econometrics

Article

Bayesian Model Averaging Using Power-Expected-Posterior Priors

Dimitris Fouskakis [1,*] and Ioannis Ntzoufras [2]

[1] Statistics Lab, Department of Mathematics, National Technical University of Athens, Zografou Campus, 15780 Athens, Greece

[2] Computational and Bayesian Statistics Lab, Department of Statistics, Athens University of Economics and Business, 10434 Athens, Greece; ntzoufras@aueb.gr

[*] Correspondence: fouskakis@math.ntua.gr

Received: 18 February 2020; Accepted: 6 May 2020; Published: 11 May 2020

Abstract: This paper focuses on the Bayesian model average (BMA) using the power–expected–posterior prior in objective Bayesian variable selection under normal linear models. We derive a BMA point estimate of a predicted value, and present computation and evaluation strategies of the prediction accuracy. We compare the performance of our method with that of similar approaches in a simulated and a real data example from economics.

Keywords: Bayesian model averaging; Bayesian variable selection; expected–posterior priors; imaginary training samples; power–expected–posterior priors

1. Introduction

We consider the variable–selection problem for normal regression models. Let us denote the model space by $\mathcal{M}$, consisting of all combinations of available covariates. Then, for every model $M_\ell \in \mathcal{M}$, the likelihood is specified by

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

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is a multivariate random variable expressing the response for each subject, X_ℓ is a $n \times k_\ell$ 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 k_ℓ with the effects of each covariate on the response data $\boldsymbol{Y}$, and σ^2_ℓ is the error variance of the model.

Under the Bayesian model choice perspective, we need to introduce priors on the model space and on the parameters of each competing model. With respect to the prior distribution on the parameters in each model, because we are not confident about any given set of regressors as explanatory variables, little prior information on their regression coefficients can be expected. This argument alone justifies the need for an objective model choice approach in which vague prior information is assumed. Hence, within each model, we consider default prior distributions on the regression coefficients and error variance. Default priors for normal regression parameters are improper and thus cannot be used, since they lead to an indeterminate Bayes factor (Berger and Pericchi 2001). This has urged the objective Bayesian community to develop various methodologies to overcome the problem of prior specification in model–selection problems. One of the proposed approaches is the expected–posterior prior (EPP) of Pérez and Berger (2002). Starting from a baseline (typically improper) prior $\pi^N_\ell(\boldsymbol{\theta}_\ell)$ of parameters

$\boldsymbol{\theta}_\ell = (\boldsymbol{\beta}_\ell, \sigma_\ell)$ of model M_ℓ, the approach relies on the utilization of the device of "imaginary training samples". If we denote by $\boldsymbol{y}^*$ the imaginary training sample of size n^*, the EPP is defined as

$$\pi_\ell^{EPP}(\boldsymbol{\theta}_\ell) = \int \pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*)\, m^*(\boldsymbol{y}^*)\, d\boldsymbol{y}^*\,, \tag{1}$$

where $\pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*)$ is the posterior distribution of $\boldsymbol{\theta}_\ell$ for model M_ℓ using the baseline prior $\pi_\ell^N(\boldsymbol{\theta}_\ell)$ and data $\boldsymbol{y}^*$. A usual choice of m^* is $m^*(\boldsymbol{y}^*) = m_0^N(\boldsymbol{y}^*) \equiv f(\boldsymbol{y}^*|M_0)$, i.e., the marginal likelihood, evaluated at $\boldsymbol{y}^*$, for a *reference* model M_0 under the baseline prior $\pi_0^N(\boldsymbol{\theta}_0)$. Then, for $\ell = 0$, it is straightforward to show that $\pi_0^{EPP}(\boldsymbol{\theta}_0) = \pi_0^N(\boldsymbol{\theta}_0)$. Under the variable–selection problem, the usual choice is to consider M_0 to be the "null" model with only the intercept; this is the choice considered in this paper in the last two experimental sections. In a more general setting, we can assume that the response variable is known to be explained by k_0 variables (including the intercept) that form the reference model M_0, and, by some subset of p, other explanatory variables that form models under comparison. Thus, in the rest of the paper, we assume that M_0 is nested to all other models under comparison. Under this more general case, we denote by $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \sigma_0)$ the parameters of M_0, and by X_0, its design matrix (assumed to be of full rank). Since M_0 is nested in every other competing model M_1, with parameters $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1, \sigma_1)$ and design matrix X_1 (again assumed to be of full rank), we can henceforth assume that $\mathrm{X}_1 = [\mathrm{X}_0|\mathrm{X}_{e_1}]$ and $\boldsymbol{\beta}_1 = \left(\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_{e_1}^T\right)^T$, so that $\boldsymbol{\beta}_0$ is a "common" parameter between the two models, and $\boldsymbol{\beta}_{e_1}$ is a model–specific. The use of a "common" parameter $\boldsymbol{\beta}_0$ in nested model comparison is often made to justify the employment of the same, potentially improper, prior on $\boldsymbol{\beta}_0$ across models. This usage is becoming standard–see, for example, Bayarri et al. (2012); Consonni et al. (2018). It can be justified if, without essential loss of generality, we assume that the model has been parametrized in an orthogonal fashion, so that $\mathrm{X}_0^T\mathrm{X}_1 = 0$. When M_0 is the "null" model, the above assumption can be easily justified, if we assume that, again without loss of generality, the columns of design matrix X of the full model, containing all p available explanatory variables, have been centred on their corresponding means, this makes the covariates orthogonal to the intercept, and gives the intercept an interpretation that is "common" to all models.

When comparing models M_0 and M_1, under the EPP methodology, imaginary design matrices X^*, with n^* rows, should also be introduced; $k_1 + 1 \leq n^* \leq n$. In what follows, we denote by X_0^* and $\mathrm{X}_1^* = \left[\mathrm{X}_0^*|\mathrm{X}_{e_1}^*\right]$ those imaginary design matrices under models M_0 and M_1 respectively. As before, we assume that those matrices are of full rank. Furthermore we denote by $\mathrm{P}_0^* = \mathrm{X}_0^*\left(\mathrm{X}_0{}^{*T}\mathrm{X}_0^*\right)^{-1}\mathrm{X}_0^{*T}$. The selection of minimal training sample size n^* was proposed by Berger and Pericchi (2004) to make information content of the prior as small as possible, and this is an appealing idea. Then, X^* can be extracted from original design matrix X by randomly selecting n^* from the n rows.

To diminish the effect of training samples, Fouskakis et al. (2015) generalized the EPP approach by introducing the power–expected–posterior (PEP) priors, combining ideas from the power–prior approach of Ibrahim and Chen (2000) and the unit–information–prior approach of Kass and Wasserman (1995). As a first step, the likelihoods involved in the EPP formula are raised to the $1/\delta$ ($\delta \geq 1$) power and are then density–normalized. This power parameter δ could be then set equal to the size of training sample n^* to represent information equal to one data point. Fouskakis et al. (2015) further set $n^* = n$; this choice gives rise to significant advantages, for example, when covariates are available, it results in automatic choice $\mathrm{X}^* = \mathrm{X}$; therefore, the selection of a training sample and its effects on the posterior model comparison are avoided while still holding prior information content at one data point. In the last two sections of this paper, this recommended setup ($\delta = n^* = n$) was used.

Specifically, the PEP prior under model M_1 is defined as

$$\pi_1^{PEP}(\boldsymbol{\theta}_1|\delta) \equiv \pi_1^{PEP}(\boldsymbol{\theta}_1) = \int \pi_1^N(\boldsymbol{\theta}_1|\boldsymbol{y}^*, \delta) m^*(\boldsymbol{y}^*|\delta) d\boldsymbol{y}^*, \tag{2}$$

with

$$\begin{aligned}\pi_1^N(\boldsymbol{\theta}_1|\boldsymbol{y}^*,\delta) &\propto f(\boldsymbol{y}^*|\boldsymbol{\theta}_1,\delta,M_1)\pi_1^N(\boldsymbol{\theta}_1)\\ f(\boldsymbol{y}^*|\boldsymbol{\theta}_1,\delta,M_1) &= \frac{f(\boldsymbol{y}^*|\boldsymbol{\theta}_1,M_1)^{1/\delta}}{\int f(\boldsymbol{y}^*|\boldsymbol{\theta}_1,M_1)^{1/\delta}d\boldsymbol{y}^*}\,.\end{aligned}$$

As before, we choose

$$m^*(\boldsymbol{y}^*|\delta) = m_0^N(\boldsymbol{y}^*|\delta) = \int f(\boldsymbol{y}^*|\boldsymbol{\theta}_0,\delta,M_0)\pi_0^N(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0\,,$$

where

$$f(\boldsymbol{y}^*|\boldsymbol{\theta}_0,\delta,M_0) = \frac{f(\boldsymbol{y}^*|\boldsymbol{\theta}_0,M_0)^{1/\delta}}{\int f(\boldsymbol{y}^*|\boldsymbol{\theta}_0,M_0)^{1/\delta}d\boldsymbol{y}^*}.$$

Regarding the baseline prior, under model M_1, we use

$$\pi_1^N(\boldsymbol{\beta}_1,\sigma_1) = c_1\pi_1^U(\boldsymbol{\beta}_1,\sigma_1) = c_1\sigma_1^{-(1+d_1)},$$

while under model M_0 we use

$$\pi_0^N(\boldsymbol{\beta}_0,\sigma_0) = c_0\pi_0^U(\boldsymbol{\beta}_0,\sigma_0) = c_0\sigma_0^{-(1+d_0)},$$

where c_0 and c_1 are the unknown normalizing constants of $\pi_0^U(\boldsymbol{\beta}_0,\sigma_0)$ and $\pi_1^U(\boldsymbol{\beta}_1,\sigma_1)$ respectively. Usual choices for d_0 and d_1 are $d_0 = d_1 = 0$ (resulting to the reference prior) or $d_0 = k_0$ and $d_1 = k_1$ (resulting in the dependence Jeffreys prior). In the last two experimental sections of this paper, the former case was used. Under the above setup, the PEP prior of the reference model is equal to the corresponding baseline prior, that is, $\pi_0^{PEP}(\boldsymbol{\beta}_0,\sigma_0) = \pi_0^N(\boldsymbol{\beta}_0,\sigma_0)$.

One of the advantages of using PEP priors (or EPPs for $\delta = 1$) is that the impropriety of baseline priors causes no indeterminacy of the Bayes factor. More specifically, the resulting Bayes factor for comparing model M_1 to M_0 takes the form of

$$\begin{aligned}B_{10} &= \frac{\int\int f(\boldsymbol{y}|\boldsymbol{\beta}_1,\sigma_1,M_1)\pi_1^{PEP}(\boldsymbol{\beta}_1,\sigma_1)d\boldsymbol{\beta}_1 d\sigma_1}{\int\int f(\boldsymbol{y}|\boldsymbol{\beta}_0,\sigma_0,M_0)\pi_0^{PEP}(\boldsymbol{\beta}_0,\sigma_0)d\boldsymbol{\beta}_0 d\sigma_0}\\ &= \frac{\int\int f(\boldsymbol{y}|\boldsymbol{\beta}_1,\sigma_1,M_1)\left[\int \frac{f(\boldsymbol{y}^*|\boldsymbol{\beta}_1,\sigma_1,\delta,M_1)\pi_1^N(\boldsymbol{\beta}_1,\sigma_1)}{m_1^N(\boldsymbol{y}^*|\delta)} m_0^N(\boldsymbol{y}^*|\delta)d\boldsymbol{y}^*\right]d\boldsymbol{\beta}_1 d\sigma_1}{\int\int f(\boldsymbol{y}|\boldsymbol{\beta}_0,\sigma_0,M_0)\pi_0^N(\boldsymbol{\beta}_0,\sigma_0)d\boldsymbol{\beta}_0 d\sigma_0}\end{aligned} \qquad (3)$$

where, for $\ell = 0, 1$,

$$\begin{aligned}m_\ell^N(\boldsymbol{y}^*|\delta) &= \int\int f(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell,\sigma_\ell,\delta,M_\ell)\pi_\ell^N(\boldsymbol{\beta}_\ell,\sigma_\ell)d\boldsymbol{\beta}_\ell d\sigma_\ell\\ &= c_\ell\int\int f(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell,\sigma_\ell,\delta,M_\ell)\pi_\ell^U(\boldsymbol{\beta}_\ell,\sigma_\ell)d\boldsymbol{\beta}_\ell d\sigma_\ell.\\ &= c_\ell m_\ell^U(\boldsymbol{\beta}_\ell,\sigma_\ell).\end{aligned}$$

Therefore, returning back to Equation (3), we have that

$$B_{10} = \frac{\int\int\int f(\boldsymbol{y}|\boldsymbol{\beta}_1,\sigma_1,M_1)f(\boldsymbol{y}^*|\boldsymbol{\beta}_1,\sigma_1,\delta,M_1)c_1\pi_1^U(\boldsymbol{\beta}_1,\sigma_1)\frac{c_0 m_0^U(\boldsymbol{y}^*|\delta)}{c_1 m_1^U(\boldsymbol{y}^*|\delta)}d\boldsymbol{y}^* d\boldsymbol{\beta}_1 d\sigma_1}{\int\int f(\boldsymbol{y}|\boldsymbol{\beta}_0,\sigma_0,M_0)c_0\pi_0^U(\boldsymbol{\beta}_0,\sigma_0)d\boldsymbol{\beta}_0 d\sigma_0}. \qquad (4)$$

As is obvious from Equation (4), normalizing constants c_0 and c_1 are cancelled out; thus, there are no issues of indeterminacy of the Bayes factor.

Under the above setup, Fouskakis and Ntzoufras (2020) proved that PEP priors (or EPPs for $\delta = 1$) for comparing model M_0 to M_1 are given by

$$\left\{ \pi_0^{PEP}(\boldsymbol{\beta}_0, \sigma_0) = \pi_0^{N}(\boldsymbol{\beta}_0, \sigma_0), \pi_1^{PEP}(\boldsymbol{\beta}_1, \sigma_1) \right\},$$

with

$$\begin{aligned} \pi_1^{PEP}(\boldsymbol{\beta}_1, \sigma_1) &= \pi_1^{PEP}(\boldsymbol{\beta}_0, \sigma_1) \int_0^1 \pi_1^{PEP}\left(\boldsymbol{\beta}_{e_1}, t | \sigma_1, \boldsymbol{\beta}_0\right) dt \\ &\propto \sigma_1^{-(d_0+1)} \int_0^1 f_N\left(\boldsymbol{\beta}_{e_1}; \mathbf{0}, \frac{\delta\sigma_1^2}{t}\Sigma_{e_1}\right) f_B\left(t; \frac{n^*+d_0-k_1}{2}, \frac{n^*+d_1-d_0-k_1}{2}\right) dt, \end{aligned} \tag{5}$$

where $\Sigma_{e_1}^{-1} = \mathrm{X}_{e_1}^{*T}(\mathrm{I}_{n^*} - \mathrm{P}_0^*)\mathrm{X}_{e_1}^*$. In the above expression,

$$\pi_1^{PEP}\left(\boldsymbol{\beta}_{e_1}, t | \sigma_1, \boldsymbol{\beta}_0\right) = \pi_1^{PEP}\left(\boldsymbol{\beta}_{e_1} | t, \sigma_1, \boldsymbol{\beta}_0\right) \pi_1^{PEP}(t)$$

is proper and $\pi_1^{PEP}(\boldsymbol{\beta}_0, \sigma_1) \propto \sigma_1^{-(d_0+1)}$; i.e., the reference prior for the baseline model M_0.

Using Equation (5), if $g = \frac{\delta}{t}$, following the results of Fouskakis and Ntzoufras (2020), the PEP prior under model M_1 can be represented as normal scale mixture distribution:

$$\pi_1^{PEP}(\boldsymbol{\beta}_{e_1}, \boldsymbol{\beta}_0, \sigma_1) = \sigma_1^{-(d_0+1)} \int_0^{+\infty} f_{N_{k_1-k_0}}\left(\boldsymbol{\beta}_{e_1}; \mathbf{0}, g\sigma_1^2\Sigma_{e_1}\right) \pi_1(g) dg, \tag{6}$$

where $f_{N_d}(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma)$ denotes the density of d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, evaluated at $\boldsymbol{y}$, and $\pi_1(g)$ denotes the prior distribution of parameter g under model M_1. The hyper–prior $\pi_1(g)$ for g is given by

$$f(g; a, b, \delta) = \frac{\left(\frac{g-\delta}{\delta}\right)^{b-1}\left(1 + \frac{g-\delta}{\delta}\right)^{-a-b}}{\delta B(a, b)}, \ g \geq \delta, \tag{7}$$

with $a = \frac{n^*+d_0-k_1}{2}$ and $b = \frac{n^*+d_1-d_0-k_1}{2}$.

The above form of the PEP prior offers great advantages; given g, posterior distributions and marginal likelihoods can be easily derived in closed-form expressions. However, even without conditioning on g, those distributions can be written in terms of Appell hypergeometric functions, and therefore again be derived. Detailed formulas that are also used in the next two sections can be found in Fouskakis and Ntzoufras (2020). In the following, a parameter of importance is also the shrinkage w that, under the PEP prior, is equal to $\frac{g}{g+1} = \frac{\delta}{\delta+t}$; its posterior mean, used in the following sections, was analytically derived in Fouskakis and Ntzoufras (2020).

Bayesian model averaging (BMA) is a standard Bayesian approach that combines predictions or estimates of a quantity of interest over different models that are weighted according to their posterior model probabilities. BMA efficiently incorporates model uncertainty that naturally exists in all statistical problems. By handling model uncertainty via BMA, we obtain posterior distributions and posterior credible intervals that are more realistic, and we avoid single model inference that can be severely biased or overconfident in terms of uncertainty. Moreover, several authors empirically showed BMA results lead to better predictive procedures (see for example Fernandez et al. 2001a; Ley and Steel 2009 2012; Raftery et al. 1997). For more details on BMA, also see Hoeting et al. (1999), Steel (2016, 2019) for BMA implementation in economics.

In this work, we derive Bayesian model averaging estimates under the PEP prior. Furthermore, we present different computational solutions for deriving the Bayes factor and performing the Bayesian model average, which are applied in a simulated and a real-life dataset.

2. BMA Point Prediction Estimates

Let us consider a set of models $M_\ell \in \mathcal{M}$, with design matrices X_ℓ, where the covariates of the design matrix X_0 (of the reference model M_0) are included in all models. Thus. we assume as before that $\mathrm{X}_\ell = [\mathrm{X}_0|\mathrm{X}_{e_\ell}]$ and $\boldsymbol{\beta}_\ell = \left(\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_{e_\ell}^T\right)^T$. We are interested in quantifying uncertainty about the inclusion or exclusion of additional columns/covariates X_{e_ℓ} of model M_ℓ. Under this setup, for any model $M_\ell \in \mathcal{M}$, given a set of new values of explanatory variables $\mathrm{X}_\ell^{new} = \left[\mathrm{X}_0^{new}|\mathrm{X}_{e_\ell}^{new}\right]$, we are interested in estimating the corresponding posterior predictive distribution $f(\boldsymbol{y}^{new}|\boldsymbol{y}, \mathrm{X}_\ell^{new}, M_\ell)$. Following Liang et al. (2008), for each model M_ℓ, we consider the BMA prediction–point estimator, which is the optimal under squared error loss and is given by

$$\begin{aligned}
\widehat{\boldsymbol{y}}^{new}_{BMA} &= E(\boldsymbol{y}^{new}|\boldsymbol{y}, \mathrm{X}^{new}) = \sum_{M_\ell \in \mathcal{M}} E(\boldsymbol{y}^{new}|\boldsymbol{y}, \mathrm{X}_\ell^{new}, M_\ell)\pi(M_\ell|\boldsymbol{y}) \\
&= \sum_{M_\ell \in \mathcal{M}} \mathrm{X}_\ell^{new} E(\boldsymbol{\beta}_\ell|\boldsymbol{y}, M_\ell)\pi(M_\ell|\boldsymbol{y}) \\
&= \sum_{M_\ell \in \mathcal{M}} \left\{\mathrm{X}_0^{new} E(\boldsymbol{\beta}_0|\boldsymbol{y}, M_\ell) + \mathrm{X}_{e_\ell}^{new} E(\boldsymbol{\beta}_{e_\ell}|\boldsymbol{y}, M_\ell)\right\}\pi(M_\ell|\boldsymbol{y}),
\end{aligned}$$

where X^{new} is the given set of new values of all explanatory variables. Detailed derivations of posterior means $E(\boldsymbol{\beta}_0|\boldsymbol{y}, M_\ell)$ and $E(\boldsymbol{\beta}_{e_\ell}|\boldsymbol{y}, M_\ell)$ are provided in Section 4.1 of Fouskakis and Ntzoufras (2020). If we now further assume that $\mathrm{X}_0^T \mathrm{X}_\ell = 0$, then the posterior means of the coefficients are considerably simplified to

$$E(\boldsymbol{\beta}_0|\boldsymbol{y}, M_\ell) = \widehat{\boldsymbol{\beta}}_0 \text{ and } E(\boldsymbol{\beta}_{e_\ell}|\boldsymbol{y}, M_\ell) = E\left(\tfrac{g}{g+1}\middle|\boldsymbol{y}, M_\ell\right)\widehat{\boldsymbol{\beta}}_{e_\ell},$$

where $\widehat{\boldsymbol{\beta}}_0 = (\mathrm{X}_0^T\mathrm{X}_0)^{-1}\mathrm{X}_0^T\boldsymbol{y}$ and $\widehat{\boldsymbol{\beta}}_{e_\ell} = (\mathrm{X}_{e_\ell}^T\mathrm{X}_{e_\ell})^{-1}\mathrm{X}_{e_\ell}^T\boldsymbol{y}$. Hence, assuming in a similar fashion that $(\mathrm{X}_0^{new})^T\mathrm{X}_\ell^{new} = 0$, the posterior predictive mean, under model M_ℓ, is now reduced to

$$\widehat{\boldsymbol{y}}^{new}_{M_\ell} = E(\boldsymbol{y}^{new}|\boldsymbol{y}, \mathrm{X}_\ell^{new}, M_\ell) = \mathrm{X}_0^{new}\widehat{\boldsymbol{\beta}}_0 + \mathrm{X}_{e_\ell}^{new} E\left(\tfrac{g}{g+1}\middle|\boldsymbol{y}, M_\ell\right)\widehat{\boldsymbol{\beta}}_{e_\ell} \tag{8}$$

and the corresponding BMA point prediction estimate is now given by

$$\widehat{\boldsymbol{y}}^{new}_{BMA} = \mathrm{X}_0^{new}\widehat{\boldsymbol{\beta}}_0 + \sum_{M_\ell \in \mathcal{M}} \mathrm{X}_{e_\ell}^{new} E\left(\tfrac{g}{g+1}\middle|\boldsymbol{y}, M_\ell\right)\widehat{\boldsymbol{\beta}}_{e_\ell}\pi(M_\ell|\boldsymbol{y}). \tag{9}$$

The expected value of the posterior distribution of $w = g/(g+1)$ in Equation (9) is given by

$$E(w|\boldsymbol{y}, M_\ell) = \frac{\delta}{\delta+1} \times \frac{F_1\left(b,\, \frac{n+d_0-k_0}{2},\, -\frac{n+d_0-k_\ell}{2}+1,\, \frac{k_\ell-k_0}{2}+a+b;\, \frac{1}{1+\delta R_{\ell 0}},\, \frac{1}{\delta+1}\right)}{F_1\left(b,\, \frac{n+d_0-k_0}{2},\, -\frac{n+d_0-k_\ell}{2},\, \frac{k_\ell-k_0}{2}+a+b;\, \frac{1}{1+\delta R_{\ell 0}},\, \frac{1}{\delta+1}\right)}, \tag{10}$$

while posterior model probabilities $\pi(M_\ell|\boldsymbol{y}) \propto f(\boldsymbol{y}|M_\ell)\pi(M_\ell)$ in Equation (9) can be calculated using the closed–form expression of the marginal likelihood

$$\begin{aligned}
f(\boldsymbol{y}|M_\ell) &= f(\boldsymbol{y}|M_0) \times \frac{B\left(\frac{k_\ell-k_0}{2}+a,\, b\right)}{B\left(a,\, b\right)} \times (\delta+1)^{\frac{n+d_0-k_\ell}{2}}\left(1+\delta R_{\ell 0}\right)^{-\frac{n+d_0-k_0}{2}} \\
&\quad \times F_1\left(b,\, \frac{n+d_0-k_0}{2},\, -\frac{n+d_0-k_\ell}{2},\, \frac{k_\ell-k_0}{2}+a+b;\, \frac{1}{1+\delta R_{\ell 0}},\, \frac{1}{\delta+1}\right),
\end{aligned} \tag{11}$$

where $F_1(a', b_1', b_2', c'; x, y)$ is the hypergeometric function of two variables or Appell hypergeometric function, $a = \frac{n^*+d_0-k_\ell}{2}$, $b = \frac{n^*+d_1-d_0-k_\ell}{2}$ and $R_{\ell 0} = \frac{1-R_\ell^2}{1-R_0^2}$, with R_ℓ^2 and R_0^2 being the coefficients of determination of models M_ℓ and M_0, respectively. For the detailed derivation of Equations (10) and (11), see Fouskakis and Ntzoufras (2020, Sections 4.2 and 5.2, respectively).

3. Computation and Evaluation of BMA Prediction

For small–to–moderate sample spaces, it is straightforward to implement full enumeration of the model space and calculate the marginal likelihoods by using Equation (11) and the Appell hypergeometric function. This function is available, for example, in the R package tolerance; see Function F1. Alternatively, it can be calculated using standard methods for the computation of one–dimensional integrals.

In our applications of Sections 4 and 5, we did not observe any "precision" problems with the calculation of the Appell hypergeometric function. Nevertheless, in the case of overflow issues in the implementation of this approach (e.g., for large n), using a simple Laplace approximation (preferably on $\log(g)$) can be an effective and relatively precise alternative.

Another way to estimate each marginal likelihood is by using simple Monte Carlo schemes. For example, we can simply generate values of g from its posterior distribution (see Fouskakis and Ntzoufras 2020), or some good approximations of the posterior distribution of g, and then calculate the final marginal likelihood as the mean of the conditional marginal likelihoods, which can be easily derived in closed-form expressions over all sampled values of g.

In the case of large model space, when full enumeration is not computationally feasible, we can implement a Markov chain Monte Carlo (MCMC) algorithm that could be considered as a simple extension of MC^3 (Madigan and York 1995) with two steps, since all needed quantities are analytically available given g. In the first step, we update the model indicator by using a simple Metropolis step where acceptance probability is a simple function of the posterior model odds; in the second step, we generate g from the marginal posterior distribution of g.

Using any of the above computational approaches, and assuming that M_0 is the "null" model, we can obtain a BMA prediction estimate by using Equation (9) and implementing the following procedure:

1. For every model $M_\ell \in \mathcal{M}$:
   (a) Obtain the least-squares estimates $\widehat{\boldsymbol{\beta}}_\ell = (\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_{e_\ell})$ of the regression coefficients using centered covariates.
   (b) Calculate the posterior expected value of w from Equation (10).
   (c) Calculate $\widehat{\boldsymbol{y}}^{new}_{M_\ell}$ from Equation (8).
2. Implement Equation (9) to calculate $\widehat{\boldsymbol{y}}^{new}_{BMA}$, as the weighted average of $\widehat{\boldsymbol{y}}^{new}_{M_\ell}$ over all models $M_\ell \in \mathcal{M}$ using posterior model probabilities as weights.

Evaluation of prediction accuracy or goodness of fit can be achieved by the square root of the mean of the squares (RMSE) between the observed and the fitted/predicted values or the corresponding mean absolute deviation (MAD) given by

$$RMSE(M_\ell) \quad = \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}^{new}_{i,M_\ell}\right)^2} \text{ and } MAD(M_\ell) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}^{new}_{i,M_\ell}\right|,$$

respectively. In the illustrated example of Section 4, we present the RMSE and the MAD for the maximum–a–posteriori (MAP) model and the median–probability (MP) model, as well as for the full BMA (with all models), the BMA for the 10 highest a–posteriori models, and for the models with posterior odds versus the MAP model of at least 1/3.

In addition, in Section 5, we compare the predictive performance of PEP with that of other mixtures of g-priors. For the application considered in Section 5, we randomly partition the sample B times in modelling and validation subsamples of a fixed size. Then, we calculate the BMA–log predictive score (BMA–LPS); see, for example, Fernandez et al. (2001b). Specifically, for each partition, we denote by $\mathbb{M} = \{\boldsymbol{y}^{\mathbb{M}}, \mathrm{X}^{\mathbb{M}}\}$ the modelling subsample of size $n^{\mathbb{M}}$, and by $\mathbb{V} = \{\boldsymbol{y}^{\mathbb{V}}, \mathrm{X}^{\mathbb{V}}\}$ the validation subsample of size $n^{\mathbb{V}}$, where $n = n^{\mathbb{M}} + n^{\mathbb{V}}$. The BMA–LPS then is given by

$$\text{BMA–LPS} = -\frac{1}{n^{\mathbb{V}}} \sum_{i=1}^{n^{\mathbb{V}}} \log f(y_i^{\mathbb{V}} | \boldsymbol{y}^{\mathbb{M}}, \mathrm{X}^{\mathbb{V}}), \tag{12}$$

where

$$\begin{aligned} f(y_i^{\mathbb{V}} | \boldsymbol{y}^{\mathbb{M}}, \mathrm{X}^{\mathbb{V}}) &= \sum_{M_\ell \in \mathcal{M}} f(y_i^{\mathbb{V}} | \boldsymbol{y}^{\mathbb{M}}, \mathrm{X}^{\mathbb{V}}, M_\ell) \pi(M_\ell | \boldsymbol{y}^{\mathbb{M}}) \\ &= \sum_{M_\ell \in \mathcal{M}} \frac{f(y_i^{\mathbb{V}}, \boldsymbol{y}^{\mathbb{M}} | M_\ell)}{f(\boldsymbol{y}^{\mathbb{M}} | M_\ell)} \pi(M_\ell | \boldsymbol{y}^{\mathbb{M}}) \\ &= \frac{\sum_{M_\ell \in \mathcal{M}} f(y_i^{\mathbb{V}}, \boldsymbol{y}^{\mathbb{M}} | M_\ell) \pi(M_\ell)}{\sum_{M_\ell \in \mathcal{M}} f(\boldsymbol{y}^{\mathbb{M}} | M_\ell) \pi(M_\ell)}, \end{aligned} \tag{13}$$

with $\pi(M_\ell | \boldsymbol{y}^{\mathbb{M}})$ and $\pi(M_\ell)$ denoting the posterior (given the data in the modelling subsample) and prior probabilities of model M_ℓ, respectively. Smaller values of BMA-LPS indicate better performance.

Concerning the computation of Equation (12), it is obvious that, when full enumeration is feasible, we can calculate the BMA–LPS by using Equation (13) for all models under consideration; for the evaluation of marginal likelihoods in the numerator and denominator, we use Equation (11). In the case where the number of predictors (and thus the number of induced models) does not allow full enumeration, there are three direct computational approaches that we may use:

1. Model search using MC^3 algorithm (Madigan and York 1995): this approach can be used since the marginal likelihood is readily available, but it is not very efficient, especially for large model spaces, since both the numerator and the denominator in Equation (13) are greatly affected by the number of visited models, and hence by the number of iterations of the algorithm.
2. g–conditional MC^3 algorithm: hyper–parameter g is generated by its marginal posterior distribution; then, we use the conditional on g marginal likelihood to move through the model space; this is the approach used by Ley and Steel (2012). Under this setup, $f(y_i^{\mathbb{V}} | \boldsymbol{y}^{\mathbb{M}}, \mathrm{X}^{\mathbb{V}})$ is estimated by
$$\widehat{f}(y_i^{\mathbb{V}} | \boldsymbol{y}^{\mathbb{M}}, \mathrm{X}^{\mathbb{V}}) = \frac{1}{T} \sum_{t=1}^{T} \frac{f(y_i^{\mathbb{V}}, \boldsymbol{y}^{\mathbb{M}} | g^{(t)}, M^{(t)})}{f(\boldsymbol{y}^{\mathbb{M}} | g^{(t)}, M^{(t)})},$$
where T is the total number of MCMC iterations, $g^{(t)}$ is the generated value of g at iteration t, and $M^{(t)}$ is the visited model at iteration t.
3. Fully Bayesian variable–selection MCMC: density $f(y_i^{\mathbb{V}} | \boldsymbol{y}^{\mathbb{M}}, \mathrm{X}^{\mathbb{V}})$ is estimated by the MCMC average of the sampling–density function of each visited model $M^{(t)}$, evaluated at $y_i^{\mathbb{V}}$, for each generated set of the model parameters. This is the approach we used in Section 5. More specifically, we implemented the Gibbs variable–selection approach of Dellaportas et al. (2002).

4. Simulation Study

In this section, we illustrate the proposed methodology in simulated data. We compare the performance of the PEP prior and the intrinsic prior, the latest as presented in Fouskakis and Ntzoufras (2020) and in Womack et al. (2014), by calculating the RMSE and the MAD under different BMA setups, as explained in Section 3.

We considered 100 datasets of $n = 50$ observations with $k = 15$ covariates. We ran two different scenarios. Under Scenario 1 (independence), all covariates were generated from multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix I_{15}, while the response is generated from

$$Y_i \sim N\big(4 + 2X_{i,1} - X_{i,5} + 1.5X_{i,7} + X_{i,11} + 0.5X_{i,13}\,,\; 2.5^2\big), \tag{14}$$

for $i = 1, \ldots, 50$. Under Scenario 2 (collinearity), the response was again generated from Equation (14), but this time, only the first 10 covariates were generated from multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix I_{10}, while

$$X_{ij} \sim N\big(0.3X_{i,1} + 0.5X_{i,2} + 0.7X_{i,3} + 0.9X_{i,4} + 1.1X_{i,5}\,,\; 1\big), \tag{15}$$

for $j = 11, \ldots, 15$; $i = 1, \ldots, 50$.

With $k = 15$ covariates, there are only 32,768 models to compare; we were able to conduct full enumeration of the model space, obviating the need for a model–search algorithm in this example.

Regarding the prior on the model space, we considered the uniform prior on the model space (uni), as well as the uniform prior on model size (BB), as a special case of the beta–binomial prior (Scott and Berger 2010); thus, in what follows, we compare the following methods: PEP–BB, PEP–Uni, I–BB and I–Uni; the first two names denote the PEP prior under the uniform prior on the model space and the uniform prior on model size, respectively; the last two names denote the intrinsic prior under the uniform prior on the model space and the uniform prior on model size, respectively.

Figure 1 presents the RMSE and the MAD under Scenario 1. The uniform prior on the model space (PEP–Uni and I–Uni) supported MAP models with better predictive abilities. Similar was the picture when we implemented BMA with any of the three approaches. PEP–BB behaved slightly worse than the rest of the methods, suggesting that the BB prior is possibly undesirable for PEP, since it over–shrank effects to zero.

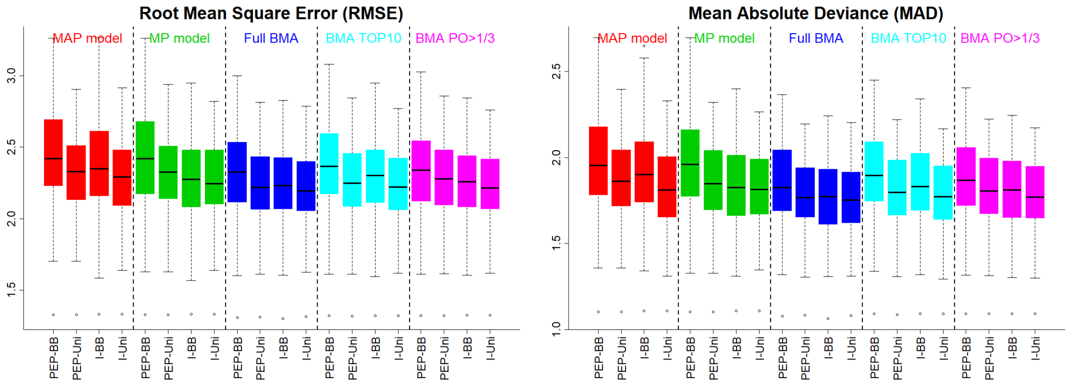

Figure 1. Simulation scenario 1. Predictive measures for maximum–a–posteriori (MAP) and median–probability (MP) models and Bayesian model averaging (BMA) using all models (full) and highest a–posteriori models [best 10 models and models with posterior odds (PO) versus the MAP model of at least $> 1/3$].

In Figure 2, we present the RMSE and the MAD under Scenario 2. The pattern was the same as that under the independence case scenario.

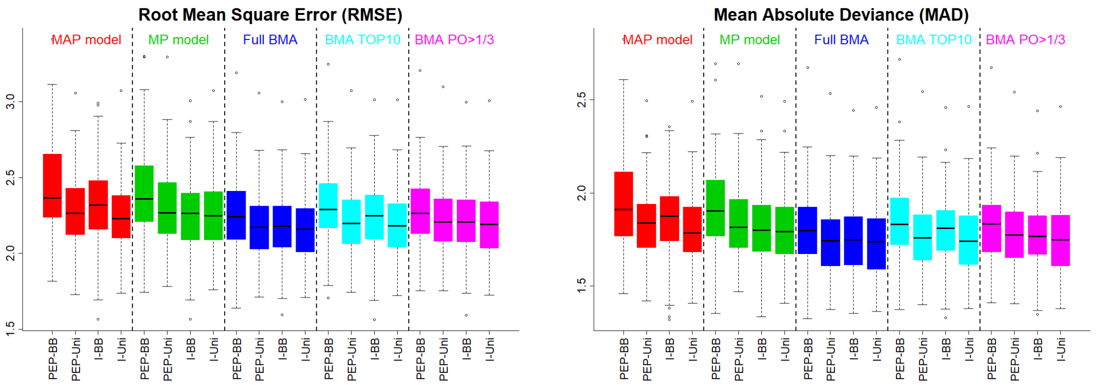

Figure 2. Simulation scenario 2. Predictive measures for maximum–a–posteriori (MAP) and median–probability (MP) models and Bayesian model averaging (BMA) using all models (full) and highest a–posteriori models [best 10 models and models with posterior odds (PO) versus the MAP model of at least $> 1/3$].

5. FLS Dataset: Cross-Country Growth GDP Study

In this section, we consider the dataset of Fernandez et al. (2001b) (also known as the FLS dataset) that contains $k = 41$ potential regressors for modelling average per capita growth over the period of 1960–1992 for a sample of $n = 72$ countries. More details on the dataset can be found in Fernandez et al. (2001b).

Emphasis was given to the posterior mean model size, to the posterior distributions of g and w, and to the comparison of the predictive performance of PEP with that of other mixtures of g–priors using the BMA–LPS as presented in Section 3. To calculate the BMA–LPS, we randomly partitioned the sample $B = 50$ times in modelling and validation subsamples of fixed sizes $n^{\mathbb{M}} = 62$ and $n^{\mathbb{V}} = 10$, respectively, as in Ley and Steel (2012).

Regarding the prior on the model space, we considered the uniform prior on model space (uni), the uniform prior on model size (BB), and the beta–binomial prior with elicitation (BBE) using the recommended value of $m = 7$, as in Ley and Steel (2009); Ley and Steel (2012).

Results under the PEP prior were compared to the ones obtained under (a) the hyper–g/n prior, with the recommended value of $a_h = 3$, as in Liang et al. (2008); and (b) the benchmark prior, with the recommended value of $c_b = 0.01$, as in Ley and Steel (2012). Further comparisons with other mixtures of g–priors could be made using the results of Section 8.1 in Ley and Steel (2012). Thus, in what follows, we make 9 comparisons in total; we use labels PEP–uni, PEP–BB, PEP–BBE, Hyper–g/n–uni, Hyper–g/n–BB, Hyper–g/n–BBE, Benchmark–uni, Benchmark–BB, and Benchmark–BBE to denote the PEP, the hyper–g/n, and the benchmark priors (under the recommended values of their hyper–parameters), respectively, combined with the different priors on the model space, i.e., the uniform on model space, the uniform on model size, and the beta–binomial with elicitation ($m = 7$), respectively.

A fully Bayesian variable selection MCMC algorithm was used, as described in Section 3; we used MCMC chains of 10,000 length after a burn-in of 1,000, which was found to be sufficient for the convergence of the quantities of interest here.

In Figure 3 we present box plots of BMA–LPS values over the 50 validation subsamples. Hyper–g/n and benchmark priors seemed to perform slightly better, but in general, we could infer that no noticeable differences were observed regarding the predictive performance of each combination of priors on the model parameters and model space.

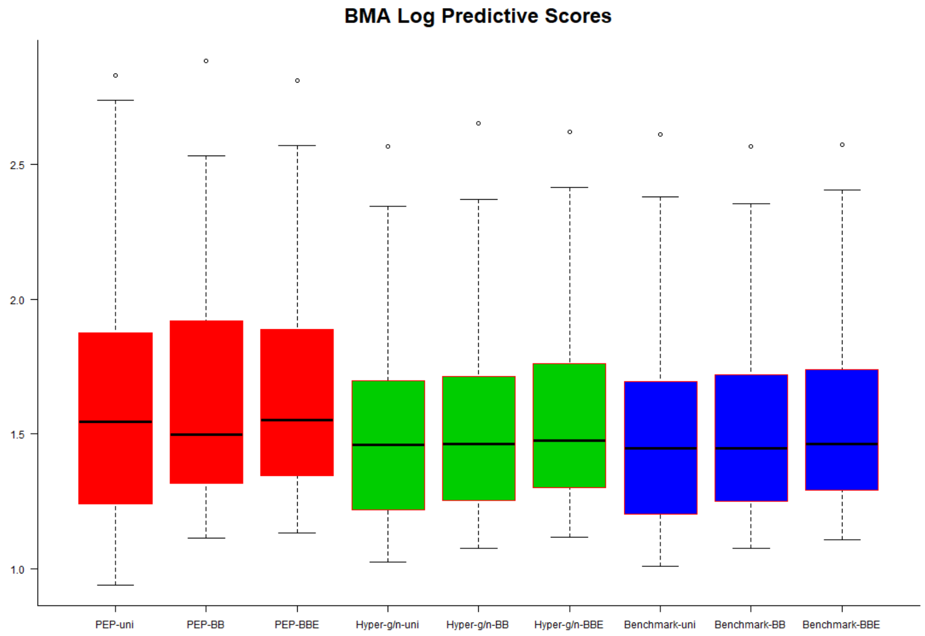

Figure 3. FLS dataset: box plots of BMA log predictive scores over 50 prediction subsamples.

In Figure 4, we present box plots of the posterior mean model size of visited models over the 50 modelling subsamples. When all three priors (PEP, hyper–g/n and benchmark) were combined with the beta–binomial prior (with and without elicitation) on model space we end up visiting models, with considerably lower, on average, size. The pattern is the same under all three priors used for the model parameters; the size, on average, of the visiting models is higher under the uniform prior on model space, followed by the beta–binomial without elicitation and the beta–binomial with elicitation. Regarding the sampling variability of the posterior mean model size, this is higher (for all three priors used on model parameters) when the beta–binomial without elicitation is used, followed by the beta–binomial with elicitation and the uniform prior on model space. The hyper–g/n and benchmark priors produced results that almost coincide when combined with the same prior on the model space. On the other hand, the PEP prior, seems to result in an approach that is more parsimonious, in contrast to the approaches resulting under the hyper–g/n and the benchmark priors when comparisons are made under the same prior on model space; the differences are sharper when the beta–binomial (with and without elicitation) prior on model space is used.

Box plots of the posterior medians of g (on a log scale) over the 50 modelling subsamples are provided in Figure 5. For all three priors used on model parameters, posterior medians of g are slightly smaller under the uniform prior on model space, followed by the beta–binomial prior without elicitation and the beta–binomial prior with elicitation. Additionally, posterior medians of g are smallest for the hyper–g/n and benchmark priors and largest under the PEP prior using any of the three priors on model space. Furthermore, we observe that the sampling variability of posterior medians across the 50 subsamples under the PEP prior was very small.

This behavior was expected, since the PEP prior induced a lower bound on g that was equal to n. In addition, in Figure 6, we present box plots of the posterior medians of the shrinkage factor w over the 50 modelling subsamples. Findings, as expected, were similar as the ones from Figure 5; a–posteriori, the median of the global shrinkage factor w under the PEP prior was close to the value of 1, implying that the induced method was more parsimonious across models, and the prior was generally noninformative (and less informative compared to the hyper–g/n and benchmark priors) within each model, since most of the information was taken from the data.

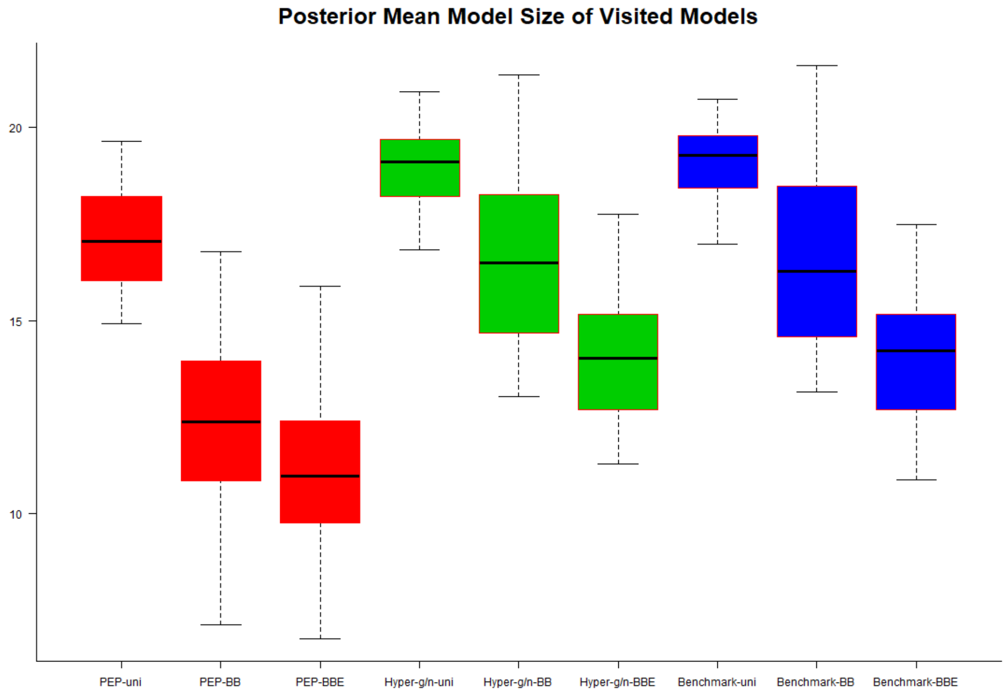

Figure 4. FLS dataset: box plots of posterior mean model size of visited models over 50 prediction subsamples.

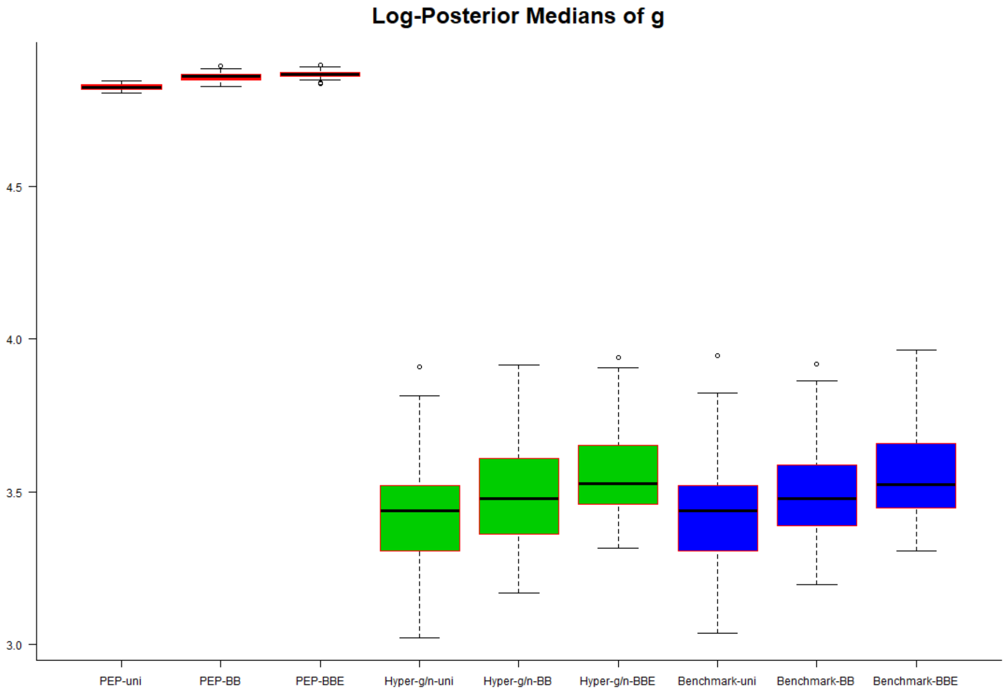

Figure 5. FLS dataset: box plots of log posterior medians of g over 50 prediction subsamples.

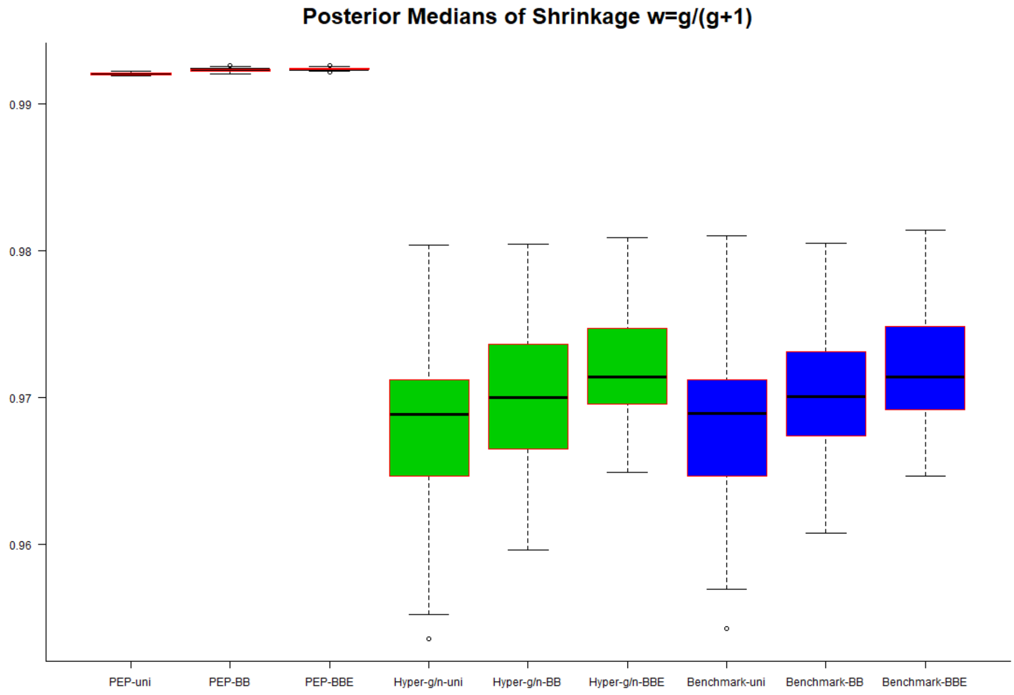

Figure 6. FLS dataset: box plots of posterior medians of the shrinkage factor w over 50 prediction subsamples.

Following the comment of a referee stating that the posterior values of g under the PEP prior are concentrated at the lower bound (which is equal to n), thus implying that the PEP prior degenerates to a fixed prior choice of g, we further present the histograms of the posterior medians and the posterior standard deviations of g for the FLS data; see Figures 7 and 8, respectively. From these histograms, it is obvious that the posterior medians were not concentrated at the left bound.

The range of posterior medians of g (across the 50 modelling subsamples of size $n^{\mathbb{M}} = 62$) was actually from 122 to 128 for PEP–Uni and slightly higher for PEP–BB and PEP–BBE; see Figure 7. Moreover, posterior standard deviations were in the range of 17.5–21 for PEP–Uni and slightly higher for PEP–BB and PEP–BBE; see Figure 8. Clearly, PEP priors were not concentrated at the lower bound (which was equal to $n^{\mathbb{M}} = 62$), and standard deviations were large enough to allow for posterior uncertainty on g. On the other hand, both hyper–g/n and benchmark priors supported posterior medians of g in the range of 20–55 that, in some subsamples, was much lower than that of the sample size of $n^{\mathbb{M}} = 62$. This raised the question of whether all model parameters are over–shrunk toward the prior mean for specific subsamples. Moreover, posterior standard deviations under both hyper–g/n and benchmark priors for g fell in the range of 10–30 under the Uni and BB priors on model space, and even higher under the BB prior (10–80 and 10–90 for hyper–g/n and benchmark priors respectively) and under the BBE prior (15–50 and 10–110 for hyper–g/n and benchmark priors respectively). We raise two points for discussion here. First, the variability of the posterior distribution of g across the 50 modelling subsamples was high, although all datasets were subsamples from the same larger dataset. Second, for some modelling subsamples, standard deviation was very high (compared to the subsample size) that, in combination with the low posterior medians, may result in a “waste” of valuable posterior probability in informative prior choices (within each model), and to the inflation of the posterior probability of irrelevant models with low practical usefulness.

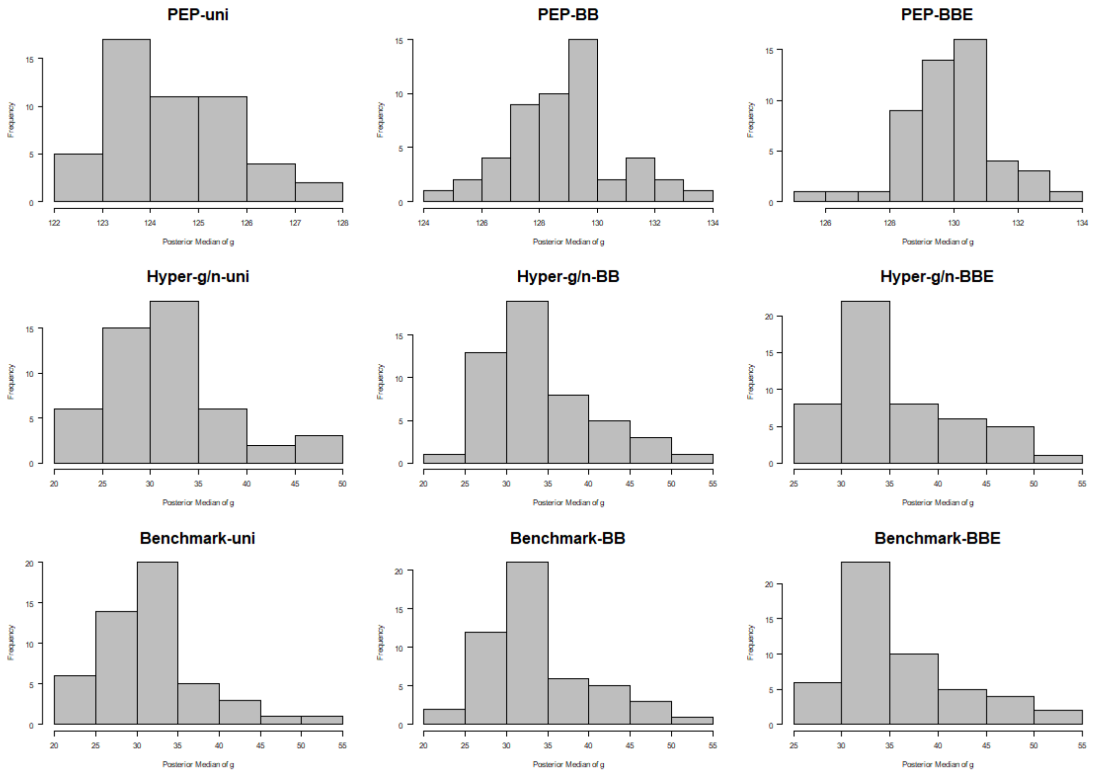

Figure 7. Histograms of posterior medians of g for all methods under comparison for the FLS data.

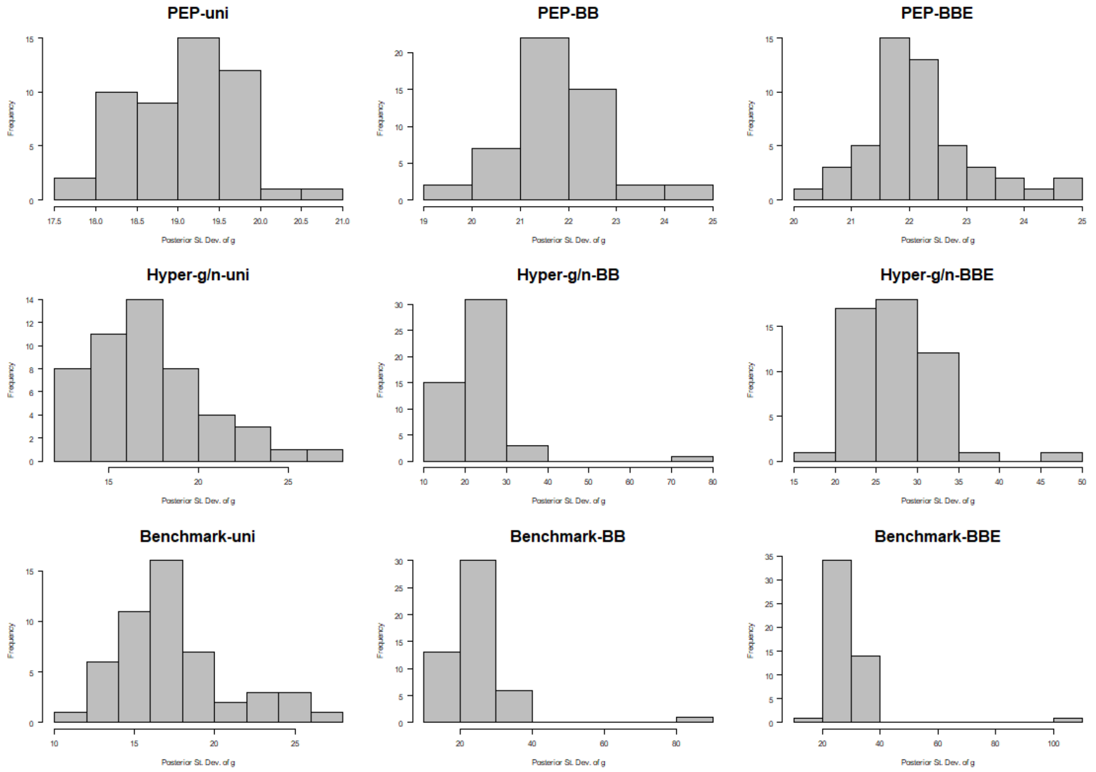

Figure 8. Histograms of posterior standard deviations of g for all methods under comparison for the FLS data.

6. Discussion

In this article, we derived a Bayesian model average (BMA) estimate of a predicted value on a variable–selection problem in normal linear models using the power–expected–posterior (PEP) prior. Furthermore, we presented computation and evaluation strategies of the prediction accuracy, and compared the performance of our method with that of similar approaches in a simulated and a real data example from economics.

An interesting point of discussion is the fact that the lower bound imposed on g seemed to drive the final results under the PEP prior. Of course, we could still specify the PEP prior with smaller values of δ in order to consider different weighting of the imaginary data. By this way, the bound (via the choice of δ) could be lower; thus, we might leave g to take values in a wider range. Nevertheless, under the recommended PEP prior specification, detailed analysis with the FLS data did demonstrate that the posterior medians of g across the 50 modelling subsamples were far away from the lower bound. Furthermore, posterior standard deviations were high enough to allow for satisfactorily posterior uncertainty for g. On the other hand, using other hyper–priors for g, like the hyper–g or the hyper–g/n, which do not restrict the range of values for g, resulted in high posterior standard deviations of g that, in combination with low posterior medians, may result in a "waste" of valuable posterior probability in informative prior choices (within each model), and to the inflation of the posterior probability of irrelevant models with low practical usefulness. This behavior has two side effects: (a) the posterior probability of the MAP model was considerably lower than the one obtained by methods with fixed prior choices for g, and (b) the posterior inclusion probabilities for the nonimportant covariates would be inflated towards 0.5; see Dellaportas et al. (2012) for an empirical illustration within the hyper–g prior setup.

Our results implied that the PEP prior was more parsimonious than its competitors. We do not claim that this property is always the best practice in variable–selection problems. The choice of parsimony or sparsity depends on the problem at hand. When we have a sparse dataset where important covariates are very few, then the PEP prior probably acts in a better way than other competitors that may spend a large portion of the posterior probability to models that are impractical in terms of dimension and sparsity.

Author Contributions: The two authors have equally contributed in the development of the ideas, the theoretical developments, the implementation of the proposed methodology, the data related illustrations and the writing of the article.

Funding: The two authors were indirectly funded by their institutions (National Technical University of Athens and Athens University of Economics). No further funding was available for this research.

Acknowledgments: We wish to thank the academic editor and the two reviewers for the comments that greatly strengthened the paper. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

Bayarri, Maria J., James O. Berger, Anabel Forte, and Gonzalo García-Donato. 2012. Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics* 40: 1550–77. [CrossRef]

Berger, James O., and Luis R. Pericchi. 2001. Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection. Institute of Mathematical Statistics Lecture Notes*. Monograph Series 38. Beachwood: IMS, pp. 135–207.

Berger, James O., and Luis R. Pericchi. 2004. Training samples in objective model selection. *Annals of Statistics* 32: 841–69. [CrossRef]

Consonni, Guido, Dimitris Fouskakis, Brunero Liseo, and Ioannis Ntzoufras. 2018. Prior distributions for objective Bayesian analysis. *Bayesian Analysis* 13: 627–79. [CrossRef]

Dellaportas, Petros, Jonathan J. Forster, and Ioannis Ntzoufras. 2002. On Bayesian model and variable selection using MCMC. *Statistics and Computing* 12: 27–36. [CrossRef]

Dellaportas, Petros, Jonathan J. Forster, and Ioannis Ntzoufras. 2012. Joint specification of model space and parameter space prior distributions. *Statistical Science* 27: 232–46. [CrossRef]

Fernandez, Carmen, Eduardo Ley, and Mark F. J. Steel. 2001a. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100: 381–427. [CrossRef]

Fernandez, Carmen, Eduardo Ley, and Mark F. J. Steel. 2001b. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16: 563–576. [CrossRef]

Fouskakis, Dimitris, and Ioannis Ntzoufras. 2020. Power-Expected-Posterior priors as mixtures of g-priors. *arXiv* arXiv:2002.05782. [CrossRef]

Fouskakis, Dimitris, Ioannis Ntzoufras, and David Draper. 2015. Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis* 10: 75–107. [CrossRef]

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. Bayesian Model Averaging: A Tutorial. *Statistical Science* 14: 382–417.

Ibrahim, Joseph G., and Ming-Hui Chen. 2000. Power prior distributions for regression models. *Statistical Science* 15: 46–60.

Kass, Robert E., and Larry Wasserman. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90: 928–34. [CrossRef]

Ley, Eduardo, and Mark F. J. Steel. 2009. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* 24: 651–74. [CrossRef]

Ley, Eduardo, and Mark F. J. Steel. 2012. Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics* 171, 251–66. [CrossRef]

Liang, Feng, Rui Paulo, German Molina, Merlise A. Clyde, and Jim O. Berger. 2008. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* 103: 410–23. [CrossRef]

Madigan, David, and Jeremy York. 1995. Bayesian graphical models for discrete data. *International Statistical Review* 63: 215–32. [CrossRef]

Pérez, José M., and James O. Berger. 2002. Expected-posterior prior distributions for model selection. *Biometrika* 89: 491–511. [CrossRef]

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92: 179–91. [CrossRef]

Scott, James G., and James O. Berger. 2010. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* 38: 2587–619. [CrossRef]

Steel, Mark F. J. 2016. Bayesian model averaging. In *Wiley StatsRef: Statistics Reference Online*. Edited by Balakrishnan, Narayanaswamy, Theodore Colton, Brian Everitt, Walter Piegorsch, Fabrizio Ruggeri and Jef Teugels. New Jersey: John Wiley & Sons, Ltd., pp. 1–7.

Steel, Mark F. J. 2019. Model Averaging and its Use in Economics. *arXiv* arxiv:1709.08221. [CrossRef]

Womack, Andrew J., Luis León-Novelo, and George Casella. 2014. Inference from intrinsic Bayes' procedures under model selection and uncertainty. *Journal of the American Statistical Association* 109: 1040–53. [CrossRef]