Bayesian Model Averaging with Non-conjugate Priors

Anastasios E. Tasiopoulos[a,b,*], Efthymios G. Tsionas[c], Nikolaos D. Vlastakis[d]

[a] *Hellenic Parliamentary Budget Office, Athens 10671, Greece*
[b] *Department of Economics, Athens University of Economics and Business, Athens, Greece*
[c] *Lancaster University Management School, Lancaster LA1 4YX, United Kingdom*
[d] *Centre for Competition Policy, Norwich Business School, University of East Anglia, Norwich NR4 7TJ, United Kingdom*

Abstract

This paper considers the case of non-conjugate prior distributions for Bayesian model averaging (BMA). Although the natural conjugate setting is the default choice for BMA, mainly for reasons of analytical tractability, it has come under considerable criticism due to its unrealistic assumptions about prior information, among others. In this study, we extend the literature by considering two special cases of the multivariate Student-t distribution. We obtain closed-form solutions using Laplace approximations and apply our techniques to a controlled numerical experiment and cross-country growth regressions. Our results show that, under fine tuning of the hyperparameters, the proposed approach has similar performance to the conjugate alternatives on synthetic datasets, whereas in real data it favors, on average, more parsimonious models than the conjugate alternatives and also exhibits superior predictive performance.

Keywords: Bayesian model averaging, Model selection, Non-conjugate priors, Laplace approximation, Growth regressions

JEL: C11, C15, O40

*Corresponding author.
Email addresses: a.tasiopoulos@parliament.gr (Anastasios E. Tasiopoulos), m.tsionas@lancaster.ac.uk (Efthymios G. Tsionas), N.Vlastakis@uea.ac.uk (Nikolaos D. Vlastakis)

1. Introduction

Model uncertainty is an issue that has attracted considerable attention from econometricians, with solutions ranging from stepwise regression to model averaging techniques. The Bayesian approach to the problem is to treat the model itself as a random variable and calculate posterior probabilities, which can then be used as weights to obtain an average (or the best) model, a technique known as Bayesian Model Averaging (BMA) (or in the case of model selection, BMS). Since the pioneering work of Leamer (1978), Raftery (1988), and Raftery et al. (1997), a voluminous literature has emerged around BMA and its applications to various problems.[1] The seminal papers by Fernandez et al. (2001a,b) are among the most influential studies in this subject. Fernandez et al. (2001a) propose benchmark priors for BMA and model selection based on a natural conjugate specification, whereas Fernandez et al. (2001b) apply the methodology to the study of cross-country growth regressions and showcase its superior performance. Sala-i Martin et al. (2004) re-examine the determinants of long-term growth using Bayesian Averaging of Classical Estimates (BACE). Still in the context of growth regressions, Ley and Steel (2007) examine the case of jointness in variable selection, Salimans (2012) extends the methodology of BMA to variable and functional form uncertainty, whereas Mirestean and Tsangarides (2016) introduce a modified BMA method, the limited - information BMA (LIBMA) dealing with issues of model uncertainty, dynamics and endogeneity. In the context of a dynamic setting, Raftery et al. (2010) introduce

[1]For a recent review of the literature, see Steel (2020). The excellent tutorial by Hoeting et al. (1999) is also worth mentioning.

Dynamic Model Averaging (DMA) and Koop and Korobilis (2012) adapt DMA and showcase its superior performance in forecasting inflation.

A problem that is common in Bayesian analysis, but becomes even more important for BMA is the selection of prior distribution. Indeed, posterior model probabilities are more sensitive to prior assumptions than is posterior analysis for a given model, see the discussion in Steel (2020, Section 3.4). Fernandez et al. (2001a) examine the effect of different choices for the prior parameters in the natural conjugate setting, with Ley and Steel (2009) and Eicher et al. (2011) expanding the envelope of choices. Ley and Steel (2012) examine hyperprior assumptions and conclude that they can have a big influence on model complexity.

Nevertheless, despite its obvious virtues in terms of analytical convenience, the decision to adopt a conjugate prior setting has itself received criticism. One important reason is that a conjugate prior embodies the assumption that prior information comes from a previous sample of the same process, which means that it does not distinguish between prior and sample information (Leamer, 1978). Moreover, the priors of slopes and scale parameters in the natural conjugate setting are not independent, which is not always a credible assumption. In addition, when one sets the degrees of freedom and the precision matrix equal to zero to obtain the limiting distribution of the normal-(inverse) gamma prior, the resulting ignorant prior is different to the usual diffuse prior and the posterior distribution has different degrees of freedom (Judge et al., 1985). Another important criticism is that, as Dawid (1987) posits, the natural conjugate prior encompasses an implicit determinism, a belief that one may accomplish arbitrarily high pre-

dictive accuracy by adding a sufficiently high number of variables to a model. Therefore, using a conjugate prior can lead to overfitting and poor predictive performance (Fang and Dawid, 2002).

In this article we address these issues and extend the BMA literature by considering cases where the prior information is assumed to be non-conjugate. Leamer (1978) and Judge et al. (1985) suggest a prior proposed by Dickey (1975), which assumes a multivariate-t prior for the slope and inverted-gamma prior for the scale parameter. Motivated by this suggestion, we consider two non-conjugate priors for the slope parameters which can be deemed as special cases of a multivariate student-t distribution. In the first case, we assume that the prior degrees of freedom tend to infinity, where the resulting prior is a non-conjugate multivariate normal distribution and in the second case, that they are equal to one, where the resulting prior is a multivariate dependent Cauchy distribution. This setting addresses all of the criticism on conjugate priors as we treat the slope and scale parameters independently and implicitly assume prior and sample information to be different. Moreover, as Fang and Dawid (2002) point out, the use of non-conjugate priors does not incorporate determinism and can avoid the problem of overfitting.

One caveat of our setting is that we cannot obtain analytical expressions namely, for the marginal likelihood of the data, the predictive density and the moments of the parameters under consideration. One way to deal with this issue is through the use of numerical methods, but these come with a heavy computational cost, which would make the proposed setting impractical. A more practical solution is to apply approximated methods, as the $MC\,(3)$

algorithm (*Markov Chain Monte Carlo Model Composition*) works faster, when the marginal likelihood can be deployed in a closed (or approximated in our case) form. We therefore adapt the method of Laplace approximation as proposed by Tierney and Kadane (1986) and Tierney et al. (1989) to a multivariate setting.[2]

We undertake a controlled simulation experiment similar to Fernandez et al. (2001a) and show that our non-conjugate priors perform as well as the conjugate alternatives (comparing favorably to them under certain measures) on synthetic data. We also apply these techniques to the cross-country growth regressions data set of Fernandez et al. (2001b) and find that, under fine tuning of the hyper parameters, the non-conjugate case results to (on average) more parsimonious models compared with the conjugate case. By estimating the predictive posteriors for the two cases under consideration we show that our non-conjugate specification outperforms the conjugate alternative in terms of predictive performance. A series of robustness checks (including different model-space priors and different model sampling procedures) verify that our setting is robust to variations of our empirical design.

The paper is organized as follows: Section 2 discusses the BMA setting, Section 3 presents the non-conjugate setting, Section 4 presents our simulation experiment, Section 5 presents empirical results on the growth dataset and Section 6 concludes the paper.

[2]We derive a multivariate version of Theorem 3, of Tierney et al. (1989) for the estimation of the moments.

2. Bayesian Model Averaging and Selection

Assume the model Space $\mathcal{M} = \{M_1, \ldots, M_{2^K}\}$, where K is the number of covariates. Our task is to choose the best subset among these K covariates from the following set, $\mathcal{X} = \{X_1, \ldots, X_K\}$. In situations where the highest-posterior model has very low posterior probability and the mass is spread over many competing models, it is natural and, under standard predictive loss functions, optimal to use Bayesian model averaging instead of conditioning on a single model, even if that model has the highest posterior probability.[3] In BMA methodology the averaging over models arises in a natural way since we use as weights the probability of each model. Assume Δ is a quantity of interest and $p(\Delta|\mathbf{y},M_j)$ its posterior probability, then by averaging across all potential models

$$p(\Delta|\mathbf{y}) = \sum_{j=1}^{2^K} p(\Delta|\mathbf{y},M_j)\, p(M_j|\mathbf{y}) \tag{1}$$

where $p(M_j|\mathbf{y})$ is the model's j posterior probability,

$$p(M_j|\mathbf{y}) = \frac{p(\mathbf{y}|M_j)\, p(M_j)}{\sum_{j=1}^{2^K} p(\mathbf{y}|M_j)\, p(M_j)}$$

and $p(\mathbf{y}|M_j)$ is the marginal likelihood of the data, which is equal to

$$p(\mathbf{y}|M_j) = \int p(\mathbf{y}|\boldsymbol{\theta}_j,M_j)\, p(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j$$

[3] See the discussion in Fernandez et al. (2001a), p.383, first paragraph and the references therein.

$\boldsymbol{\theta}_j$ is the vector of parameters included in model j.[4] The posterior moments of Δ are

$$E\left(\Delta|\mathbf{y}\right) = \sum_{j=1}^{2^K} E\left(\Delta|\mathbf{y},M_j\right) p\left(M_j|\mathbf{y}\right) \tag{2}$$

$$Var\left(\Delta|\mathbf{y}\right) = \sum_{j=1}^{2^K} E\left(\Delta^2|\mathbf{y},M_j\right) p\left(M_j|\mathbf{y}\right) - E\left(\Delta|\mathbf{y}\right)^2 \tag{3}$$

In practice, it is impossible to construct all 2^K models, so we apply MC(3) iterations (model draws). The most popular algorithm for BMA is the Markov Chain Monte Carlo Model Composition or $MC\,(3)$ by Madigan and York (1995). This can be regarded as a Metropolis-Hastings algorithm applied to the discrete model space $\mathcal{M} = \{M_1, \ldots, M_{2^K}\}$, from which we obtain sample draws (model indices) from a total number of S simulations. The model space of the selected models $\mathcal{M}_{S_0} = \{M_1, \ldots, M_{S_0}\} \subset \mathcal{M}$ and $S_0 \leq S$. One of the advantages of this algorithm is that we do not need to consider all 2^K models but only a small portion of them, since the algorithm samples from regions with high posterior probability. For example, if the posterior probability of model M^* is higher than that of model M', then model draws of M' are not as many as those of M^*. To focus on the effect of the non-conjugate priors on the model selection process, we apply a Random Walk (or Birth - Death) Metropolis-Hastings algorithm, following Fernandez et al. (2001a,b).

The algorithm works as follows: Assume that $M^{(s-1)}$ is the current model in the chain, then choose randomly one of the K potential explanatory variables. If it is already in the model, delete it (Death step) else, add it (Birth

[4]Throughout this paper, we prefer the notation $f\left(\mathbf{y}|M_j\right)$ to $p\left(\mathbf{y}|M_j\right)$ to denote the marginal likelihood of the data.

step) and this is how the candidate model M^* is constructed. The acceptance is based on the posterior odds, so

$$\alpha\left(M^{(s-1)}, M^*\right) = \min\left(1, \frac{p\left(M^*|\mathbf{y}\right)}{p\left(M^{(s-1)}|\mathbf{y}\right)}\right)$$

If S is the number of simulations and J is the number of M_j models drawn from $\mathcal{M}$ then J/S converges to $p\left(M_j\left|y\right.\right)$. As a measure of convergence FLS (2001a, 2001b) propose the correlation between simulated and theoretical model posterior probabilities to be close to 1. The algorithm is also useful to estimate the Marginal Posterior Probability of Inclusion variable X_i : $p\left(\gamma_i = 1\left|y\right.\right)$, where γ_i is a Bernoulli variable taking the value of one when the i^{th} variable is included in model M_j and zero otherwise. It is straightforward to realize that the sum of the posterior probabilities of all models that contain this covariate is the marginal posterior probability of including this variable in a model, so

$$p\left(\gamma_i = 1\left|\mathbf{y}\right.\right) = \sum_{j=1}^{2^K} p\left(M_j\left|\mathbf{y}, X_i \in M_j\right.\right) \tag{4}$$

In the next sections we assume that the prior model probabilities are equal for all models, so the simulations are based solely on the Bayes factor.[5]

[5]By default, we assign equal prior probability for a variable to be included or not, that means $p\left(\gamma_i = 1\right) = 0.5$. In Appendix F.2, we relax this assumption by considering alternative model-space priors.

3. The Non-conjugate Setting

3.1. Posterior Inference

In this section we estimate the marginal likelihood of the linear regression model assuming two different non-conjugate priors for the slopes. In general the models we consider have the following structure:

$$\begin{aligned}
\mathbf{y} &= \alpha\boldsymbol{\iota}_n + \mathbf{Z}\boldsymbol{\beta} + \mathbf{u} && \text{(5a)}\\
\mathbf{u} &\sim N\left(0, \sigma^2\mathbf{I}_n\right) && \text{(5b)}\\
p(\alpha) &\propto 1 && \text{(5c)}\\
p(\sigma) &= 2\frac{(b_0/2)^{a_0/2}}{\Gamma(a_0/2)}\sigma^{-(a_0+1)}\exp\left(-\frac{b_0}{2\sigma^2}\right) && \text{(5d)}\\
p(\boldsymbol{\beta}) &= \frac{\Gamma\left((k+\underline{v})/2\right)}{\Gamma(\underline{v}/2)\left(\underline{v}\pi\right)^{k/2}\left|\omega^2\overline{\mathbf{A}}\right|^{1/2}\left(1+\frac{1}{\underline{v}}\boldsymbol{\beta}'\left(\omega^2\overline{\mathbf{A}}\right)^{-1}\boldsymbol{\beta}\right)^{\frac{\underline{v}+k}{2}}} && \text{(5e)}
\end{aligned}$$

The above mean that the constant term obtains a flat prior, which is a standard approach in the BMA literature (Fernandez et al., 2001a,b). This seems realistic, since there is not much information for the intercept relative to the slopes and the fact that the vector of $\mathbf{y}$ is untransformed. The regressors $\mathbf{Z}$ $(n \times k, k \leq K)$ are demeaned,[6] a standard practice in Bayesian Model Selection procedure (for instance, Fernandez et al., 2001a,b), and $\mathbf{X} = \begin{pmatrix} \boldsymbol{\iota} & \vdots & \mathbf{Z} \end{pmatrix}$ where $\boldsymbol{\iota}$ is a vector of ones. It is also reasonable to use

[6] We suppress the index j for ease of exposition. Notice that k is the number of regressors $\mathbf{Z}$, where the formal notation should be k_j and $\mathbf{Z}_j$ respectively. By "demeaned", we mean that each column of $\mathbf{Z}$ is centered by subtracting its sample mean. This is a standard normalization that separates the intercept from the regressors and does not affect model comparisons or BMA results, indicator (dummy) variables can be treated in the same way (or left uncentered), with only the intercept reparameterized.

non informative priors for parameters that are common in all models, although for σ we assume an *Inverse-Gamma*(a_0, b_0) type prior like in (5d), which becomes proportional to the flat prior of dispersion (σ^{-1}), when the hyper-parameters tend to zero. Equation (5e) is a general form of the non-conjugate priors we consider in this paper. In what follows we assume two special cases of $p(\boldsymbol{\beta})$, namely case one, when $\underline{v}$ tends to infinity, where we obtain the non-conjugate normal prior and case two, when $\underline{v}$ is equal to one, where we obtain the multivariate dependent Cauchy distribution.

In the non-conjugate case the marginal likelihood of the data (or integrated likelihood) cannot be computed analytically. We can either integrate for the slopes (regression coefficients) or the scale parameter, but not for both of them. Since no analytical solutions are available, we estimate the marginal likelihood by applying the method of Laplace Approximation, by Tierney and Kadane (1986) and Tierney et al. (1989). Specifically, we apply the following approximations:

Univariate Integral Approximation

$$\int_{\mathbb{R}} \exp\left(\lambda h_\lambda\left(x\right)\right) dx \approx \sqrt{2\pi/\lambda\left(-h''_\lambda\left(x^*\right)\right)} \exp\left(\lambda h_\lambda\left(x^*\right)\right) \tag{6}$$

Multivariate Integral Approximation

$$\int_{\mathbb{R}^k} \exp\left(\lambda h_\lambda\left(\mathbf{x}\right)\right) d\mathbf{x} \approx \left(2\pi/\lambda\right)^{k/2} \left|\mathbf{H}^*\right|^{-1/2} \exp\left(\lambda h_\lambda\left(\mathbf{x}^*\right)\right) \tag{7}$$

where $h_\lambda(x)$ or $h_\lambda(\mathbf{x})$ is the function of interest and λ is a large number. Finally, $x^* = \arg\max h_\lambda(x)$ (or $\mathbf{x}^* = \arg\max h_\lambda(\mathbf{x})$) and $h''_\lambda(x^*)$ is the second derivative at the maximum. In the multivariate case $\mathbf{H}^*$ is the negative

Hessian evaluated at the maximum.

For the approximation of the moments of vector $\boldsymbol{\beta}$ we extend Tierney et al. (1989) by deriving a multivariate version of their theorem 3. The moment generating function (hereafter MGF) can be approximated by:

$$\widehat{M}_g(s) \approx \frac{\widehat{b}_N \left|\mathbf{H}_N\right|^{-\frac{1}{2}} \exp\left(-n\widehat{h}_N\right)}{\widehat{b}_D \left|\mathbf{H}_D\right|^{-\frac{1}{2}} \exp\left(-n\widehat{h}_D\right)} \tag{8}$$

We use the full exponential form where, $b_N = b_D$ and in particular $b_N = b_D = 1$. The moments are obtained through the multivariate versions of Theorems 2a and 3 of Tierney et al. (1989):

Theorem 1 *(Multivariate Version of Theorem 2a of Tierney et al. (1989))*

$$\widehat{E}\left(g_l\left(\boldsymbol{\beta}\right)|\mathbf{y}, M_j\right) = g_l\left(\widehat{\boldsymbol{\beta}}_D\right) - \frac{1}{2}\left.\frac{\partial \ln\left|\mathbf{H}_s\right|}{\partial s}\right|_{s=0}$$

where $g_l(\boldsymbol{\beta})$ is a general function of the parameter vector, $\widehat{\mathbf{H}}_i = \left.\frac{\partial^2 h_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_i}$, $\widehat{\boldsymbol{\beta}}_i = \arg\max\left(-h_i\right)$ for $i = s, N, D$, and $h_s \equiv h_N$, $h \equiv h_D$, $h_s\left(\boldsymbol{\beta}\right) = h\left(\boldsymbol{\beta}\right) - \frac{1}{n} s g_l\left(\boldsymbol{\beta}\right)$, for $l = 1, \ldots, k$ and $nh\left(\boldsymbol{\beta}\right)$ is the negative log-posterior. Next, we proceed to the multivariate version of theorem 3 of Tierney et al. (1989):

Theorem 2 *(Multivariate Version of Theorem 3 of Tierney et al. (1989))*

$$\widehat{E}\left(g_l\left(\boldsymbol{\beta}\right)|\mathbf{y}, M_j\right) = g_l\left(\widehat{\boldsymbol{\beta}}_D\right) - \frac{1}{2n}\left(\left(\frac{\partial \widehat{g}_l}{\partial\boldsymbol{\beta}}\right)' \widehat{\mathbf{H}}^{-1}\left(\frac{\partial \widehat{\mathbf{H}}}{\partial\boldsymbol{\beta}}\right)' - vec\left(\mathbf{G}_l\right)'\right) vec\left(\widehat{\mathbf{H}}^{-1}\right)$$

where, $\frac{\partial \widehat{g}_l}{\partial\boldsymbol{\beta}} \equiv \left.\frac{\partial g_l}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$, $\widehat{\mathbf{H}} \equiv \left.\frac{\partial^2 h(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$, $\widehat{\mathbf{G}}_l = \left.\frac{\partial^2 g_l(\boldsymbol{\beta})}{\partial\beta\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$.

The derivatives that arise are based on matrix differential calculus, and

the proofs are presented in the Appendix, section Appendix C.

3.1.1. Non-Conjugate Normal Prior

Given the aforementioned criticism about the use of conjugate priors, a reasonable choice for prior of $\boldsymbol{\beta}$ should be (5e) when $\underline{v} \to \infty$, i.e. the non-conjugate normal prior:

$$p(\boldsymbol{\beta}) = \left(\frac{1}{2\pi\omega^2}\right)^{-k/2} \left|\overline{\mathbf{A}}\right|^{-1/2} \exp\left(-\frac{\boldsymbol{\beta}'\overline{\mathbf{A}}^{-1}\boldsymbol{\beta}}{2\omega^2}\right) \tag{9}$$

We can set $\overline{\mathbf{A}} = \mathbf{I}_k$ without loss of generality. The joint posterior of all parameters is $p(\alpha, \boldsymbol{\beta}, \sigma \mid \mathbf{y}, M_j) \propto \mathcal{L}(\alpha, \boldsymbol{\beta}, \sigma \mid \mathbf{y}, M_j) p(\boldsymbol{\beta}) p(\sigma) p(\alpha)$ and the resulting marginal likelihood is $f(\mathbf{y}; M_j) \propto \int_{\sigma>0} \int_{\mathbb{R}^k} \int_{\mathbb{R}} \mathcal{L}(\alpha, \boldsymbol{\beta}, \sigma \mid \mathbf{y}, M_j) p(\boldsymbol{\beta}) p(\sigma) p(\alpha) d\alpha d\boldsymbol{\beta} d\sigma$. It is obvious that we cannot derive an analytical expression for marginal likelihood with non-conjugate priors. We can analytically integrate either the location parameter or the scale parameter and then use an approximation method for the remaining one. The integration of the slopes has the advantage that we can use a univariate approximation for the scale parameter. Under these circumstances, we apply the Laplace approximation (6) to the marginal posterior of σ after the transformation $\tau = \ln \sigma$.[7] The resulting log-marginal likelihood (or log-evidence) is

$$\ln f(\mathbf{y}; M_j) \propto \ln c_{\sigma^*} - \frac{1}{2} \ln \left| \frac{\mathbf{X}'\mathbf{X}}{\sigma^{*2}} + \frac{\mathbf{I}_{k+1}}{\omega^2} \right| - \frac{1}{2} \frac{Q_{\sigma^*} + b_0}{\sigma^{*2}} \tag{10}$$

[7] This transformation improves the approximation; it is explained in Section 3.3.

where $c_{\sigma^*} = \omega^{-k}\left(-h_n^{N\prime\prime}(\sigma^*)\right)^{-1/2} g_2(\sigma^*)^{1/2}\sigma^{*-(n+a_0)}$ is a component of the normalising constant, and $g_2(\sigma) = \frac{n\omega^2+\sigma^2}{\omega^2}$, with $h_n^{N\prime\prime}(\sigma^*)$ denoting the second derivative of the transformed log-marginal posterior of σ, evaluated at its maximiser σ^* ($\tau^* = \ln\sigma^*$), $Q_{\sigma^*} = (\mathbf{y} - \overline{y}\boldsymbol{\iota}_n)'\,\mathbf{M}_{\sigma^*}\,(\mathbf{y} - \overline{y}\boldsymbol{\iota}_n)$, and $\mathbf{M}_{\sigma^*} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X} + \frac{\sigma^{*2}}{\omega^2}\mathbf{I}_{k+1}\right)^{-1}\mathbf{X}'$. The analytical derivations of the marginal likelihood are provided in Appendix Appendix B.1.

3.1.2. Non-Conjugate Dependent Cauchy Prior

The other prior we consider is the multivariate dependent Cauchy distribution or the multivariate Student's t-distribution with one degree of freedom (i.e., $\underline{v} = 1$), which simplifies equation (5e) to the following:

$$p(\boldsymbol{\beta}) = \Gamma\left((k+1)/2\right)\pi^{-(k+1)/2}\omega^{-k}\left|\overline{\mathbf{A}}\right|^{-1/2}\left(1 + \boldsymbol{\beta}'\left(\omega^2\overline{\mathbf{A}}\right)^{-1}\boldsymbol{\beta}\right)^{-\frac{1+k}{2}} \quad (11)$$

Related default Cauchy priors in regression go back at least to Zellner and Siow (1980), who use a multivariate Cauchy prior (scaled using an information-matrix-type choice) to derive posterior odds ratios, Bayes factors for exclusion restrictions and to screen variables for inclusion. Our use differs in purpose and structure: we work in a BMA setting and, in line with the motivation for non-conjugate priors, treat slope and scale parameters a priori independently. Incorporating an information-matrix-based scaling within (11) would be a natural extension, but is not pursued here. As in the previous case, we assume that $\overline{\mathbf{A}} = \mathbf{I}_k$. By marginalizing over the intercept α from the joint posterior distribution of $(\alpha, \boldsymbol{\beta}, \sigma)$, we derive the joint posterior distribution for $\boldsymbol{\beta}$ and σ. In this case, we are able to integrate out only σ, and then apply the multivariate version of the Laplace approximation for multivariate integrals to integrate over $\boldsymbol{\beta}$, as stated in equation (7). The

marginal posterior of $\boldsymbol{\beta}$ is:

$$p\left(\boldsymbol{\beta}|\mathbf{y},M_j\right)\propto\left(vs_{b_0}^2\right)^{-(v+k)/2}\left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-\frac{1+k}{2}}\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-\frac{v+k}{2}} \tag{12}$$

where $vs_{b_0}^2=\underline{\widehat{\mathbf{u}}}'\underline{\widehat{\mathbf{u}}}+b_0$, $\widehat{\boldsymbol{\beta}}=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y}$, $\underline{\widehat{\mathbf{u}}}=\left(\mathbf{y}-\overline{y}\boldsymbol{\iota}_n-\mathbf{Z}\widehat{\boldsymbol{\beta}}\right)$ and $v=n+a_0-(k+1)$. This distribution kernel is a member of a wider class of distributions known as poly-t distributions. Poly-t distribution kernels are products, ratios or products and ratios of student-t distribution kernels, see Dreze (1977) and Richard and Tompa (1980). In general the normalizing constant and the moments of the poly-t distribution are not known so we proceed to the integration of (12) through (7). There are also many other approximations one can apply for poly-t distribution, see Box and Tiao (1992), Press (1982), Guttman and Menzefricke (1983), Broemeling and Abdullah (1984), but as discussed in the latter paper, in the case of poly-t 2/0 distribution when the location parameters of the student-t distributions are "close" then the distribution is unimodal.[8]

Therefore, the marginal log-likelihood with dependent Cauchy priors is

$$\ln f\left(\mathbf{y};M_j\right)\propto\ln c_{\widetilde{\boldsymbol{\beta}}}-\frac{v+k}{2}\ln\left(vs_{b_0}^2\right)-\frac{1}{2}\ln|\mathbf{H}_n^*|+nh_n^C\left(\widetilde{\boldsymbol{\beta}}\right) \tag{13}$$

where $c_{\widetilde{\boldsymbol{\beta}}}=n^{-(k+1)/2}\left(2/\omega^2\right)^{k/2}\Gamma((k+1)/2)$ is a component of the normalising constant, $\widetilde{\boldsymbol{\beta}}=\arg\max_{\boldsymbol{\beta}\in\mathbb{R}^k}h_n^C(\boldsymbol{\beta})$, with $h_n^C(\boldsymbol{\beta})$ denoting the marginal log-posterior of $\boldsymbol{\beta}$, and $\mathbf{H}_n^*$ is the negative Hessian evaluated at $\widetilde{\boldsymbol{\beta}}$. The

[8]In cases where multiple isolated maxima exist, a possible solution would be to use a multi-modal ("multi-Laplace") approximation by applying the multivariate Laplace expansion around each mode and summing the resulting contributions; see, e.g., Chapter 4 of De Bruijn (1981).

analytical derivations of equation 13 can be found in Appendix B.2.

3.2. Predictive Inference

In this part we estimate the predictive posterior distribution of the models with non-conjugate prior to assess the performance on out of sample data. According to Fernandez et al. (see 2001a, p. 396), the purpose of the researcher is to provide a prediction for the observable rather than to reveal its structure, thus our goal is to estimate the posterior distribution of the unobserved variable y_f. Assuming that y_f is unobserved response at the time the sample was collected, then the predictive posterior is equal to

$$p\left(y_f|\mathbf{y},\mathbf{z}_f,M_j\right)=\int\limits_{\theta\in\boldsymbol{\Theta}} p\left(y_f\left|\mathbf{y},\mathbf{z}_f,M_j,\boldsymbol{\theta}\right.\right)p\left(\boldsymbol{\theta}|\boldsymbol{y},M_j\right)d\boldsymbol{\theta}$$

$$p\left(y_f|\mathbf{y},\mathbf{z}_f,M_j\right)=\int\limits_{\theta\in\boldsymbol{\Theta}} \frac{p\left(y_f\left|\mathbf{y},\mathbf{z}_f,M_j,\boldsymbol{\theta}\right.\right)\mathcal{L}\left(\boldsymbol{\theta}|\mathbf{y},M_j\right)p\left(\boldsymbol{\theta}\right)}{f\left(\mathbf{y};M_j\right)}d\boldsymbol{\theta} \qquad (14)$$

where, $p\left(y_f\left|\mathbf{y},\mathbf{z}_f,M_j,\boldsymbol{\theta}\right.\right)$ is the conditional predictive posterior of y_f, $p\left(\boldsymbol{\theta}|\mathbf{y},M_j\right)$ is the posterior distribution of the parameters vector $\boldsymbol{\theta}'=\left(\alpha,\boldsymbol{\beta},\sigma\right)$, $\boldsymbol{\Theta}=\mathbb{R}^{k+1}\times\mathbb{R}_+^*$, $\mathbf{z}_f$ is a $k\times 1$ vector of future observations of the exogenous variables. Following Fernandez et al. (2001a,b), we assume that the conditional density of a new observation is

$$p\left(y_f\left|\mathbf{y},\mathbf{z}_f,M_j,\boldsymbol{\theta}\right.\right)\equiv\left(2\pi\sigma^2\right)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\left(y_f-a-\mathbf{z}_f'\boldsymbol{\beta}\right)'\left(y_f-a-\mathbf{z}_f'\boldsymbol{\beta}\right)\right)$$

Since we obtain the conditional density of the new observation, next we are going to estimate the predictive posterior of the non-conjugate cases under consideration.

3.2.1. Non-Conjugate Normal Prior

Assuming the non-conjugate normal prior (9) and the assumptions stated in subsection 3.1.1, the predictive posterior (14) becomes,

$$p\left(y_f|\mathbf{y},\mathbf{z}_f,M_j\right) \propto \int_{\mathbb{R}_+^*} \sigma^{-(n+a_0+1)} \left|\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2}+\frac{\mathbf{I}_k}{\omega^2}+\frac{n\mathbf{z}_f\mathbf{z}_f'}{\overline{n}}\right|^{-\frac{1}{2}} \exp\left(-\frac{Q_\sigma^f+b_0}{2\sigma^2}\right) d\sigma \tag{15}$$

where, $Q_\sigma^f = \left(\left(y_f-\overline{y}-\mathbf{z}_f'\boldsymbol{\beta}^*/\sigma^2\right)^2 / \left(1+n^{-1}+\sigma^{-2}\mathbf{z}_f'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2}+\frac{\mathbf{I}_k}{\omega^2}\right)^{-1}\mathbf{z}_f\right)\right) + Q_\sigma$, $\boldsymbol{\beta}^* = \left(\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2}+\frac{\mathbf{I}_k}{\omega^2}\right)^{-1}\mathbf{Z}'\mathbf{y}$, recall that $Q_\sigma = (\mathbf{y}-\overline{y}\boldsymbol{\iota})'\mathbf{M}_\sigma(\mathbf{y}-\overline{y}\boldsymbol{\iota})$ and $\overline{n} = n+1$. We apply Laplace approximation to (15), by setting $\sigma = \exp(\tau)$. After the approximation is employed, we replace $\sigma_f^* = \exp\left(\tau_f^*\right)$, where $\tau_f^* = \arg\max h_{\overline{n}f}^N(\tau)$. The replacement is valid due to the continuity of the transformation. In appendix Appendix D we show that by applying the Sherman-Morrison formula and Matrix Determinant Lemma, we obtain

$$p\left(y_f|\mathbf{y},\mathbf{z}_f,M_j\right) \propto \sigma_f^{*-(n+a_0+1)} g_2\left(\sigma_f^*\right)^{0.5} g_1\left(\sigma_f^*\right)^{-0.5} \left|\frac{\mathbf{X}'\mathbf{X}}{\sigma_f^{*2}}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right|^{-\frac{1}{2}} \exp\left(-\frac{Q_{\sigma_f^*}^f+b_0}{2\sigma_f^{*2}}\right) \tag{16}$$

where $g_1(\sigma) = 1+(n/\overline{n})\,\sigma^{-2}\mathbf{z}_f'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2}+\frac{\mathbf{I}_k}{\omega^2}\right)^{-1}\mathbf{z}_f$. Finally, by plugging the marginal likelihood implied by (10) into (14), equation (16) becomes

$$p\left(y_f|\mathbf{y},\mathbf{z}_f,M_j\right)=c_{\sigma_f^*}\left(\frac{h_n^{*\prime\prime N}}{h_{\overline{n}f}^{*\prime\prime N}}\right)^{0.5}\frac{\sigma_f^{*-(\overline{n}+a_0)}}{\sigma^{*-(n+a_0)}}\frac{g_2\left(\sigma_f^*\right)^{0,5}}{g_2\left(\sigma^*\right)^{0.5}}\frac{\left|\frac{\mathbf{X}'\mathbf{X}}{\sigma_f^{*2}}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right|^{-\frac{1}{2}}}{\left|\frac{\mathbf{X}'\mathbf{X}}{\sigma^{*2}}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right|^{-\frac{1}{2}}}\frac{\exp\left(-\frac{Q_{\sigma_f^*}^f+b_0}{2\sigma_f^{*2}}\right)}{\exp\left(-\frac{Q_{\sigma^*}+b_0}{2\sigma^{*2}}\right)} \tag{17}$$

where, $c_{\sigma_f^*}=n/\overline{n}\left(2\pi\right)^{0.5}g_1\left(\sigma_f^*\right)^{-0.5}$ is a part of the normalising constant.[9]

3.2.2. Non-Conjugate Dependent Cauchy Prior

Assuming the dependent Cauchy prior (equation 11), and after integrating over the constant and scale parameters, the predictive posterior (equation 14) is obtained as

$$p\left(y_f|\mathbf{y},\mathbf{z}_f,M_j\right)\propto\delta\left(y_f\right)\int_{\boldsymbol{\beta}\in\mathbb{R}^k}\left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-\frac{1+k}{2}}\left(1+\frac{1}{v_0}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_f\right)'\frac{\mathbf{D}_{\boldsymbol{z}}}{s_{f,b_0}^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_f\right)\right)^{-\frac{v_0+k}{2}}d\boldsymbol{\beta} \tag{18}$$

where, $\delta\left(y_f\right)=\left(v_0s_{f,b_0}^2\right)^{-(v_0+k)/2}$,
$v_0s_{f,b_0}^2=\left(\left(y_f-\overline{y}-\mathbf{z}_f'\widehat{\boldsymbol{\beta}}\right)^2/\left(1+n^{-1}+\mathbf{z}_f'\mathbf{D}_{\mathbf{z}}^{-1}\mathbf{z}_f\right)\right)+\widehat{\mathbf{u}}'\widehat{\mathbf{u}}+b_0$, $v_0=n+a_0-k$, $\mathbf{D}_{\mathbf{z}}=\mathbf{Z}'\mathbf{Z}+\frac{n}{\overline{n}}\mathbf{z}_f\mathbf{z}_f'$, $\widehat{\boldsymbol{\beta}}_f=\mathbf{D}_{\mathbf{z}}^{-1}\left(\mathbf{Z}'\mathbf{y}+\frac{n}{\overline{n}}\mathbf{z}_f\left(y_f-\overline{y}\right)\right)$. Notice that $\widehat{\mathbf{u}}$ is the residual vector of the regression of $\mathbf{y}$ on $\mathbf{X}$. The integral above cannot be solved analytically, so we employ the multivariate Laplace approximation (7) to proceed. Finally, by plugging the marginal likelihood implied by (13)

[9]The analytical derivations of equation 17 can be found in section Appendix D.1 of the Appendix.

into (14), equation (18) becomes

$$p\left(y_f|\mathbf{y},\mathbf{z}_f,M_j\right) = c_{\widetilde{\beta}_f}\frac{\left(v_0 s^2_{f,b_0}\right)^{-(v+k)/2}\left|\mathbf{H}^*_{f\overline{n}}\right|^{-1/2}\exp\left(\overline{n}h^{C_f}_{\overline{n}}\left(\widetilde{\boldsymbol{\beta}_f}\right)\right)}{\left(vs^2_{b_0}\right)^{-(v+k)/2}\left|\mathbf{H}^*_{n}\right|^{-1/2}\exp\left(nh^{C}_{n}\left(\widetilde{\boldsymbol{\beta}}\right)\right)} \tag{19}$$

where $c_{\widetilde{\boldsymbol{\beta}}_f} = \frac{1}{\pi^{1/2}}\frac{\Gamma((\overline{n}+a_0-1)/2)}{\Gamma((n+a_0-1)/2)}\left(\frac{n}{\overline{n}}\right)^{(k+1)/2}$ is a component of the normalising constant, $h^{C_f}_{\overline{n}}(\boldsymbol{\beta})$ denotes the log-integrand in (18), $\widetilde{\boldsymbol{\beta}}_f = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^k} h^{C_f}_{\overline{n}}(\boldsymbol{\beta})$, and $\mathbf{H}^*_{f\overline{n}}$ is the negative Hessian evaluated at $\widetilde{\boldsymbol{\beta}}_f$.[10]

3.3. Laplace approximation error evaluation

There are three Laplace approximation templates used throughout this paper.[11] (i) Univariate Laplace: For marginal likelihoods like in eq. 6 (e.g., under the non-conjugate normal prior), we approximate a *one-dimensional* integral with respect to the transformed variance parameter $\tau = \log\sigma$. The approximation is obtained via a one-dimensional Laplace expansion about the mode $\hat{\tau}$. (ii) Multivariate Laplace: For marginal likelihoods of the form in 7 (e.g., under the independent Cauchy-inverse-gamma prior), we integrate out σ analytically. The remaining marginal likelihood is a *multivariate* integral over the k non-constant regression coefficients, which we approximate using a multivariate Laplace expansion around the posterior mode $\hat{\boldsymbol{\beta}}$. (iii) Ratios of Laplace approximations: Posterior moments and predictive quantities are evaluated by expressing them as ratios of integrals and applying Laplace methods to the numerator and denominator. In particular, poste-

[10] The analytical derivations of equation 19 can be found in Appendix D.2.

[11] In the discussion in this sub-section, we suppress the subscript j (denoting model j) in k, τ, and $\boldsymbol{\beta}$ for ease of exposition.

rior moments are obtained from ratios derived from the moment-generating function (MGF, see equation 8) using fully exponential Laplace expansions, while predictive quantities are expressed as ratios of integrals of the form 6 or 7, following Tierney and Kadane (1986) and Tierney et al. (1989).

For regular parametric models with *fixed* parameter dimension, Laplace approximations to integrals of the form 6 and 7 are accurate to relative order $O(n^{-1})$, and that *ratios* of Laplace-approximated integrals typically enjoy higher-order cancellation, see e.g. Tierney and Kadane (1986), Kass and Vaidyanathan (1992), and Kass and Raftery (1995). In particular, Tierney and Kadane (1986) show that the leading error terms cancel in Laplace ratios, yielding $O(n^{-2})$ accuracy for posterior expectations under standard regularity conditions, and Tierney et al. (1989) obtain $O(n^{-2})$ accuracy for expectations and $O(n^{-3})$ for variances using the fully exponential form (see e.g. Tierney and Kadane, 1986 and Tierney et al., 1989). In our setting, these results apply directly to the predictive density and moment approximations developed in Theorems 1 and 2. Rue et al. (2009) further connect these expansions to the Laplace approximations used in Integrated Nested Laplace Approximations (INLA) and emphasize that they rely on the same local Gaussian expansion around the posterior mode. They report that, under normalization, the Laplace approximation to marginal likelihoods and posterior marginals has relative error of order $O(n^{-3/2})$.

When the dimension of the integral increases with sample size, the accuracy of Laplace approximations can deteriorate. Shun and McCullagh (1995) develop high-dimensional Laplace expansions and show that the standard Laplace formula remains reliable only when the dimension grows sufficiently

slowly relative to n (their expansions make the dependence on k explicit). In particular, their results suggest that the usual Laplace approximation is asymptotically valid only if k grows slowly enough, roughly if $k^3/n \to 0$, and they provide examples in which the approximation error becomes non-negligible when this condition is violated. To address this breakdown, they reorganize the expansion on the log integral scale and place the leading correction term in the exponent (a fully exponential Laplace form), which restores $o(1)$ relative error in regimes where the standard multiplicative correction fails as k increases. For example, in Appendix E we show that for the growth regression application ($n = 72$), posterior mass is concentrated on models with moderate k, the largest benchmarked models have $k \leq 16$, placing the empirical application well within the low-to-moderate dimensional regime where Laplace diagnostics can be meaningfully assessed numerically.

Laplace approximations are local expansions around posterior modes and are most reliable when the likelihood-prior product is *unimodal* and the log-posterior is well approximated by a quadratic in the region that carries most posterior mass. In particular, the key regularity condition in the fully exponential Laplace framework is that the likelihood times prior be unimodal, and the usual Laplace approximation "will produce reasonable results as long as the posterior is unimodal or at least dominated by a single mode" (Tierney and Kadane, 1986). Two standard remedies help when local Gaussianity is threatened. First, *reparameterisation* can reduce skewness and boundary effects for positive scale parameters: in the non-conjugate normal case we integrate in $\tau = \log \sigma$, mapping the support to $\mathbb{R}$ and typically yielding a more symmetric local shape. This is consistent with the well known sensitiv-

ity of Laplace methods to parameterisation. Tierney et al. (1989) explicitly note that the parameterisation in which Laplace/MGF approximations are applied is very important, and demonstrate sizeable accuracy gains from simple log-transformations of parameters. Second, *multimodality* is handled diagnostically: because Laplace methods assume (effective) unimodality, one should check for multiple modes by running the mode-finder from dispersed starting points and verifying that the optimiser converges to the same solution. If several well separated modes contribute non-negligibly, a pragmatic extension is to approximate the integral by summing separate Laplace contributions computed at the dominant modes (a multimodal/mixture Laplace, see De Bruijn, 1981), and to validate key cases using simulation based benchmarks (e.g., bridge sampling) on a manageable subset of models. This emphasis on checking (and not assuming) unimodality is also echoed in the Laplace/INLA literature, where concerns are raised that priors can induce multimodality even in simple variance-component settings (see the discussion in Rue et al., 2009).

In order to evaluate the transformation $\sigma = \exp(\tau)$ to the Laplace approximation we apply a simple exercise where we show that the approximation error with the transformation is smaller than it is without the transformation: we apply the Laplace approximation to estimate the normalizing constant of (5d); recall that prior (5d) has a constant term equal to $\int_{\sigma>0} \sigma^{-(a_0+1)} \exp\left(-\frac{b_0}{2\sigma^2}\right) d\sigma = \frac{\Gamma(a_0/2)}{2(b_0/2)^{a_0/2}}$. In Figure 1 we examine the following three cases:[12] 1) Integration with truncated (half) normal on $\sigma > 0$

[12]For this example, we set $b = \frac{1}{1000}$; however, it can be easily shown that the third case yields the smallest error for any positive value of b.

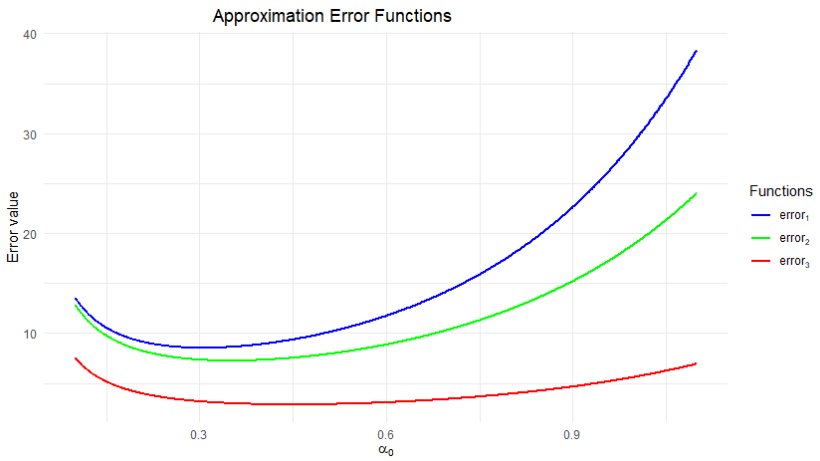

Figure 1: Error Graph

(Blue Line), 2) Integration with normal (Green Line), and 3) Use the transformation $\sigma = \exp(\tau)$ and integrate with normal (Red Line). In terms of the approximation error, it is easy to see that case (3) obtains the smallest one. Given this result, we apply the Laplace approximation to estimate the marginal likelihood using the transformed scale parameter. This approach also applies to the approximation of predictive density and moments.

4. Simulation experiment

4.1. Introduction

In this section we study the behavior of the non-conjugate priors in a controlled Monte Carlo experiment in the style of Fernandez et al. (2001a). We contrast the performance of the non-conjugate priors with the two best performing priors from their experiment (their priors a and i). We focus on their Model 1, which has a sparse coefficient vector with five non–zero coefficients and $K = 15$ potential regressors. The experiment is used to compare

the four priors in terms of model selection and predictive performance.

4.2. Data generating process

For each replication we generate data from the normal linear regression model given in (5a). The design matrix Z is obtained in two steps. First we construct an $n \times 15$ "raw" regressor matrix $R = (r_1, \ldots, r_{15})$, which, for $j = 1, \ldots, 10$, the j-th column r_j has independent $\mathcal{N}(0,1)$ entries. To induce strong correlation among some regressors, we define for each observation $i = 1, \ldots, n$, $u_i = 0.3\, r_{i1} + 0.5\, r_{i2} + 0.7\, r_{i3} + 0.9\, r_{i4} + 1.1\, r_{i5}$, and set $r_{ij} = u_i + e_{ij}$, $j = 11, \ldots, 15$, where $e_{ij} \sim \mathcal{N}(0,1)$ are independent of all other random quantities.

Second, each column of R is centred to have sample mean zero, and the resulting matrix is denoted by $Z = (z_1, \ldots, z_{15})$.The true coefficient vector β^* is sparse with five active regressors, so that the conditional mean can be written explicitly as

$$\mathbb{E}(y_i \mid z_i) = 4 + 2z_{i1} - z_{i5} + 1.5z_{i7} + z_{i,11} + 0.5z_{i,13}, \qquad i = 1, \ldots, n. \quad (20)$$

The remaining ten coefficients are equal to zero. The disturbance standard deviation is fixed at $\sigma^* = 2.5$. Hence the true DGP consists of a moderate signal in a subset of regressors, combined with sizeable noise and substantial collinearity between some of the candidate covariates. We consider three sample sizes, $n \in \{50, 100, 1000\}$, so that the experiment covers a range between a small sample and a situation where the likelihood dominates the prior more strongly. For each (n, prior) combination we generate independent pairs (y, Z) from (5a) and (20) and treat the corresponding model with five

active regressors as the true model, $M_\star$.

4.3. Priors and model space

The set of candidate models $\mathcal{M} = \{M_1, \ldots, M_{2^K}\}$ is defined by all subsets of the K regressors, as in the general BMA setup in Section 2. For the prior over model space we adopt the Uniform specification used by Fernandez et al. (2001a): $P(M_j) = 2^{-K}$. We compare the following four priors for the regression parameters:

1. *Fernandez et al. (2001a) prior a (FLS a) (conjugate g-prior).* This is the standard g-prior of Zellner, applied to the slopes in M_j with $g = n$ and the usual improper prior for (α, σ^2) used in Fernandez et al. (2001a).
2. *Fernandez et al. (2001a) prior i (FLS i)(conjugate g-prior).* This is the prior which, together with prior a, had the best overall performance in the Fernandez et al. (2001a) experiment. For this prior $g = 1/K^2$.
3. *Non–conjugate Normal prior (NCN).* The first non–conjugate prior is the dependent Normal prior for the slopes discussed in Section 3.1 (equation 9).
4. *Non–conjugate Cauchy prior (NCC).* The second non–conjugate prior is the multivariate dependent Cauchy prior, as discussed in section 3.2 (equation 11).

4.4. Simulation design and performance measures

For each prior and each sample size n, the simulation proceeds as follows:

1. Generate a design matrix Z and response vector y from (20).

2. Compute the marginal likelihood of each visited model under the prior under consideration using the analytical expressions for the conjugate priors and the Laplace approximations for the NCN and NCC priors.
3. Explore the model space using the MC^3 algorithm of Madigan and York, with a single-variable flip proposal.[13] The chain is run for a burn-in period of 20,000 iterations and then for a sampling phase of 50,000 iteration, identical to the Fernandez et al. (2001a) setting.

This procedure is repeated independently for 100 replications for each (n, prior) combination. All summary statistics reported in the next subsection are based on averages (and selected quantiles) over these Monte Carlo replications. For each replication we calculate the following measures related to Model-selection accuracy: 1) The posterior probability of the true model: $p_{\text{true}} = p(M_\star \mid y)$, 2) The ratio between the posterior probability of the true model and that of the best incorrect model: $R = \frac{p(M_\star|y)}{\max_{M_j \neq M_\star} p(M_j|y)}$, which is a measure of separation of the true model from the closest competitor, 3) The number of models visited, 4) The posterior inclusion probabilities of each regressor, from which we derive the expected number of false positives (irrelevant regressors included) and false negatives (true regressors omitted), 5) The Mean squared error of the coefficient vector relative to the true DGP coefficients $MSE(\beta)$, and 6) the ℓ_2 norm (Euclidean norm) of the estimated coefficients on irrelevant ("noise") regressors, which measures how well a prior shrinks noise.

[13] A single-variable flip proposal means that, at each MC^3 iteration, one covariate is randomly selected and its inclusion indicator is switched, so the proposed model differs from the current model by adding or removing exactly one variable.

To evaluate predictive performance we follow Fernandez et al. (2001a) and construct a set of $q = 19$ out–of–sample design points $z_f^{(1)}, \ldots, z_f^{(q)}$ from the same distribution as the rows of Z. For each design point z_f we consider future observations

$$y_f \mid z_f \sim \mathcal{N}\big(\alpha^* + z_f'\beta^*, (\sigma^*)^2\big)\,,$$

and generate, for each replication, v independent draws $y_f^{(1)}, \ldots, y_f^{(v)}$ from this conditional distribution. Predictive ability is summarised by the log predictive score (LPS),[14] evaluated at each design point z_f:

$$\text{LPS}(z_f) \;=\; -\frac{1}{v}\sum_{i=1}^{v} \log p\big(y_{f_i} \mid z_f, \mathbf{y}\big), \tag{21}$$

where $p(y_f \mid z_f, \mathbf{y})$ is the BMA predictive density under the prior under consideration. For the non-conjugate priors, the estimation of $p\left(y_f|\mathbf{y}\right)$ comes from (1), by setting $\Delta = y_f$, so that

$$p\left(y_f|\mathbf{y}\right) = \sum_{j=1}^{2^k} p\left(y_f|\mathbf{y}, \mathbf{z}_{fj}, M_j\right) p\left(M_j|\mathbf{y}\right)$$

[14] Assume two probability distributions P and Q, the Kullback–Leibler divergence is a measure of the distance of how much an approximating probability distribution Q is different from a true probability distribution P, and defined as $KL\left(p\|\, q\right) = E_p\left(\ln p\right) - E_p\left(\ln q\right)$, where the first term of the RHS equation is the negative Entropy and the second term is the theoretical counterpart of LPS. It is reasonable to prefer distributions Q for which $KL\left(p|\, q\right)$ is minimized. When comparing two different distances from the same origin, say $KL\left(p|\, q_a\right)$ against $KL\left(p|\, q_b\right)$, it is easy to show that the comparison reduces to minimizing $-E_p\left(\ln q_i\right)$, $i = a, b$. In empirical cases the comparison can be done through the LPS.

Notice that, LPS is the sample analog of the expected loss with a logarithmic rule. Smaller values of LPS, mean that $p\left(y_f|\mathbf{y}\right)$ is closer to the true predictive distribution. For comparability with Fernandez et al. (2001a) we pay particular attention to three design points: $z_{\min}$, z_{med} and $z_{\max}$, corresponding to the smallest, median and largest values of the true conditional mean $\alpha^* + z_f'\beta^*$ among the 19 design points. In addition, we compute summary measures of predictive performance that aggregate the LPS over all 19 design points.

4.5. Simulation Results

In this section we report Monte Carlo summaries of the model selection and predictive criteria described above for the four priors and the three sample sizes. This provides a controlled comparison of the conjugate and non-conjugate specifications. As in Fernandez et al. (2001a), we begin by examining the posterior probability assigned to the data generating model, $p_{\text{true}} = p(M_\star \mid y)$ (Table 1). In absolute terms, p_{true} is small for $n = 50$ and $n = 100$ under all priors, which is unsurprising given the size of the model space ($2^{15} = 32{,}768$ models), the sizable noise variance, and the induced collinearity among candidate regressors. Nevertheless, the posterior probability of the true model is orders of magnitude above its prior probability 2^{-15}, and increases sharply with sample size for all priors.

A more nuanced pattern emerges across priors. For small and moderate samples ($n = 50, 100$), FLS prior i (FLS-i) places the highest average mass on the exact true model, consistent with its strong small sample parsimony. For $n = 1000$, however, the ranking changes: the non-conjugate Normal prior (NCN) yields the highest p_{true} (0.568), followed by FLS-a (0.505), while NCC

assigns less mass to the exact DGP model (0.407) and FLS-i remains the lowest (0.288). Thus, in this DGP the NCN prior becomes clearly favorable in terms of posterior concentration on $M_\star$ once the likelihood is sufficiently informative. At the same time, the heavier-tailed NCC prior remains competitive but appears to spread posterior probability more across close competitors, which mechanically lowers p_{true} even when the true model is strongly favored in relative terms.

Table 1: Means and Stds of posterior probability of the true model (Prior 1 = FLS a, Prior 2 = FLS i, Prior 3 = NC Normal, Prior 4 = NC Cauchy)

	$n = 50$		$n = 100$		$n = 1000$	
Prior	Mean	Std	Mean	Std	Mean	Std
1	0.0122	0.0167	0.0729	0.0735	0.5049	0.1735
2	0.0190	0.0316	0.1019	0.0956	0.2878	0.1310
3	0.0136	0.0194	0.0849	0.0893	0.5682	0.1786
4	0.0086	0.0115	0.0481	0.0472	0.4072	0.1611

Since a low p_{true} can still be compatible with correct model ranking (if posterior mass is spread over near equivalent competitors), Table 2 reports quartiles of the separation ratio. Across all priors and sample sizes, the quartiles lie comfortably above one, indicating that the true model is typically favored over the best incorrect alternative even when its posterior mass is diluted over many close substitutes. For $n = 50$ and $n = 100$ the ratios remain moderate, reflecting the fact that, under strong collinearity, several models can approximate the true conditional mean nearly equally well. This is also consistent with the inclusion probability patterns below: regressors with relatively weaker signals and/or high correlation are intrinsically harder to identify in small samples, so posterior mass is naturally distributed across

models that swap correlated regressors.

For $n = 1000$, separation becomes strong for all priors, with particularly large ratios under NCN and FLS-a (median $R \approx 21$ and $R \approx 18$, respectively). NCC also yields clear separation (median $R \approx 14$), while FLS-i exhibits markedly weaker separation (median $R \approx 9$), which aligns with its lower p_{true} at $n = 1000$: rather than concentrating on the exact DGP model, it assigns non-negligible probability to close competitors, especially models that augment the true specification with additional regressors.

Table 2: Quartiles of ratio of posterior probabilities (true model vs best incorrect)

	$n = 50$			$n = 100$			$n = 1000$		
Prior	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
1	1.398	3.029	6.501	1.835	5.797	10.427	8.716	18.430	28.810
2	1.808	4.826	11.696	2.248	6.096	12.670	4.232	9.125	13.997
3	1.155	3.520	6.145	1.848	5.694	10.452	10.454	21.205	31.733
4	1.261	2.826	4.534	2.361	4.503	6.698	5.194	14.291	20.573

Table 3 provides additional insight into posterior concentration through the number of distinct models visited by the MC^3 chain. For $n = 50$ and $n = 100$, the chain explores a large subset of the model space under all priors, reflecting weak sample information. In this regime, FLS-i visits the fewest models, consistent with its stronger shrinkage toward parsimonious specifications and its relatively higher p_{true} in small samples. NCC, by contrast, visits substantially more models when n is small, suggesting a flatter posterior landscape and more posterior uncertainty about model composition, an effect that is in line with its heavier-tailed slope prior.

For $n = 1000$ the situation changes: posterior concentration becomes much tighter, but the degree of concentration varies. NCN visits the fewest

models on average (about 118), followed by FLS-a (about 147), whereas NCC and FLS-i explore noticeably larger sets (about 234 and 305, respectively). Taken together with Tables 1 and 2, this indicates that NCN delivers the sharpest posterior concentration around the exact DGP model in large samples, while the heavier tails in NCC (and the relatively weak penalty implied by FLS-i) tend to keep more models in play.

Table 3: Means and Stds of number of models visited (Prior 1 = FLS a, Prior 2 = FLS i, Prior 3 = NC Normal, Prior 4 = NC Cauchy)

	$n = 50$		$n = 100$		$n = 1000$	
Prior	Mean	Std	Mean	Std	Mean	Std
1	2367	643	1170	325	147	42
2	1173	349	713	203	305	67
3	2226	718	984	307	118	35
4	3428	1100	1690	491	234	67

Turning to variable level inference, Table 4 confirms that the strongest signals are learned quickly under all priors: regressors 1 and 7 have posterior inclusion probabilities near one even at $n = 50$. The more subtle differences across priors emerge for the weaker (or more collinear) true regressors (notably regressors 5 and 13) and for the noise regressors. In small samples, FLS-i tends to downweight noise variables most aggressively, but it also assigns somewhat lower inclusion probability to the weaker true regressors, which is reflected in Table 5: FLS-i achieves the lowest false positives for $n = 50$ and $n = 100$, but at the cost of the highest false negatives. The non–conjugate Normal prior (NCN) occupies an intermediate, and arguably attractive, position in finite samples: relative to FLS-a, it yields fewer false positives and a smaller ℓ_2 norm of posterior means on noise coefficients (Ta-

ble 6), while maintaining broadly similar inclusion probabilities for the key signals.

The non–conjugate Cauchy prior (NCC) exhibits a different tradeoff in small samples: it produces the lowest false negatives (i.e., it is most reluctant to omit true regressors), but it pays for this with the highest false positives when $n = 50$ and $n = 100$. At $n = 1000$, all priors essentially eliminate false negatives, but they continue to differ in how aggressively they exclude noise regressors. In particular, NCN now achieves the lowest false positives (0.55 on average) and the smallest noise ℓ_2 norm (0.043), with FLS-a close behind. NCC is somewhat less parsimonious in this configuration, which is consistent with its lower p_{true} and larger number of visited models.

The coefficient accuracy measure in Table 7 reinforces this interpretation. In small and moderate samples ($n = 50, 100$), NCC yields the lowest MSE for β, consistent with the idea that heavier tails can reduce shrinkage-induced bias for non-zero coefficients. For $n = 1000$, however, NCN attains the lowest MSE (0.00237), with FLS-a close behind; NCC and FLS-i are somewhat higher, though differences are small in absolute terms once the likelihood dominates.

Finally, Tables 8 and 9 report predictive performance via the log predictive score. Here, differences are generally modest, and all priors approach the entropy lower bound as n becomes large, implying that the predictive distributions become very close to the true Gaussian sampling distribution. For $n = 50$ and $n = 100$, NCN achieves the best overall LPS (albeit by a small margin), indicating that the additional flexibility of the non-conjugate Normal prior can translate into slightly improved predictive density fit in

Table 4: Posterior inclusion probabilities by regressor († indicates true regressors. (Prior 1 = FLS a, Prior 2 = FLS i, Prior 3 = NC Normal, Prior 4 = NC Cauchy)

Reg.	Prior 1		Prior 2		Prior 3		Prior 4	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$n = 50$								
$\beta_1^\dagger$	0.988	0.038	0.979	0.060	0.986	0.049	0.990	0.031
β_2	0.264	0.190	0.179	0.192	0.230	0.154	0.309	0.183
β_3	0.255	0.172	0.172	0.173	0.222	0.153	0.305	0.163
β_4	0.276	0.179	0.190	0.177	0.240	0.147	0.332	0.158
$\beta_5^\dagger$	0.488	0.279	0.389	0.293	0.476	0.276	0.540	0.246
β_6	0.207	0.157	0.122	0.133	0.180	0.143	0.244	0.143
$\beta_7^\dagger$	0.948	0.124	0.917	0.167	0.944	0.143	0.951	0.114
β_8	0.212	0.147	0.126	0.127	0.186	0.124	0.251	0.143
β_9	0.220	0.139	0.129	0.116	0.198	0.124	0.257	0.138
β_{10}	0.201	0.110	0.115	0.105	0.174	0.092	0.242	0.106
$\beta_{11}^\dagger$	0.788	0.260	0.752	0.306	0.759	0.278	0.792	0.252
β_{12}	0.238	0.162	0.152	0.164	0.184	0.154	0.244	0.151
$\beta_{13}^\dagger$	0.397	0.269	0.313	0.287	0.341	0.261	0.406	0.259
β_{14}	0.217	0.119	0.135	0.127	0.165	0.102	0.228	0.113
β_{15}	0.213	0.147	0.131	0.125	0.161	0.129	0.221	0.137
$n = 100$								
$\beta_1^\dagger$	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
β_2	0.193	0.168	0.154	0.162	0.171	0.160	0.245	0.170
β_3	0.216	0.167	0.168	0.162	0.186	0.124	0.272	0.160
β_4	0.223	0.171	0.169	0.162	0.191	0.158	0.295	0.164
$\beta_5^\dagger$	0.668	0.280	0.622	0.293	0.656	0.278	0.721	0.242
β_6	0.176	0.138	0.135	0.127	0.150	0.127	0.221	0.145
$\beta_7^\dagger$	0.988	0.057	0.986	0.066	0.987	0.063	0.991	0.046
β_8	0.172	0.173	0.128	0.137	0.149	0.161	0.215	0.171
β_9	0.153	0.114	0.111	0.110	0.130	0.108	0.196	0.124
β_{10}	0.156	0.112	0.115	0.112	0.133	0.100	0.202	0.120
$\beta_{11}^\dagger$	0.945	0.147	0.942	0.155	0.945	0.147	0.957	0.128
β_{12}	0.142	0.101	0.100	0.093	0.106	0.074	0.162	0.084
$\beta_{13}^\dagger$	0.461	0.310	0.385	0.308	0.461	0.310	0.545	0.297
β_{14}	0.142	0.093	0.096	0.083	0.125	0.095	0.190	0.117
β_{15}	0.138	0.119	0.092	0.090	0.128	0.126	0.193	0.142
$n = 1000$								
$\beta_1^\dagger$	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
β_2	0.059	0.093	0.107	0.103	0.052	0.091	0.085	0.100
β_3	0.078	0.103	0.134	0.127	0.073	0.101	0.115	0.122
β_4	0.062	0.079	0.112	0.098	0.065	0.082	0.105	0.102
$\beta_5^\dagger$	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
β_6	0.065	0.089	0.116	0.118	0.054	0.080	0.087	0.106
$\beta_7^\dagger$	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
β_8	0.071	0.100	0.126	0.123	0.059	0.092	0.095	0.111
β_9	0.055	0.052	0.104	0.081	0.044	0.043	0.076	0.065
β_{10}	0.063	0.074	0.115	0.105	0.051	0.064	0.085	0.088
$\beta_{11}^\dagger$	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
β_{12}	0.063	0.065	0.117	0.095	0.044	0.050	0.076	0.074
$\beta_{13}^\dagger$	1.000	0.001	1.000	0.000	1.000	0.001	1.000	0.001
β_{14}	0.079	0.121	0.132	0.141	0.059	0.107	0.091	0.125
β_{15}	0.065	0.103	0.114	0.114	0.048	0.099	0.076	0.105

Table 5: Means and Stds of expected number of false negatives (missed true regressors) and false positives (included noise regressors) (lower is better

	False Negatives						False Positives					
	$n = 50$		$n = 100$		$n = 1000$		$n = 50$		$n = 100$		$n = 1000$	
Prior	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	1.391	0.460	0.865	0.478	0.000	0.001	2.302	0.488	1.783	0.445	0.660	0.320
2	1.649	0.481	0.977	0.515	0.000	0.000	1.454	0.453	1.370	0.456	1.178	0.418
3	1.505	0.443	0.951	0.490	0.000	0.001	1.994	0.435	1.531	0.461	0.550	0.295
4	1.321	0.416	0.785	0.448	0.000	0.001	2.633	0.455	2.191	0.524	0.890	0.395

Table 6: Means and Stds of ℓ_2 norm of posterior mean on noise coefficients (lower is better)

	$n = 50$		$n = 100$		$n = 1000$	
Prior	Mean	Std	Mean	Std	Mean	Std
1	0.618	0.351	0.362	0.238	0.048	0.059
2	0.503	0.366	0.319	0.268	0.070	0.066
3	0.573	0.350	0.338	0.269	0.043	0.057
4	0.613	0.321	0.389	0.259	0.059	0.065

Table 7: Means and Stds of coefficient MSE (MSE(β) (lower is better)

	$n = 50$		$n = 100$		$n = 1000$	
Prior	Mean	Std	Mean	Std	Mean	Std
1	0.1212	0.06383	0.05956	0.04404	0.002422	0.001663
2	0.1241	0.06477	0.06157	0.04493	0.002670	0.001813
3	0.1195	0.06226	0.06059	0.04434	0.002372	0.001636
4	0.1125	0.05608	0.05665	0.04087	0.002555	0.001767

finite samples. NCC performs worse in small samples, consistent with a predictive distribution that is somewhat too diffuse when information is limited: by $n = 1000$ the gap narrows substantially, though NCC still remains slightly behind the other priors in overall LPS. Overall, this controlled experiment suggests that the non-conjugate priors, and especially NCN, can deliver (i) predictive performance comparable to benchmark conjugate priors and (ii) at least as good, and in some cases better, behavior in terms of posterior concentration and parsimony. NCC continues to provide a distinct small sample tradeoff (fewer missed signals but more noise inclusion), but in this configuration it is somewhat less concentrated than NCN in large samples.

Table 8: Overall predictive performance (LPS all19) (lower is better)

Prior	$n = 50$	$n = 100$	$n = 1000$
1	2.4298	2.3839	2.3360
2	2.4324	2.3860	2.3362
3	2.4257	2.3829	2.3356
4	2.4870	2.4146	2.3389

Table 9: Conditional predictive performance (LPS at $z_{\min}$, z_{med}, $z_{\max}$) (lower is better)

	$n = 50$			$n = 100$			$n = 1000$		
Prior	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$
1	2.4851	2.4381	2.4433	2.4169	2.3844	2.4199	2.3438	2.3497	2.3308
2	2.5109	2.4452	2.4484	2.4178	2.3878	2.4208	2.3420	2.3482	2.3312
3	2.4923	2.4356	2.4450	2.4189	2.3855	2.4197	2.3441	2.3496	2.3306
4	2.5336	2.4977	2.5026	2.4549	2.4188	2.4459	2.3470	2.3529	2.3338

It is worth emphasizing that this is a deliberately well behaved synthetic setting: the DGP is linear and Gaussian, so conjugate priors are naturally well aligned with the likelihood. In that sense, the relatively small predictive

differences (and the broadly similar qualitative behavior across priors) can be interpreted as a conservative benchmark: even in a setting that is favorable to conjugate analysis, the proposed non-conjugate priors remain competitive and, in the case of NCN, exhibit clear and interpretable advantages in posterior concentration and noise control.

5. Empirical Results on Growth Data

5.1. Posterior Analysis

In this section we present estimation results on the benchmark growth dataset of Sala-i Martin (1997) based on the $MC\,(3)$ algorithm. The dataset is a cross-sectional collection of economic, social, and political variables designed to identify robust determinants of long-term economic growth. It covers 140 countries, with average per capita GDP growth computed over the period 1960-1992 used as the dependent variable. We follow the design of Fernandez et al. (2001b) and retain the same 41 explanatory variables in our empirical design. We compare the estimation output of the non-conjugate with the conjugate case. For the latter case we replicate the results of Fernandez et al. (2001b). We conduct two million MC(3) iterations/model draws (the chain is over model indices only), using a burn-in sample of one million for the estimation of (marginal) probabilities of inclusion (4), BMA moments of the regressor coefficients (mean (2) and standard deviation, the square root of (3)). The algorithm converges since the correlation between the approximated (analytical) numerical model probabilities, for both of the non-conjugate cases is greater than 99.5%.

We find that in the non-conjugate cases the average model size is between 6

to 7 variables, with acceptance rate 16% for the normal and 13% for the Cauchy priors, while the acceptance rate of the conjugate normal is about 17%. The conjugate prior is favorable to larger models on average, since the expected number of covariates is between 10 to 11. These results are in line with most researchers in regard to the (average) number of the determinants in a growth regression e.g. see Sala-i Martin (1997) and Sala-i Martin et al. (2004). It is clear that the non-conjugate setting results to more parsimonious models. These results are in line with Dawid (1987) and Fang and Dawid (2002) who promulgate the view that the non-conjugate setting does not suffer from determinism which may induce overfitting.

Tables 10 and 11 present the 10 best determinants of growth, according to the marginal probability of inclusion of each case under consideration. Table 10 reports the results of conjugate (CN) against the non-conjugate (NCN) normal and Table 11 the output of conjugate normal against non-conjugate Cauchy (NCC). It is obvious that the conjugate and the non-conjugate case share only 5 common variables among the 10 best. These are the *GDP level in 1960, Fraction Confucian*, *Equipment Investment*, *Sub-Saharan Africa* and *No. Years Open Economy.* The conjugate case also includes *Life expectancy*, *Fraction Muslim*, *Rule of Law*, *Degree of Capitalism*, *Fraction of GDP in mining.* On the other hand the non-conjugate cases share 8 common variables among the 10 best. The non-conjugate normal includes also *Fraction Protestant*, *Population Growth*, *Non-Equipment Investment*, *Public Education Share*, *Fraction Buddhist.* In the Cauchy case *Population Growth* and *Public Education Share* are replaced by *Fraction Muslim* and *Life Expectancy.*

Table 10: Ten Best Determinants - Conjugate agst. Non-Conjugate Normal

	CN		NCN	
Ranking	Growth Determinant	PIP	Growth Determinant	PIP
1	GDP level in 1960	0.998800	Equipment Investment	0.995750
2	Fraction Confucian	0.989000	Fraction Confucian	0.966590
3	Life expectancy	0.931500	No. Years Open Economy	0.513320
4	Equipment Investment	0.922600	Fraction Protestant	0.467870
5	Sub-Saharan Africa	0.735400	GDP level in 1960	0.429240
6	Fraction Muslim	0.645200	Population Growth	0.428830
7	No. Years Open Economy	0.514600	Sub-Saharan Africa	0.385850
8	Rule of Law	0.490000	Non-equipment Investment	0.361280
9	Fraction GDP in Mining	0.459200	Public Education Share	0.334550
10	Degree of Capitalism	0.455400	Fraction Buddhist	0.265820

Table 11: Ten Best Determinants - Conjugate agst. Non-Conjugate Cauchy

	CN		NCC	
Ranking	Growth Determinant	PIP	Growth Determinant	PIP
1	GDP level in 1960	0.998800	Equipment Investment	0.978890
2	Fraction Confucian	0.989000	Fraction Confucian	0.977540
3	Life expectancy	0.931500	No. Years Open Economy	0.601550
4	Equipment Investment	0.922600	GDP level in 1960	0.527490
5	Sub-Saharan Africa	0.735400	Fraction Protestant	0.456510
6	Fraction Muslim	0.645200	Non-equipment Investment	0.427980
7	No. Years Open Economy	0.514600	Sub-Saharan Africa	0.419610
8	Rule of Law	0.490000	Life expectancy	0.242090
9	Fraction GDP in Mining	0.459200	Fraction Muslim	0.230790
10	Degree of Capitalism	0.455400	Fraction Buddhist	0.216390

Table 12 reports the results ranked by the probability of inclusion of the non-conjugate normal case. The hyperparameters are set to $(a_0, b_0, \omega^{-2}) = (0.01, 0.001, 10)$ and $(a_0, b_0, \omega^{-2}) = (0.01, 0.001, 3)$ for the normal and Cauchy case, respectively. The results are robust to alternative values of the hyper-parameters and the prior elicitation procedure is based on the fact that the following assumptions should hold simultaneously: The acceptance rate of a new variable is roughly $15\% - 20\%$, the total probability of the total number models visited should be roughly 7% to 10% and finally the convergence criterion should be greater than 99%. The choice of these criteria is made so as to be close to the conjugate case, for the results to be comparable.

Table 12 shows that, there are just three common variables between the best models of conjugate and non-conjugate cases (conjugate and non-conjugate cases includes 10 and 4 variables, respectively). These are *Equipment Investment*, *Fraction Comfucian* and *Fraction Protestant.* Despite the fact that the *GDP level in 1960* has almost a 50% probability of inclusion in the non-conjugate case, it is not included in the best model. *Life expectancy*, a variable with very large probability of inclusion according to Fernandez et al. (2001b) and Sala-i Martin (1997) is not only missing from the list but also obtains a very low probability in the non-conjugate case (5.5% for the Normal case and 24.2% for Cauchy). Instead, the *number of years of open economy* is included in both non-conjugate models but this is not the case for the conjugate one, despite the fact that in all cases (conjugate included) the marginal probability of inclusion is very close. One may also note that in the non-conjugate normal case, there are 15 variables with probability of inclusion greater than 10% , 17 for the NCC and 20 for the conjugate normal

case.

tabularxtabularx

Table 12: Marginal Posterior Probabilities of Inclusion and Best Models*

Index	Regressor	NCN	BM-NCN	NCC	BM-NCC	CN	BM-CN
1	Equipment Investment	0.996	1	0.979	1	0.923	1
2	Fraction Confucian	0.967	1	0.978	1	0.989	1
3	No. Years Open Economy	0.513	1	0.602	1	0.515	0
4	Fraction Protestant	0.468	1	0.457	1	0.451	1
5	GDP level in 1960	0.429	0	0.527	0	0.999	1
6	Population Growth	0.429	0	0.173	0	0.037	0
7	Sub-Saharan Africa	0.386	0	0.420	0	0.735	1
8	Non-equipment Investment	0.361	0	0.428	0	0.442	1
9	Public Education Share	0.335	0	0.204	0	0.028	0
10	Fraction Buddhist	0.266	0	0.216	0	0.198	0
11	Primary School Enroll.	0.196	0	0.176	0	0.204	0
12	Higher Education Enroll.	0.188	0	0.130	0	0.045	0
13	Fraction Muslim	0.163	0	0.231	0	0.645	1
14	Rule of Law	0.140	0	0.195	0	0.490	1
15	Fraction GDP in Mining	0.109	0	0.200	0	0.459	0

Continued on next page

Table 12 – Continued

Index	Regressor	NCN	BM-NCN	NCC	BM-NCC	CN	BM-CN
16	Latin America	0.083	0	0.107	0	0.207	0
17	Fraction Catholic	0.071	0	0.076	0	0.131	0
18	Fraction Hindu	0.059	0	0.074	0	0.127	0
19	Primary Exports	0.058	0	0.064	0	0.100	0
20	Ratio Workers to Population	0.056	0	0.041	0	0.044	0
21	Life expectancy	0.055	0	0.242	0	0.931	1
22	Fraction Jewish	0.052	0	0.041	0	0.034	0
23	Degree of Capitalism	0.048	0	0.077	0	0.455	1
24	Revolutions and Coups	0.028	0	0.026	0	0.031	0
25	% of Pop. Speaking English	0.027	0	0.023	0	0.069	0
26	Ethnolinguistic Fraction.	0.026	0	0.021	0	0.057	0
27	Spanish Colony	0.025	0	0.021	0	0.057	0
28	Black Market Premium	0.023	0	0.035	0	0.179	0
29	Outward Orientation	0.020	0	0.017	0	0.038	0
30	Civil Liberties	0.016	0	0.020	0	0.131	0
31	War	0.016	0	0.016	0	0.076	0
32	% of Pop. Speak. For. Lang.	0.015	0	0.021	0	0.070	0
33	French Colony	0.015	0	0.016	0	0.052	0

Continued on next page

Table 12 – Continued

Index	Regressor	NCN	BM-NCN	NCC	BM-NCC	CN	BM-CN
34	British Colony	0.012	0	0.010	0	0.037	0
35	Political Rights	0.010	0	0.012	0	0.097	0
36	Latitude	0.001	0	0.005	0	0.044	0
37	Age	0.001	0	0.003	0	0.086	0
38	Exchange Rate Distortions	0.001	0	0.003	0	0.082	0
39	S.D. of Black Market Prem.	0.000	0	0.001	0	0.050	0
40	Area	0.000	0	0.000	0	0.030	0
41	Size of Labor Force	0.000	0	0.000	0	0.080	0

* The regressors are ranked according to Normal Non-Conjugate probabilities of inclusion. The values of the hyper-parameters of the Non-Conjugate Normal and Non-Conjugate Cauchy are $(a_0, b_0, \omega^{-2}) = (0.01, 0.001, 10)$ and $(a_0, b_0, \omega^{-2}) = (0.01, 0.001, 3)$, respectively. The estimation is based on 2 million draws, after discarding 1 million draws as burn-in sample. BM stands for Best Model, i.e. the model with the highest posterior probability.

Table 13 displays the BMA-Moments and the variables included in at least one of the 10 best models. One can see that the conjugate case includes, at least one time, 15 variables, while in the non-conjugate there are 13 variables. In NCN and NCC cases all 13 variables that are included at least once in the ten best models are identical. The fact that the conjugate case obtains 10-11 variables on average in the 10 best models, while the non-conjugate cases about 6-7 and in conjunction with the total number of variables that appear at least once in the 10 best models, means that the models in the non-conjugate case are diversified. This evidence also supports our claim that the non-conjugate case is favorable to more parsimonious models.

tabularxtabularx

Table 13: BMA Posterior Mean, SD and Variables Included in 10 Best Models

		NCN			NCC			CN		
Index	Regressor	M	SD	BM*	M	SD	BM*	M	SD	BM*
1	Equipment Investment	0.243	0.245	1	0.216	0.069	1	0.160	0.069	1
2	Fraction Confucian	0.065	0.067	1	0.063	0.020	1	0.057	0.015	1
3	No. Years Open Economy	0.009	0.015	1	0.010	0.009	1	0.007	0.008	1
4	GDP level in 1960	-0.005	0.010	1	-0.007	0.007	1	-0.016	0.003	1
5	Fraction Protestant	-0.009	0.016	1	-0.009	0.010	1	-0.006	0.007	1
6	Population Growth	-0.044	0.120	1	-0.001	0.065	1	0.005	0.047	0
7	Sub-Saharan Africa	-0.006	0.013	1	-0.007	0.009	1	-0.011	0.009	1
8	Non-equipment Investment	0.021	0.042	1	0.024	0.033	1	0.025	0.032	1
9	Public Education Share	-0.002	0.055	1	0.003	0.058	1	0.001	0.024	0
10	Fraction Buddhist	0.006	0.014	1	0.004	0.009	1	0.003	0.006	1
11	Primary School Enroll.	0.005	0.016	1	0.004	0.011	1	0.004	0.009	0
12	Fraction Muslim	0.002	0.008	1	0.003	0.006	1	0.009	0.008	1
13	Higher Education Enroll.	-0.008	0.027	1	-0.005	0.019	1	-0.002	0.011	0

Continued on next page

Table 13 – Continued

		NCN			NCC			CN		
Index	Regressor	M	SD	BM*	M	SD	BM*	M	SD	BM*
14	Rule of Law	0.002	0.008	0	0.003	0.007	0	0.007	0.008	1
15	Fraction GDP in Mining	0.003	0.011	0	0.007	0.016	0	0.019	0.023	1
16	Latin America	-0.001	0.004	0	-0.001	0.003	0	-0.002	0.004	0
17	Fraction Catholic	-0.001	0.003	0	-0.001	0.003	0	0.000	0.003	1
18	Life expectancy	0.000	0.000	0	0.000	0.000	0	0.001	0.000	1
19	Fraction Hindu	-0.001	0.005	0	-0.001	0.006	0	-0.004	0.012	1
20	Primary Exports	-0.001	0.004	0	-0.001	0.003	0	-0.001	0.004	0
21	Ratio Workers to Population	-0.001	0.004	0	0.000	0.003	0	0.000	0.002	0
22	Degree of Capitalism	0.000	0.001	0	0.000	0.001	0	0.001	0.001	1
23	Fraction Jewish	0.000	0.003	0	0.000	0.004	0	0.000	0.003	0
24	Black Market Premium	0.000	0.001	0	0.000	0.002	0	-0.001	0.003	0
25	Revolutions and Coups	0.000	0.001	0	0.000	0.001	0	0.000	0.001	0
26	% of Pop. Speaking English	0.000	0.001	0	0.000	0.001	0	0.000	0.002	0
27	Ethnolinguistic Fraction.	0.000	0.001	0	0.000	0.001	0	0.000	0.002	0
28	Spanish Colony	0.000	0.001	0	0.000	0.001	0	0.000	0.002	0
29	Outward Orientation	0.000	0.001	0	0.000	0.001	0	0.000	0.001	0
30	Civil Liberties	0.000	0.001	0	0.000	0.000	0	0.000	0.001	0
31	% of Pop. Speak. For. Lang.	0.000	0.001	0	0.000	0.001	0	0.000	0.001	0
32	French Colony	0.000	0.000	0	0.000	0.001	0	0.000	0.001	0
33	War	0.000	0.001	0	0.000	0.001	0	0.000	0.001	0
34	Political Rights	0.000	0.000	0	0.000	0.000	0	0.000	0.001	0
35	British Colony	0.000	0.000	0	0.000	0.000	0	0.000	0.001	0
36	Latitude	0.000	0.000	0	0.000	0.000	0	0.000	0.000	0
37	Exchange Rate Distortions	0.000	0.000	0	0.000	0.000	0	0.000	0.000	0
38	Age	0.000	0.000	0	0.000	0.000	0	0.000	0.000	0
39	S.D. of Black Market Prem.	0.000	0.000	0	0.000	0.000	0	0.000	0.000	0

Continued on next page

Table 13 – Continued

		NCN			NCC			CN		
Index	Regressor	M	SD	BM*	M	SD	BM*	M	SD	BM*
40	Area	0.000	0.000	0	0.000	0.000	0	0.000	0.000	0
41	Size of Labor Force	0.000	0.000	0	0.000	0.000	0	0.000	0.000	0

* M stands for BMA Posterior Mean and SD for BMA Posterior St.Dev., see (2) and (3) in conjunction with Theorems 1 and 2 and Appendix Appendix C. BM* stands for the 10 best models with highest posterior probability and equals 1 if the regressor is included in at least one of the 10 best models.

5.2. Predictive performance

We measure predictive performance using the Log Predictive Score, (equation 22) defined in section 4.4. Following Fernandez et al. (2001a,b), we split the sample in two parts $n = n_{ins} + n_{oos}$, where n_{ins} is regarded as the in-sample observations where the posterior analysis is conducted and n_{oos} is regarded as the unobserved future observations (see Zellner, 1971, p. 28), necessary to evaluate the out-of-sample performance of the model:

$$LPS = -\frac{1}{n_{oos}} \sum_{f=1}^{n_{oos}} \ln p\left(y_f|\mathbf{y}\right) \tag{22}$$

We set $n_{oos} = n/4$, where $n = 72$. In the empirical study we proceed to 20 random partitions of the original sample to in-sample and out of sample observations. Finally, in Table 14, we present the mean, the minimum and maximum values of the 20 scores. It is clear that the non-conjugate prior cases obtain better predictive performance in terms of LPS. NCN has the best overall performance, with the lowest LPS minimum, mean and maximum values, followed by the NCC, with the conjugate prior in the last place.

These results are somewhat different to the ones produced by the numerical experiment, where conjugate and non-conjugate priors had roughly similar predictive performance. This fact indicates that in real datasets with a large number of diverse types of covariates that are not necessarily Normally distributed, the flexibility of the non-conjugate priors allows them to obtain superior predictive results, in line with the thesis in this paper.

Table 14: Logarithmic Posterior Score*

	Min	Mean	Max
Conjugate Normal	-3.335	-2.559	-1.217
Non-Conjugate Normal	-6.225	-5.105	-4.340
Non-Conjugate Cauchy	-3.716	-3.543	-3.407

* We proceed to 20 different partitions of the original sample, where $n_{\text{oos}} = 18$ observations. We report the minimum, maximum and mean values of the estimated 20 logarithmic predictive scores.

5.3. Approximation and robustness diagnostics

In the Appendix we provide a detailed analysis of the Laplace approximation error (in Appendix E), as well as other variations of our empirical design for robustness purposes (in Appendix F). The main conclusions are as follows. First, a bridge sampling benchmark confirms that the Laplace approximations used for the non-conjugate marginal likelihoods deliver stable posterior model summaries.[15] For the non-conjugate Normal prior, the mean log-evidence discrepancy between Laplace and bridge sampling is modest,

[15]Bridge sampling is a simulation-based method for estimating ratios of normalising constants, and hence marginal likelihoods, by linking the target posterior density to a proposal density through a bridge identity (see, for example Meng and Wong, 1996).

and posterior summaries are almost unchanged: the posterior mean model size changes only from 5.65 to 5.76, the top-10 posterior mass is virtually identical, and eight of the ten highest-probability models are common across the two methods. For the non-conjugate Cauchy prior, the raw log-evidence discrepancy is larger, but it is mainly an additive shift that has limited impact on Bayes factors and posterior odds; the posterior mean model size changes from 5.75 to 6.06, and the top-10 model overlap is again eight out of ten. The resulting posterior inclusion probabilities remain close to the Laplace-based estimates, with maximum absolute differences below 0.05 for the non-conjugate Normal prior and below 0.10 for the non-conjugate Cauchy prior. Second, the empirical conclusions are robust to alternative MC(3) proposal mechanisms. An extended sampler that combines single-variable Birth–Death moves with grouped-covariate and swap moves produces virtually the same posterior output as the baseline Birth–Death sampler. For the non-conjugate Normal prior, the posterior mean model size and top-10 posterior mass are identical across the two samplers; for the non-conjugate Cauchy prior, they are almost unchanged. Posterior inclusion probabilities differ by less than one percentage point, and the ranking of the main growth determinants is unaffected. Third, the results are not driven by the baseline uniform model-space prior. Replacing it with Beta-Binomial, loss-based, and Poisson model-size priors affects the overall degree of sparsity, as expected, but does not alter the main ranking across parameter priors. Under all model-space priors considered, the two non-conjugate specifications select more parsimonious models than the conjugate Normal benchmark, while also delivering better predictive performance. The non-conjugate Normal prior remains the

best-performing specification, followed by the non-conjugate Cauchy prior, with the conjugate Normal prior performing worst.

6. Conclusion

In this paper we highlight the effects of prior assumptions on model uncertainty in the linear regression model, by employing non-conjugate priors for the parameters of the explanatory variables in a BMA setting. We use two different priors which can be regarded as two limiting cases of the multivariate student-t distribution. In the first case we assume that the degrees of freedom tend to infinity so, a non-conjugate multivariate normal arises and in the second case we set the degrees of freedom equal to one, where the resulting prior is a multivariate dependent Cauchy distribution. We estimate the approximated marginal likelihoods, necessary for the efficiency of the $MC\,(3)$, the BMA moments and predictive posterior. We apply the proposed methods to a controlled numerical experiment, as well as the growth regressions data of Fernandez et al. (2001b). Our results indicate that, while in the context of a synthetic dataset the non-conjugate priors perform similarly to the conjugates, while outperforming them in some metrics, they clearly outperform the conjugates in the growth dataset. Specifically, we find some striking differences regarding the factors that determine growth. The non-conjugate approach is more stringent, in the sense that the average number of covariates is less than that of the conjugate case (this holds for the best model, as well as the 10 best models), which suggests that the non-conjugate case is in favor of selecting more parsimonious models. In line with criticisms of the natural conjugate setting, we find that the predictive

criteria support the superiority of the non-conjugate choice vs the conjugate alternative for the growth data. The superior predictive performance of the proposed setting, in conjunction with the efficiency of the adopted approximation methods, make the results presented in the paper relevant not only to academics, but to practitioners as well.

Appendix A. General Rules

In the appendix, we derive the proofs of the algebraic results presented in the main text. First, we expose some propositions.[16]

Proposition 1. *Assume the following invertible matrices* $\mathbf{A} = \mathbf{A}'$, $\mathbf{B} = \mathbf{B}'$. *Define a function on* $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ *and the matrix* $\mathbf{C}_\tau = \varphi(\tau)\,\mathbf{A} + \lambda\mathbf{I}$. *Then the following derivatives hold:*

$$\frac{\partial tr\ \mathbf{C}_\tau^{-1}\mathbf{B}}{\partial \tau} = -\varphi'(\tau)\, tr\left(\ \mathbf{C}_\tau^{-1}\mathbf{B}\ \mathbf{C}_\tau^{-1}\mathbf{A}\right) \qquad \text{(Appendix A.1)}$$

$$\frac{\partial \ln|\mathbf{C}_\tau|}{\partial \tau} = \varphi'(\tau)\, tr\left(\ \mathbf{C}_\tau^{-1}\mathbf{A}\right) \qquad \text{(Appendix A.2)}$$

$$\frac{\partial tr\ \mathbf{C}_\tau^{-1}\mathbf{B}\ \mathbf{C}_\tau^{-1}\mathbf{A}}{\partial \tau} = -2\varphi'(\tau)\, tr\ \mathbf{C}_\tau^{-1}\mathbf{B}\left(\ \mathbf{C}_\tau^{-1}\mathbf{A}\right)^2 \qquad \text{(Appendix A.3)}$$

[16] For the application of the necessary matrix calculus techniques, the reader may refer to Turkington (2002) and Magnus and Neudecker (2007).

Proof:

$$\begin{aligned}
\frac{\partial tr\ \mathbf{C}_\tau^{-1}\mathbf{B}}{\partial \tau} &= \frac{\partial vec\left(\mathbf{C}_\tau\right)}{\partial \tau}\frac{\partial vec\ \mathbf{C}_\tau^{-1}}{\partial vec\left(\mathbf{C}_\tau\right)}\frac{\partial tr\ \mathbf{C}_\tau^{-1}\mathbf{B}}{\partial vec\ \mathbf{C}_\tau^{-1}} \\
&= \varphi'\left(\tau\right) vec\left(\mathbf{A}\right)' \left(-\ \mathbf{C}_\tau^{-1} \otimes\ \mathbf{C}_\tau^{-1}\right) vec\left(\mathbf{B}\right) \\
&= -\varphi'\left(\tau\right) tr\ \mathbf{C}_\tau^{-1}\mathbf{B}\ \mathbf{C}_\tau^{-1}\mathbf{A}
\end{aligned}$$

$$\frac{\partial \ln\left|\mathbf{C}_\tau\right|}{\partial \tau} = \varphi'\left(\tau\right) vec\left(\mathbf{A}\right)' vec\ \mathbf{C}_\tau^{-1} = \varphi'\left(\tau\right) tr\ \mathbf{C}_\tau^{-1}\mathbf{A}$$

$$\begin{aligned}
\frac{\partial tr\ \mathbf{C}_\tau^{-1}\mathbf{B}\ \mathbf{C}_\tau^{-1}\mathbf{A}}{\partial \tau} &= \frac{\partial vec\left(\mathbf{C}_\tau\right)}{\partial \tau}\frac{\partial vec\ \mathbf{C}_\tau^{-1}}{\partial vec\left(\mathbf{C}_\tau\right)}\frac{\partial tr\ \mathbf{C}_\tau^{-1}\mathbf{B}\ \mathbf{C}_\tau^{-1}\mathbf{A}}{\partial vec\ \mathbf{C}_\tau^{-1}} \\
&= \varphi'\left(\tau\right) vec\left(\mathbf{A}\right)' \left(-\ \mathbf{C}_\tau^{-1} \otimes\ \mathbf{C}_\tau^{-1}\right) \left(2vec\mathbf{A}\ \mathbf{C}_\tau^{-1}\mathbf{B}\right) \\
&= -2\varphi'\left(\tau\right) vec\left(\mathbf{A}\right)' \left(\ \mathbf{C}_\tau^{-1} \otimes\ \mathbf{C}_\tau^{-1}\right) vec\left(\mathbf{A}\ \mathbf{C}_\tau^{-1}\mathbf{B}\right) \\
&= -2\varphi'\left(\tau\right) tr\ \mathbf{C}_\tau^{-1}\mathbf{B}\ \mathbf{C}_\tau^{-1}\mathbf{A}\ \mathbf{C}_\tau^{-1}\mathbf{A} \\
&= -2\varphi'\left(\tau\right) tr\ \mathbf{C}_\tau^{-1}\mathbf{B}\left(\ \mathbf{C}_\tau^{-1}\mathbf{A}\right)^2
\end{aligned}$$

Proposition 2. *Assume that* $\mathbf{A}_k = \mathbf{A}_k'$ *and* $f : \mathbb{R}^k \to \mathbb{R}$. *Set the function*

$$\mathbf{\Phi}\left(\boldsymbol{\beta}\right) = f\left(\boldsymbol{\beta}\right)\left(\boldsymbol{A}_k\left(\boldsymbol{\beta}-\boldsymbol{\alpha}\right)\left(\boldsymbol{\beta}-\boldsymbol{\alpha}\right)'\mathbf{A}_k'\right)$$

then:

$$D\mathbf{\Phi}\left(\boldsymbol{\beta}\right) = Df\left(\boldsymbol{\beta}\right)\otimes\left(\mathbf{A}_k\left(\boldsymbol{\beta}-\boldsymbol{\alpha}\right)\left(\boldsymbol{\beta}-\boldsymbol{\alpha}\right)'\mathbf{A}_k'\right)+f\left(\boldsymbol{\beta}\right)\left(\mathbf{A}_k\left(\boldsymbol{\beta}-\boldsymbol{\alpha}\right)\otimes\mathbf{I}+\mathbf{I}\otimes\mathbf{A}_k\left(\boldsymbol{\beta}-\boldsymbol{\alpha}\right)\right)\mathbf{A}_k$$

where, D is the matrix derivative operator.

Proof: see Turkington (2002) page 83, Table 4.1 (General Rules).

Appendix B. Marginal Likelihood Estimation

Appendix B.1. Non-Conjugate Normal Prior

In this part we derive of the algebraic results for the estimation of the marginal likelihood of the data, given the assumptions stated in section 3.1.

$$f\left(\mathbf{y};M_j\right)=\int_{\sigma>0}\int_{\mathbb{R}^k}\int_{\mathbb{R}}\mathcal{L}\left(\alpha,\boldsymbol{\beta},\sigma|\mathbf{y},M_j\right)p\left(\boldsymbol{\beta}\right)p\left(\sigma\right)p\left(\alpha\right)d\alpha d\boldsymbol{\beta}d\sigma$$

Notice that the likelihood function is

$$\mathcal{L}\left(\alpha,\boldsymbol{\beta},\sigma|\mathbf{y},M_j\right)=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{\left(\mathbf{y}-\alpha\boldsymbol{\iota}-\boldsymbol{Z\beta}\right)'\left(\mathbf{y}-\alpha\boldsymbol{\iota}-\boldsymbol{Z\beta}\right)}{2\sigma^2}\right)$$

Assuming a flat prior, we integrate over the intercept, yielding the joint posterior distribution of the regressor coefficients and the scale parameter, as shown in equation (Appendix B.1), which we rewrite below.

$$p\left(\beta,\sigma|\mathbf{y},M_j\right)\propto p\left(\boldsymbol{\beta}\right)p\left(\sigma\right)\sigma^{-n+1}\exp\left(-\frac{1}{2\sigma^2}\left(\left(\mathbf{y}-\boldsymbol{Z\beta}\right)'\left(\mathbf{y}-\boldsymbol{Z\beta}\right)-n\overline{y}^2\right)\right)$$

The marginal posterior of the scale parameter, $\sigma|\mathbf{y},M_j$, is obtained by integrating equation (Appendix B.1) with respect to the vector $\boldsymbol{\beta}$, under the assumption of a non-conjugate normal prior for $\boldsymbol{\beta}$, as specified in equation (9) of the main paper. Using the properties of the multivariate normal distribution, we derive the following expression:

$$p\left(\sigma|\mathbf{y},M_j\right)\propto\sigma^{-(n+a_0)}\left|\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2}+\frac{\mathbf{I}_k}{\omega^2}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(Q_\sigma+b_0\right)\right)$$

The following properties can be easily demonstrated:

B1. $\left|\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2}+\frac{\mathbf{I}_k}{\omega^2}\right| = \left(\frac{n\omega^2+\sigma^2}{\omega^2\sigma^2}\right)^{-1}\left|\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right|$

B2. $\mathbf{Z}'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2}+\frac{\mathbf{I}_k}{\omega^2}\right)\mathbf{Z} = \mathbf{X}'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right)\mathbf{X} - \frac{n\omega^2+\sigma^2}{\omega^2\sigma^2}\mathbf{J}_n$

Thus, the marginal posterior is given by:

$$p\left(\sigma|\mathbf{y},M_j\right) \propto g_2\left(\sigma\right)^{0.5}\sigma^{-(n+a_0+1)}\left|\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(Q_\sigma+b_0\right)\right)$$

Given the transformation $\tau = \ln\sigma$, the transformed marginal posterior of σ becomes:

$$p\left(\tau|\mathbf{y},M_j\right) \propto g_2\left(\tau\right)^{0.5}\exp\left(-\left(n+a_0\right)\tau\right)\left|\frac{\mathbf{X}'\mathbf{X}}{e^{2\tau}}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right|^{-0.5}\exp\left(-\frac{1}{2e^{2\tau}}\left(Q_\tau+b_0\right)\right)$$

According to (6) we set the following function:

$$h_n^N\left(\tau\right) = \frac{1}{2n}\ln g_2\left(\tau\right) - \frac{\left(n+a_0\right)\tau}{n} - \frac{1}{2n}\ln\left|\frac{\mathbf{X}'\mathbf{X}}{e^{2\tau}}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right| - \frac{1}{2ne^{2\tau}}\left(Q_\tau+b_0\right)$$

In order to estimate $\tau^* = \arg\max h_n^N\left(\tau\right)$, we proceed to numerical optimization procedures based in proper *MATLAB* routines. By applying proposition 1, we obtain the first and second derivatives of $h_n^N\left(\tau\right)$:

$$\begin{aligned}
\frac{dh_n^N\left(\tau\right)}{d\tau} &= \frac{1}{n}e^{-2\tau}\left(b_0+\underline{\mathbf{y}}'\underline{\mathbf{y}} - 2e^{-2\tau}tr\mathbf{C}_\tau^{-1}\mathbf{X}'\underline{\mathbf{y}\mathbf{y}}'\mathbf{X} + e^{-4\tau}tr\mathbf{C}_\tau^{-1}\mathbf{X}'\underline{\mathbf{y}\mathbf{y}}'\mathbf{X}\mathbf{C}_\tau^{-1}\mathbf{X}'\mathbf{X}\right.\\
&\quad \left.+tr\mathbf{C}_\tau^{-1}\mathbf{X}'\mathbf{X}\right) + e^{2\tau}/n\left(n\omega^2+e^{2\tau}\right) - \left(n+a_0\right)/n\\
\frac{d^2h_n^N\left(\tau\right)}{d\tau^2} &= n^{-1}tr\left(-2e^{-2\tau}\left(b_0+\underline{\mathbf{y}}'\underline{\mathbf{y}}\right)\left(\mathbf{I}_{k+1}/\left(k+1\right)\right) + 8e^{-4\tau}\mathbf{C}_\tau^{-1}\mathbf{X}'\underline{\mathbf{y}\mathbf{y}}'\mathbf{X}\right.\\
&\quad -10e^{-6\tau}\mathbf{C}_\tau^{-1}\mathbf{X}'\underline{\mathbf{y}\mathbf{y}}'\mathbf{X}\mathbf{C}_\tau^{-1}\mathbf{X}'\mathbf{X} + 4e^{-8\tau}\mathbf{C}_\tau^{-1}\mathbf{X}'\underline{\mathbf{y}\mathbf{y}}'\mathbf{X}\left(\mathbf{C}_\tau^{-\mathbf{1}}\mathbf{X}'\mathbf{X}\right)^2\\
&\quad \left.-2e^{-2\tau}\left(\mathbf{C}_\tau^{-\mathbf{1}}\mathbf{X}'\mathbf{X}\right) + 2e^{-4\tau}\left(\mathbf{C}_\tau^{-\mathbf{1}}\mathbf{X}'\mathbf{X}\right)^2\right) + 2\omega^2e^{2\tau}/\left(n\omega^2+e^{2\tau}\right)^2
\end{aligned}$$

where, $\mathbf{C}_\tau = \frac{\mathbf{X}'\mathbf{X}}{e^{2\tau}} + \frac{\mathbf{I}_{k+1}}{\omega^2}$ is a part of the normalising constant. The marginal likelihood through equation (6) of the main paper becomes

$$f(\mathbf{y}|M_j) \propto \int_{\mathbb{R}} p(\tau|\mathbf{y}, M_j)\, d\tau \approx \int_{\mathbb{R}} \exp n h_n^N(\tau)\, d\tau \approx \sqrt{2\pi/n\left(-h_n^{N\prime\prime}(\tau^*)\right)} \exp\left(n h_n^N(\tau^*)\right)$$

and after the proper replacements we obtain

$$f(\mathbf{y}|M_j) \propto c_{\tau^*} \left| \frac{\mathbf{X}'\mathbf{X}}{e^{2\tau^*}} + \frac{\mathbf{I}_{k+1}}{\omega^2} \right|^{-1/2} \exp\left(\left(-\frac{1}{2e^{2\tau^*}} \left(Q_{\tau^*} + b_0\right) - \left(n + a_0\right)\tau^* \right) \right)$$

where, $c_{\tau^*} = \omega^{-k}\left(-h_n^{N\prime\prime}(\tau^*)\right)^{-0.5} g_2(\tau^*)^{0.5}$ is a part of the normalising constant.[17]. Taking the log of this expression gives equation (10) of the main paper.

Appendix B.2. Non-Conjugate Dependent Cauchy Prior

The other prior we consider is the multivariate dependent Cauchy distribution, equivalently a multivariate Student's t distribution with one degree of freedom (i.e., $v = 1$), which simplifies as (5e) in the main paper. By marginalizing over the intercept α from the joint posterior distribution of $(\alpha, \boldsymbol{\beta}, \sigma)$, we obtain the joint posterior distribution of $(\boldsymbol{\beta}, \sigma)$, given by:

$$p(\boldsymbol{\beta}, \sigma|\mathbf{y}, M_j) \propto p(\boldsymbol{\beta})\, p(\sigma)\, \sigma^{-n+1} \exp\left(-\frac{1}{2\sigma^2} \left((\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) - n\bar{y}^2 \right) \right) \quad \text{(Appendix B.1)}$$

[17] Throughout this paper, after computing the functions of interest, we report the remaining constant term, which is not necessarily the integrating constant. We include this because part of the constant depends on the specific model j, and it is essential to account for this in numerical computations.

The marginal posterior of $\boldsymbol{\beta}$ can easily be derived after the integration of (Appendix B.1) with respect to σ. Using the properties of inverse gamma distribution, we obtain (13) of the main paper.

$$p\left(\boldsymbol{\beta}|\mathbf{y}, M_j\right) \propto \left(vs_{b_0}^2\right)^{-(v+k)/2} \left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-\frac{1+k}{2}} \left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)' \frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-\frac{v+k}{2}}$$

According to (7), we set the following function:

$$h_n^C\left(\boldsymbol{\beta}\right) = n^{-1}\ln\left(\left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-\frac{1+k}{2}} \left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)' \frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{\frac{v+k}{2}}\right)$$

we also estimate, numerically, $\widetilde{\boldsymbol{\beta}} = \arg\max h_n^C\left(\boldsymbol{\beta}\right)$ using the suitable *MATLAB* routines. From the application of proposition 1, we obtain the first and second derivatives of $h_n^C\left(\boldsymbol{\beta}\right)$.

$$\begin{aligned}
\frac{\partial h_n^C\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} &= -\frac{v+k}{nv}\frac{\mathbf{Z}_s\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{Z}_s\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)} - \frac{1+k}{n\omega^2}\frac{\boldsymbol{\beta}}{1+\boldsymbol{\beta}'\boldsymbol{\beta}/\omega^2} \\
\frac{\partial^2 h_n^C\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= -\frac{v+k}{nv}\frac{\mathbf{Z}_s}{1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{Z}_s\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)} + 2\frac{v+k}{nv^2}\frac{\mathbf{Z}_s\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{Z}_s}{\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\mathbf{Z}_s\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^2} \\
&\quad -\frac{1+k}{n\omega^2}\frac{\mathbf{I}_k}{1+\boldsymbol{\beta}'\boldsymbol{\beta}/\omega^2} + 2\frac{1+k}{n\omega^4}\frac{\boldsymbol{\beta}\boldsymbol{\beta}'}{\left(1+\boldsymbol{\beta}'\boldsymbol{\beta}/\omega^2\right)^2}
\end{aligned}$$

where $\mathbf{Z}_s = \frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}$. Finally, we obtain the marginal likelihood, whose logarithm is given in (13).

$$f\left(\mathbf{y}; M_j\right) \propto c_{\widetilde{\boldsymbol{\beta}}}\left(vs_{b_0}^2\right)^{-(v+k)/2}\left|\mathbf{H}_n^*\right|^{-1/2}\exp\left(nh_n^C\left(\widetilde{\boldsymbol{\beta}}\right)\right)$$

Appendix C. Moments Estimation

Appendix C.1. MGF Approximation and Approximated Posterior Moments

Our task is to approximate the moments of vector $\boldsymbol{\beta}$ (see section 3.1 of the paper), following Tierney et al. (1989). The moment generating function (hereafter MGF) can be approximated by

$$\widehat{M_g}(s) = \frac{\widehat{b}_N \left|\mathbf{H}_N\right|^{-\frac{1}{2}} \exp\left(-n\widehat{h}_N\right)}{\widehat{b}_D \left|\mathbf{H}_D\right|^{-\frac{1}{2}} \exp\left(-n\widehat{h}_D\right)} + O\left(n^{-2}\right)$$

We use the full exponential form where, $b_N = b_D$ and in particular $b_N = b_D = 1$, so we obtain that

$$\widehat{M_g}(s) \approx \frac{\left|\mathbf{H}_N\right|^{-\frac{1}{2}} \exp\left(-n\widehat{h}_N\right)}{\left|\mathbf{H}_D\right|^{-\frac{1}{2}} \exp\left(-n\widehat{h}_D\right)}$$

Theorem 1 *(Multivariate Version of Theorem 2a of Tierney et al., 1989)*

$$\widehat{E}\left(g_l(\boldsymbol{\beta})\left|\mathbf{y}, M_j\right.\right) = g_l\left(\widehat{\boldsymbol{\beta}}_D\right) - \frac{1}{2}\left.\frac{\partial \ln\left|\mathbf{H}_s\right|}{\partial s}\right|_{s=0} \qquad \text{(Appendix C.1)}$$

where $\widehat{\mathbf{H}}_i = \left.\frac{\partial^2 h_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_i}$, $\widehat{\boldsymbol{\beta}}_i = \arg\max\left(-h_i\right)$ for $i = s, N, D$, and $h_s \equiv h_N$, $h \equiv h_D$, $h_s(\boldsymbol{\beta}) = h(\boldsymbol{\beta}) - \frac{1}{n} s g_l(\boldsymbol{\beta})$, for $l = 1, \ldots, k$ and $nh(\boldsymbol{\beta})$ is the negative log-posterior.

Proof: Like Tierney et al. (1989), we proceed as follows:

$$\ln \widehat{M_g}(s) = \ln\left|\mathbf{H}_s\right|^{-\frac{1}{2}} - nh_s(\boldsymbol{\beta}) - \ln\left|\mathbf{H}\right|^{-\frac{1}{2}} + nh(\boldsymbol{\beta})$$

then

$$\begin{aligned}
\widehat{E}\left(g_l\left(\boldsymbol{\beta}\right)|\mathbf{y},M_j\right) &= \left.\frac{\partial \ln \widehat{M_g}(s)}{\partial s}\right|_{s=0} \\
&= \left.\frac{\partial\left(\ln\left|\mathbf{H}_s\right|^{-\frac{1}{2}} - nh_s\left(\boldsymbol{\beta}_s\right)\right)}{\partial s}\right|_{s=0} \\
&= -n\left.\frac{\partial h_s\left(\boldsymbol{\beta}_s\right)}{\partial s}\right|_{s=0} - \frac{1}{2}\left.\frac{\partial \ln\left|\mathbf{H}_s\right|}{\partial s}\right|_{s=0} \\
&= g_l\left(\widehat{\boldsymbol{\beta}}_D\right) - \frac{1}{2}\left.\frac{\partial \ln\left|\mathbf{H}_s\right|}{\partial s}\right|_{s=0}
\end{aligned}$$

since,

$$\begin{aligned}
\left.\frac{\partial h_s\left(\boldsymbol{\beta}_s\right)}{\partial s}\right|_{s=0} &= \left.\left(\frac{\partial h\left(\boldsymbol{\beta}_s\right)}{\partial \boldsymbol{\beta}_s}\frac{\partial \boldsymbol{\beta}_s}{\partial s} - \frac{s}{n}\frac{\partial g_l\left(\boldsymbol{\beta}_s\right)}{\partial \boldsymbol{\beta}_s}\frac{\partial \boldsymbol{\beta}_s}{\partial s} - \frac{1}{n}g_l\left(\boldsymbol{\beta}_s\right)\right)\right|_{s=0} \\
\boldsymbol{\beta}_{s=0} &= \widehat{\boldsymbol{\beta}}_D = \arg\max\left(-h_D\right) \\
\left.\frac{\partial h_s\left(\boldsymbol{\beta}_s\right)}{\partial s}\right|_{s=0} &= -\frac{1}{n}g_l\left(\widehat{\boldsymbol{\beta}}_D\right)
\end{aligned}$$

Theorem 2 *(Multivariate Version of Theorem 3 of Tierney et al., 1989)*

$$\widehat{E}\left(g_l\left(\boldsymbol{\beta}\right)|\mathbf{y},M_j\right) = g_l\left(\widehat{\boldsymbol{\beta}}_D\right) - \frac{1}{2n}\left(\left(\frac{\partial \widehat{g}_l}{\partial \boldsymbol{\beta}}\right)'\widehat{\mathbf{H}}^{-1}\left(\frac{\partial \widehat{\mathbf{H}}}{\partial \boldsymbol{\beta}}\right)' - vec\left(\mathbf{G}_l\right)'\right)vec\left(\widehat{\mathbf{H}}^{-1}\right) \qquad \text{(Appendix C.2)}$$

where, $\frac{\partial \widehat{g}_l}{\partial \boldsymbol{\beta}} \equiv \left.\frac{\partial g_l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$, $\widehat{\mathbf{H}} \equiv \left.\frac{\partial^2 h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$, $\widehat{\mathbf{G}}_l = \left.\frac{\partial^2 g_l(\boldsymbol{\beta})}{\partial \beta \partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$

Proof:

From theorem 1 we have that $\widehat{E}\left(g_l\left(\boldsymbol{\beta}\right)|\mathbf{y},M_j\right) = g_l\left(\widehat{\beta}_D\right) - \frac{1}{2}\left.\frac{\partial \ln|\mathbf{H}_s|}{\partial s}\right|_{s=0}$,

then the following holds

$$\frac{\partial \ln |\mathbf{H}_s|}{\partial s} = \left(\frac{\partial vec\mathbf{H}_s}{\partial s} \right)' vec\left(\mathbf{H}_s^{-1}\right)$$

Taking the second derivative of h_s with respect to $\boldsymbol{\beta}$

$$\begin{aligned}
\frac{\partial^2 h_s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= \frac{\partial^2 h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} - \frac{s}{n} \frac{\partial^2 g_l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \\
\mathbf{H}_s &= \mathbf{H} - \frac{s}{n} \mathbf{G}_l \\
\mathbf{G}_l &= \frac{\partial^2 g_l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}
\end{aligned}$$

So,

$$\begin{aligned}
\left. \frac{\partial vec\mathbf{H}_s}{\partial s} \right|_{s=0} &= \left. \frac{\partial}{\partial s} \left(vec\mathbf{H}_s - \frac{s}{n} vec(\mathbf{G}_l) \right) \right|_{s=0} \\
&= \left. \left(\frac{\partial \mathbf{H}_s}{\partial \boldsymbol{\beta}_s} \frac{\partial \boldsymbol{\beta}_s}{\partial s} - \frac{vec(\mathbf{G}_l)}{n} \right) \right|_{s=0}
\end{aligned}$$

Notice that

$$\begin{aligned}
\left. \frac{\partial h_s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_s} &= 0 \\
\left. \left(\frac{\partial}{\partial s} \left(\frac{\partial h_s(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) - \frac{s}{n} \frac{\partial g_l}{\partial s} \right) \right|_{s=0} &= 0 \\
\left. \left(\frac{\partial^2 h_s}{\partial \boldsymbol{\beta}_s \partial \boldsymbol{\beta}_s'} \frac{\partial \boldsymbol{\beta}_s}{\partial s} - \frac{1}{n} \frac{\partial g_l}{\partial \boldsymbol{\beta}_s} \right) \right|_{s=0} &= 0 \\
\left. \frac{\partial \boldsymbol{\beta}_s}{\partial s} \right|_{s=0} &= \left. \left(\frac{1}{n} \mathbf{H}_s^{-1} \frac{\partial g_l}{\partial \boldsymbol{\beta}_s} \right) \right|_{s=0} \\
\left. \frac{\partial \boldsymbol{\beta}_s}{\partial s} \right|_{s=0} &= \frac{1}{n} \widehat{\mathbf{H}}^{-1} \frac{\partial \widehat{g}_l}{\partial \boldsymbol{\beta}}
\end{aligned}$$

finally,

$$\left.\frac{\partial vec\mathbf{H}_s}{\partial s}\right|_{s=0} = \frac{1}{n}\left(\underbrace{\frac{\partial \widehat{\mathbf{H}}}{\partial \boldsymbol{\beta}}}_{k^2\times k}\underbrace{\widehat{\mathbf{H}}^{-1}}_{k\times k}\underbrace{\frac{\partial \widehat{g}_l}{\partial \boldsymbol{\beta}}}_{k\times 1} - \underbrace{vec\left(\widehat{\mathbf{G}}_l\right)}_{k^2\times 1}\right)$$

thus,

$$\widehat{E}\left(g_l\left(\boldsymbol{\beta}\right)|\mathbf{y}, M_j\right) = g_l\left(\widehat{\boldsymbol{\beta}}_D\right) - \frac{1}{2n}\left(\left(\frac{\partial \widehat{g}_l}{\partial \boldsymbol{\beta}}\right)' \widehat{\mathbf{H}}^{-1}\left(\frac{\partial \widehat{\mathbf{H}}}{\partial \boldsymbol{\beta}}\right)' - vec\left(\widehat{\mathbf{G}}_l\right)'\right) vec\left(\widehat{\mathbf{H}}^{-1}\right)$$

In the subsequent section, we estimate the moments for the $\boldsymbol{\beta}$ parameters given two assumptions about the prior distributions. For this purpose, we construct the auxiliary variable-vector $\boldsymbol{\lambda}_l$, for the estimation of the expectation of the function $g_l(\boldsymbol{\beta})$, for $l = 1, \ldots, k$, which is not necessarily positive.

For this purpose, $g_l(\boldsymbol{\beta})$ should either be equal to $\boldsymbol{\lambda}_l'\boldsymbol{\beta} = \beta_l$ for the approximation of the mean, or $\boldsymbol{\beta}'\boldsymbol{\lambda}_l\boldsymbol{\lambda}_l'\boldsymbol{\beta} = \beta_l^2$ for the approximation of the variance. The auxiliary vector $\boldsymbol{\lambda}_l$ is defined as:

$$\boldsymbol{\lambda}_l = \begin{cases} 1 & \text{at position } l \\ 0 & \text{at all other positions} \end{cases} \qquad \text{for } l = 1, \ldots, k.$$

Finally, for the calculation of $\frac{\partial \widehat{\mathbf{H}}}{\partial \boldsymbol{\beta}}$, we apply Proposition 2 from section Appendix A.

Appendix C.2. Application to Non-Conjugate Normal Prior

In order to estimate the moments for the non-conjugate normal case we integrate the scale parameter from (Appendix B.1), assuming the normal non-conjugate prior (9), and obtain the joint posterior distribution of vector

$\boldsymbol{\beta}$

$$p\left(\boldsymbol{\beta}|\mathbf{y},M_j\right)\propto\left(vs_{b_0}^2\right)^{-(v+k)/2}\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-\frac{v+k}{2}}\exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\omega^2}\right)$$

$$p\left(\boldsymbol{\beta}\,|\mathbf{y},M_j\,\right)=\frac{\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-(v+k)/2}\exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\omega^2}\right)}{\int_{\mathbb{R}^k}\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-(v+k)/2}\exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\omega^2}\right)d\boldsymbol{\beta}}$$

According to Tierney et al. (1989), we obtain:

$$\begin{aligned}
\widehat{M}_g\left(s\right) &= \frac{\int_{\mathbb{R}^k}\exp\left(sg_l\left(\boldsymbol{\beta}\right)\right)\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-(v+k)/2}\exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\omega^2}\right)d\boldsymbol{\beta}}{\int_{\mathbb{R}^K}\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{s_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-(v+k)/2}\exp\left(-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\omega^2}\right)d\boldsymbol{\beta}} \\
&= \frac{\widehat{b}_N\left|\mathbf{H}_N\right|^{-\frac{1}{2}}\exp\left(-n\widehat{h}_N\right)}{\widehat{b}_D\left|\mathbf{H}_D\right|^{-\frac{1}{2}}\exp\left(-n\widehat{h}_D\right)}+O\left(n^{-2}\right) \\
&= \frac{\left|\mathbf{H}_N\right|^{-\frac{1}{2}}}{\left|\mathbf{H}_D\right|^{-\frac{1}{2}}}\frac{\exp\left(n\left(\frac{sg_l\left(\widehat{\boldsymbol{\beta}}_N\right)}{n}-\frac{\widehat{\boldsymbol{\beta}}_N'\widehat{\boldsymbol{\beta}}_N}{2n\omega^2}-\frac{v+k}{2n}\ln\left(1+\frac{1}{v}\left(\widehat{\boldsymbol{\beta}}_N-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\widehat{\boldsymbol{\beta}}_N-\widehat{\boldsymbol{\beta}}\right)\right)\right)\right)}{\exp\left(n\left(-\frac{\widehat{\boldsymbol{\beta}}_D'\widehat{\boldsymbol{\beta}}_D}{2n\omega^2}-\frac{v+k}{2n}\ln\left(1+\frac{1}{v}\left(\widehat{\boldsymbol{\beta}}_D-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\widehat{\boldsymbol{\beta}}_D-\widehat{\boldsymbol{\beta}}\right)\right)\right)\right)}
\end{aligned}$$

In the final equation, we set $b_N=b_D=1$, since we apply the full exponential form and ignore the error term $O\left(n^{-2}\right)$.

Mean Approximation. According to the multivariate version of Theorem 3 of Tierney et al. (1989), theorem 2, and by setting $g_l\left(\boldsymbol{\beta}\right)=\boldsymbol{\lambda}_l'\boldsymbol{\beta}=\beta_l$, we obtain:

$$\widehat{E}\left(\beta_l\,|\mathbf{y},M_j\,\right)=\widehat{\beta}_{lD}-\frac{1}{2n}\boldsymbol{\lambda}_l'\widehat{\mathbf{H}}^{-1}\left(\frac{\partial\widehat{\mathbf{H}}}{\partial\boldsymbol{\beta}}\right)'vec\left(\widehat{\mathbf{H}}^{-1}\right)$$

$$h\left(\boldsymbol{\beta}\right)=\frac{\boldsymbol{\beta}'\beta}{2n\omega^{2}}+\frac{v+k}{2n}\ln\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^{2}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)$$

$$\frac{\partial h\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}=\frac{\boldsymbol{\beta}}{n\omega^{2}}+\frac{v+k}{n}\frac{\frac{1}{v}\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^{2}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^{2}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}$$

$$\begin{aligned}\mathbf{H} &= \frac{\partial^{2}h\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}=\frac{\mathbf{I}_{k}}{n\omega^{2}}+\frac{v+k}{nv}\frac{\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^{2}}\right)}{1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^{2}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}\\ &\quad -2\frac{v+k}{nv^{2}}\frac{\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^{2}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^{2}}\right)}{\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^{2}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{2}}\end{aligned}$$

and

$$\widehat{\mathbf{H}}=\left.\frac{\partial^{2}h\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{D}}$$

applying proposition 2, we obtain:

$$\frac{\partial \mathbf{H}}{\partial \boldsymbol{\beta}} = -2\frac{v+k}{nv^2}\frac{\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{\left(1+\frac{1}{v}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)^2} \otimes \left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2}\right) +$$
$$+8\frac{v+k}{nv^3}\frac{\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{\left(1+\frac{1}{v}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)^3} \otimes \left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)$$
$$-2\frac{v+k}{nv^2}\frac{\left(\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\otimes\mathbf{I}_k+\mathbf{I}_k\otimes\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)}{\left(1+\frac{1}{v}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)^2}$$

and $\frac{\partial \widehat{\mathbf{H}}}{\partial \boldsymbol{\beta}} = \frac{\partial \mathbf{H}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$

Variance Approximation. By setting $g_l(\boldsymbol{\beta}) = \boldsymbol{\beta}\boldsymbol{\lambda}_l\boldsymbol{\lambda}_l'\boldsymbol{\beta} = \beta_l^2$ the variance is equal to

$$\widehat{var}(\beta_l|\mathbf{y}, M_j) = \widehat{E}\left(\beta_l^2|\mathbf{y}, M_j\right) - \left(\widehat{E}(\beta_l|\mathbf{y}, M_j)\right)^2$$

Where from equation (Appendix C.2) we obtain:

$$\widehat{E}\left(\beta_l^2|\mathbf{y}, M_j\right) = \widehat{\beta}^2_{lD} - \frac{1}{n}\left(\widehat{\boldsymbol{\beta}}_D'\boldsymbol{\Lambda}_{ll}'\widehat{\mathbf{H}}^{-1}\left(\frac{\partial \widehat{\mathbf{H}}}{\partial \boldsymbol{\beta}}\right)' - vec(\boldsymbol{\Lambda}_{ll})'\right)vec\left(\widehat{\mathbf{H}}^{-1}\right)$$

Appendix C.3. Application to Non-Conjugate Dependent Cauchy Prior

From equation (12), we obtain the posterior of $\boldsymbol{\beta}$ parameters:

$$p(\boldsymbol{\beta}|\mathbf{y}, M_j) = \frac{\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-(v+k)/2}\left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-(1+k)/2}}{\int_{\mathbb{R}^K}\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-(v+k)/2}\left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-(1+k)/2}d\boldsymbol{\beta}}$$

$$\begin{aligned}
\widehat{M}_g(s) &= \frac{\int_{\mathbb{R}^k} \exp(sg_l(\boldsymbol{\beta})) \left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-(v+k)/2}\left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-(1+k)/2} d\boldsymbol{\beta}}{\int_{\mathbb{R}^k}\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-(v+k)/2}\left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-(1+k)/2} d\boldsymbol{\beta}} = \\
&= \frac{\widehat{b}_N\left|\mathbf{H}_N\right|^{-\frac{1}{2}}\exp\left(-n\widehat{h}_N\right)}{\widehat{b}_D\left|\mathbf{H}_D\right|^{-\frac{1}{2}}\exp\left(-n\widehat{h}_D\right)} + O\left(n^{-2}\right) \\
&\approx \frac{\left|\mathbf{H}_N\right|^{-\frac{1}{2}}}{\left|\mathbf{H}_D\right|^{-\frac{1}{2}}}\frac{\exp\left(\frac{sg_l(\widehat{\boldsymbol{\beta}}_N)}{n}-\frac{1+k}{2n}\ln\left(1+\frac{\widehat{\boldsymbol{\beta}}'_N\widehat{\boldsymbol{\beta}}_N}{\omega^2}\right)-\frac{v+k}{2n}\ln\left(1+\frac{1}{v}\left(\widehat{\boldsymbol{\beta}}_N-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\widehat{\boldsymbol{\beta}}_N-\widehat{\boldsymbol{\beta}}\right)\right)\right)}{\exp\left(-\frac{1+k}{2n}\ln\left(1+\frac{\widehat{\boldsymbol{\beta}}'_D\widehat{\boldsymbol{\beta}}_D}{\omega^2}\right)-\frac{v+k}{2n}\ln\left(1+\frac{1}{v}\left(\widehat{\boldsymbol{\beta}}_D-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\widehat{\boldsymbol{\beta}}_D-\widehat{\boldsymbol{\beta}}\right)\right)\right)}
\end{aligned}$$

Mean Approximation.

$$\widehat{E}\left(\beta_l\left|\mathbf{y},M_j\right.\right) = \widehat{\beta}_{lD} - \frac{1}{2n}\lambda'_l\widehat{\mathbf{H}}^{-1}\left(\frac{\partial\widehat{\mathbf{H}}}{\partial\boldsymbol{\beta}}\right)' vec\left(\widehat{\mathbf{H}}^{-1}\right)$$

$$h(\boldsymbol{\beta}) = \frac{1}{2n}\ln\left(1+\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)+\frac{v+k}{2n}\ln\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)$$

$$\frac{\partial h(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \frac{k+1}{n}\frac{\boldsymbol{\beta}}{\omega^2+\boldsymbol{\beta}'\boldsymbol{\beta}}+\frac{v+k}{n}\frac{\frac{1}{v}\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}$$

$$\begin{aligned}
\mathbf{H} &= \frac{\partial^2 h(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \frac{k+1}{n}\frac{\left(\omega^2+\boldsymbol{\beta}'\boldsymbol{\beta}\right)\mathbf{I}_k-2\boldsymbol{\beta}\boldsymbol{\beta}'}{\left(\omega^2+\boldsymbol{\beta}'\boldsymbol{\beta}\right)^2}+\frac{v+k}{nv}\frac{\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)}{1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)} \\
&\quad -2\frac{v+k}{nv^2}\frac{\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2}\right)}{\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}^2_{b_0}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^2}
\end{aligned}$$

and

$$\widehat{\mathbf{H}} = \left. \frac{\partial^2 h\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$$

Variance Approximation.

$$\widehat{var}\left(\beta_l \left| \mathbf{y}, M_j \right.\right) = \widehat{E}\left(\beta_l^2 \left| \mathbf{y}, M_j \right.\right) - \left(\widehat{E}\left(\beta_l \left| \mathbf{y}, M_j \right.\right)\right)^2$$

where,

$$\widehat{E}\left(\beta_l^2 \left| \mathbf{y}, \mathbf{X} \right.\right) = \widehat{\beta}_{lD}^2 - \frac{1}{n}\left(\widehat{\boldsymbol{\beta}}_D' \boldsymbol{\Lambda}_{ll}' \widehat{\mathbf{H}}^{-1} \left(\frac{\partial \widehat{\mathbf{H}}}{\partial \boldsymbol{\beta}}\right)' - vec\left(\boldsymbol{\Lambda}_{ll}\right)'\right) vec\left(\widehat{\mathbf{H}}^{-1}\right)$$

applying proposition 2, we obtain:

$$\begin{aligned}
&\tfrac{\partial \mathbf{H}}{\partial \boldsymbol{\beta}} = -2\tfrac{k+1}{n}\boldsymbol{\beta} \otimes \mathbf{I}_k + 8\tfrac{k+1}{n}\tfrac{\boldsymbol{\beta}}{(\omega^2+\boldsymbol{\beta}'\boldsymbol{\beta})^3} \otimes \boldsymbol{\beta}\boldsymbol{\beta}' - 2\tfrac{k+1}{n}\tfrac{(\boldsymbol{\beta}\otimes\mathbf{I}_k+\mathbf{I}_k\otimes\boldsymbol{\beta})}{(\omega^2+\boldsymbol{\beta}'\boldsymbol{\beta})^2} + \\
&-2\tfrac{v+k}{nv^2}\frac{\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^2} \otimes \left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right) + \\
&+8\tfrac{v+k}{nv^3}\frac{\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^3} \otimes \left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right) + \\
&-2\tfrac{v+k}{nv^2}\frac{\left(\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\otimes\mathbf{I}_k+\mathbf{I}_k\otimes\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)}{\left(1+\frac{1}{v}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{Z}'\mathbf{Z}}{\widehat{s}_{b_0}^2}\right)\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^2}
\end{aligned}$$

with

$$\frac{\partial \widehat{\mathbf{H}}}{\partial \boldsymbol{\beta}} = \left.\frac{\partial \mathbf{H}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_D}$$

Appendix D. Predictive Posterior Estimation

Appendix D.1. Non-Conjugate Normal Prior

In this section we show the algebraic derivation of equation (17) in the main paper**.** We can easily derive the joint posterior of $(y_f, \boldsymbol{\beta}, \sigma)$ after the integration of the constant term, where we obtain

$$p\left(y_f, \boldsymbol{\beta}, \sigma | \mathbf{y}, \mathbf{z}_f, M_j\right) \propto \sigma^{-n} p\left(\sigma\right) \exp\left(-\frac{Q_0}{2}\right) \qquad \text{(Appendix D.1)}$$

where,

$$\begin{aligned} Q_0 &= \frac{\left(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta}\right) + \left(y_f - \mathbf{z}_f'\boldsymbol{\beta}\right)^2 - \overline{n}\xi^2}{\sigma^2} + \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2} \\ \xi &= \left(\boldsymbol{\iota}'\mathbf{y} + y_f - \mathbf{z}_f'\boldsymbol{\beta}\right)/\overline{n} \end{aligned}$$

By completing the square, $Q_0 = \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_0\right)' \mathbf{V} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_0\right) + \frac{Q_1}{\sigma^2}$, where

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_0 &= \left(\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2} + \left(\frac{n}{\overline{n}}\right)\frac{\mathbf{z}_f\mathbf{z}_f'}{\sigma^2} + \frac{\mathbf{I}_k}{\omega^2}\right)^{-1} \left(\frac{\mathbf{Z}'\mathbf{y}}{\sigma^2} + \left(\frac{n}{\overline{n}}\right)\frac{\mathbf{z}_f}{\sigma^2}\left(y_f - \overline{y}\right)\right) \\ \mathbf{V} &= \frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2} + \left(\frac{n}{\overline{n}}\right)\frac{\mathbf{z}_f\mathbf{z}_f'}{\sigma^2} + \frac{\mathbf{I}_k}{\omega^2} \\ Q_\sigma^f &= \left(\mathbf{y} - \overline{y}\boldsymbol{\iota}\right)'\left(\mathbf{y} - \overline{y}\boldsymbol{\iota}\right) + \left(\frac{n}{\overline{n}}\right)\left(y_f - \overline{y}\right)^2 - \sigma^2\widehat{\boldsymbol{\beta}}_0'\mathbf{V}\widehat{\boldsymbol{\beta}}_0 \end{aligned}$$

After performing these algebraic manipulations, we integrate the vector $\boldsymbol{\beta}$ using the normal distribution kernel, yielding the joint posterior distribution

of (y_f, σ).

$$p\left(y_f, \sigma \mid \mathbf{y}, \mathbf{z}_f, M_j\right) \propto \sigma^{-(n+a_0+1)} \left| \frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2} + \frac{\mathbf{I}_k}{\omega^2} + \frac{n\mathbf{z}_f\mathbf{z}_f'}{\overline{n}\sigma^2} \right|^{-\frac{1}{2}} \exp\left(-\frac{Q_\sigma^f + b_0}{2\sigma^2}\right)$$

Next, we present some key properties utilized in our calculations::

D1. Sherman-Morrison formula: Suppose $\mathbf{A}$ is an invertible square matrix and $\mathbf{u}, \mathbf{v}$ are column vectors and that $1+\mathbf{v}'\mathbf{A}\mathbf{u} \neq 0$. Then the Sherman–Morrison formula states that: $\left(\mathbf{A} + \mathbf{u}\mathbf{v}'\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}$

D2. Matrix Determinant Lemma: $\left|\mathbf{A} + \mathbf{u}\mathbf{v}'\right| = \left(1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}\right)\left|\mathbf{A}\right|$

We apply property (D2) to the following determinant

$$\left| \frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2} + \frac{\mathbf{I}_k}{\omega^2} + \frac{n\mathbf{z}_f\mathbf{z}_f'}{\overline{n}\sigma^2} \right| = \left| \frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2} + \frac{\mathbf{I}_k}{\omega^2} \right| \left(1 + \frac{n}{\overline{n}\sigma^2}\mathbf{z}_f' \left(\frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2} + \frac{\mathbf{I}_k}{\omega^2}\right)^{-1} \mathbf{z}_f\right)$$

From Property (B1) in Appendix Appendix B, we obtain,

$$\left| \frac{\mathbf{Z}'\mathbf{Z}}{\sigma^2} + \frac{\mathbf{I}_k}{\omega^2} \right| = \left| \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} + \frac{\mathbf{I}_{k+1}}{\omega^2} \right| \left(\frac{\omega^2\sigma^2}{n\omega^2 + \sigma^2}\right)$$

Upon completing the square with respect to y_f, we derive (15).

$$p\left(y_f, \sigma | \mathbf{y}, \mathbf{z}_f, M_j\right) \propto \sigma^{-(\overline{n}+a_0+1)} \left(\frac{g_2\left(\sigma\right)}{g_1\left(\sigma\right)}\right)^{\frac{1}{2}} \left| \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} + \frac{\mathbf{I}_{k+1}}{\omega^2} \right|^{-\frac{1}{2}} \exp\left(-\frac{Q_\sigma^f + b_0}{2\sigma^2}\right)$$

For the approximation of predictive distribution we apply the transformation $\sigma = \exp\left(\tau\right)$, so

$$p\left(y_f, \tau | \mathbf{y}, \mathbf{z}_f, M_j\right) \propto e^{-(\overline{n}+a_0)\tau} \left(\frac{g_2\left(\tau\right)}{g_1\left(\tau\right)}\right)^{\frac{1}{2}} \left| \frac{\mathbf{X}'\mathbf{X}}{e^{2\tau}} + \frac{\mathbf{I}_{k+1}}{\omega^2} \right|^{-\frac{1}{2}} \exp\left(-\frac{Q_\tau^f + b_0}{2e^{2\tau}}\right)$$

and

$$h_{\overline{n}f}^{N}(\tau)=\frac{1}{\overline{n}}\left(\left(\overline{n}+a_0\right)\tau+\frac{1}{2}\ln\left(\frac{g_2(\tau)}{g_1(\tau)}\right)-\frac{1}{2}\ln\left|\frac{\mathbf{X}'\mathbf{X}}{e^{2\tau}}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right|-\frac{Q_\tau^f+b_0}{2e^{2\tau}}\right)$$

Notice that, we can split the term Q_τ^f, to $Q_\tau^f=Q_\tau+Q_{0\tau}^f$, where the function Q_τ corresponds to Q_σ, introduced in Section 3.1, under the transformation $\sigma=\exp(\tau)$.

$$Q_{0\tau}^f=\frac{\left(y_f-\overline{y}-\mathbf{z}_f'\boldsymbol{\beta}^*/e^{2\tau}\right)^2}{1+n^{-1}+e^{-2\tau}\mathbf{z}_f'\left(\frac{\mathbf{Z}'\mathbf{Z}}{e^{2\tau}}+\frac{\mathbf{I}_k}{\omega^2}\right)^{-1}\mathbf{z}_f}$$

The function $h_{\overline{n}f}^{N}(\tau)$ can be decomposed in two parts, so $h_{\overline{n}f}^{N}(\tau)=h_{\overline{n}}^{N}(\tau)+\varphi_{\overline{n}f}^{N}(\tau)$ where,

$$\begin{aligned}
h_{\overline{n}}^{N}(\tau) &= \frac{1}{2\overline{n}}\ln g_2(\tau)-\frac{\left(\overline{n}+a_0\right)\tau}{\overline{n}}-\frac{1}{2\overline{n}}\ln\left|\frac{\mathbf{X}'\mathbf{X}}{e^{2\tau}}+\frac{\mathbf{I}_{k+1}}{\omega^2}\right|-\frac{1}{2\overline{n}e^{2\tau}}\left(Q_\tau+b_0\right)\\
\varphi_{\overline{n}f}^{N}(\tau) &= -\frac{1}{2\overline{n}}\left(\ln g_1(\tau)+e^{-2\tau}Q_{0\tau}^f\right)
\end{aligned}$$

notice that $h_{\overline{n}}^{N}(\tau)$ share the same functional form with $h_n^N(\tau)$, with respect to τ and in case of n replace with $\overline{n}$, so

$$\begin{aligned}
\frac{\partial h_{\overline{n}f}^{N}(\tau)}{\partial\tau} &= \frac{\partial h_{\overline{n}}^{N}(\tau)}{\partial\tau}+\frac{\partial\varphi_{\overline{n}f}^{N}(\tau)}{\partial\tau}\\
\frac{\partial^2 h_{\overline{n}f}^{N}(\tau)}{\partial\tau^2} &= \frac{\partial^2 h_{\overline{n}}^{N}(\tau)}{\partial\tau^2}+\frac{\partial^2\varphi_{\overline{n}f}^{N}(\tau)}{\partial\tau^2}
\end{aligned}$$

Since the derivatives of $h_{\overline{n}}^{N}(\tau)$ are the same with those of $h_n^N(\tau)$, by replacing n with $\overline{n}$, we just have to calculate the derivatives of $\varphi_{\overline{n}f}^{N}(\tau)$. We replace $Q_{0\tau}^f$

with the following equivalent expression:

$$Q_{0\tau}^{f}=\frac{\left(y_{f}-\overline{y}-e^{-2\tau}tr\left(\frac{\mathbf{Z}'\mathbf{Z}}{e^{2\tau}}+\frac{\mathbf{I}_{k}}{\omega^{2}}\right)^{-1}\mathbf{Z}'\mathbf{y}\mathbf{z}_{f}'\right)^{2}}{\frac{\overline{n}}{n}g_{1}\left(\tau\right)}$$

With this substitution, estimating the derivative becomes more straightforward by applying the rules of standard calculus and matrix calculus, as outlined in Propositions 1 and 2. We proceed by defining the following functions., $\varphi_{0}\left(\tau\right)=e^{-2\tau}$, $f\left(\tau\right)=\left(y_{f}-\overline{y}-\varphi_{0}\left(\tau\right)tr\mathbf{C}_{2\tau}^{-1}\mathbf{A}_{1}\right)$, $\mathbf{C}_{1\tau}=\varphi_{0}\left(\tau\right)\mathbf{B}_{0}+\omega^{-2}\mathbf{I}_{k}$, $\mathbf{C}_{2\tau}=\varphi_{0}\left(\tau\right)\mathbf{A}_{0}+\omega^{-2}\mathbf{I}_{k},\mathbf{A}_{0}=\mathbf{Z}'\mathbf{Z}$, $\mathbf{A}_{1}=\mathbf{Z}\mathbf{y}\mathbf{z}_{f}'$, $\mathbf{B}=\mathbf{z}_{f}\mathbf{z}_{f}'$ and $\mathbf{B}_{0}=\mathbf{A}_{0}+\frac{n}{\overline{n}}\mathbf{B}$. In order to obtain compact algebraic results we transform $g_{1}\left(\tau\right)$ using the Matrix Determinant Lemma, property (D2), so $g_{1}\left(\tau\right)=\frac{\left|\mathbf{C}_{1\tau}\right|}{\left|\mathbf{C}_{2\tau}\right|}$ and $\ln g_{1}\left(\tau\right)=\ln\left|\mathbf{C}_{1\tau}\right|-\ln\left|\mathbf{C}_{2\tau}\right|$ where

$$\begin{aligned}e^{-2\tau}Q_{0\tau}^{f} &= \frac{\varphi_{0}\left(\tau\right)f\left(\tau\right)^{2}}{\frac{\overline{n}}{n}g_{1}\left(\tau\right)}\\ \varphi_{\overline{n}f}^{N}\left(\tau\right) &= -\frac{1}{2\overline{n}}\left(\ln\left|\mathbf{C}_{1\tau}\right|-\ln\left|\mathbf{C}_{2\tau}\right|+\frac{\overline{n}}{n}\left|\mathbf{C}_{2\tau}\right|\left|\mathbf{C}_{1\tau}^{-1}\right|\varphi_{0}\left(\tau\right)f\left(\tau\right)^{2}\right)\end{aligned}$$

To simplify the analysis, we define the following univariate functions. However, their differentiation requires the use of matrix differential calculus. The first two derivatives, which are needed for applying the Laplace approximation, are derived as a function of the derivatives of $\varphi_{1},\ldots,\varphi_{4}$ and f, which are presented below. The proof is based on Appendix A (General Rules). So we

define: $\varphi_1(\tau) = \ln|\mathbf{C}_{1\tau}|, \varphi_2(\tau) = \ln|\mathbf{C}_{2\tau}|, \varphi_3(\tau) = |\mathbf{C}_{2\tau}|, \varphi_4(\tau) = \left|\mathbf{C}_{1\tau}^{-1}\right|$.

$$\begin{aligned}
\varphi_{\overline{n}f}^{N}(\tau) &= -\frac{1}{2\overline{n}}\left(\varphi_1(\tau) - \varphi_2(\tau) + \varphi_3(\tau)\varphi_4(\tau)\varphi_0(\tau) f(\tau)^2\right) \\
\frac{\partial\varphi_{\overline{n}f}^{N}(\tau)}{\partial\tau} &= -\frac{1}{2\overline{n}}\left(\varphi_1'(\tau) - \varphi_2'(\tau) + \varphi_3'(\tau)\varphi_4(\tau)\varphi_0(\tau) f(\tau)^2\right. \\
&\quad +\varphi_3(\tau)\varphi_4'(\tau)\varphi_0(\tau) f(\tau)^2 + \varphi_3(\tau)\varphi_4(\tau)\varphi_0'(\tau) f(\tau)^2 \\
&\quad \left. + 2\varphi_3(\tau)\varphi_4(\tau)\varphi_0(\tau) f(\tau) f'(\tau)\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2\varphi_{\overline{n}f}^{N}(\tau)}{\partial\tau^2} &= -\frac{1}{2\overline{n}}\left(\varphi_1''(\tau) - \varphi_2''(\tau) + \varphi_3''(\tau)\varphi_4(\tau)\varphi_0(\tau) f(\tau)^2 + \right. \\
&\quad +\varphi_3'(\tau)\varphi_4'(\tau)\varphi_0(\tau) f(\tau)^2 + \varphi_3'(\tau)\varphi_4(\tau)\varphi_0'(\tau) f(\tau)^2 + \\
&\quad +2\varphi_3'(\tau)\varphi_4(\tau)\varphi_0(\tau) f(\tau) f'(\tau) + \varphi_3'(\tau)\varphi_4'(\tau)\varphi_0(\tau) f(\tau)^2 \\
&\quad +\varphi_3(\tau)\varphi_4''(\tau)\varphi_0(\tau) f(\tau)^2 + \varphi_3(\tau)\varphi_4'(\tau)\varphi_0'(\tau) f(\tau)^2 + \\
&\quad +2\varphi_3(\tau)\varphi_4'(\tau)\varphi_0(\tau) f(\tau) f'(\tau) + \varphi_3'(\tau)\varphi_4(\tau)\varphi_0'(\tau) f(\tau)^2 \\
&\quad +\varphi_3(\tau)\varphi_4'(\tau)\varphi_0'(\tau) f(\tau)^2 + \varphi_3(\tau)\varphi_4(\tau)\varphi_0''(\tau) f(\tau)^2 \\
&\quad +2\varphi_3(\tau)\varphi_4(\tau)\varphi_0'(\tau) f(\tau) f'(\tau) + 2\varphi_3'(\tau)\varphi_4(\tau)\varphi_0(\tau) f(\tau) f'(\tau) \\
&\quad +2\varphi_3(\tau)\varphi_4'(\tau)\varphi_0(\tau) f(\tau) f'(\tau) + 2\varphi_3(\tau)\varphi_4(\tau)\varphi_0'(\tau) f(\tau) f'(\tau) \\
&\quad \left. +2\varphi_3(\tau)\varphi_4(\tau)\varphi_0(\tau) f'(\tau)^2 + 2\varphi_3(\tau)\varphi_4(\tau)\varphi_0(\tau) f(\tau) f''(\tau)\right)
\end{aligned}$$

where,

$$
\begin{aligned}
\varphi_1'(\tau) &= \varphi_0'(\tau)\ \mathrm{tr}\left(\mathbf{C}_{1\tau}^{-1}\mathbf{B}_0\right) \\
\varphi_2'(\tau) &= \varphi_0'(\tau)\ \mathrm{tr}\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right) \\
\varphi_3'(\tau) &= \varphi_0'(\tau)\left|\mathbf{C}_{2\tau}\right|\ \mathrm{tr}\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right) \\
\varphi_4'(\tau) &= \varphi_0'(\tau)\left|\mathbf{C}_{1\tau}^{-1}\right|\ \mathrm{tr}\left(\mathbf{C}_{1\tau}\mathbf{B}_0\right) \\
f'(\tau) &= -\varphi_0'(\tau)\ \mathrm{tr}\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_1\right) + \varphi_0(\tau)\,\varphi_0'(\tau)\ \mathrm{tr}\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_1\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right) \\
\varphi_1''(\tau) &= \varphi_0''(\tau)\ \mathrm{tr}\left(\mathbf{C}_{1\tau}^{-1}\mathbf{B}_0\right) - \left(\varphi_0'(\tau)\right)^2\ \mathrm{tr}\left(\mathbf{C}_{1\tau}^{-1}\mathbf{B}_0\right)^2 \\
\varphi_2''(\tau) &= \varphi_0''(\tau)\ \mathrm{tr}\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right) - \left(\varphi_0'(\tau)\right)^2\ \mathrm{tr}\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\varphi_3''(\tau) &= \varphi_0''(\tau)\left|\mathbf{C}_{2\tau}\right| tr\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right) + \left|\mathbf{C}_{2\tau}\right|\left(\varphi_0'(\tau)\, tr\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right)\right)^2 - \left(\varphi_0'(\tau)\right)^2\left|\mathbf{C}_{2\tau}\right| tr\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right)^2 \\
\varphi_4''(\tau) &= \varphi_0''(\tau)\left|\mathbf{C}_{1\tau}^{-1}\right| tr\left(\mathbf{C}_{1\tau}\mathbf{B}_0\right) + \left|\mathbf{C}_{1\tau}^{-1}\right|\left(\varphi_0'(\tau)\, tr\left(\mathbf{C}_{1\tau}\mathbf{B}_0\right)\right)^2 - \left(\varphi_0'(\tau)\right)^2\left|\mathbf{C}_{1\tau}^{-1}\right| tr\left(\mathbf{C}_{1\tau}\mathbf{A}_0\right)^2 \\
f''(\tau) &= -\varphi_0''(\tau)\, tr\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_1\right) + \left(\varphi_0'(\tau)\right)^2 + 2\varphi_0'(\tau)\, tr\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_1\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right) \\
&\quad -2\left(\varphi_0'(\tau)\right)^3 tr\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_1\left(\mathbf{C}_{2\tau}^{-1}\mathbf{A}_0\right)^2\right)
\end{aligned}
$$

Where, $\tau_f^* = \arg\max h_{\overline{n}}^N(\tau)$, then by continuity we obtain (16), and finally we obtain (17) in the main paper. The expression $4b_0^{a_0/2}/\overline{n}2^{(\overline{n}+a_0)/2}\omega^k\pi^{(\overline{n}-2)/2}\Gamma(a_0/2)$ is part of the normalising constant.

Appendix D.2. Non-Conjugate Dependent Cauchy Prior

Based on the assumptions and definitions outlined in Subsection 3.1.2 and Section 3.2 of the paper, we now proceed to derive the predictive posterior, equation (19) in the paper. After integrating out the constant term, we obtain the joint posterior distribution of $(y_f, \boldsymbol{\beta}, \sigma)$ as follows:

$$p\left(y_f, \boldsymbol{\beta}, \sigma \mid \mathbf{y}, \mathbf{z}_f, M_j\right) \propto \sigma^{-n} p\left(\boldsymbol{\beta}, \sigma\right) \exp\left(-\frac{Q_C}{2}\right)$$

$$Q_C = \frac{\left(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\right)' \left(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\right) - \left(y_f - \mathbf{z}_f'\boldsymbol{\beta}\right)^2 - \overline{n}\xi^2}{\sigma^2}$$

Next, we integrate out the scale parameter σ by applying the properties of the Inverse Gamma distribution:

$$p\left(y_f, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{z}_f, M_j\right) \propto p\left(\boldsymbol{\beta}\right) \left(Q_C\right)^{-\frac{v_0+k}{2}}$$

By completing the square, we express Q_C as:

$$Q_C = \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_f\right)' \mathbf{D}_{\mathbf{z}} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_f\right) + v_0 s_{f,b_0}^2$$

Substituting this into the equation above, we proceed with the integration of equation (18). Since this integral does not have a closed-form solution, we employ the multivariate version of the Laplace approximation, similar to previous cases.

We define $\widetilde{\boldsymbol{\beta}}_f = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^k} h_{\overline{n}}^{C_f}\left(\boldsymbol{\beta}\right)$, where $\mathbf{H}_{f\overline{n}}^*$ is the negative Hessian evaluated at $\widetilde{\boldsymbol{\beta}}_f$. The function $h_{\overline{n}}^{C_f}\left(\boldsymbol{\beta}\right)$ is given by:

$$h_{\overline{n}}^{C_f}\left(\boldsymbol{\beta}\right) = \overline{n}^{-1} \ln\left(\left(1 + \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\omega^2}\right)^{-\frac{1+k}{2}} \left(1 + \frac{1}{v_0}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_f\right)' \frac{\mathbf{D}_{\mathbf{z}}}{s_{f,b_0}^2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_f\right)\right)^{-\frac{v_0+k}{2}}\right)$$

Notice that the functional form of $h_{\overline{n}}^{C_f}\left(\boldsymbol{\beta}\right)$ is the same as that used for

the estimation of the moments $h_D(\boldsymbol{\beta})$, implying that the first and second derivatives are identical. Thus, we have:

$$p(y_f \mid \mathbf{y}, \mathbf{z}_f, M_j) \propto \left(v_0 s^2_{f,b_0}\right)^{-\frac{v_0+k}{2}} \left|\mathbf{H}^*_{f\overline{n}}\right|^{-1/2} \exp\left(\overline{n} h^{C_f}_{\overline{n}}\left(\widetilde{\boldsymbol{\beta}}_f\right)\right) \tag{Appendix D.2}$$

Finally, after making the necessary substitutions and dividing equation (Appendix D.2) by the marginal likelihood implied by equation (13) of the paper, we obtain equation (19) of the paper. Additionally, the expression $\frac{\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{\overline{n}+a_0-1}{2}\right) b_0^{a_0/2} 2^{k/2}}{\pi^{\overline{n}/2}\omega^k \overline{n}^{\frac{k+1}{2}}\Gamma\left(\frac{a_0}{2}\right)}$ is part of the normalising constant.

Appendix E. Assessing approximation error

Appendix E.1. Benchmark design.

After running the MC^3 algorithm using Laplace-based log-marginal likelihoods in the Metropolis-Hastings acceptance ratios, we record, for each *distinct* model visited post burn-in, its inclusion pattern γ_j, its Laplace log marginal $\log \widehat{f}_j^{\text{Lap}}(\mathbf{y})$ and its size k_j. Computing bridge sampling estimates for every visited model is computationally infeasible, so we benchmark Laplace on a representative subset of visited models: (i) the $N_{\text{high}} = 500$ models with highest Laplace-implied posterior mass under a uniform model prior, and (ii) among the remaining visited models, up to 100 additional models sampled without replacement within each model size k_j. This yields 1,619 benchmarked models for the normal prior and 1,765 for the Cauchy prior.

For each selected model we compute a bridge sampling log marginal $\log \widehat{f}_j^{\text{Br}}(\mathbf{y})$: for the normal prior we bridge sample the one-dimensional latent

$\tau = \log \sigma$ posterior, for the Cauchy prior we bridge sample the k_j-dimensional $\boldsymbol{\beta}_j$ posterior after integrating out σ. In both cases, posterior draws are generated by long random walk Metropolis chains and the bridge proposal q_j is Gaussian, centered at the Laplace mode and scaled using local curvature (mode based variance in the univariate case and a mode based Gaussian covariance proxy in the multivariate case). We define the model-wise log evidence error:

$$\Delta_j = \log \widehat{f}_j^{\mathrm{Lap}}(\mathbf{y}) - \log \widehat{f}_j^{\mathrm{Br}}(\mathbf{y}), \qquad \text{(Appendix E.1)}$$

summarize Δ_j over the benchmark subset, and examine its dependence on model size k_j.

Appendix E.2. Normal prior benchmark.

Table Appendix E.1 reports overall error summaries for both priors. For the non-conjugate normal prior, Δ_j is concentrated and mildly negative, with mean -0.5405, median -0.4819, standard deviation 0.2530 and range $[-1.815, -0.111]$. Figure Appendix E.1 shows a smooth, monotone pattern in the *average* error by model size: the mean error decreases from about -0.18 at $k_j = 2$ to about -0.96 at $k_j = 13$, with fewer sampled models in the largest size classes. Despite these non-zero errors, posterior summaries computed from the bridge-based weights remain close to those based on Laplace: on the benchmark subset, the posterior mean model size changes from 5.65 (Laplace) to 5.76 (bridge), and the maximum absolute difference in posterior inclusion probabilities across regressors is below 0.05.

Appendix E.3. Cauchy prior benchmark.

For the independent Cauchy prior, the Laplace log marginals display a larger negative shift: mean $\Delta_j = -4.579$, median -4.352, standard deviation 1.071 and range $[-9.708, -2.795]$. Figure Appendix E.2 shows that the mean error becomes more negative with k_j in this case as well. Since model comparison depends on *differences* in log marginal likelihoods, it is useful to decompose $\log \widehat{f}_j^{\mathrm{Lap}}(\mathbf{y}) = \log f_j(\mathbf{y}) + c + \varepsilon_j$ into an additive shift c and model-specific residual errors ε_j. The constant component cancels exactly in all Bayes factors and Metropolis-Hastings acceptance ratios, so only the residual differences $\varepsilon_j - \varepsilon_{j'}$ can affect posterior odds. To illustrate the relative magnitude of the residual component, we also report mean-centred errors $\widetilde{\Delta}_j = \Delta_j - \bar{\Delta}$ (where $\bar{\Delta}$ is the benchmark-sample mean) and plot $\widetilde{\Delta}_j$ against k_j in Figure Appendix E.3. The centered plot confirms that the dominant discrepancy is an approximately additive downward shift, while the remaining k_j dependent component is materially smaller in magnitude. Consistent with this, posterior summaries remain stable: on the benchmark subset, the posterior mean model size changes from 5.75 (Laplace) to 6.06 (bridge), and the maximum absolute inclusion probability difference across regressors is below 0.10.

Appendix E.4. Implications for MC^3 and the $S \to \infty$ limit.

Because MC^3 uses $\log \widehat{f}_j^{\mathrm{Lap}}(\mathbf{y})$ in the acceptance probability, the chain targets the *Laplace-induced* model posterior $\widetilde{\Pr}(M_j \mid \mathbf{y}) \propto \widehat{f}_j^{\mathrm{Lap}}(\mathbf{y}) \Pr(M_j)$, not the exact $\Pr(M_j \mid \mathbf{y}) \propto f_j(\mathbf{y}) \Pr(M_j)$. As $S \to \infty$, Monte Carlo error vanishes but the approximation error in $\widehat{f}_j^{\mathrm{Lap}}(\mathbf{y})$ persists. The bridge

Table Appendix E.1: Bridge-sampling benchmark summary for the growth-regression application. Errors are $\Delta_j = \log \widehat{f}_j^{\mathrm{Lap}}(y) - \log \widehat{f}_j^{\mathrm{Br}}(y)$ over the selected benchmark subset. Mean model size $\bar{k}$ excludes the intercept.

Prior	n_{sub}	mean(Δ)	sd(Δ)	min(Δ)	max(Δ)	$\bar{k}^{\mathrm{Lap}}$	$\bar{k}^{\mathrm{Br}}$	P_{10}^{Lap}	P_{10}^{Br}	$\lvert\mathcal{T}_{10}^{\mathrm{Lap}} \cap \mathcal{T}_{10}^{\mathrm{Br}}\rvert$
Normal (non-conjugate)	1619	−0.5405	0.2530	−1.8150	−0.1113	5.6506	5.7596	0.17077	0.17084	8
Cauchy (non-conjugate)	1765	−4.5790	1.0713	−9.7078	−2.7953	5.7525	6.0588	0.22405	0.17883	8

Note: For each prior, n_{sub} is the number of models in the benchmark subset. The log-evidence discrepancy is $\Delta_j = \log \widehat{f}_j^{\mathrm{Lap}}(y) - \log \widehat{f}_j^{\mathrm{Br}}(y)$, and the table reports its empirical mean, standard deviation, minimum, and maximum across the benchmarked models. $\bar{k}^{\mathrm{Lap}}$ and $\bar{k}^{\mathrm{Br}}$ are posterior mean model sizes (number of regressors, excluding the intercept) computed on the benchmark subset using, respectively, Laplace-based and bridge-based posterior model probabilities under a uniform prior over the benchmarked models. P_{10}^{Lap} and P_{10}^{Br} are the posterior masses of the top-10 models under the Laplace-based and bridge-based posteriors (restricted to the benchmark subset). $\mathcal{T}_{10}^{\mathrm{Lap}}$ and $\mathcal{T}_{10}^{\mathrm{Br}}$ denote the corresponding top-10 sets; the last column reports the size of their intersection (how many models appear in both top-10 lists).

benchmark directly quantifies this discrepancy: the observed model-wise errors are systematic but smooth in k_j, and the resulting changes in posterior inclusion probabilities, mean model size and top model posterior mass are small on the benchmark set (Table Appendix E.1). This evidence supports the numerical reliability of the BMA summaries reported in the paper. In principle, one could embed a fast approximation within a delayed-acceptance MCMC scheme (Christen and Fox, 2005). In large model spaces, however, repeatedly computing high-accuracy "exact" marginal likelihoods is typically prohibitive, motivating the diagnostic benchmark approach adopted here.

Appendix F. Robustness

Appendix F.1. Extended $MC(3)$ sampler with grouped covariate moves

In very large model spaces a one at a time Birth-Death $MC(3)$ sampler may, in principle, move slowly between regions in which groups of covariates tend to enter or exit jointly. Min and Sun (2016) address this by specifying

Table Appendix E.2: Normal prior: error summaries by model size k_j (benchmark subset).

k_j	mean(Δ_j)	sd(Δ_j)	count
2	−0.1758	0.0361	5
3	−0.2170	0.0361	62
4	−0.2648	0.0513	136
5	−0.3434	0.0721	204
6	−0.4195	0.0979	266
7	−0.5022	0.1253	228
8	−0.5491	0.1350	153
9	−0.6044	0.1473	105
10	−0.6808	0.1659	100
11	−0.7568	0.1980	100
12	−0.8411	0.1861	100
13	−0.9624	0.2107	100
14	−0.9643	0.2221	51
15	−0.8879	0.2197	8
16	−0.7513	0.0000	1

Note: k_j is model size (number of included regressors excluding the intercept). For each k_j, the table reports the mean and standard deviation of $\Delta_j = \log \widehat{f}_j^{\mathrm{Lap}}(y) - \log \widehat{f}_j^{\mathrm{Br}}(y)$ across benchmarked models of that size, along with the number of benchmarked models (count) in that size class.

grouped g-priors in normal linear models, where regressors are partitioned into prespecified groups and model selection proceeds over group inclusion indicators with group-level marginal likelihoods. Chen et al. (2017) extend this idea in an econometric application on the determinants of the 2008 financial crisis, introducing a hierarchical prior with separate indicators for groups and individual variables and a group-wise Gibbs sampler that updates group and within group inclusion states jointly. In both papers, grouped covariates are built directly into the prior and the posterior model probabilities.

Our approach is complementary. We retain the prior structure and Laplace

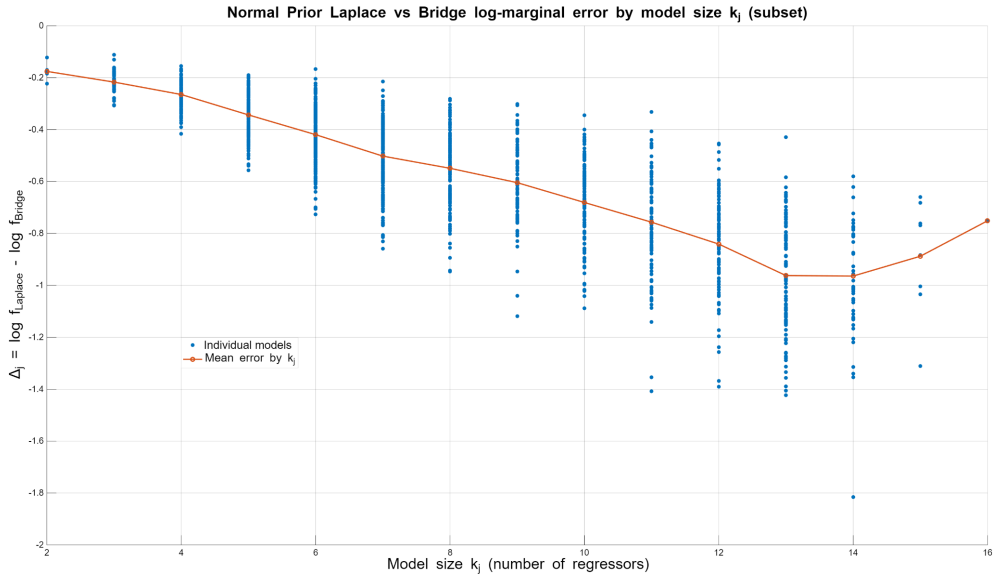

Figure Appendix E.1: Normal prior: Δ_j against model size k_j for the benchmark subset. Points show individual models; the line connects mean errors by k_j (Table Appendix E.2).

approximations of Section 3.1, so the posterior model probabilities underlying our BMA results are unchanged. Grouped covariates are used only at the proposal level, by enriching the $MC(3)$ kernel on model space. Let X denote the $n \times K$ matrix of demeaned regressors used in the $MC(3)$ implementation ($n = 72$, $K = 41$). We compute the pairwise correlation matrix of the columns of X and identify connected components at the threshold $|\rho| > 0.80$, which yields three empirical groups of highly correlated regressors, $G_1 = \{5, 6, 11, 21, 30, 35, 36\}$, $G_2 = \{18, 41\}$ and $G_3 = \{16, 27\}$, where indices refer to the column numbers of X (and to the variable numbering in Table 12). These groups define additional proposal moves in the $MC(3)$ sampler.

The extended sampler mixes three symmetric move types: (i) single variable Birth-Death moves, which randomly flip one inclusion indicator as in

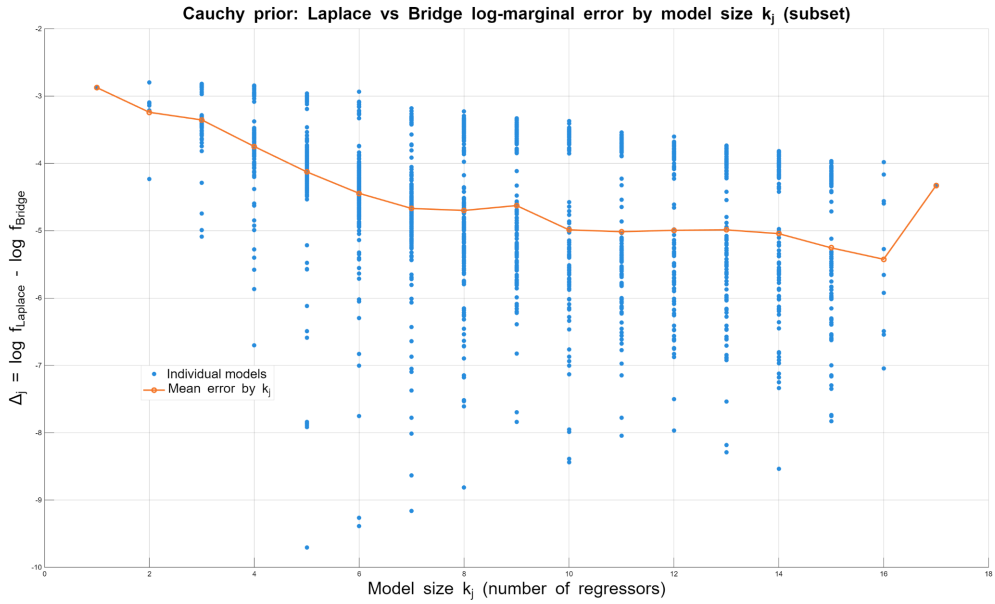

Figure Appendix E.2: Cauchy prior: raw log-evidence errors Δ_j against model size k_j for the benchmark subset.

the baseline algorithm, (ii) group moves, which randomly select one G_g and jointly toggle the inclusion indicators of all regressors in that group, and (iii) swap moves, which randomly select one included and one excluded regressor and swap their inclusion states. Let $(p_{\text{single}}, p_{\text{group}}, p_{\text{swap}})$ denote the probabilities of proposing a single, group or swap move. In the growth application we take $(0.6, 0.3, 0.1)$ whenever at least one group is available and $(0.9, 0, 0.1)$ otherwise. Each move type is constructed to be symmetric, and the mixture weights do not depend on the current model, so the overall proposal is symmetric. The Metropolis-Hastings acceptance probability therefore remains

$$\alpha\big(M^{(s-1)}, M^*\big) = \min\left\{1, \frac{p(M^* \mid \mathbf{y})}{p(M^{(s-1)} \mid \mathbf{y})}\right\},$$

and the stationary distribution of the extended chain is identical to that of

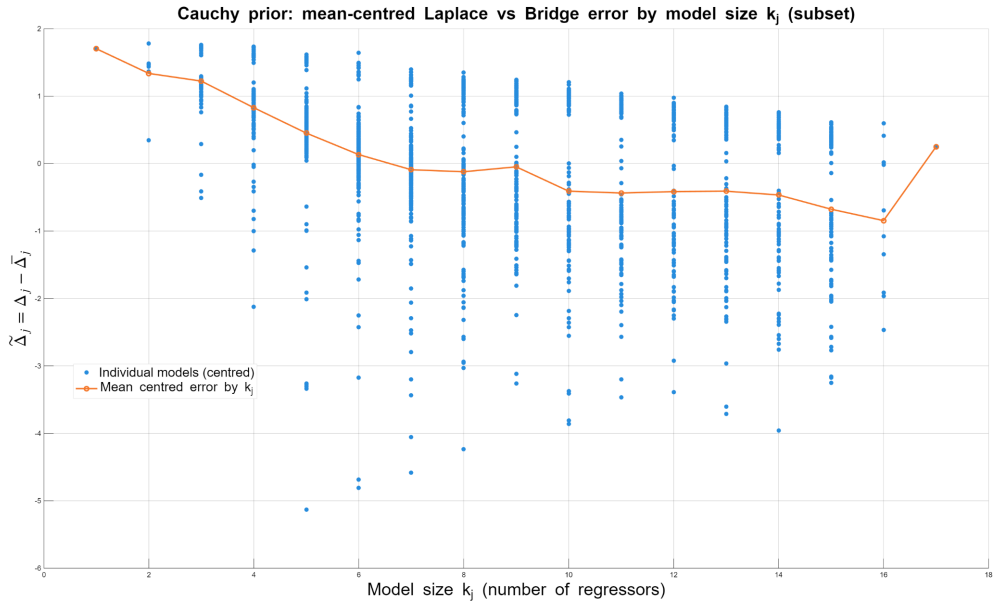

Figure Appendix E.3: Cauchy prior: mean-centred errors $\widetilde{\Delta}_j = \Delta_j - \bar{\Delta}$ against model size k_j for the benchmark subset.

the original Birth-Death sampler; any differences in Monte Carlo output are due solely to mixing behaviour.

We keep the same simulation design as in the baseline analysis, with a burn-in of one million iterations and $S_0 = 2 \times 10^6$ post-burn draws. Results are presented in table Appendix F.1. For the non-conjugate normal prior (NCN), the original Birth-Death sampler yields an acceptance rate of 15.6% and a posterior mean model size of 7.66 regressors. The extended sampler has an overall acceptance rate of 10.9%, with move-specific rates of 15.6% (single), 1.5% (group) and 11.1% (swap), an almost identical mean model size of 7.66, and a top-10 posterior mass of 7.7% in both implementations. For the non-conjugate Cauchy prior (NCC), the extended sampler exhibits an overall acceptance rate of 9.1%, a mean model size of 7.9 regressors, a top-10

mass of 11.4%, and a correlation of 0.997 between analytical and numerical top-10 model probabilities. Across both priors, posterior inclusion probabilities from the extended sampler are virtually identical to those from the original Birth-Death sampler, with maximum absolute differences below one percentage point and an unchanged ranking of the most important growth determinants. Since BMA summaries (posterior means, variances and predictive quantities) are simple functionals of the posterior model probabilities and model-specific moments, this close agreement in model posterior output implies that the corresponding BMA moments and predictive distributions would also be numerically indistinguishable. For this reason we report BMA results only once, based on the original Birth-Death sampler. Overall, the diagnostics indicate that, even in the presence of empirical group structures of the type emphasised by Min and Sun (2016) and Chen et al. (2017), the one at a time Birth-Death $MC(3)$ sampler explores well the regions of high posterior probability and our empirical conclusions are not driven by deficiencies of the proposal distribution.

Appendix F.2. Sensitivity to the choice of Model Space Prior

In the baseline specification we follow the conventional BMA literature on growth regressions and impose a uniform prior over the model space, which corresponds to independent Bernoulli inclusion with $w = 1/2$ for each of the $K = 41$ regressors and thus a prior mean model size of $E(|M|) = wK = K/2$ (see Steel (2020)). To assess the sensitivity of our results to this choice, we compare it with three alternative priors on model space that are standard in the literature discussed by Steel (2020) and the references therein.

Let M_j denote a model including k_j regressors. Our baseline prior is

$$p(M_j \mid w) \propto w^{k_j}(1-w)^{K-k_j}, \qquad w = \tfrac{1}{2},$$

which we denote by Fixed-w. As noted by Steel (2020), values $w > 1/2$ ($w < 1/2$) favor larger (smaller) models. We then consider: (i) a Beta-Binomial prior (Betabinom) obtained by placing a Beta(a, b) prior on w, with $a = 1$ and b chosen so that the implied prior mean model size is approximately $K/4$, (ii) a loss-based prior (Loss) of the form $p(M_j) \propto \exp(-ck_j)$, calibrated to yield a similar prior mean model size, and (iii) a Poisson model size prior (Poisson) where $|M_j|$ has a truncated Poisson distribution with parameter $\lambda = K/2$, following Womack, Fuentes and Taylor-Rodriguez (2015). All four priors enter the MC^3 sampler used in the paper by adding the corresponding log prior-odds term in the Metropolis-Hastings step; marginal likelihoods for the non-conjugate priors are still computed using the Laplace approximations described in Section 3.1.

Table Appendix F.2 reports the posterior mean model size for each combination of parameter prior (conjugate Normal, non-conjugate Normal, non-conjugate Cauchy) and model space prior. As expected, priors centered on smaller models (Betabinom, Loss) yield noticeably more parsimonious posteriors than Fixed-w and Poisson, which are centered on $k/2$. More importantly, for *every* model space prior, both non-conjugate specifications lead to smaller posterior model sizes than the conjugate Normal prior, confirming the parsimony result obtained under the baseline uniform prior.

Table Appendix F.3 shows posterior inclusion probabilities (PIPs) for selected regressors under the two non-conjugate priors across the four model-

space priors. The model prior mainly shifts the overall level of PIPs: priors favouring larger models (Fixed-w, Poisson) increase PIPs, whereas sparsity-inducing priors (Betabinom, Loss) reduce them, especially for borderline variables. The ranking of regressors, however, is remarkably stable, and key variables retain very high PIPs under all priors for both non-conjugate specifications.

Finally, Table Appendix F.4 reports the mean log predictive score (LPS). Within a given parameter prior, changing the model space prior affects LPS only modestly (by at most a few tenths), whereas differences across parameter priors are much larger. For all four model space priors, the non-conjugate Normal prior yields the best (highest) average LPS, the non-conjugate Cauchy prior comes second, and the conjugate Normal prior performs worst. Thus, while alternative priors on model space predictably influence posterior model size and PIPs, our main conclusions, that non-conjugate priors produce more parsimonious models and dominate the conjugate Normal prior in terms of predictive performance, are robust across all model space priors considered.

Table Appendix E.3: Cauchy prior: error summaries by model size k_j (benchmark subset). Mean-centred errors are $\widetilde{\Delta}_j = \Delta_j - \bar{\Delta}$, where $\bar{\Delta}$ is the overall mean of Δ_j on the benchmark subset.

k_j	mean(Δ_j)	mean($\widetilde{\Delta}_j$)	sd(Δ_j)	count
1	−2.8750	1.7040	0.0000	1
2	−3.2411	1.3379	0.4556	7
3	−3.3544	1.2245	0.4517	70
4	−3.7494	0.8296	0.5576	140
5	−4.1257	0.4533	0.8690	190
6	−4.4448	0.1341	0.7627	225
7	−4.6691	−0.0901	0.9007	202
8	−4.6988	−0.1199	1.0666	181
9	−4.6260	−0.0470	1.0089	144
10	−4.9871	−0.4081	1.2157	107
11	−5.0140	−0.4350	1.1290	101
12	−4.9936	−0.4146	1.1219	100
13	−4.9871	−0.4081	1.1379	100
14	−5.0434	−0.4644	1.1087	100
15	−5.2542	−0.6752	1.1242	86
16	−5.4237	−0.8447	1.0775	10
17	−4.3282	0.2508	0.0000	1

Note: k_j is model size (number of included regressors excluding the intercept). For each k_j, the table reports the mean and standard deviation of $\Delta_j = \log \widehat{f}_j^{\mathrm{Lap}}(y) - \log \widehat{f}_j^{\mathrm{Br}}(y)$ across benchmarked models of that size, along with the number of benchmarked models (count) in that size class. mean($\widetilde{\Delta}_j$) reports the within-class mean of centred errors $\widetilde{\Delta}_j = \Delta_j - \bar{\Delta}$, where $\bar{\Delta}$ is the overall mean error across the full benchmark subset. The centering removes any approximately constant additive shift in log-evidences, isolating the model-specific component that affects Bayes factors and Metropolis-Hastings acceptance ratios.

Table Appendix F.1: MC(3) sampler diagnostics for the growth data under non-conjugate priors.

Prior	Sampler	Overall acc. (%)	Single (%)	Group (%)	Swap (%)	Mean model size	Top-10 mass	Corr.
NCN	Birth–Death	15.61	15.61	–	–	6.66	0.077	0.996
NCN	Single/Group/Swap	10.87	15.55	1.48	11.06	6.66	0.077	0.996
NCC	Birth–Death	12.92	12.92	–	–	6.87	0.113	0.998
NCC	Single/Group/Swap	9.08	12.95	1.35	9.02	6.88	0.114	0.997

Note: Overall acc. is the unconditional (marginal) acceptance rate over all proposed moves. Single/Group/Swap report move-specific acceptance rates (in percent). Top-10 mass is the estimated posterior probability mass covered by the ten highest posterior probability models, computed as the fraction of post-burn-in iterations in which the chain is in one of these models. Corr. is the correlation between analytical and numerical posterior probabilities for these top models.

Table Appendix F.2: Posterior mean model size by parameter prior and model-space prior.

Model space prior	Conjugate Normal	Non-conjugate Normal	Non-conjugate Cauchy
Fixed-w ($w = 1/2$)	10.45	6.66	6.87
Beta-Binomial	6.04	2.89	2.57
Loss-based	9.80	6.14	6.10
Poisson	12.40	9.09	10.11

Note: Entries report the posterior mean number of regressors ("mean model size") for each combination of parameter prior (columns) and prior over model space (rows). The fixed-w prior with $w = 1/2$ corresponds to equal prior probabilities on all models. The Beta-Binomial and loss-based priors are calibrated to favor smaller models, while the Poisson prior is centered on larger model sizes, as discussed in Section Appendix F.2

Table Appendix F.3: Posterior inclusion probabilities for selected covariates under non-conjugate priors across model-space priors.

Prior	Covariate	Fixed-w ($w = 1/2$)	Beta–Binomial	Loss-based	Poisson
NCN	x_5	0.429	0.020	0.350	0.748
NCN	x_3	0.513	0.132	0.491	0.550
NCN	x_1	0.996	0.996	0.996	0.991
NCN	x_7	0.386	0.053	0.328	0.579
NCN	x_2	0.967	0.983	0.971	0.953
NCN	x_4	0.468	0.201	0.484	0.408
NCC	x_5	0.533	0.010	0.400	0.909
NCC	x_3	0.593	0.100	0.575	0.573
NCC	x_1	0.980	0.994	0.985	0.951
NCC	x_7	0.427	0.037	0.343	0.686
NCC	x_2	0.973	0.982	0.979	0.970
NCC	x_4	0.455	0.138	0.479	0.411

Note: Entries are posterior inclusion probabilities (PIPs) for selected covariates under the two non–conjugate priors for the regression coefficients (NCN: non–conjugate Normal; NCC: non–conjugate Cauchy), for different priors over model space. The subscript of each variable x_j denotes its ranking according to Table 12. The variables x_1, x_2, x_3, x_4, x_5 and x_7 are those with high PIPs in the baseline specification. For both non–conjugate priors, PIPs move up (down) when the model-space prior favours larger (smaller) models, while the ranking of these covariates remains essentially unchanged.

Table Appendix F.4: Mean log predictive score (LPS) by parameter prior and model-space prior.

Model–space prior	Conjugate Normal	Non–conjugate Normal	Non–conjugate Cauchy
Fixed-w ($w = 1/2$)	-2.559	-5.105	-3.543
Beta–Binomial	-2.843	-4.775	-3.421
Loss-based	-2.655	-5.098	-3.521
Poisson	-2.391	-5.330	-3.691

Note: Entries report the mean log predictive score (LPS) over the predictive replications for each combination of parameter prior (columns) and prior over model space (rows). Higher LPS (less negative values) indicates better out-of-sample predictive performance. For all model-space priors, the non–conjugate Normal prior yields the best predictive performance, followed by the non–conjugate Cauchy prior and then the conjugate Normal prior.

References

Box, G.E.P., Tiao, G.C., 1992. Bayesian Inference in Statistical Analysis. Wiley, New York.

Broemeling, L.D., Abdullah, M.Y., 1984. An approximation to the poly-t distribution. Communications In Statistics 13, 1407–1422.

Chen, R.B., Chen, Y.C., Chu, C.H., Lee, K.J., 2017. On the determinants of the 2008 financial crisis: A bayesian approach to the selection of groups and variables. Studies in Nonlinear Dynamics & Econometrics 21, 20160107.

Christen, J.A., Fox, C., 2005. Markov chain monte carlo using an approximation. Journal of Computational and Graphical statistics 14, 795–810.

Dawid, A.P., 1987. The infinite regress and its conjugate analysis. University College London. Department of Statistical Science.

De Bruijn, N.G., 1981. Asymptotic methods in analysis. volume 4. Courier Corporation.

Dickey, J.M., 1975. Bayesian alternatives to the f-test and least-squares estimate in the normal linear model. Studies in Bayesian econometrics and statistics , 515–554.

Dreze, J., 1977. Bayesian regression using poly-t densities. Journal of Econometrics 6, 329–354.

Eicher, T.S., Papageorgiou, C., Raftery, A.E., 2011. Default priors and predictive performance in bayesian model averaging, with application to growth determinants. Journal of Applied Econometrics 26, 30–55.

Fang, B., Dawid, A., 2002. Nonconjugate bayesian regression on many variables. Journal of statistical planning and inference 103, 245–261.

Fernandez, C., Ley, E., Steel, M.F.J., 2001a. Benchmark priors for bayesian model averaging. Journal of Econometrics 100, 381–427.

Fernandez, C., Ley, E., Steel, M.F.J., 2001b. Model uncertainty in cross-country growth regressions. Journal of Applied Econometrics 16, 563–576.

Guttman, I., Menzefricke, U., 1983. Bayesian inference in multivariate regression with missing observations on the response variables. Journal of Business & Economic Statistics 1, 239–248.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. Statistical Science 14, 382–401.

Judge, G., Griffiths, W., Hill, C., Lee, T., 1985. The Theory and Practice of Econometrics. John Wiley and Sons, New York.

Kass, R., Raftery, A., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Kass, R.E., Vaidyanathan, S.K., 1992. Approximate bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. Journal of the Royal Statistical Society Series B: Statistical Methodology 54, 129–144.

Koop, G., Korobilis, D., 2012. Forecasting inflation using dynamic model averaging. International Economic Review 53, 867–886.

Leamer, E.E., 1978. Specification Searches: Ad Hoc Inference with Non Experimental Data. Wiley, New York.

Ley, E., Steel, M.F., 2012. Mixtures of g-priors for bayesian model averaging with economic applications. Journal of Econometrics 171, 251–266.

Ley, E., Steel, M.F.J., 2007. Jointness in bayesian variable selection with applications to growth regression. Journal of Macroeconomics 29, 476–493.

Ley, E., Steel, M.F.J., 2009. On the effect of prior assumptions in bayesian model averaging with applications to growth regression. Journal of Applied Econometrics 24, 651–674.

Madigan, D., York, J., 1995. Bayesian graphical models for discrete data. International Statistical Review 63, 215–232.

Magnus, J.R., Neudecker, H., 2007. Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, Chichester, UK.

Sala-i Martin, X., 1997. I just ran 2 million regressions. American Economic Review 87, 178–83.

Sala-i Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. American Economic Review 94, 813–835.

Meng, X.L., Wong, W.H., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica , 831–860.

Min, X., Sun, D., 2016. Bayesian model selection for a linear model with grouped covariates. Annals of the Institute of Statistical Mathematics 68, 877–903.

Mirestean, A., Tsangarides, C.G., 2016. Growth determinants revisited using limited-information bayesian model averaging. Journal of Applied Econometrics 31, 106–132.

Press, J.S., 1982. Applied Multivariate Analysis. Robert E. Kreiger Publishing Co., Malabar, Florida.

Raftery, A.E., 1988. Approximate Bayes factors for generalized linear models. University of Washington, Department of Statistics.

Raftery, A.E., Kárnỳ, M., Ettler, P., 2010. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. Technometrics 52, 52–66.

Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for regression models. Journal of the American Statistical Association 92, 179–191.

Richard, J.F., Tompa, H., 1980. On the evaluation of poly-t density functions. Journal of Econometrics 12, 335–351.

Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology 71, 319–392.

Salimans, T., 2012. Variable selection and functional form uncertainty in cross-country growth regressions. Journal of Econometrics 171, 267–280.

Shun, Z., McCullagh, P., 1995. Laplace approximation of high dimensional integrals. Journal of the Royal Statistical Society Series B: Statistical Methodology 57, 749–760.

Steel, M.F., 2020. Model averaging and its use in economics. Journal of Economic Literature 58, 644–719.

Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, Theory and Methods 81, 82–86.

Tierney, L., Kass, R.E., Kadane, J.B., 1989. Fully exponential laplace approximations to expectations and variances of nonpositive functions. Journal of the American Statistical Association, Theory and Methods 84, 710–716.

Turkington, D.A., 2002. Matrix Calculus & Zero-One Matrices. Cambridge University Press, New York.

Zellner, A., 1971. An Introduction to Bayesian Inference in Econometrics. John Wiley & Sons, New York.

Zellner, A., Siow, A., 1980. Posterior odds ratios for selected regression hypotheses. Bayesian Statistics 1, 585–603.