Computation for Intrinsic Variable Selection in Normal Regression Models via Expected-Posterior Prior

D. Fouskakis,[*] and I. Ntzoufras[†]

Summary: 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 model spaces, a model search algorithm based on MC^3 is constructed. The proposed methodology is implemented in two real life examples, already used in the relevant literature of objective variable selection. In both illustrated examples, uncertainty over different training samples is also considered.

Keywords: Bayesian variable selection; Expected posterior priors; Imaginary data; Intrinsic priors; Jeffreys prior; Objective model selection methods; Normal regression models.

1 Introduction

In this paper we focus on the variable selection problem for normal regression models. Let us denote by $\mathcal{M}$ the model space, consisting of all combinations of the available covariates, then for every $m_\ell \in \mathcal{M}$ with parameters $(\boldsymbol{\beta}_\ell, \sigma^2)$ the likelihood is specified by

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

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. One of the proposed approached is the intrinsic Bayes factors (IBF), introduced by Berger & Pericchi (1996). In order to provide a full Bayesian interpretation of IBFs, they also define intrinsic prior (IP) distributions. The intrinsic prior methodology has been applied for objective variable selection problems in normal regression

[*] D. Fouskakis is with the Department of Mathematics, National Technical University of Athens, Zografou Campus, Athens 15780 Greece; email fouskakis@math.ntua.gr

[†] I. Ntzoufras is with the Department of Statistics, Athens University of Economics and Business, 76 Patision Street, Athens 10434 Greece; email ntzoufras@aueb.gr

models, by Casella & Moreno (2006), Moreno & Girón (2008), Girón, Moreno & Martínez (2006) and Casella, Girón, Martínez & Moreno (2009).

Intrinsic priors are closely related with the expected-posterior prior distributions of Pérez & Berger (2002) which have nice interpretation based on imaginary training data coming from prior predictive distributions. In this paper we implement the expected-posterior prior methodology on variable selection problems in normal regression models. By this way, 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; by this way the uncertainty over different training samples is considered.

2 Expected posterior priors

Pérez & Berger (2002) provided a different viewpoint of the intrinsic priors. They have defined the expected posterior prior (EPP) as the posterior distribution of a 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 & Berger 2002, def. 1, p. 493). Hence the EPP for the parameters of any model $m_\ell \in \mathcal{M}$ is given by

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

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 reference prior $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma^2)$ 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 $\pi_0^N(\boldsymbol{\beta}_\ell, \sigma^2)$. For the reference model ($m_\ell = m_0$) this prior degenerates to $\pi_0^N(\boldsymbol{\beta}_\ell, \sigma^2)$.

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 $m_0^N(\boldsymbol{y}^*)$ 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 & Pericchi (1996).

A question that naturally arises is which model must 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 Jeffreys prior as the baseline prior distribution. Hence for $m_\ell \in \mathcal{M}$, where $\mathcal{M}$ is the model space, we have

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

where c_ℓ is an unknown normalizing constant.

Under the above setup, for every $m_\ell \in \mathcal{M}$, if we suppose we have 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:

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

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.

$$\begin{aligned} m_\ell^N(\boldsymbol{y}^*|\,\mathrm{X}_\ell^*\,) &= \int\!\!\int f(\boldsymbol{y}^*|\,\boldsymbol{\beta}_\ell\,,\sigma^2,m_\ell\,;\mathrm{X}_\ell^*\,)\pi_\ell^N(\boldsymbol{\beta}_\ell\,,\sigma^2)d\boldsymbol{\beta}_\ell\,d\sigma^2 \\ &= c_\ell\,(\pi)^{\frac{d_\ell-n^*}{2}}\,|\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*|^{-\frac{1}{2}}\,\frac{\Gamma\left(\frac{n^*-d_\ell}{2}\right)}{RSS_\ell^{*\frac{n^*-d_\ell}{2}}} \end{aligned} \tag{3}$$

with

$$RSS_\ell^* = (\boldsymbol{y}^* - \mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell^*)^T(\boldsymbol{y}^* - \mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell^*) = \boldsymbol{y}^{*T}\big(\mathrm{I}_{n^*} - \mathrm{X}_\ell^*(\mathrm{X}_\ell{}^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell{}^{*T}\big)\boldsymbol{y}^* \tag{4}$$

being the residual sum of squares using $(\boldsymbol{y}^*,\mathrm{X}_\ell^*)$ as data and $\widehat{\boldsymbol{\beta}}_\ell^* = (\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*T}\boldsymbol{y}^*$. For a detailed derivation of the marginal likelihood see at the Appendix of this article.

4 Computation of the Posterior Distribution

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

$$\begin{aligned} \pi_\ell^I(\boldsymbol{\beta}_\ell,\sigma^2|\boldsymbol{y};\mathrm{X}_\ell\,,\mathrm{X}_\ell^*) &\propto f(\boldsymbol{y}|\,\boldsymbol{\beta}_\ell\,,\sigma^2,m_\ell\,;\mathrm{X}_\ell^*)\pi_\ell^I(\boldsymbol{\beta}_\ell,\sigma^2|\,\mathrm{X}_\ell^*) \\ &\propto \int f(\boldsymbol{y}|\boldsymbol{\beta}_\ell\,,\sigma^2,m_\ell\,;\mathrm{X}_\ell)f(\boldsymbol{\beta}_\ell\,,\sigma^2|\boldsymbol{y}^*,m_\ell\,;\,\mathrm{X}_\ell^*)m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)d\boldsymbol{y}^* \\ &\propto \int f(\boldsymbol{\beta}_\ell\,,\sigma^2|\boldsymbol{y},\boldsymbol{y}^*,m_\ell\,;\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*;\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)d\boldsymbol{y}^* \\ &\propto \int f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell\,;\,\widetilde{\boldsymbol{\beta}}^N,\,\widetilde{\Sigma}^N\sigma^2\big)f_{IG}(\sigma^2\,;\,\widetilde{a}_\ell^N,\widetilde{b}_\ell^N)m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*;\mathrm{X}_\ell,\mathrm{X}_\ell^*)m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)d\boldsymbol{y}^* \end{aligned}$$

where

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}^N &= \widetilde{\Sigma}^N\big(\mathrm{X}_\ell^T\boldsymbol{y} + \mathrm{X}_\ell{}^{*T}\boldsymbol{y}^*\big),\;\; \widetilde{\Sigma}^N = \left\{\mathrm{X}_\ell{}^{*T}\mathrm{X}_\ell^* + \mathrm{X}_\ell^T\mathrm{X}_\ell\right\}^{-1}, \\ \widetilde{a}_\ell^N &= n/2 + n^*/2 - d_\ell/2, \text{ and } \widetilde{b}_\ell^N = RSS_\ell^N/2 + RSS_\ell^*/2 \end{aligned}$$

with

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

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

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

Thus, we can write the following MCMC scheme:

1. Generate $\boldsymbol{y}^*$ from
$$f(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*) \propto \frac{m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y};\, \mathrm{X}_\ell, \mathrm{X}_\ell^*) m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell, \mathrm{X}_\ell^*)} m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*).$$

2. Generate σ^2 from $IG(\widetilde{a}_\ell^N, \widetilde{b}_\ell^N)$.

3. Generate $\boldsymbol{\beta}_\ell$ from $N_{d_\ell}\big(\widetilde{\boldsymbol{\beta}}^N,\ \widetilde{\Sigma}^N\sigma^2\big)$.

We can generate the imaginary data $\boldsymbol{y}^*$ by using a Metropolis-Hastings algorithm with proposal

$$q(\boldsymbol{y}^{*'}) = m_\ell^N(\boldsymbol{y}^{*'}|\boldsymbol{y},\, \mathrm{X}_\ell, \mathrm{X}_\ell^*) = f_{St_{n^*}}\left(\boldsymbol{y}^{*'}\,;\, n-d_\ell\,, \mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell\,, \frac{RSS_\ell}{n-d_\ell}\big(\mathrm{I}_{n^*} + \mathrm{X}_\ell^*(\mathrm{X}_\ell^T\mathrm{X}_\ell)^{-1}\mathrm{X}_\ell^{*^T}\big)\right) \tag{5}$$

and acceptance probability

$$\begin{aligned} \alpha &= \min\left\{1, \frac{m_0^N(\boldsymbol{y}^{*'}|\mathrm{X}_0^*) m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*)}{m_\ell^N(\boldsymbol{y}^{*'}|\mathrm{X}_\ell^*) m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*)}\right\} \\ &= \min\left\{1, \left(\frac{RSS_\ell^{*'}}{RSS_\ell^*}\right)^{(n^*-d_\ell)/2} \times \left(\frac{RSS_0^{*'}}{RSS_0^*}\right)^{-(n^*-d_0)/2}\right\} \end{aligned} \tag{6}$$

where $RSS_\ell = (\boldsymbol{y} - \mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell)^T(\boldsymbol{y} - \mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell)$ being the residual sum of squares using $(\boldsymbol{y}, \mathrm{X}_\ell)$ as data and $RSS_\ell^{*'}$ is given in (4) using $(\boldsymbol{y}^{*'}, \mathrm{X}_\ell^*)$ as data.

5 Variable Selection Computation

In this Section we provide two alternative approaches for the evaluation of the models under consideration. In Section 5.1 we construct an efficient Monte Carlo scheme for the estimation of the marginal likelihood for any given training sample X*, while in Section 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

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

$$
\begin{aligned}
&= \int \left\{ \int \int f(\boldsymbol{y}^* | \boldsymbol{\beta}_\ell \,, \sigma^2, m_\ell \,; \mathrm{X}^*_\ell \,) f(\boldsymbol{\beta}_\ell, \sigma^2 | \boldsymbol{y}, m_\ell \,; \mathrm{X}_\ell) d\boldsymbol{\beta}_\ell \, d\sigma^2 \right\} m^N_\ell(\boldsymbol{y}|\mathrm{X}_\ell \,) \frac{m^N_0(\boldsymbol{y}^*|\mathrm{X}^*_0 \,)}{m^N_\ell(\boldsymbol{y}^*|\mathrm{X}^*_\ell \,)} d\boldsymbol{y}^* \\
&= \int m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,) m^N_\ell(\boldsymbol{y}|\mathrm{X}_\ell \,) \frac{m^N_0(\boldsymbol{y}^*|\mathrm{X}^*_0 \,)}{m^N_\ell(\boldsymbol{y}^*|\mathrm{X}^*_\ell \,)} d\boldsymbol{y}^* \\
&= \int m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,) m^N_\ell(\boldsymbol{y}|\mathrm{X}_\ell \,) \frac{m^N_0(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}^*_0 \,) m^N_0(\boldsymbol{y}|\mathrm{X}_0 \,)/m^N_0(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_0, \mathrm{X}^*_0 \,)}{m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,) m^N_\ell(\boldsymbol{y}|\mathrm{X}_\ell)/m^N_\ell(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,)} d\boldsymbol{y}^* \\
&= m^N_0(\boldsymbol{y}|\mathrm{X}_0 \,) \int \frac{m^N_0(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}^*_0 \,)/m^N_0(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_0, \mathrm{X}^*_0 \,)}{m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,)/m^N_\ell(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,)} m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,) d\boldsymbol{y}^* \ . \qquad (7)
\end{aligned}
$$

In the above expression we can calculate the predictive densities $m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,)$ and $m^N_\ell(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,)$ for any model ℓ, under the baseline prior $\pi^N_\ell(\boldsymbol{\beta}_\ell, \sigma^2)$, which are given by

$$
m^N_\ell(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,) = f_{St_n}\left(\boldsymbol{y} \,;\, n^* - d_\ell, \mathrm{X}_\ell \widehat{\boldsymbol{\beta}}^*_\ell, \left(\mathrm{I}_n + \mathrm{X}_\ell ({\mathrm{X}^*_\ell}^T \mathrm{X}^*_\ell)^{-1} \mathrm{X}^T_\ell\right) \widehat{\sigma}^{*2}_U\right) \qquad (8)
$$

and

$$
m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,) = f_{St_{n^*}}\left(\boldsymbol{y}^* \,;\, n - d_\ell, \mathrm{X}^*_\ell \widehat{\boldsymbol{\beta}}_\ell, \left(\mathrm{I}_{n^*} + \mathrm{X}^*_\ell (\mathrm{X}^T_\ell \mathrm{X}_\ell)^{-1} \mathrm{X}^{*T}_\ell\right) \widehat{\sigma}^2_U\right) \ , \qquad (9)
$$

where $\widehat{\boldsymbol{\beta}}^*_\ell$ is the maximum likelihood estimate of $\boldsymbol{\beta}_\ell$, $\widehat{\sigma}^{*2}_U = RSS^*_\ell/(n^* - d_\ell)$ 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; $\widehat{\boldsymbol{\beta}}_\ell$, $\widehat{\sigma}^2_U$ and RSS_ℓ are the corresponding measures using $(\boldsymbol{y}, \mathrm{X}_\ell)$ as data. The quantity $m^N_0(\boldsymbol{y}|\mathrm{X}^*_0 \,)$ denotes the marginal likelihood of the reference model m_0, as derived in (3). The appearance of this quantity in (7) does not cause any problem in our setup since it is common in all marginal likelihoods and it is cancelled out when we compare models using Bayes factors, posterior model odds or probabilities.

We can now use (7) to setup a Monte Carlo scheme and estimate the marginal likelihood up to the common constant $m^N_0(\boldsymbol{y}|\mathrm{X}_0 \,)$. Thus we generate $\boldsymbol{y}^{*(t)}$, $t = 1, \ldots, T$, from $m^N_\ell(\boldsymbol{y}^*|\, \boldsymbol{y}; \mathrm{X}_\ell \,, \mathrm{X}^*_\ell \,)$ given in (9) and estimate the unnormalized marginal likelihood

$$
m^{IU}_\ell(\boldsymbol{y}|\mathrm{X}_\ell \,, \mathrm{X}^*_\ell \,) = m^I_\ell(\boldsymbol{y}|\mathrm{X}_\ell \,, \mathrm{X}^*_\ell \,)/m^N_0(\boldsymbol{y}|\mathrm{X}_0 \,) \qquad (10)
$$

by

$$
\widehat{m}^{IU}_\ell(\boldsymbol{y}|\mathrm{X}_\ell \,, \mathrm{X}^*_\ell \,, \delta) = \frac{1}{T} \sum_{t=1}^{T} \frac{m^N_\ell(\boldsymbol{y}|\, \boldsymbol{y}^{*(t)}; \mathrm{X}_\ell, \mathrm{X}^*_\ell \,)}{m^N_0(\boldsymbol{y}|\, \boldsymbol{y}^{*(t)}; \mathrm{X}_0, \mathrm{X}^*_0)} \frac{m^N_0(\boldsymbol{y}^{*(t)}|\, \boldsymbol{y}; \mathrm{X}_0, \mathrm{X}^*_0 \,)}{m^N_\ell(\boldsymbol{y}^{*(t)}|\, \boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell)} \ . \qquad (11)
$$

For small model spaces it is easy to estimate the unnormalized marginal likelihoods (10) for all models under consideration using the above sampling scheme. For large spaces, it is possible to implement an MC^3 algorithm (Madigan & York 1995, Kass & Raftery 1995) by estimating (10) for each model that is evaluated for the first time within the iterative scheme.

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 set of covariates). If we denote by X the $(n \times d)$ design/data matrix of the full model, the algorithm can be summarized by

- For $k = 1, \dots, K$ (training samples):

  1. Randomly consider a submatrix X^* of X with dimension $(n^* \times d)$.
  2. For $t = 1, \dots, T$ (iterations):
     (a) For a given model m_ℓ, propose $\boldsymbol{y}^*$ from (5) and accept it with probability (6).
     (b) For $j = 1, \dots, p$, 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 Section 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 model space.

6.1 Hald's data

We consider the Hald's cement data (Montgomery & Peck 1982) to illustrate the proposed approach. This dataset consists of $n = 13$ observations and $p = 4$ covariates and it was previously used by Girón et al. (2006) for illustrating objective variable selection methods. The response variable $\boldsymbol{y}$ 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 these data is that variables X_1 and X_3 are highly correlated ($corr(X_1,\, X_3) = -0.824$), as well as the variables X_2 and X_4 (with $corr(X_1, X_4) = -0.975$).

Table 1 presents posterior model probabilities summaries for the best models over 100 different training sub-samples, after performing a full enumeration search, together with their corresponding Bayes factors. For estimating the marginal likelihood we have used 1000 iterations. Figures 1 and 2 provide a pictorial representation of the distributions of posterior model probabilities (for the five best models) and the marginal inclusion probabilities, respectively, across different training samples. We note that the MAP model is $X_1{+}X_2$ with posterior probabilities ranging in (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 probability values correspond to median Bayes factors. The boxplots of the inclusion probabilities indicate that covariates X_1 and

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

		Posterior Model Probabilities					
	Model				Percentiles		Median Based
m_j	Formula	Mean	Median	SD	2.5%	97.5%	Bayes Factors
1	$X_1 + X_2$	0.411	0.405	0.074	0.309	0.554	1.000
2	$X_1 + X_4$	0.167	0.160	0.040	0.107	0.278	2.529
3	$X_1 + X_2 + X_4$	0.132	0.138	0.034	0.061	0.183	2.930
4	$X_1 + X_2 + X_3$	0.128	0.132	0.033	0.062	0.185	3.061
5	$X_1 + X_3 + X_4$	0.105	0.106	0.027	0.051	0.148	3.807

X_2 should be included in the model formulation with posterior probabilities clearly above 0.5 for all training samples.

We have also performed the same task with 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 Section 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 samples which may 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 Figures 1 and 2.

Figure 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 Hald's data (Example 6.1).

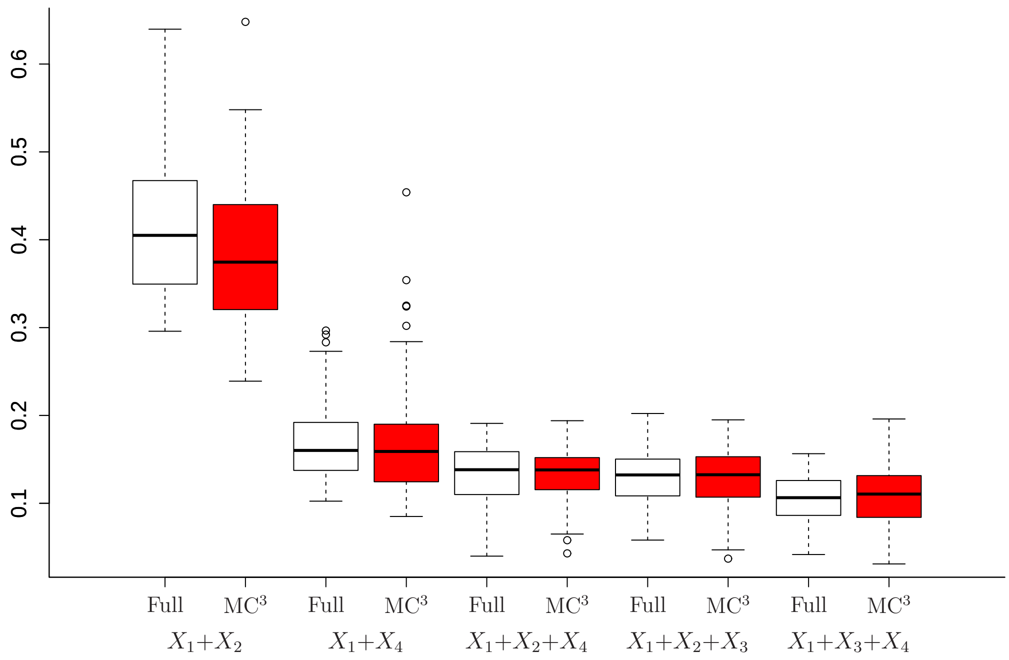
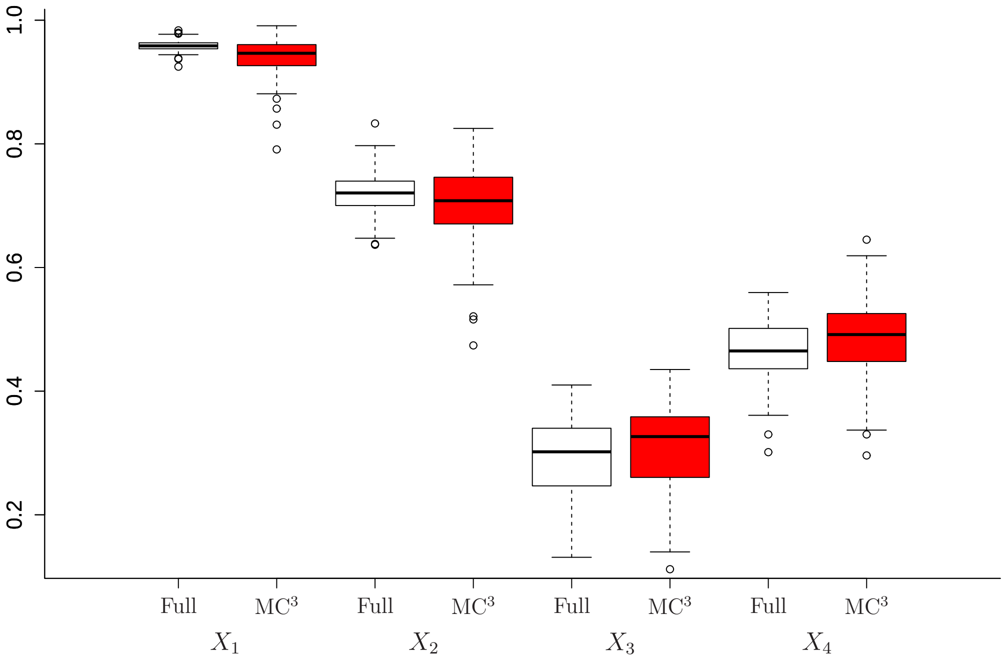


Figure 2: Boxplots comparing the posterior marginal inclusion probabilities over 100 training samples for the full enumeration (Full) using Monte Carlo estimates and the MC^3 for Hald's data (Example 6.1).



6.2 Prostate cancer data

In this Section, we present results of our methodology in the prostate cancer data (Stamey, Kabakin, McNeal, Johnstone, Freiha, Redwine & Yang 1989), which was also used by Girón et al. (2006) and Moreno & Girón (2008) to illustrate their approach. This dataset consists of $n = 97$ observations and $p = 8$ covariates. The response variable $\boldsymbol{y}$ 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).

Table 2: Summaries of posterior model probabilities for the best models over 100 different training sub-samples together with Bayes factors of the MAP for prostate cancer data (Example 6.2).

		Posterior Model Probabilities					
	Model				Percentiles		Median Based
m_j	Formula	Mean	Median	SD	2.5%	97.5%	Bayes Factors
1	$X_1 + X_2 + X_5$	0.299	0.296	0.062	0.199	0.446	1.000
2	$X_1 + X_2 + X_4 + X_5$	0.107	0.104	0.036	0.054	0.177	2.845
3	$X_1 + X_2 + X_3 + X_5$	0.076	0.069	0.032	0.035	0.148	4.300
4	$X_1 + X_2 + X_5 + X_8$	0.067	0.066	0.021	0.033	0.110	4.472

The structure as well as the results of this illustration are similar as in Section 6.1. Results are summarized in Table 2 and Figures 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 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 Section 5.2 for 2000 iterations, considering 100 different training samples. Results from the two methods were equivalent; see Figures 3 and 4 for a graphical comparison.

Figure 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 prostate cancer data (Example 6.2).

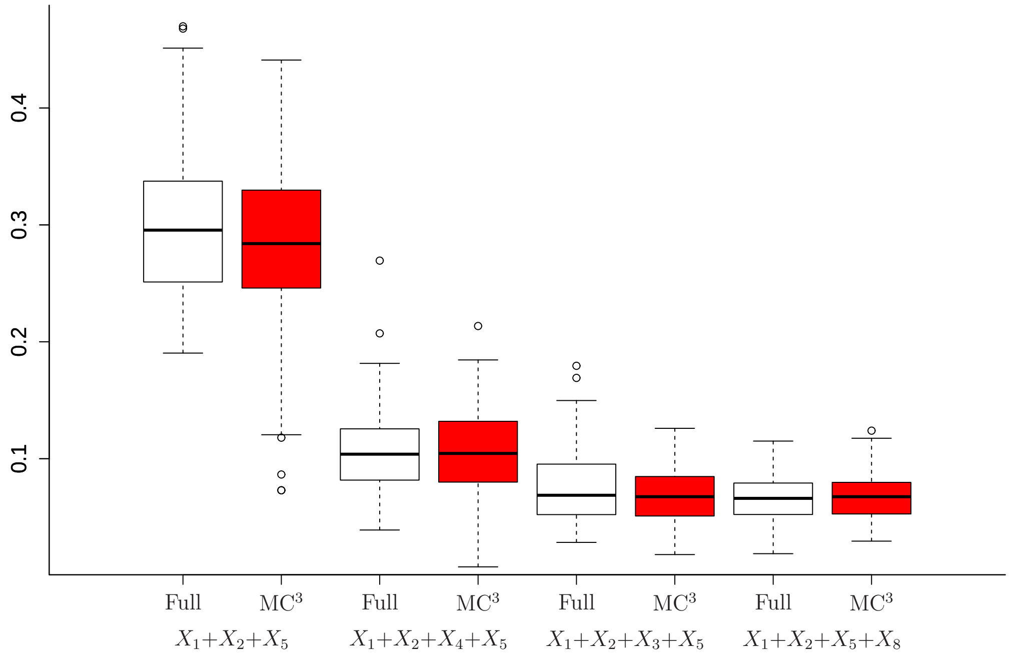
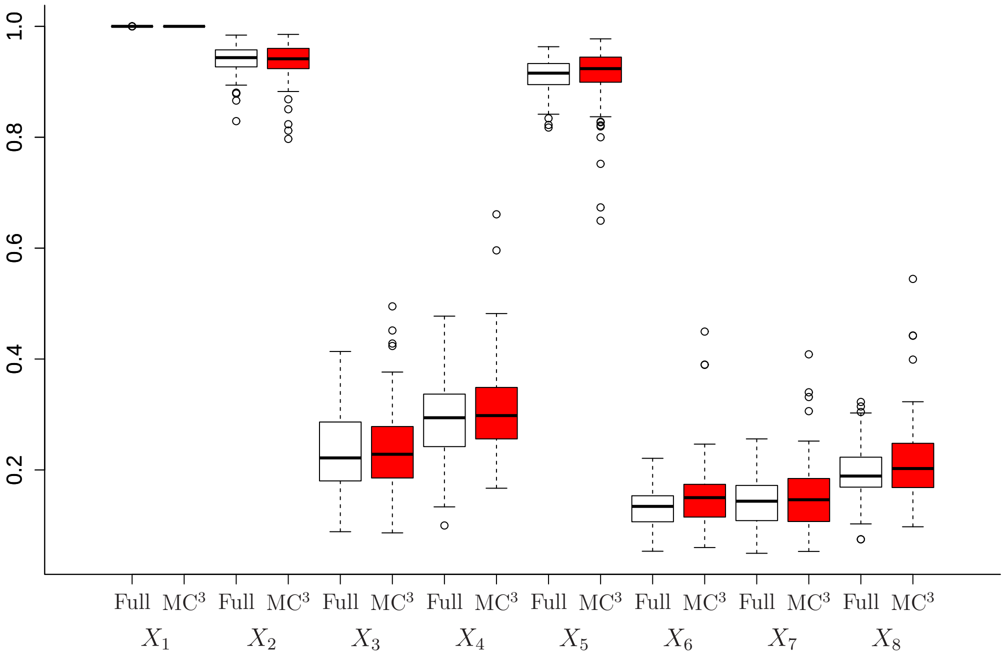


Figure 4: Boxplots comparing the posterior marginal inclusion probabilities over 100 training samples for the full enumeration (Full) using Monte Carlo estimates and the MC^3 for prostate cancer data (Example 6.2).



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 are presented over different training samples, in contrast to relevant research work, where uncertainty due to the training sample selection is ignored. Selection of "good" models can be based on the posterior inclusion probabilities which are more robust across training sample, in comparison to posterior model probabilities.

References

Berger, J. & Pericchi, L. (1996), “The intrinsic Bayes factor for model selection and prediction”, Journal of the American Statistical Association **91**, 109–122.

Casella, G., Girón, F., Martínez, M. & Moreno, E. (2009), ‘Consistency of bayesian procedures for variable selection’, Annals of Statistics **37**, 1207–1228.

Casella, G. & Moreno, E. (2006), ‘Objective bayesian variable selection’, Journal of the American Statistical Association **101**, 157–167.

Girón, F., Moreno, E. & Martínez, M. (2006), “An objective Bayesian procedure for variable regression in regression”, in N. Balakrishnan, E. Castillo and J.M. Sarabia, eds., *Advances on distribution theory, order statistics and inference*, Birkihäuser, Boston, pp. 393–408.

Kass, R. & Raftery, A. (1995), “Bayes factors”, Journal of the American Statistical Association **90**, 773–795.

Madigan, D. & York, J. (1995), ‘Bayesian graphical models for discrete data’, International Statistical Review **63**, 215–232.

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

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

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

Stamey, T., Kabakin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. & Yang, N. (1989), ‘Prostate-specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: radical prostatectomy treated patients’, Journal of Urology **16**, 1076–1083.

Appendix

Calculation of the Marginal Likelihood

$$
\begin{aligned}
m_\ell^N(\boldsymbol{y}^*|\,\mathrm{X}_\ell^*\,) &= \int\int f(\boldsymbol{y}^*|\,\boldsymbol{\beta}_\ell\,,\sigma^2, m_\ell\,;\mathrm{X}_\ell^*\,)\pi_\ell^N(\boldsymbol{\beta}_\ell\,,\sigma^2)d\boldsymbol{\beta}_\ell\,d\sigma^2 \\
&= c_\ell \int\int (2\pi\sigma^2)^{-\frac{n^*}{2}}\, exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}^* - \mathrm{X}_\ell^*\,\boldsymbol{\beta}_\ell)^T(\boldsymbol{y}^* - \mathrm{X}_\ell^*\,\boldsymbol{\beta}_\ell)\right\}\frac{1}{\sigma^2}d\boldsymbol{\beta}_\ell\,d\sigma^2 \\
&= c_\ell\,(2\pi)^{-\frac{n^*}{2}}\int\left(\sigma^2\right)^{-\frac{n^*+2}{2}} exp\left\{-\frac{{\boldsymbol{y}^*}^T\boldsymbol{y}^*}{2\sigma^2}\right\}\int exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\beta}_\ell^T{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*\boldsymbol{\beta}_\ell - 2\boldsymbol{\beta}_\ell^T{\mathrm{X}_\ell^*}^T\boldsymbol{y}^*\right]\right\}d\boldsymbol{\beta}_\ell\,d\sigma^2 \\
&= c_\ell\,(2\pi)^{-\frac{n^*}{2}}\int\left(\sigma^2\right)^{-\frac{n^*+2}{2}} exp\left\{-\frac{{\boldsymbol{y}^*}^T\boldsymbol{y}^*}{2\sigma^2} + \frac{1}{2\sigma^2}\widehat{\boldsymbol{\beta}}_\ell^T{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell\right\} \\
&\qquad\qquad \times\int exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}_\ell - \widehat{\boldsymbol{\beta}}_\ell^*\right)^T{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*\left(\boldsymbol{\beta}_\ell - \widehat{\boldsymbol{\beta}}_\ell^*\right)\right\}d\boldsymbol{\beta}_\ell\,d\sigma^2 \\
&= c_\ell\,(2\pi)^{-\frac{n^*}{2}}\int\left(\sigma^2\right)^{-\frac{n^*+2}{2}} exp\left\{-\frac{{\boldsymbol{y}^*}^T\boldsymbol{y}^*}{2\sigma^2} + \frac{1}{2\sigma^2}\widehat{\boldsymbol{\beta}}_\ell^T{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell\right\}(2\pi)^{\frac{d_\ell}{2}}|{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*|^{-\frac{1}{2}}\left(\sigma^2\right)^{\frac{d_\ell}{2}}d\sigma^2 \\
&= c_\ell\,(2\pi)^{\frac{d_\ell-n^*}{2}}\,|{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*|^{-\frac{1}{2}}\int\left(\sigma^2\right)^{-\left(\frac{n^*-d_\ell}{2}+1\right)} exp\left\{-\frac{\frac{{\boldsymbol{y}^*}^T\boldsymbol{y}^*-\widehat{\boldsymbol{\beta}}_\ell^T{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell}{2}}{\sigma^2}\right\}d\sigma^2 \\
&= c_\ell\,(2\pi)^{\frac{d_\ell-n^*}{2}}\,|{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{n^*-d_\ell}{2}\right)}{\left(\frac{{\boldsymbol{y}^*}^T\boldsymbol{y}^*-\widehat{\boldsymbol{\beta}}_\ell^T{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*\widehat{\boldsymbol{\beta}}_\ell}{2}\right)^{\frac{n^*-d_\ell}{2}}} \\
&= c_\ell\,(\pi)^{\frac{d_\ell-n^*}{2}}\,|{\mathrm{X}_\ell^*}^T\mathrm{X}_\ell^*|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{n^*-d_\ell}{2}\right)}{{RSS_\ell^*}^{\frac{n^*-d_\ell}{2}}}.
\end{aligned}
$$