0, as derived in (4). The presence of this quantity in (8) does not cause any problem in our setup; $m_0^N(\boldsymbol{y}|\mathrm{X}_0)$ is common in all marginal likelihoods and therefore cancels out when we compare models using Bayes factors, posterior model odds or probabilities.

We use (8) to setup a Monte Carlo scheme and to estimate the marginal likelihood up to the common constant $m_0^N(\boldsymbol{y}|\mathrm{X}_0)$. We generate $\boldsymbol{y}^{*(t)}$, $t=1,\ldots,T$, from $m_\ell^N(\boldsymbol{y}^*|\,\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)$ given in (10) and estimate the unnormalized marginal likelihood (i.e. the integral involved in (8))

$$m_\ell^{IU}\left(\boldsymbol{y}|\mathrm{X}_\ell,\mathrm{X}_\ell^*\right) = m_\ell^I\left(\boldsymbol{y}|\mathrm{X}_\ell,\mathrm{X}_\ell^*\right)/m_0^N(\boldsymbol{y}|\mathrm{X}_0), \tag{11}$$

by

$$\widehat{m}_\ell^{IU}\left(\boldsymbol{y}|\mathrm{X}_\ell,\mathrm{X}_\ell^*,\delta\right) = \frac{1}{T}\sum_{t=1}^{T}\frac{m_\ell^N(\boldsymbol{y}|\,\boldsymbol{y}^{*(t)},\mathrm{X}_\ell,\mathrm{X}_\ell^*)}{m_0^N(\boldsymbol{y}|\,\boldsymbol{y}^{*(t)},\mathrm{X}_0,\mathrm{X}_0^*)}\frac{m_0^N(\boldsymbol{y}^{*(t)}|\,\boldsymbol{y},\mathrm{X}_0,\mathrm{X}_0^*)}{m_\ell^N(\boldsymbol{y}^{*(t)}|\,\boldsymbol{y},\mathrm{X}_\ell,\mathrm{X}_\ell^*)}. \tag{12}$$

For small model spaces it is easy to estimate the unnormalized marginal likelihoods (11) for all models under consideration using the above sampling scheme. For large model spaces, it is possible to implement an MC^3 algorithm (Madigan and York 1995; Kass and Raftery 1995) by estimating (11) for each model that is evaluated for the first time within the iterative scheme. The estimator presented in this section is compatible with the importance sampling estimator of Pérez (1998, Sect. 3.4.1); here we use the conditional posterior predictive distributions which under the baseline prior (2) are multivariate Student distributions and therefore can be calculated directly.

5.2 Computation of the posterior model weights over different training samples

An alternative approach is to use an MC^3 scheme by generating $\boldsymbol{y}^*$ for the given model m_ℓ and then move to a model $m_{\ell'}$ (by proposing to add or delete a specific covariate). If X is the $(n\times d)$ design/data matrix of the full model, the algorithm can be summarized by

- For $k=1,\ldots,K$ (training samples):
  1. Randomly consider a submatrix X^* of X with dimension $(n^*\times d)$.
  2. For $t=1,\ldots,T$ (iterations):
     (a) For a given model m_ℓ, generate a proposed $\boldsymbol{y}^*$ from (6) and accept it with probability (7).
     (b) For $j=1,\ldots,p$, propose with probability one to move to model $m_{\ell'}$ by changing the status of the j covariate and accept the proposed model with probability $\alpha=\min\{1,A\}$, where
     $$A = \frac{f(\boldsymbol{y}^*|\boldsymbol{y},m_{\ell'})}{f(\boldsymbol{y}^*|\boldsymbol{y},m_\ell)}\times\frac{f(m_{\ell'})}{f(m_\ell)} = \frac{m_{\ell'}^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_\ell,\mathrm{X}_\ell^*)}{m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*,\mathrm{X}_\ell,\mathrm{X}_\ell^*)}\times\frac{f(m_{\ell'})}{f(m_\ell)}$$
     and $f(m_\ell)$ is the prior probability of model m_ℓ.
  3. Calculate the posterior weights for each training sample k.

From the above MC^3 scheme we can produce summaries of the posterior model weights over the K different training samples. This might be more efficient in large model spaces since we avoid implementing the Monte Carlo computation presented in Sect. 5.1 for each newly visited model. Nevertheless, in such cases, the number of iterations T within each training sample must be increased to ensure that the model space is satisfactorily explored for each training sample.

6 Experimental results

In this section the proposed methodology is illustrated on two real life examples. In both examples we use a uniform prior on the model space.

6.1 Hald's data

We consider the Hald's cement data (Montgomery and Peck 1982) to illustrate the proposed approach. This dataset consists of $n=13$ observations and $p=4$ covariates and has been previously used by Girón et al. (2006) for illustrating objective variable selection methods. The response variable Y is the heat evolved in a cement mix and the explanatory variables are the tricalcium aluminate (X_1), the tricalcium silicate (X_2), the tetracalcium alumino ferrite (X_3) and the dicalcium silicate (X_4). An important feature of Hald's cement data is that variables X_1 and X_3 and variables X_2 and X_4 are highly correlated ($corr(X_1,X_3)=-0.824$ and $corr(X_1,X_4)=-0.975$).

Table 1 presents posterior model probability summaries together with median based posterior odds across 100 different training sub-samples, for the best models after performing a full enumeration search. For estimating the marginal likelihood we used 1000 iterations. Figures 1 and 2 provide

Table 1 Summaries of posterior model probabilities for the best models over 100 different training sub-samples together with median based posterior odds of the MAP (m_1) vs. $m_j < 5$ for the Hald's cement data (Example 6.1)

m_j	Model formula	Expected posterior prior						g-Prior[a]	
		Posterior model probabilities					Median based PO[c]	$f(m_j\|\mathbf{y})$[b]	PO[c]
		Mean	Median	SD	Percentiles				
					2.5 %	97.5 %			
1	$X_1 + X_2$	0.411	0.405	0.074	0.309	0.554	1.000	0.325	1.000
2	$X_1 + X_4$	0.167	0.160	0.040	0.107	0.278	2.529	0.225	1.444
3	$X_1 + X_2 + X_4$	0.132	0.138	0.034	0.061	0.183	2.930	0.109	2.980
4	$X_1 + X_2 + X_3$	0.128	0.132	0.033	0.062	0.185	3.061	0.109	2.990
5	$X_1 + X_3 + X_4$	0.105	0.106	0.027	0.051	0.148	3.807	0.102	3.185

[a] As implemented by Liang et al. (2008)

[b] $f(m_j|\mathbf{y})$: Posterior probability of model m_j

[c] PO: Posterior odds of MAP model vs. each model

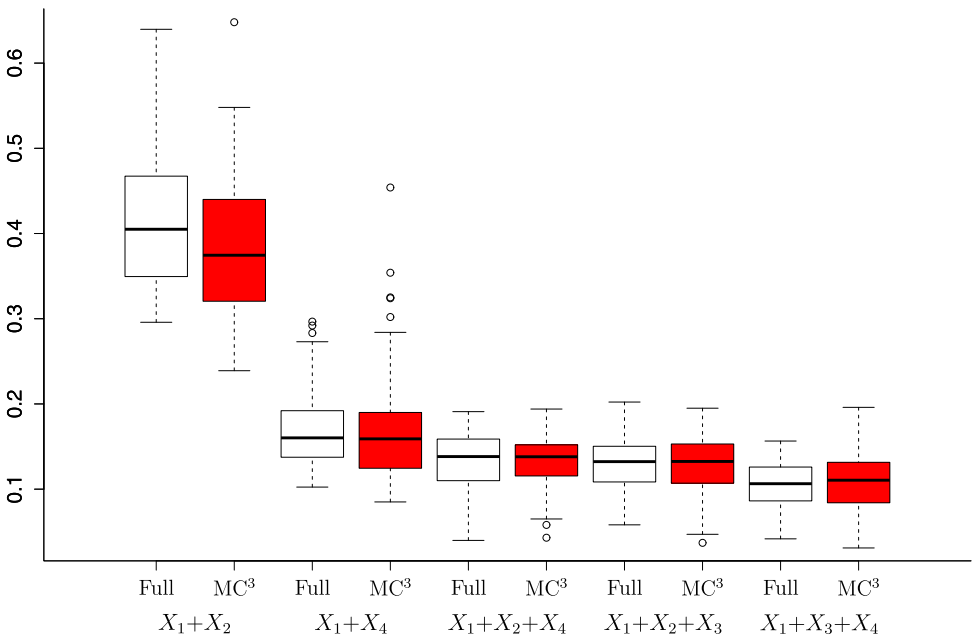


Fig. 1 Boxplots comparing the posterior model probabilities over 100 training samples for the full enumeration (Full) using Monte Carlo estimates and the MC^3 for the Hald's cement data (Example 6.1)

a graphical representation of the distributions of the posterior model probabilities (for the five best models) and of the posterior marginal inclusion probabilities respectively, across different training samples. We note that the maximum a posteriori (MAP) model is $X_1 + X_2$ with posterior probabilities within (0.31, 0.55) and median value equal to 0.40, which is 2.5 times the corresponding value from the second best model. Averages of posterior model probabilities were very close to the median values presented in Table 1; the latter is preferred since median posterior model probabilities correspond to median Bayes factors. The boxplots of the posterior marginal inclusion probabilities indicate that covariates X_1 and X_2 should be included in the model formulation with posterior probabilities clearly above 0.5 for all different training samples.

We also performed the same task using 1000 different training samples, instead of 100; results were almost identical. Furthermore, for illustrative reasons and in order to evaluate the efficiency of our approach, we implemented the proposed MC^3 scheme of Sect. 5.2 for 1000 iterations, considering 100 different training samples. Results were very similar to the ones from the full enumeration run, with some increased variability across different samples that could be eliminated by increasing the number of iterations. Graphical comparison of the results obtained using MC^3 and the Monte Carlo full enumeration results are presented in Figs. 1 and 2.

For comparison purposes, we also performed full enumeration using the Zellner's (1986) g-prior with $g = n = 13$. The original Zellner's g-prior formulation provided totally

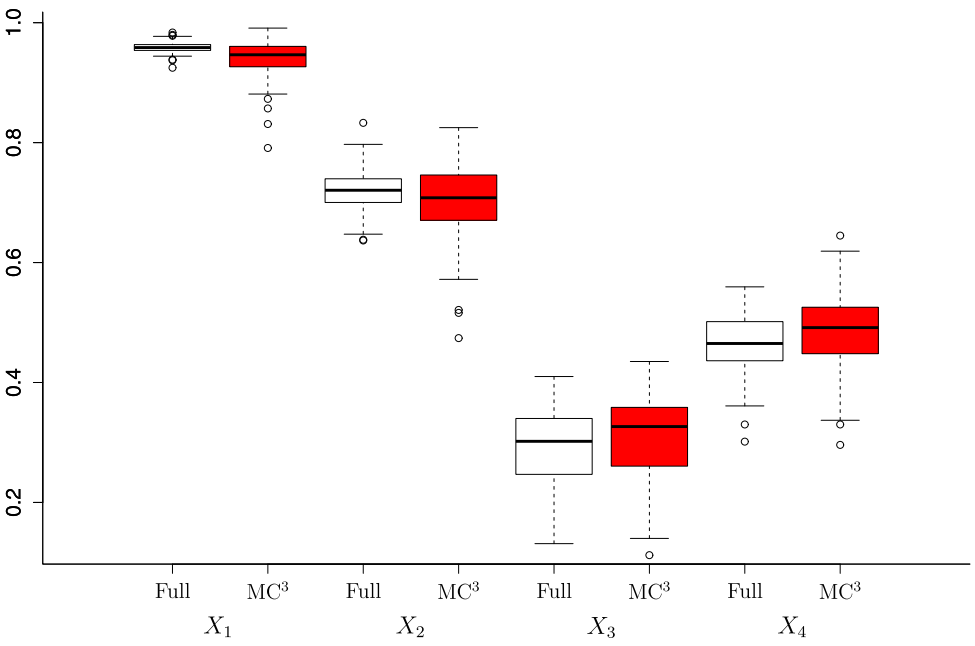


Fig. 2 Boxplots comparing the posterior marginal inclusion probabilities over 100 training samples for the full enumeration (Full) using Monte Carlo estimates and the MC^3 for the Hald's cement data (Example 6.1)

Table 2 Posterior marginal inclusion probabilities for the Hald's cement data (Example 6.1)

Method	X_1	X_2	X_3	X_4
Monte Carlo EPP[a]	0.958	0.721	0.302	0.465
MC^3 EPP[a]	0.946	0.708	0.326	0.492
Liang et al. (2008) g-prior[b]	0.900	0.636	0.340	0.564
g-Prior[c] with standardized data	0.908	0.644	0.338	0.558
g-Prior[c] with unstandardized data	0.319	0.340	0.265	0.350

[a] Averages across 100 training samples; EPP: expected posterior prior

[b] g-Prior as implemented by Liang et al. (2008)

[c] Original g-prior as introduced by Zellner (1986)

different answers supporting the null model with all posterior marginal inclusion probabilities being lower than 0.36. This is due to the large intercept in combination with the small sample size which makes this prior informative here; see in Ntzoufras (2010) and Dellaportas et al. (2012) for similar illustrations. This problem can be solved by either standardizing the data, or by using the modified version of g-prior as in Liang et al. (2008). In the last two columns of Table 1 we present results under this modified g-prior setup; results for the original g-prior with standardized data were almost identical and therefore have been omitted. Posterior marginal inclusion probabilities for all these different prior setups are presented in Table 2. There is strong agreement among the posterior marginal inclusion probabilities from the modified g-prior, the original g-prior with standardized data and the EPP; the two different versions of the g-prior support slightly less covariates X_1 and X_2 and support slightly more covariate X_4. Additional results can also be found in Pérez (1998, Sect. 3.4.2) including posterior model probabilities using the Zellner and Siow (1980) prior (ZS), arithmetic intrinsic Bayes factors and results after applying the EPP methodology using three different baseline priors (with the one termed reference to be the closer to the one using in this article). All his EPP based results are close to the ones presented in Table 1; minor differences are due to the different definition of the baseline prior. Under different baseline prior, similar results have been also obtained by Girón et al. (2006), using the approach of Casella and Moreno (2006) for handling the presence of multiple different covariate training samples.

Finally, we calculated the BMA leave-one-out cross-validatory log-score

$$\text{CV-log-score} = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{y}_{\setminus i})$$

$$\text{with } f(y_i|\boldsymbol{y}_{\setminus i}) = \sum_{m_\ell \in \mathcal{M}} f(y_i|\boldsymbol{y}_{\setminus i}, m_\ell) f(m_\ell|\boldsymbol{y}_{\setminus i}),$$

Table 3 Summaries of posterior model probabilities for the best models over 100 different training sub-samples together with median based posterior odds of the MAP (m_1) vs. $m_j < 5$ for the prostate cancer data (Example 6.2)

m_j	Model formula	Expected posterior prior						g-Prior[a]	
		Posterior model probabilities					Median based PO[c]	$f(m_j\|\boldsymbol{y})$[b]	PO[c]
		Mean	Median	SD	Percentiles				
					2.5 %	97.5 %			
1	$X_1 + X_2 + X_5$	0.299	0.296	0.062	0.199	0.446	1.000	0.374	1.000
2	$X_1 + X_2 + X_4 + X_5$	0.107	0.104	0.036	0.054	0.177	2.845	0.101	3.696
3	$X_1 + X_2 + X_3 + X_5$	0.076	0.069	0.032	0.035	0.148	4.300	0.071	5.278
4	$X_1 + X_2 + X_5 + X_8$	0.067	0.066	0.021	0.033	0.110	4.472	0.062	5.981

[a] As implemented by Liang et al. (2008)

[b] $f(m_j|\boldsymbol{y})$: Posterior probability of model m_j

[c] PO: Posterior odds of MAP model vs. each model

where $\boldsymbol{y}_{\setminus i}$ is the vector of data $\boldsymbol{y}$ without observation i, $f(m_\ell|\boldsymbol{y}_{\setminus i})$ is the posterior probability of model m_ℓ given data $\boldsymbol{y}_{\setminus i}$ and $f(y_i|\boldsymbol{y}_{\setminus i}, m_\ell)$ is the posterior predictive ordinate of model m_ℓ evaluated at y_i having observed $\boldsymbol{y}_{\setminus i}$. For 100 different training samples, the BMA leave-one-out cross-validatory log-scores under the EPP approach ranged between -33.2 and -35.9, with mean -34.5 and standard deviation 0.53. On the contrary, the BMA leave-one-out cross-validatory log-scores under the 3 different g-prior setups (modified version of g-prior as in Liang et al. (2008), original g-prior with standardized data and original g-prior with unstandardized data) were -40.3, -39.8 and -62.5 respectively, indicating, for this illustration, a clearly better predictive performance under the EPP approach.

6.2 Prostate cancer data

In this section, we present results of our methodology for the prostate cancer data (Stamey et al. 1989). This dataset has been also used by Girón et al. (2006) and Moreno and Girón (2008) to illustrate their approach. It consists of $n = 97$ observations and $p = 8$ covariates. The response variable Y is the level of prostate-specific antigen, and the covariates are the logarithm of cancer volume (X_1), the logarithm of prostate weight (X_2), the age of the patient (X_3), the logarithm of the amount of benign prostatic hyperplasia (X_4), the seminal vesicle invasion (X_5), the logarithm of capsular penetration (X_6), the Gleason score (X_7) and the percent of Gleason scores 4 and 5 (X_8).

The structure of this section is similar to that of Sect. 6.1. Results, over 100 different training sub-samples, are summarized in Table 3 and Figs. 3 and 4. To be more specific, the MAP model includes covariates X_1, X_2 and X_5 with posterior probabilities taking values in (0.20, 0.45) and median value equal to 0.3, which is 2.8 times the corresponding value from the second best model. The boxplots of the posterior marginal inclusion probabilities indicate that the same covariates should be included in the model formulation with posterior probabilities clearly above 0.5 for all training samples. Furthermore, we implemented the proposed MC^3 algorithm of Sect. 5.2 for 2000 iterations, considering 100 different training samples. Results from the two methods were equivalent; see Figs. 3 and 4 for a graphical comparison.

Posterior model probabilities for the three different versions of the g-prior as described in Sect. 6.1 with $g = n = 97$ were also calculated. They were all very similar, thus results only for the Liang et al. (2008) setup appear in Table 3. We also notice strong agreement with the results obtained by the EPP approach; the latter supports slightly less the MAP model (average posterior model probability of 0.30 versus 0.37 for the g-prior). Posterior marginal inclusion probabilities for all different prior setups are given in Table 4; results are similar for all priors used with the non-important covariates to be supported with slightly lower posterior weight when using the g-prior. The same MAP model, with posterior probability 0.28, has been also identified by Leng et al. (2010) using Bayesian adaptive lasso methods. Their method generally supported more complicated models, attributing higher posterior marginal inclusion probabilities for most of the covariates.

Finally, the BMA leave-one-out cross-validatory log-scores for EPP, over 30 different training samples, and averaged over models with posterior probabilities higher than 0.01, were calculated. They ranged between -138.8 and -71.9 with mean -105.8 and standard deviation 13.3. The corresponding log-scores under the three different g-prior setups (modified version of g-prior as in Liang et al. (2008), original g-prior with standardized data and original g-prior with unstandardized data) were found equal to -108.6, -108.1 and -113.8 respectively, indicating, for this illustration, a better predictive performance on average for the EPP approach.

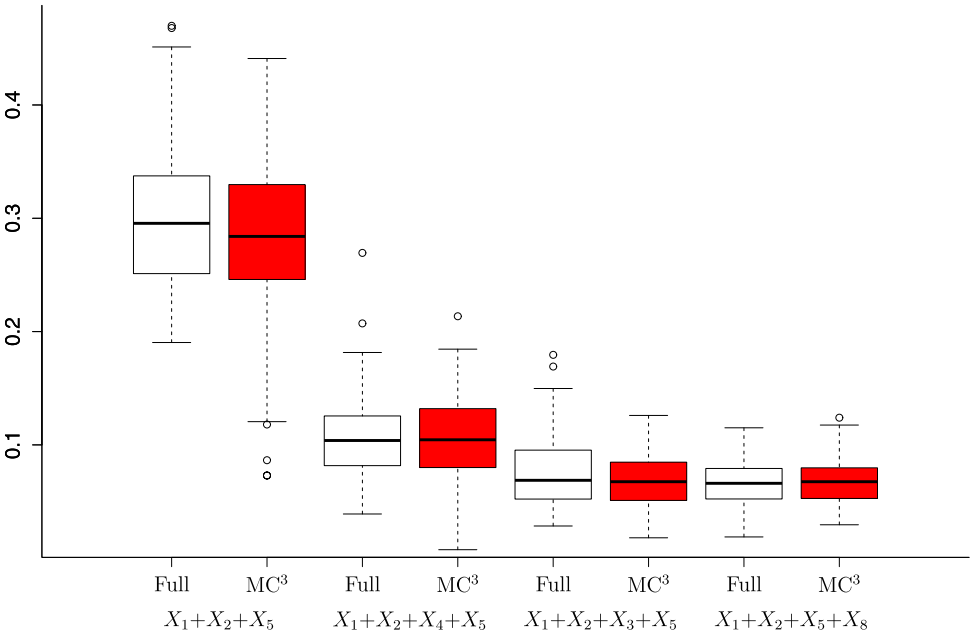


Fig. 3 Boxplots comparing the posterior model probabilities over 100 training samples for the full enumeration (Full) using Monte Carlo estimates and the MC^3 for the prostate cancer data (Example 6.2)

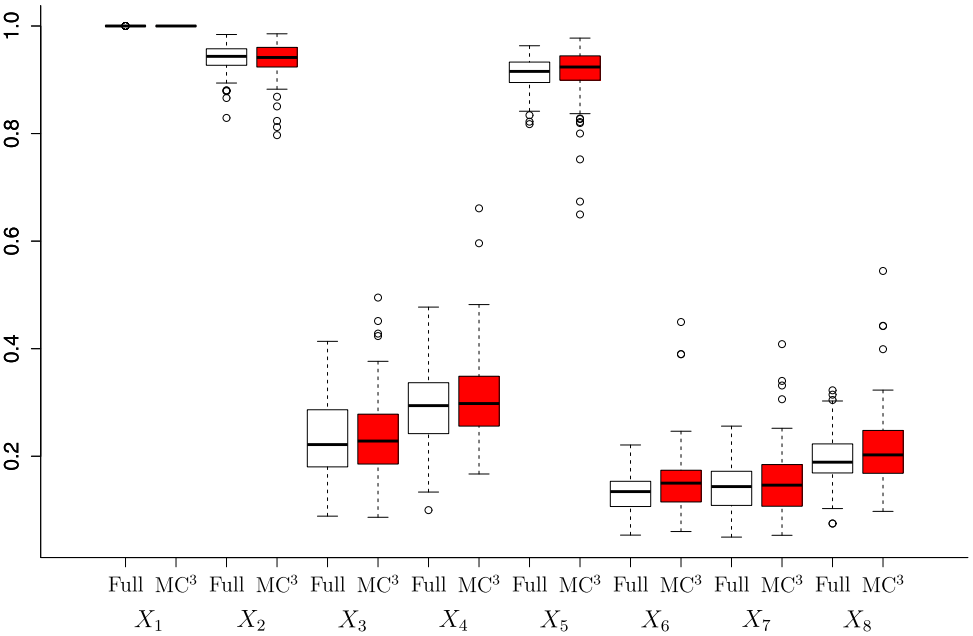


Fig. 4 Boxplots comparing the posterior marginal inclusion probabilities over 100 training samples for the full enumeration (Full) using Monte Carlo estimates and the MC^3 for the prostate cancer data (Example 6.2)

Table 4 Posterior marginal inclusion probabilities for the prostate cancer data (Example 6.2)

Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Monte Carlo EPP[a]	1.000	0.939	0.232	0.293	0.911	0.132	0.145	0.196
MC^3 EPP[a]	1.000	0.936	0.234	0.309	0.910	0.152	0.153	0.215
Liang et al. (2008) g-prior[b]	1.000	0.946	0.193	0.254	0.917	0.110	0.125	0.162
g-Prior[c] with standardized data	1.000	0.948	0.195	0.255	0.920	0.110	0.125	0.163
g-Prior[c] with unstandardized data	1.000	0.924	0.175	0.241	0.875	0.109	0.121	0.159

[a]Averages across 100 training samples; EPP: expected posterior prior

[b]g-Prior as implemented by Liang et al. (2008)

[c]Original g-prior as introduced by Zellner (1986)

7 Discussion

We have presented a computational approach for variable selection in normal regression models, based on the expected-posterior prior methodology. We have constructed efficient MCMC schemes for the estimation of the parameters within each model, based on data-augmentation of the imaginary data, coming from the prior predictive distribution of a reference model. Exploiting this data-augmentation scheme, we have also constructed an efficient Monte Carlo estimate of the marginal likelihood of each competing model. Variable selection is then attained by estimating posterior model weights in the full space, or by considering an alternative MC^3 scheme. The proposed methodology has been implemented on two real life examples.

All results have been presented over different training samples, in contrast to relevant research work, where uncertainty due to the training sample selection has been ignored. On large model spaces where accurate estimation of posterior model probabilities is computationally demanding (if not infeasible), selection of "good" models can be based on posterior marginal inclusion probabilities (Barbieri and Berger 2004), which can be estimated in an easier fashion and more accurately from an MCMC output as suggested by Berger and Molina (2005); also see in Clyde et al. (2011) for an efficient variable selection method based on posterior marginal inclusion probabilities.

Appendix

Using efficient and optimized R code running under Windows on an i5-2430M processor at 2.40 GHz, we estimate that the clock time for performing the full enumeration search, with 1000 iterations for estimating the marginal likelihood and 100 different training sub-samples, for the Hald's cement data would be approximately 100 s, while for the prostate cancer data would be approximately 3500 s (about 2.4 days). On the contrary the clock time for the full enumeration R code using the Zellner's g prior was approximately 1 s for each illustration. The R programs are available upon request.

References

Barbieri, M., Berger, J.: Optimal predictive model selection. Ann. Stat. **32**, 870–897 (2004)

Berger, J., Molina, G.: Posterior model probabilities via path-based pairwise priors. Stat. Neerl. **59**, 3–15 (2005)

Berger, J., Pericchi, L.: The intrinsic Bayes factor for model selection and prediction. J. Am. Stat. Assoc. **91**, 109–122 (1996)

Casella, G., Girón, F., Martínez, M., Moreno, E.: Consistency of Bayesian procedures for variable selection. Ann. Stat. **37**, 1207–1228 (2009)

Casella, G., Moreno, E.: Objective Bayesian variable selection. J. Am. Stat. Assoc. **101**, 157–167 (2006)

Celeux, G., El Anbari, M., Marin, J.-M., Robert, C.P.: Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation. Bayesian Anal. (forthcoming), arXiv:1010.0300

Clyde, M., Ghosh, J., Littman, M.: Bayesian adaptive sampling for variable selection and model averaging. J. Comput. Graph. Stat. **20**, 80–101 (2011)

Dellaportas, P., Forster, J., Ntzoufras, I.: Joint specification of model space and parameter space prior distributions. Statist. Sci. (2012, forthcoming). Currently available at http://www.stat-athens.aueb.gr/~jbn/papers/paper24.htm

Fernandez, C., Ley, E., Steel, M.: Benchmark priors for Bayesian model averaging. J. Econom. **100**, 381–427 (2001)

Girón, F., Moreno, E., Martínez, M.: An objective Bayesian procedure for variable regression in regression. In: Balakrishnan, N., Castillo, E., Sarabia, J.M. (eds.) Advances on Distribution Theory, Order Statistics and Inference, pp. 393–408. Birkhäuser, Boston (2006)

Kass, R., Raftery, A.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)

Leng, C., Tran, M.N., Nott, D.: Bayesian adaptive lasso (2010), arXiv:1009.2300. Available at http://adsabs.harvard.edu/abs/2010arXiv1009.2300L

Liang, F., Paulo, R., Molina, G., Clyde, M., Berger, J.: Mixtures of g priors for Bayesian variable selection. J. Am. Stat. Assoc. **103**, 410–423 (2008)

Madigan, D., York, J.: Bayesian graphical models for discrete data. Int. Stat. Rev. **63**, 215–232 (1995)

Montgomery, D., Peck, E.: Introduction to Linear Regression Analysis. Wiley, New York (1982)

Moreno, E., Girón, F.: Comparison of Bayesian objective procedures for variable selection in linear regression. Test **17**, 472–490 (2008)

Ntzoufras, I.: Bayesian analysis of the normal regression model. In: Bocker, K. (ed.) Rethinking Risk Measurement and Reporting: Uncertainty, Bayesian Analysis and Expert Judgment: Volume I, pp. 69–106 (2010). ISBN-10:1-906348-40-5, ISBN-13:978-1-906348-40-3: Risk Books

Pérez, J.: Development of expected posterior prior distribution for model comparisons. Ph.D. thesis, Department of Statistics, Purdue University, USA (1998)

Pérez, J., Berger, J.: Expected-posterior prior distributions for model selection. Biometrika **89**, 491–511 (2002)

Stamey, T., Kabakin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., Yang, N.: Prostate-specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II: radical prostatectomy treated patients. J. Urol. **16**, 1076–1083 (1989)

Zellner, A.: On assessing prior distributions and Bayesian regression analysis using g-prior distributions. In: Goel, P., Zellner, A. (eds.) Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, pp. 233–243. North-Holland, Amsterdam (1986)

Zellner, A., Siow, A.: Posterior odds ratios for selected regression hypothesis (with discussion). In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (eds.) Bayesian Statistics, vol. 1, pp. 585–606 & 618–647 (discussion). Oxford University Press, Oxford (1980).