Power-Expected-Posterior Priors for Variable Selection in Gaussian Linear Models

D. Fouskakis*, I. Ntzoufras† and D. Draper‡

June 19, 2012

Summary: Imaginary training samples are often used in Bayesian statistics to develop prior distributions, with appealing interpretations, for use in model comparison. Expected-posterior priors are defined via imaginary training samples coming from a common underlying predictive distribution m^*, using an initial baseline prior distribution. These priors can have subjective and also default Bayesian implementations, based on different choices of m^* and of the baseline prior. One of the main advantages of the expected-posterior priors is that impropriety of baseline priors causes no indeterminacy of Bayes factors; but at the same time they strongly depend on the selection and the size of the training sample. Here we combine ideas from the power-prior and the unit-information prior methodologies to greatly diminish the effect of training samples on a Bayesian variable-selection problem using the expected-posterior prior approach: we raise the likelihood involved in the expected-posterior prior distribution to a power that produces a prior information content equivalent to one data point. The result is that in practice our *power-expected-posterior* (PEP) methodology is sufficiently insensitive to the size n^* of the training sample that one may take n^* equal to the full-data sample size and dispense with training samples altogether; this promotes stability of the resulting Bayes factors, removes the arbitrariness arising from individual training-sample selections, and greatly increases computational speed, allowing many more models to be compared within a fixed CPU budget. Here we focus on Gaussian linear models and develop our method under two different baseline prior choices: the independence Jeffreys prior and the Zellner g-prior. The method's performance is compared, in simulation studies and a real example involving prediction of air-pollutant concentrations from meteorological covariates, with a variety of previously-defined variants on Bayes factors for variable selection. We find that the variable-selection procedure using our PEP prior (1) is systematically more parsimonious than the original expected-posterior prior with minimal training sample, while sacrificing no desirable performance characteristics to achieve this parsimony; (2) is robust to the size of the training sample, thus enjoying the advantages described above arising from the avoidance of training samples altogether; and (3) identifies maximum-a-posteriori models that achieve good out-of-sample predictive performance.

Keywords: Bayesian variable selection; Bayes factors; Expected-posterior priors; Gaussian linear models; Power-prior; Training samples; Unit-information prior.

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

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

‡D. Draper is with the Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, 1156 High Street, Santa Cruz CA 95064 USA; email `draper@ams.ucsc.edu`

1 Introduction

A leading approach to Bayesian variable selection in regression models is based on posterior model probabilities and the corresponding posterior model odds, which are functions of Bayes factors. Often little information is available, external to the present data set, about the values of the parameters of the models under comparison, motivating a desire to use proper-but-diffuse or improper prior distributions for the parameters. This leads immediately to well-known difficulties: if proper prior distributions with large variances are used, the resulting Bayes factors can be highly sensitive to the chosen prior variances, but improper priors do not work either, since the resulting normalizing constants involved in Bayes-factor computations cannot be determined.

In the case of Gaussian regression models, on which we focus in this paper, the most commonly used proper prior distribution for the parameters of the models under comparison is the conjugate Normal-Inverse-Gamma distribution (or variations of it), leading to analytical computations of the marginal likelihoods and their corresponding ratios. As a special case, g-priors (Zellner, 1986) have been a leading choice for variable selection in Gaussian regression models, due to their ease of interpretation and nice properties. For example, Kass and Wasserman (1995) introduce the unit-information prior, which is a special case of the g-prior and corresponds (for large sample sizes) to variable selection using BIC (Schwarz, 1978). Moreover g-priors can be derived as power priors (Ibrahim and Chen, 2000), using as imaginary data a response vector of zeros and the same design matrix as employed in the current sample (Zellner, 1986, Ibrahim and Chen, 2000, Liang, Paulo, Molina, Clyde and Berger, 2008). Different choices of the hyper-parameter g have been examined thoroughly by Fernandez, Ley and Steel (2001), while George and Foster (2000) demonstrated how different values of g lead to different model-comparison criteria. Liang et al. (2008) proposed to use mixtures of g-priors, leading to a hyper-g-prior in which a Beta hyper-prior is placed on the shrinkage factor $\frac{g}{g+1}$. In a similar manner, Zellner and Siow (1980) considered a mixture of g-priors with an Inverse-Gamma hyper-prior on g. Other proper-prior choices include independent Normal priors on the regression coefficients (which is equivalent to implementing ridge regression), the spike-and-slab prior (Mitchell and Beauchamp, 1988, Ishwaran and Rao, 2005) and the Normal mixture prior by George and McCulloch (1993, 1997).

Recently, Bayesian methods based on the *lasso* (Tibshirani, 1996) and other shrinkage methods have been considered. In these methods, proper priors (highly peaked at zero) are used, such as the Double-Exponential prior (Park and Casella, 2008), which gives estimates equivalent to those provided by the lasso method. Balakrishnan and Madigan (2010) used a Normal-Exponential mixture prior, Hans (2009, 2010) directly imposed the Double-Exponential prior on the lasso regression coefficients and a Gamma prior on the shrinkage parameter, while Griffin and Brown (2010) discussed the implementation of a Normal-Gamma prior for model coefficients, which is a generalization of the Bayesian lasso and has adaptive properties in terms of the shrinkage imposed on the coefficients. A Bayesian version of the *elastic net* (Zou and Hastie, 2005) has also been introduced by Li and Lin (2010), using a prior distribution that is a compromise between a Normal and a Double-Exponential distribution; the penalty parameters are chosen through an empirical method that maximizes the data marginal likelihood. Fahrmeir, Kneib and Konrath (2010) proposed a mixture of Normal and Inverse-Gamma (NMIG) prior for the model coefficients and a spike-and-slab prior for the corresponding variances, resulting in a method that includes the Bayesian lasso as a special case. Finally, shrinkage priors, such as the horse-shoe prior (Carvalho, Polson and Scott, 2010) and the Double-Generalized-Pareto (Armagan, Dunson and Lee, 2012), aim to over-shrink small coefficients and leave large ones as unaffected as possible, in an attempt

to retain some consistency properties.

Another active area of research has emerged from attempts to use improper prior distributions in Bayesian variable selection; leading contributions include a variety of Bayes-factor variants (*posterior*, *fractional* and *intrinsic*: see, e.g., Aitkin (1991), O'Hagan (1995), and Berger and Pericchi 1996*a*, 1996*b*, respectively). An important part of this work is focused on *objective model selection methods* (Casella and Moreno, 2006, Moreno and Girón, 2008, Casella, Girón, Martínez and Moreno, 2009), having their source in the intrinsic priors originally introduced by Berger and Pericchi (1996*b*); these methods attempt to provide an approximate proper Bayesian interpretation for intrinsic Bayes factors. Intrinsic priors can be considered as special cases of the *expected-posterior* prior distributions of Pérez and Berger (2002), which have an appealing interpretation based on imaginary training data coming from prior predictive distributions. Expected-posterior prior distributions can accommodate improper *baseline* priors as a starting point, and the marginal likelihoods for all models are calculated up to the same normalizing constant; this overcomes the problem of indeterminacy of the Bayes factors, since the unknown normalizing constant cancels out in the marginal likelihood ratios. However, the approach is based on one or more *training samples* chosen from the data, and this raises two new questions: how large should such training samples be, and how should they be chosen?

In this paper we greatly diminish the effect of training samples on the expected-posterior-prior methodology, by combining ideas from the power-prior method of Ibrahim and Chen (2000) and the unit-information-prior approach of Kass and Wasserman (1995): we raise the likelihood involved in the expected-posterior prior distribution to the power $\frac{1}{n}$, to produce a prior information content equivalent to one data point. In this manner the effect of the imaginary/training sample is small with even modest n. Moreover, as will become clear in Section 6, in practice our *power-expected-posterior* (PEP) prior methodology is sufficiently insensitive to the size n^* of the training sample that one may take $n = n^*$ and dispense with training samples altogether; this both removes the instability arising from the random choice of training samples and greatly reduces computing time.

The PEP prior approach can be implemented under any baseline prior choice, proper or improper. In this paper, results are presented for two different prior baseline choices: the g-prior and the independence Jeffreys prior. The conjugacy structure of the first permits computation of the first two moments (see Appendix A) of the resulting PEP prior, which offers flexibility in situations in which non-diffuse parametric prior information is available. When (on the other hand) little information about the parameters in the competing models is available, the PEP prior with the independence Jeffreys baseline prior can be viewed as an *objective model-selection* technique. Moreover, with either choice of baseline prior, simple but efficient Monte-Carlo schemes for the estimation of the marginal likelihoods can be constructed in a straightforward manner.

The plan of the paper is as follows. In the next three sub-sections, to fix notation, we provide some preliminary details on the expected-posterior prior approach and we highlight difficulties that arise when implementing it. Our PEP prior methodology is described in detail in Section 2, and the resulting prior and posterior distributions are presented analytically under the two different baseline prior choices mentioned above. In Section 3, we give an MCMC algorithm to sample from the posterior, and in Section 4 Monte-Carlo estimates of the marginal likelihood are provided. In Section 5 we give a description of model-search strategies when the model space is large and full enumeration of all models is not possible. Section 6 presents illustrations of our method, under both baseline prior choices, in a simulation experiment and in a real-data example involving the prediction of atmospheric ozone levels from meteorological covariates. Finally, in Section 7 we conclude the paper with a brief discussion and some ideas for further research.

1.1 Expected-posterior priors

Pérez and Berger (2002) developed priors for use in model comparison, through utilization of the device of "imaginary training samples" (Spiegelhalter and Smith, 1988, Iwaki, 1997, Good, 2004). They defined the expected-posterior prior (EPP) as the posterior distribution of a parameter vector for the model under consideration, averaged over all possible imaginary samples $\boldsymbol{y}^*$ coming from a "suitable" predictive distribution $m^*(\boldsymbol{y}^*)$. Hence the EPP for the parameters of any model $m_\ell \in \mathcal{M}$, with $\mathcal{M}$ denoting the model space, is

$$\begin{aligned} \pi_\ell^E(\boldsymbol{\theta}_\ell) &= \int f(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, m_\ell)\, m^*(\boldsymbol{y}^*)\, d\boldsymbol{y}^* \\ &= \int \pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*)\, m^*(\boldsymbol{y}^*)\, d\boldsymbol{y}^* \end{aligned} \tag{1}$$

where $\pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*)$ is the posterior of $\boldsymbol{\theta}_\ell$ for model m_ℓ using a baseline prior $\pi_\ell^N(\boldsymbol{\theta}_\ell)$ and data $\boldsymbol{y}^*$.

A question that naturally arises when using EPPs is which predictive distribution m^* to use for the imaginary data $\boldsymbol{y}^*$ in (1). Pérez and Berger (2002) discussed several different choices for m^*. An attractive choice, leading to the so-called *base-model approach*, arises from selecting a "reference" or "base" model m_0 for the training sample and defining $m^*(\boldsymbol{y}^*) = m_0^N(\boldsymbol{y}^*) \equiv f(\boldsymbol{y}^*|m_0)$ to be the prior predictive distribution, evaluated at $\boldsymbol{y}^*$, for the reference model m_0 under the baseline prior $\pi_0^N(\boldsymbol{\theta}_0)$. Then, for the reference model (i.e., when $m_\ell = m_0$), (1) reduces to $\pi_0^E(\boldsymbol{\theta}_0) = \pi_0^N(\boldsymbol{\theta}_0)$. Intuitively, the reference model should be at least as simple as the other competing models, and therefore a reasonable choice is to take m_0 to be a common sub-model of all $m_\ell \in \mathcal{M}$. This interpretation is close to the skeptical-prior approach described by Spiegelhalter, Abrams and Myles (2004, Section 5.5.2), in which a tendency toward the null hypothesis can be a-priori supported by centering the prior around values assumed by this hypothesis when no other information is available. In the variable-selection problem that we consider in this paper, the constant model (with no predictors) is clearly a good reference model that is nested in all the models under consideration.

An alternative selection for m_0 was made by Casella and Moreno (2006); they used the full model (with all available predictors) as the reference model. The latter approach was named *variable selection from above* (VSA) by Casella et al. (2009), in contrast to the constant-model approach, which they called *variable selection from below* (VSB). Moreno and Girón (2008) showed that the two approaches provide similar orderings of the linear models for finite sample size n. Other choices of m^* include the empirical distribution of the actual data, a suitable choice when it is not easy to determine a simple base-model (e.g., with non-nested models), or m^* can be viewed as arising from judgments as to how a training sample should behave, allowing for the possibility of eliciting m^* based on expert knowledge about the problem. In this paper, on the basis of the *parsimony principle*, we adopt the base-model approach, with the constant model as our reference model. This selection makes calculations simpler, and additionally (as discussed in the next section) makes the expected-posterior prior approach essentially equivalent to the intrinsic Bayes factor approach of Berger and Pericchi (1996*a*).

One of the advantages of using EPPs is that impropriety of baseline priors causes no indeterminacy. There is no problem with the use of an improper baseline prior $\pi_\ell^N(\boldsymbol{\theta}_\ell)$ in (1); the arbitrary constants cancel out in the calculation of any Bayes factor. Impropriety in m^* also does not cause indeterminacy, because m^* is common to the EPPs for all models. For example, under the base-model approach, if we assume that the baseline prior is improper and rewrite equation

(1) including the unknown normalizing constants, we have

$$\begin{aligned}\pi_\ell^E(\boldsymbol{\theta}_\ell) &= \int \left\{ \frac{f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell) c_\ell \, \pi_\ell^U(\boldsymbol{\theta}_\ell)}{\int f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell) \, c_\ell \, \pi_\ell^U(\boldsymbol{\theta}_\ell) \, d\boldsymbol{\theta}_\ell} \right\} \left\{ \int f(\boldsymbol{y}^*|\boldsymbol{\theta}_0, m_0) \, c_0 \, \pi_0^U(\boldsymbol{\theta}_0) \, d\boldsymbol{\theta}_0 \right\} d\boldsymbol{y}^* \\ &= c_0 \pi_\ell^U(\boldsymbol{\theta}_\ell) \int \int \frac{f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell) f(\boldsymbol{y}^*|\boldsymbol{\theta}_0, m_0)}{\int f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell) \, \pi_\ell^U(\boldsymbol{\theta}_\ell) \, d\boldsymbol{\theta}_\ell} \pi_0^U(\boldsymbol{\theta}_0) \, d\boldsymbol{\theta}_0 \, d\boldsymbol{y}^* \, ; \end{aligned} \tag{2}$$

here $\pi_\ell^U(\boldsymbol{\theta}_\ell)$ is the un-normalized kernel of $\pi_\ell^N(\boldsymbol{\theta}_\ell)$, i.e., $\pi_\ell^N(\boldsymbol{\theta}_\ell) = c_\ell \, \pi_\ell^U(\boldsymbol{\theta}_\ell)$, and c_ℓ is the corresponding normalizing constant for model $m_\ell \in \mathcal{M}$. From (2) we notice that $\pi_\ell^E(\boldsymbol{\theta}_\ell)$ depends only on the normalizing constant of the reference model c_0; this unknown normalizing constant is therefore common in all EPPs and will cancel out when calculating Bayes factors or posterior model probabilities. When a proper prior is used as a baseline prior, c_ℓ denotes the normalizing prior constant, which is directly influenced by the prior variance of $\boldsymbol{\theta}_\ell$. Therefore, in the case of proper baseline priors, the EPP and the corresponding Bayes factors will be relatively insensitive to large values of the prior variance of $\boldsymbol{\theta}_\ell$.

1.2 Connection between expected-posterior priors and intrinsic priors

Intrinsic prior (IP) distributions were introduced by Berger and Pericchi (1996*b*) to provide a fully-Bayesian justification for intrinsic Bayes factors (IBFs). Berger and Pericchi define the IP as the prior that "would yield Bayes factors that are approximately equal to IBFs, in an asymptotic case" (Berger and Pericchi, 2004).

Consider two models under comparison, m_0 and m_ℓ, with parameter vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_\ell$, respectively, where m_0 is nested in m_ℓ. Berger and Pericchi (1996*b*) proposed as intrinsic priors the following:

$$\begin{aligned}\pi_0^I(\boldsymbol{\theta}_0) &= \pi_0^N(\boldsymbol{\theta}_0) \, , \\ \pi_\ell^I(\boldsymbol{\theta}_\ell|\boldsymbol{\theta}_0) &= \pi_\ell^N(\boldsymbol{\theta}_\ell) E_{(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell)} \left[\frac{f(\boldsymbol{y}^*|\boldsymbol{\theta}_0, m_0)}{\int f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell) \, \pi_\ell^N(\boldsymbol{\theta}_\ell) \, d\boldsymbol{\theta}_\ell} \right] \\ &= \pi_\ell^N(\boldsymbol{\theta}_\ell) \int \frac{f(\boldsymbol{y}^*|\boldsymbol{\theta}_0, m_0) f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell)}{\int f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell) \, \pi_\ell^N(\boldsymbol{\theta}_\ell) \, d\boldsymbol{\theta}_\ell} \, d\boldsymbol{y}^* \, , \end{aligned} \tag{3}$$

where $\boldsymbol{y}^*$ is a hypothetical training sample of size n^*, $\pi_\ell^N(\boldsymbol{\theta}_\ell)$ $(\ell = 0, 1)$ is the improper prior distribution for model m_ℓ used as a starting point (the baseline prior) and $\pi_\ell^I(\boldsymbol{\theta}_\ell)$ $(\ell = 0, 1)$ is the resulting intrinsic prior distribution for model m_ℓ ; also see Casella and Moreno (2006).

If we use Bayes's Theorem in (3) to replace the corresponding likelihood-prior product of model m_ℓ (over its corresponding integral) with $\pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*)$ and write the marginal likelihood $m_0^N(\boldsymbol{y}^*)$ as an integral of the likelihood over the prior of the parameters of the reference model, we end up with the EPP as defined in equation (1), under the base-model approach.

1.3 Expected-posterior prior for Bayesian variable selection in Gaussian regression models

In what follows, we examine variable-selection problems in Gaussian regression models. We consider two models m_ℓ for $\ell = 0, 1$ with parameters $\boldsymbol{\theta}_\ell = (\boldsymbol{\beta}_\ell \, , \sigma_\ell^2)$ and likelihood specified by

$$(\boldsymbol{Y}|\mathrm{X}_\ell, \boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell) \sim N_n(\mathrm{X}_\ell \, \boldsymbol{\beta}_\ell \, , \sigma_\ell^2 \mathrm{I}_n) \, , \tag{4}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is a vector containing the responses for all subjects, X_ℓ is an $n \times d_\ell$ design matrix containing the values of the explanatory variables in its columns, I_n is the $n \times n$ identity matrix, $\boldsymbol{\beta}_\ell$ is a vector of length d_ℓ summarizing the effects of the covariates on the response $\boldsymbol{Y}$ and σ^2_ℓ is the error variance for model m_ℓ. Variable selection based on EPP was originally presented by Pérez (1998), while additional computational details have recently appeared in Fouskakis and Ntzoufras (2012).

Suppose we have imaginary/training data $\boldsymbol{y}^*$, of size n^*, and design matrix X^* of size $n^* \times p$, where p denotes the total number of available covariates. Then the EPP distribution, given by (1), will depend on X^* but not on $\boldsymbol{y}^*$, since the latter is integrated out. The selection of a *minimal training sample* has been proposed, to make the information content of the prior as small as possible, and this is an appealing idea. However, even the definition of *minimal* turns out to be open to question, since it is problem-specific (which models are we comparing?) and data-specific (how many variables are we considering?). For example, if we define "minimal" in terms of the largest model in every pairwise comparison, then the prior will change in every comparison, making the overall variable-selection procedure incoherent. Another idea is to let the size of the full model specify the minimal training sample; this choice makes inference within the current data set coherent, but what happens if some additional variables are included later in the study? In such cases, the size of the training sample and hence the prior must be changed, and the overall inference is again incoherent. Moreover, when the sample size is small in comparison to the number of covariates, working with a minimal training sample can result in an influential prior. Additionally, if the data derive from a highly structured situation, such as a complete randomized-blocks experiment, any choice of a small part of the data to act as a training sample would be somewhat untypical.

Even if the minimal-training-sample idea is accepted, the problem of choosing such a subset of the full data set still remains. A natural solution involves computing the arithmetic mean (or some other summary of distributional center) of the Bayes factors over all possible training samples, but this approach can be computationally infeasible, especially when the number n of observations is much larger than the number p of covariates; for example, with $(n, p) = (100, 50)$ and $(500, 100)$ there are about 10^{29} and 10^{107} possible training samples, respectively, over which to average. An obvious choice at this point is to take a random sample from the set of all possible minimal training samples, but this adds an extraneous layer of Monte-Carlo noise to the model-comparison process. These difficulties have been well-documented in the literature, but the quest for a fully satisfactory solution is still on-going; for example, Berger and Pericchi (2004) note that they "were unable to define any type of 'optimal' training sample."

2 Power-expected-posterior (PEP) prior methodology

In this paper, under the EPP methodology, we combine ideas from the power-prior approach of Ibrahim and Chen (2000) and the unit-information-prior approach of Kass and Wasserman (1995). As a first step, the likelihoods involved in the EPP distribution (see equation (2)) are raised to the power $\frac{1}{\delta}$ and density-normalized. Then we set the power parameter δ equal to n^*, to represent information equal to one data point; in this way the prior corresponds to a sample of size one with the same sufficient statistics as the observed data. Regarding the size of the training sample, n^*, this could be any integer from $(p + 2)$ (the minimal training sample size) to n. As will become clear below, we have found that significant advantages (and no disadvantages) arise from the choice

$n^* = n$, from which $\mathrm{X}^* = \mathrm{X}$. In this way we completely avoid the selection of a training sample and its effects on the posterior model comparison, while still holding the prior information content at one data point. Sensitivity analysis for different choices of n^* is performed as part of the first set of experimental results below (see Section 6.1).

For any $m_\ell \in \mathcal{M}$, we denote by $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\mathrm{X}_\ell^*)$ the baseline prior for model parameters $\boldsymbol{\beta}_\ell$ and σ_ℓ^2. Then the *power-expected-posterior* (PEP) prior $\pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \, \mathrm{X}_\ell^* \, , \delta)$ takes the following form:

$$\pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 \, | \, \mathrm{X}_\ell^* \, , \delta) = \pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\mathrm{X}_\ell^*) \int \frac{m_0^N(\boldsymbol{y}^* | \, \mathrm{X}_0^* \, , \delta)}{m_\ell^N(\boldsymbol{y}^* | \, \mathrm{X}_\ell^* \, , \delta)} \, f(\boldsymbol{y}^* | \, \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell^* \, , \delta) \, d\boldsymbol{y}^* \, , \tag{5}$$

where $f(\boldsymbol{y}^* | \, \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell^* \, , \delta) \propto f(\boldsymbol{y}^* | \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell^*)^{\frac{1}{\delta}}$ is the likelihood raised to the power of $\frac{1}{\delta}$ and density-normalized, i.e.,

$$\begin{aligned} f(\boldsymbol{y}^* | \, \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell^* \, , \delta) &= \frac{f(\boldsymbol{y}^* | \boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell^*)^{\frac{1}{\delta}}}{\int f(\boldsymbol{y}^* | \boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell^*)^{\frac{1}{\delta}} d\boldsymbol{y}^*} = \frac{f_{N_{n^*}}(\boldsymbol{y}^* \, ; \, \mathrm{X}_\ell^* \boldsymbol{\beta}_\ell \, , \sigma_\ell^2 \mathrm{I}_{n^*})^{\frac{1}{\delta}}}{\int f_{N_{n^*}}(\boldsymbol{y}^* \, ; \, \mathrm{X}_\ell^* \boldsymbol{\beta}_\ell \, , \sigma_\ell^2 \mathrm{I}_{n^*})^{\frac{1}{\delta}} d\boldsymbol{y}^*} \\ &= f_{N_{n^*}}(\boldsymbol{y}^* \, ; \, \mathrm{X}_\ell^* \boldsymbol{\beta}_\ell \, , \delta \, \sigma_\ell^2 \mathrm{I}_{n^*}) \, ; \end{aligned} \tag{6}$$

here $f_{N_d}(\boldsymbol{y} \, ; \, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of the d-dimensional Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, evaluated at $\boldsymbol{y}$.

The distribution $m_\ell^N(\boldsymbol{y}^* | \, \mathrm{X}_\ell^* \, , \delta)$ appearing in (5) is the prior predictive distribution (or the marginal likelihood), evaluated at $\boldsymbol{y}^*$, of model m_ℓ with the power likelihood defined in (6) under the baseline prior $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma_\ell^2 \, | \, \mathrm{X}_\ell^*)$, i.e.,

$$\begin{aligned} m_\ell^N(\boldsymbol{y}^* | \, \mathrm{X}_\ell^* \, , \delta) &= \iint f(\boldsymbol{y}^* \, | \, \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell^* \, , \delta) \, \pi_\ell^N(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2 \, | \, \mathrm{X}_\ell^*) \, d\boldsymbol{\beta}_\ell \, d\sigma_\ell^2 \\ &= \iint f_{N_{n^*}}(\boldsymbol{y}^* \, ; \, \mathrm{X}_\ell^* \boldsymbol{\beta}_\ell \, , \delta \, \sigma_\ell^2 \mathrm{I}_{n^*}) \, \pi_\ell^N(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2 \, | \, \mathrm{X}_\ell^*) \, d\boldsymbol{\beta}_\ell \, d\sigma_\ell^2 \, . \end{aligned} \tag{7}$$

Under the PEP prior distribution (5), the posterior distribution of the model parameters $(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2)$ is

$$\begin{aligned} \pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) &\propto f(\boldsymbol{y} | \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell) \, \pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \mathrm{X}_\ell^*, \delta) \\ &\propto \int f(\boldsymbol{y} | \boldsymbol{\beta}_\ell \, , \sigma_\ell^2, m_\ell \, ; \mathrm{X}_\ell) \, f(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2 | \boldsymbol{y}^*, m_\ell \, ; \, \mathrm{X}_\ell^* \, , \delta) \, m_0^N(\boldsymbol{y}^* | \mathrm{X}_0^* \, , \delta) \, d\boldsymbol{y}^* \\ &= \int f(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \boldsymbol{y}, \boldsymbol{y}^*, m_\ell \, ; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) \, m_\ell^N(\boldsymbol{y} | \boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}_\ell^* \, , \delta) \\ &\qquad \times m_0^N(\boldsymbol{y}^* | \mathrm{X}_0^* \, , \delta) \, d\boldsymbol{y}^* \, , \end{aligned} \tag{8}$$

where $f(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \boldsymbol{y}, \boldsymbol{y}^*, m_\ell \, ; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)$ and $m_\ell^N(\boldsymbol{y} | \boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}_\ell^* \, , \delta)$ are the posterior distribution of $(\boldsymbol{\beta}_\ell, \sigma_\ell^2)$ and the marginal likelihood of model m_ℓ, respectively, using data $\boldsymbol{y}$ and design matrix X_ℓ under prior $f(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2 | \boldsymbol{y}^*, m_\ell \, ; \, \mathrm{X}_\ell^* \, , \delta)$, i.e., the posterior of $(\boldsymbol{\beta}_\ell \, , \sigma_\ell^2)$ with power Normal likelihood (6) and baseline prior $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \mathrm{X}_\ell^*)$.

In what follows we present results for the PEP prior using two specific baseline choices: the usual independence Jeffreys prior (improper) and the g-prior (proper).

2.1 PEP-prior methodology with the Jeffreys baseline prior

Here we use the independence Jeffreys prior (or reference prior) as the baseline prior distribution. Hence for $m_\ell \in \mathcal{M}$ we have

$$\pi_\ell^N(\boldsymbol{\beta}_\ell\,, \sigma^2 \,|\, \mathrm{X}_\ell^*) = \frac{c_\ell}{\sigma_\ell^2}, \tag{9}$$

where c_ℓ is an unknown normalizing constant.

2.1.1 Prior setup

Following (5) for the baseline prior (9) and the power likelihood specified in (6), the PEP prior, for any model m_ℓ, now becomes

$$\begin{aligned}
\pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \mathrm{X}_\ell^*\,, \delta) &= \int f(\boldsymbol{\beta}_\ell\,, \sigma_\ell^2 | \boldsymbol{y}^*, m_\ell\,;\, \mathrm{X}_\ell^*\,, \delta)\, m_0^N(\boldsymbol{y}^* | \mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^* \\
&= \int f_{N_{d_\ell}}\Big[\boldsymbol{\beta}_\ell\,;\, \widehat{\boldsymbol{\beta}}_\ell^*, \delta\,(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\sigma_\ell^2\Big]\, f_{IG}\Big(\sigma_\ell^2\,;\, \frac{n^*-d_\ell}{2}, \frac{RSS_\ell^*}{2\delta}\Big) \\
&\qquad \times m_0^N(\boldsymbol{y}^* | \mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^*\,,
\end{aligned} \tag{10}$$

where $f_{IG}\,(y\,;\, a, b)$ is the density of the Inverse-Gamma distribution with parameters a and b and mean $\frac{b}{a-1}$, evaluated at y. Here $\widehat{\boldsymbol{\beta}}_\ell^* = (\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*T}\boldsymbol{y}^*$ is the MLE with outcome vector $\boldsymbol{y}^*$ and design matrix X_ℓ^*, and $RSS_\ell^* = \boldsymbol{y}^{*T}\big[\mathrm{I}_{n^*} - \mathrm{X}_\ell^*(\mathrm{X}_\ell{}^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell{}^{*T}\big]\boldsymbol{y}^*$ is the residual sum of squares using $(\boldsymbol{y}^*, \mathrm{X}_\ell^*)$ as data. The prior predictive distribution of any model m_ℓ with power likelihood defined in (6) under the baseline prior (9) is given by

$$m_\ell^N(\boldsymbol{y}^* \,|\, \mathrm{X}_\ell^*\,, \delta) = c_\ell\, \pi^{\frac{1}{2}(d_\ell - n^*)}\, |\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*|^{-\frac{1}{2}}\, \Gamma\left(\frac{n^*-d_\ell}{2}\right) RSS_\ell^{*\,-\left(\frac{n^*-d_\ell}{2}\right)}\,. \tag{11}$$

2.1.2 Posterior distribution

For the PEP prior (10), the posterior distribution of the model parameters $(\boldsymbol{\beta}_\ell\,, \sigma_\ell^2)$ is given by (8) with $f(\boldsymbol{\beta}_\ell, \sigma_\ell^2 | \boldsymbol{y}, \boldsymbol{y}^*, m_\ell\,; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)$ and $m_\ell^N(\boldsymbol{y} | \boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)$ as the posterior distribution of $(\boldsymbol{\beta}_\ell, \sigma_\ell^2)$ and the marginal likelihood of model m_ℓ, respectively, using data $\boldsymbol{y}$, design matrix X_ℓ, and the Normal-Inverse-Gamma distribution appearing in (10) as prior. Hence

$$\begin{aligned}
f(\boldsymbol{\beta}_\ell | \sigma_\ell^2, \boldsymbol{y}, \boldsymbol{y}^*, m_\ell\,; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) &= f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell\,;\, \widetilde{\boldsymbol{\beta}}^N,\ \widetilde{\Sigma}^N\sigma_\ell^2\big) \text{ and} \\
f(\sigma_\ell^2 | \boldsymbol{y}, \boldsymbol{y}^*, m_\ell\,; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) &= f_{IG}(\sigma_\ell^2\,;\, \widetilde{a}_\ell^N, \widetilde{b}_\ell^N)\,,
\end{aligned} \tag{12}$$

with

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^N &= \widetilde{\Sigma}^N(\mathrm{X}_\ell^T\boldsymbol{y} + \delta^{-1}\mathrm{X}_\ell^{*T}\boldsymbol{y}^*),\ \widetilde{\Sigma}^N = \Big[\mathrm{X}_\ell^T\mathrm{X}_\ell + \delta^{-1}\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*\Big]^{-1} \text{ and} \\
\widetilde{a}_\ell^N &= \frac{n+n^*-d_\ell}{2}\,,\ \widetilde{b}_\ell^N = \frac{SS_\ell^N + \delta^{-1}RSS_\ell^*}{2} + b_\ell\,;
\end{aligned} \tag{13}$$

here

$$SS_\ell^N = \big(\boldsymbol{y} - \mathrm{X}_\ell\,\widehat{\boldsymbol{\beta}}_\ell^*\big)^T\Big[\mathrm{I}_n + \delta\,\mathrm{X}_\ell(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^T\Big]^{-1}\big(\boldsymbol{y} - \mathrm{X}_\ell\,\widehat{\boldsymbol{\beta}}_\ell^*\big) \tag{14}$$

and

$$m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) = f_{St_n}\left\{\boldsymbol{y}\,;\, n^* - d_\ell,\, \mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell^*,\, \frac{RSS_\ell^*}{\delta(n^*-d_\ell)}\left[\mathrm{I}_n + \delta\,\mathrm{X}_\ell(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^T\right]\right\}, \qquad (15)$$

in which $St_n(\cdot\,; d, \boldsymbol{\mu}, \Sigma)$ is the multivariate Student distribution in n dimensions with d degrees of freedom, location $\boldsymbol{\mu}$ and scale Σ. Thus the posterior distribution of the model parameters $(\boldsymbol{\beta}_\ell\,, \sigma_\ell^2)$ under the PEP prior (10) is

$$\begin{aligned}\pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) \;\; \propto \;\; & \int f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell\,;\, \widetilde{\boldsymbol{\beta}}^N,\; \widetilde{\Sigma}^N\sigma_\ell^2\big)\, f_{IG}(\sigma_\ell^2\,;\, \widetilde{a}_\ell^N, \widetilde{b}_\ell^N) \\ & \times\, m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*;\, \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^*\,, \qquad (16)\end{aligned}$$

with $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)$ given in (11). A detailed MCMC scheme for sampling from this distribution is presented in Section 3.

2.2 PEP-prior methodology with the g-prior as baseline

Here we use the g-prior as the baseline prior distribution; in other words, for any $m_\ell \in \mathcal{M}$

$$\pi_\ell^N(\boldsymbol{\beta}_\ell|\sigma_\ell^2\,;\, \mathrm{X}_\ell^*) = f_{N_{d_\ell}}\left[\boldsymbol{\beta}_\ell\,;\, \mathbf{0}, g\,(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\sigma_\ell^2\right] \text{ and } \pi_\ell^N(\sigma_\ell^2) = f_{IG}\left(\sigma_\ell^2\,;\, a_\ell, b_\ell\right). \qquad (17)$$

2.2.1 Prior setup

For any model m_ℓ, under the baseline prior setup (17) and the power likelihood (6), the prior predictive distribution is a multivariate Student distribution with $2\,a$ degrees of freedom, mean vector $\mathbf{0}$ and scale parameter $\Sigma_\ell = \frac{b_\ell}{a_\ell}[\delta\,\mathrm{I}_{n^*} + g\,\mathrm{X}_\ell^*\left(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*\right)^{-1}\mathrm{X}_\ell^{*T}]$. Hence

$$m_\ell^N(\boldsymbol{y}^*\,|\,\mathrm{X}_\ell^*\,, \delta) = f_{St_{n^*}}\left(\boldsymbol{y}^*\,; 2\,a_\ell, \mathbf{0}, \frac{b_\ell}{a_\ell}\Lambda_\ell^{*-1}\right), \qquad (18)$$

where

$$\Lambda_\ell^{*-1} = \delta\left[\mathrm{I}_{n^*} - \frac{g}{g+\delta}\mathrm{X}_\ell^*\left(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*\right)^{-1}\mathrm{X}_\ell^{*T}\right]^{-1} = \delta\,\mathrm{I}_{n^*} + g\,\mathrm{X}_\ell^*\left(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*\right)^{-1}\mathrm{X}_\ell^{*T}. \qquad (19)$$

In the special case of the constant model, (19) simplifies to $\left(\delta\,\mathrm{I}_{n^*} + \frac{g}{n}\mathbf{1}_{n^*}\mathbf{1}_{n^*}^T\right)$, where $\mathbf{1}_{n^*}$ is a vector of length n^* with all elements equal to one.

Following (5) for the baseline prior (17) and the power likelihood specified in (6), the PEP prior, for any model m_ℓ, now becomes

$$\begin{aligned}\pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\mathrm{X}_\ell^*\,, \delta) \;\; &= \;\; \int f(\boldsymbol{\beta}_\ell\,, \sigma_\ell^2|\boldsymbol{y}^*, m_\ell\,;\, \mathrm{X}_\ell^*\,, \delta)\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^* \\ &= \;\; \int f_{N_{d_\ell}}\left[\boldsymbol{\beta}_\ell\,;\, w\,\widehat{\boldsymbol{\beta}}_\ell^*, w\,\delta\,(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\sigma_\ell^2\right] f_{IG}\Big(\sigma_\ell^2\,;\, a_\ell + \frac{n^*}{2}, b_\ell + \frac{SS_\ell^*}{2}\Big) \\ & \qquad \times\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^*\,. \qquad (20)\end{aligned}$$

Here $w = \frac{g}{g+\delta}$ is the shrinkage weight, $\widehat{\boldsymbol{\beta}}_\ell^* = (\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*T}\boldsymbol{y}^*$ is the MLE with outcome vector $\boldsymbol{y}^*$ and design matrix X_ℓ^*, and $SS_\ell^* = \boldsymbol{y}^{*T}\Lambda_\ell^*\,\boldsymbol{y}^*$ is the posterior sum of squares.

The PEP marginal prior distributions can then be expressed as

$$\begin{aligned}
\pi_\ell^{PE}(\boldsymbol{\beta}_\ell|\mathrm{X}_\ell^*\,,\delta) &= \int \pi_\ell^{PE}(\boldsymbol{\beta}_\ell,\sigma_\ell^2|\mathrm{X}_\ell^*\,,\delta)\,d\sigma_\ell^2 \qquad (21)\\
&= \int\left\{\int f_{N_{d_\ell}}\left[\boldsymbol{\beta}_\ell\,;\,w\,\widehat{\boldsymbol{\beta}}_\ell^*,\delta\,w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\sigma_\ell^2\right] f_{IG}\left(\sigma_\ell^2\,;\,a_\ell+\frac{n^*}{2},b_\ell+\frac{SS_\ell^*}{2}\right)d\sigma_\ell^2\right\}\\
&\qquad\times m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\\
&= \int f_{St_{d_\ell}}\left[\boldsymbol{\beta}_\ell\,;\,2\,a_\ell+n^*,\,w\,\widehat{\boldsymbol{\beta}}_\ell^*,\,\delta\,w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\frac{b_\ell+\frac{SS_\ell^*}{2}}{a_\ell+\frac{n^*}{2}}\right] m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*
\end{aligned}$$

and

$$\begin{aligned}
\pi_\ell^{PE}(\sigma_\ell^2|\mathrm{X}_\ell^*\,,\delta) &= \int \pi_\ell^{PE}(\boldsymbol{\beta}_\ell,\sigma_\ell^2|\mathrm{X}_\ell^*\,,\delta)\,d\boldsymbol{\beta}_\ell\\
&= \int\left\{\int f_{N_{d_\ell}}\left[\boldsymbol{\beta}_\ell\,;\,w\,\widehat{\boldsymbol{\beta}}_\ell^*,\delta\,w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\sigma_\ell^2\right]d\boldsymbol{\beta}_\ell\right\}\\
&\qquad\times f_{IG}\left(\sigma_\ell^2\,;\,a_\ell+\frac{n^*}{2},b_\ell+\frac{SS_\ell^{\,*}}{2}\right) m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\\
&= \int f_{IG}\left(\sigma_\ell^2\,;\,a_\ell+\frac{n^*}{2},b_\ell+\frac{SS_\ell^*}{2}\right) m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\,. \qquad (22)
\end{aligned}$$

The prior mean vector and covariance matrix of $\boldsymbol{\beta}_\ell$, and the prior mean and variance of σ_ℓ^2, can be calculated analytically from these expressions; details are available in Theorems 1 and 2 in Appendix A.

2.2.2 Posterior distribution

The distributions $f(\boldsymbol{\beta}_\ell,\sigma_\ell^2|\boldsymbol{y},\boldsymbol{y}^*,m_\ell\,;\mathrm{X}_\ell,\mathrm{X}_\ell^*,\delta)$ and $m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*;\mathrm{X}_\ell,\mathrm{X}_\ell^*\,,\delta)$ involved in the calculation of the posterior distribution (8) are now the posterior distribution of $(\boldsymbol{\beta}_\ell,\sigma_\ell^2)$ and the marginal likelihood of model m_ℓ, respectively, using data $\boldsymbol{y}$, design matrix X_ℓ, and $f(\boldsymbol{\beta}_\ell\,,\sigma_\ell^2|\boldsymbol{y}^*,m_\ell\,;\,\mathrm{X}_\ell^*\,,\delta)$ as a prior density (which is the Normal-Inverse-Gamma distribution appearing in (20)). Therefore the posterior distribution of the model parameters $(\boldsymbol{\beta}_\ell\,,\sigma_\ell^2)$ under the PEP prior (20) is given by

$$\begin{aligned}
\pi_\ell^{PE}(\boldsymbol{\beta}_\ell,\sigma_\ell^2|\boldsymbol{y};\mathrm{X}_\ell,\mathrm{X}_\ell^*,\delta) \;\propto\; &\int f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell\,;\,\widetilde{\boldsymbol{\beta}}^N,\;\widetilde{\Sigma}^N\sigma_\ell^2\big)\,f_{IG}(\sigma_\ell^2\,;\,\widetilde{a}_\ell^N,\widetilde{b}_\ell^N)\\
&\times\, m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*;\,\mathrm{X}_\ell,\mathrm{X}_\ell^*\,,\delta)\,m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\,, \qquad (23)
\end{aligned}$$

with

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^N &= \widetilde{\Sigma}^N(\mathrm{X}_\ell^T\boldsymbol{y}+\delta^{-1}\mathrm{X}_\ell^{*^T}\boldsymbol{y}^*),\;\widetilde{\Sigma}^N=\left[\mathrm{X}_\ell^T\mathrm{X}_\ell+(w\,\delta)^{-1}\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*\right]^{-1}\;\text{and}\\
\widetilde{a}_\ell^N &= \frac{n+n^*}{2}+a_\ell\,,\;\widetilde{b}_\ell^N=\frac{SS_\ell^N+SS_\ell^*}{2}+b_\ell\,; \qquad (24)
\end{aligned}$$

here

$$SS_\ell^N=\big(\boldsymbol{y}-w\,\mathrm{X}_\ell\,\widehat{\boldsymbol{\beta}}_\ell^*\big)^T\left[\mathrm{I}_n+\delta\,w\,\mathrm{X}_\ell(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^T\right]^{-1}\big(\boldsymbol{y}-w\,\mathrm{X}_\ell\,\widehat{\boldsymbol{\beta}}_\ell^*\big)\,, \qquad (25)$$

while

$$m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*;\mathrm{X}_\ell,\mathrm{X}_\ell^*,\delta)=f_{St_n}\left\{\boldsymbol{y}\,;\,2\,a_\ell+n^*,w\,\mathrm{X}_\ell\widehat{\boldsymbol{\beta}}_\ell^*,\,\frac{2b_\ell+SS_\ell^*}{2\,a_\ell+n^*}\left[\mathrm{I}_n+w\,\delta\,\mathrm{X}_\ell(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^T\right]\right\}\,, \quad (26)$$

and $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)$ is given in (18). A detailed MCMC scheme for sampling from this posterior distribution is presented in Section 3.

2.2.3 Specification of hyper-parameters

The marginal likelihood for the PEP-prior methodology, using the g-prior as a baseline, depends on the selection of the hyper-parameters g, a and b. We make the following proposals for specifying these quantities, in settings in which strong prior information about the parameter vectors in the models is not available.

The parameter g in the Normal baseline prior is set to $\delta\,n^*$, so that with $\delta=n^*$ we use $g=(n^*)^2$. This choice will make the g-prior contribute information equal to one data point within the posterior $f(\boldsymbol{\beta}_\ell\,,\sigma_\ell^2|\boldsymbol{y}^*,m_\ell\,;\mathrm{X}_\ell^*\,,\delta)$. In this manner, the entire PEP prior accounts for information equal to $\left(1+\frac{1}{\delta}\right)$ data points.

We set the parameters a and b in the Inverse-Gamma baseline prior to 0.01, yielding a baseline prior mean of 1 and variance of 100 (i.e., a large amount of prior uncertainty) for the precision parameter. (If strong prior information about the model parameters is available, Theorems 1 and 2 in Appendix A can be used to guide the choice of a and b.)

2.3 Connection between the PEP-Jeffreys-prior and the PEP-g-prior distributions

By comparing the posterior distributions under the two different baseline schemes described in Sections 2.1 and 2.2, it is straightforward to prove that they coincide under the following conditions $(*)$: large g (and therefore for $w\approx 1$), $a_\ell=-\frac{d_\ell}{2}$ and $b_\ell=0$.

To be more specific, the posterior distribution in both cases takes the form of equation (16). The parameters of the Normal-Inverse-Gamma distribution (see equations (24)) involved in the posterior distribution using the g-prior as baseline become equal to the corresponding parameters for the Jeffreys baseline (see equations (13)) with parameter values $(*)$. Similarly, the conditional marginal likelihood $m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*;\,\mathrm{X}_\ell,\mathrm{X}_\ell^*\,,\delta)$ under the two baseline priors (see equations (15) and (26)) becomes the same under conditions $(*)$.

Finally, the prior predictive densities $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)$ involved in equations (16) and (23) can be written as $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\propto(2\,b_\ell\,+\,SS_\ell^*)^{-\frac{n^*+a_\ell}{2}}$ for the g-prior baseline and as $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\propto RSS_\ell^{*\,-\frac{n^*-d_\ell}{2}}$ for the Jeffreys baseline. For large values of g, $SS_\ell^*\to\delta^{-1}RSS_\ell^*$, and the two unnormalized prior predictive densities clearly become equal if we further set $a_\ell=-\frac{d_\ell}{2}$ and $b_\ell=0$. Any differences in the normalizing constants of $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)$ cancel out when normalizing the posterior distributions (16) and (23).

For these reasons, the posterior results using the Jeffreys prior as baseline can be obtained as a special (limiting) case of the results using the g-prior as baseline. This can be beneficial for the computation of the posterior distribution, which follows in Section 3, and the estimation of the marginal likelihood presented in Section 4.

3 MCMC for sampling from the posterior

To generate MCMC samples from (16) or (23) under the two baseline prior choices, we consider the following conditional distribution:

$$f(\boldsymbol{\beta}_\ell, \sigma^2_\ell, \boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta) \;\propto\; f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell\,;\, \widetilde{\boldsymbol{\beta}}^N,\; \widetilde{\Sigma}^N\sigma^2_\ell\big) f_{IG}(\sigma^2_\ell\,;\, \widetilde{a}^N_\ell, \widetilde{b}^N_\ell) \times m^N_\ell(\boldsymbol{y}|\boldsymbol{y}^*;\, \mathrm{X}_\ell, \mathrm{X}^*_\ell\,, \delta)\, m^N_0(\boldsymbol{y}^*|\mathrm{X}^*_0\,, \delta)\,. \tag{27}$$

The parameters in the above Normal-Inverse-Gamma distribution are given in (13) and (24) for the baseline Jeffreys and g-priors, respectively. From the above we have that $f(\sigma^2_\ell|\boldsymbol{y}, \boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta) = f_{IG}(\sigma^2_\ell\,;\, \widetilde{a}^N_\ell, \widetilde{b}^N_\ell)$, $f(\boldsymbol{\beta}_\ell|\sigma^2_\ell, \boldsymbol{y}, \boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta) = f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell\,;\, \widetilde{\boldsymbol{\beta}}^N, \widetilde{\Sigma}^N\sigma^2_\ell\big)$ and

$$\begin{aligned} f(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta) \;&\propto\; \iint f(\boldsymbol{\beta}_\ell, \sigma^2_\ell, \boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}^*_\ell, \delta)\, d\boldsymbol{\beta}_\ell\; d\sigma^2_\ell \\ &\propto\; \left[\iint f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell\,;\, \widetilde{\boldsymbol{\beta}}^N,\; \widetilde{\Sigma}^N\sigma^2_\ell\big)\; f_{IG}(\sigma^2_\ell\,;\, \widetilde{a}^N_\ell, \widetilde{b}^N_\ell)\, d\boldsymbol{\beta}_\ell\, d\sigma^2_\ell\right] \\ &\qquad \times\; m^N_\ell(\boldsymbol{y}|\boldsymbol{y}^*;\, \mathrm{X}_\ell, \mathrm{X}^*_\ell\,, \delta)\, m^N_0(\boldsymbol{y}^*|\mathrm{X}^*_0\,, \delta) \\ &\propto\; m^N_\ell(\boldsymbol{y}|\boldsymbol{y}^*;\, \mathrm{X}_\ell, \mathrm{X}^*_\ell\,, \delta)\, m^N_0(\boldsymbol{y}^*|\mathrm{X}^*_0\,, \delta) \\ &\propto\; m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y};\, \mathrm{X}_\ell, \mathrm{X}^*_\ell\,, \delta)\, \frac{m^N_0(\boldsymbol{y}^*|\mathrm{X}^*_0\,, \delta)}{m^N_\ell(\boldsymbol{y}^*|\mathrm{X}^*_\ell\,, \delta)}\,, \end{aligned} \tag{28}$$

with

$$m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y},\, \mathrm{X}_\ell, \mathrm{X}^*_\ell\,, \delta) \;=\; \iint f(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \sigma^2_\ell, m_\ell\,; \mathrm{X}^*_\ell\,, \delta)\, f(\boldsymbol{\beta}_\ell\,, \sigma^2_\ell|\boldsymbol{y}, m_\ell\,;\, \mathrm{X}_\ell)\, d\boldsymbol{\beta}_\ell\, d\sigma^2_\ell\,. \tag{29}$$

For the baseline g-prior, (29) becomes

$$\begin{aligned} m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y},\, \mathrm{X}_\ell, \mathrm{X}^*_\ell\,, \delta) \;&=\; \iint f_{N_{n^*}}(\boldsymbol{y}^*\,;\, \mathrm{X}^*_\ell\boldsymbol{\beta}_\ell\,, \delta\,\sigma^2_\ell\,\mathrm{I}_{n^*})\, f_{N_{d_\ell}}\left[\boldsymbol{\beta}_\ell\,;\, \frac{g}{g+1}\widehat{\boldsymbol{\beta}}_\ell\,, \frac{g}{g+1}(\mathrm{X}^T_\ell\mathrm{X}_\ell)^{-1}\sigma^2_\ell\right] \\ &\qquad \times\; f_{IG}\Big(\sigma^2_\ell\,;\, a_\ell + \frac{n}{2}, b_\ell + \frac{SS_\ell}{2}\Big)\, d\boldsymbol{\beta}_\ell\, d\sigma^2_\ell \\ &= f_{St_{n^*}}\left\{\boldsymbol{y}^*\,;\, 2\,a_\ell + n,\; \frac{g}{g+1}\mathrm{X}^*_\ell\widehat{\boldsymbol{\beta}}_\ell\,, \frac{2b_\ell + SS_\ell}{2\,a_\ell + n}\left[\delta\,\mathrm{I}_{n^*} + \frac{g}{g+1}\mathrm{X}^*_\ell(\mathrm{X}^T_\ell\mathrm{X}_\ell)^{-1}\mathrm{X}^{*^T}_\ell\right]\right\}\,, \end{aligned} \tag{30}$$

where $SS_\ell = \boldsymbol{y}^T\left[\mathrm{I}_n - \frac{g}{g+1}\mathrm{X}_\ell\,(\mathrm{X}^T_\ell\mathrm{X}_\ell)^{-1}\mathrm{X}^T_\ell\right]\boldsymbol{y}$; for the Jeffreys baseline prior the expression is the same with $\frac{g}{g+1} = 1$, $a_\ell = -\frac{d_\ell}{2}$ and $b_\ell = 0$. Therefore for the Jeffreys baseline prior, equation (29) becomes

$$m^N_\ell(\boldsymbol{y}^*|\boldsymbol{y},\, \mathrm{X}_\ell, \mathrm{X}^*_\ell\,, \delta) = f_{St_{n^*}}\left\{\boldsymbol{y}^*\,;\, n - d_\ell,\, \mathrm{X}^*_\ell\widehat{\boldsymbol{\beta}}_\ell\,, \frac{SS_\ell}{n - d_\ell}\left[\delta\,\mathrm{I}_{n^*} + \mathrm{X}^*_\ell(\mathrm{X}^T_\ell\mathrm{X}_\ell)^{-1}\mathrm{X}^{*^T}_\ell\right]\right\}\,, \tag{31}$$

with the posterior sum of squares now given by $SS_\ell = \boldsymbol{y}^T\left[\mathrm{I}_n - \mathrm{X}_\ell\,(\mathrm{X}^T_\ell\mathrm{X}_\ell)^{-1}\mathrm{X}^T_\ell\right]\boldsymbol{y}$.

Using the above expressions, we can specify the following MCMC scheme, in which the Inverse-Gamma distribution $IG(a, b)$ was previously defined in Section 2.1.1:

(1) Generate $\boldsymbol{y}^*$ from (28);

(2) Generate σ_ℓ^2 from $IG(\widetilde{a}_\ell^N, \widetilde{b}_\ell^N)$; and

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

In Step 1, we can generate the imaginary data $\boldsymbol{y}^*$ by using a Metropolis-Hastings algorithm with proposal $q(\boldsymbol{y}^{*'}) = m_\ell^N(\boldsymbol{y}^{*'}|\boldsymbol{y}, \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)$ given in (30) or (31) (for the baseline g-prior or Jeffreys prior choices, respectively) and acceptance probability

$$\alpha = \min\left[1, \frac{m_0^N(\boldsymbol{y}^{*'}|\mathrm{X}_0^*, \delta)\, m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*, \delta)}{m_\ell^N(\boldsymbol{y}^{*'}|\mathrm{X}_\ell^*, \delta)\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)}\right], \tag{32}$$

where $m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*, \delta)$ is given in (11) and (18) for the baseline Jeffreys and g-prior choices, respectively.

4 Marginal-likelihood computation

Under the PEP-prior approach, the marginal likelihood of any model $m_\ell \in \mathcal{M}$ is

$$\begin{aligned}
m_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) &= \int\!\!\int f(\boldsymbol{y}|\boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell; \mathrm{X}_\ell)\, \pi_\ell^{PE}(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\mathrm{X}_\ell^*, \delta)\, d\boldsymbol{\beta}_\ell\, d\sigma_\ell^2 \\
&= \int\!\!\int\!\!\int f(\boldsymbol{y}|\boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell; \mathrm{X}_\ell)\, f(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\boldsymbol{y}^*, m_\ell; \mathrm{X}_\ell^*, \delta)\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)\, d\boldsymbol{\beta}_\ell\, d\sigma_\ell^2\, d\boldsymbol{y}^* \\
&= \int\!\!\int\!\!\int f(\boldsymbol{y}|\boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell; \mathrm{X}_\ell)\, f(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \sigma_\ell^2, m_\ell; \mathrm{X}_\ell^*, \delta)\, \pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\mathrm{X}_\ell^*)\, \frac{m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*, \delta)}\, d\boldsymbol{\beta}_\ell\, d\sigma_\ell^2\, d\boldsymbol{y}^* \\
&= \int \left[\int\!\!\int f(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\boldsymbol{y}, \boldsymbol{y}^*, m_\ell; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)\, d\boldsymbol{\beta}_\ell\, d\sigma_\ell^2\right] m_\ell^N(\boldsymbol{y}, \boldsymbol{y}^*|\mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)\, \frac{m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*, \delta)}\, d\boldsymbol{y}^* \\
&= m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*) \int \frac{m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y}, \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*, \delta)}\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)\, d\boldsymbol{y}^*.
\end{aligned} \tag{33}$$

Note that in the above expression $m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*)$ is the marginal likelihood of model m_ℓ for the actual data under the baseline prior. Therefore, under the baseline g-prior, (17) is given by

$$m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*) = f_{St_n}\left\{\boldsymbol{y}; 2\,a_\ell, \mathbf{0}, \frac{b_\ell}{a_\ell}\left[\mathrm{I}_n + g\,\mathrm{X}_\ell\left(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*\right)^{-1}\mathrm{X}_\ell^T\right]\right\}, \tag{34}$$

while under the Jeffreys baseline prior (9) is given by equation (11) with data $(\boldsymbol{y}, \mathrm{X}_\ell)$.

In cases when the marginal likelihood in (33) is not analytically tractable, possible Monte-Carlo estimates include the following:

(1) Generate $\boldsymbol{y}^{*(t)}$ $(t = 1, \ldots, T)$ from $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)$ and estimate the marginal likelihood by

$$\widehat{m}_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta) = \frac{1}{T}\, m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*) \sum_{t=1}^{T} \frac{m_\ell^N(\boldsymbol{y}^{*(t)}|\boldsymbol{y}, \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)}{m_\ell^N(\boldsymbol{y}^{*(t)}|\mathrm{X}_\ell^*, \delta)}. \tag{35}$$

This approach was also proposed by Pérez and Berger (2002).

(2) Generate $\boldsymbol{y}^{*(t)}$ $(t = 1, \ldots, T)$ from $m_0^N(\boldsymbol{y}^*|\boldsymbol{y},\, \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)$ and estimate the marginal likelihood by

$$\widehat{m}_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell\,, \mathrm{X}_\ell^*\,, \delta) = m_0^N(\boldsymbol{y}|\mathrm{X}_0, \mathrm{X}_0^*) \left[\frac{1}{T}\sum_{t=1}^{T} \frac{m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^{*(t)}, \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}{m_0^N(\boldsymbol{y}|\boldsymbol{y}^{*(t)}, \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)}\right]. \tag{36}$$

This works because

$$\begin{aligned}
m_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell\,, \mathrm{X}_\ell^*\,, \delta) &= \int \frac{m_\ell^N(\boldsymbol{y}, \boldsymbol{y}^*|\mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*\,, \delta)}\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)\, \frac{m_0^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)}{m_0^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)}\, d\boldsymbol{y}^* \\
&= \int \frac{m_\ell^N(\boldsymbol{y}, \boldsymbol{y}^*|\mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*\,, \delta)}\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)\, \frac{m_0^N(\boldsymbol{y}|\mathrm{X}_0, \mathrm{X}_0^*)}{m_0^N(\boldsymbol{y}, \boldsymbol{y}^*|\mathrm{X}_0, \mathrm{X}_0^*\,, \delta)}\, m_0^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^* \\
&= m_0^N(\boldsymbol{y}|\mathrm{X}_0, \mathrm{X}_0^*) \int \frac{m_\ell^N(\boldsymbol{y}, \boldsymbol{y}^*|\mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}{m_0^N(\boldsymbol{y}, \boldsymbol{y}^*|\mathrm{X}_0, \mathrm{X}_0^*\,, \delta)}\, \frac{m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*\,, \delta)}\, m_0^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^* \\
&= m_0^N(\boldsymbol{y}|\mathrm{X}_0, \mathrm{X}_0^*) \int \frac{m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*, \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}{m_0^N(\boldsymbol{y}|\boldsymbol{y}^*, \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)}\, m_0^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^*\,.
\end{aligned} \tag{37}$$

(3) Generate $\boldsymbol{y}^{*(t)}$ $(t = 1, \ldots, T)$ from $m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y},\, \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)$ and estimate the marginal likelihood by

$$\widehat{m}_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell\,, \mathrm{X}_\ell^*\,, \delta) = m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*) \left[\frac{1}{T}\sum_{t=1}^{T} \frac{m_0^N(\boldsymbol{y}^{*(t)}|\mathrm{X}_0^*\,, \delta)}{m_\ell^N(\boldsymbol{y}^{*(t)}|\mathrm{X}_\ell^*\,, \delta)}\right], \tag{38}$$

since

$$\begin{aligned}
m_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell\,, \mathrm{X}_\ell^*\,, \delta) &= \int \frac{m_\ell^N(\boldsymbol{y}, \boldsymbol{y}^*|\mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*\,, \delta)}\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^* \\
&= \int \frac{m_\ell^N(\boldsymbol{y}^*|\,\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta) m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*\,, \delta)}\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)\, d\boldsymbol{y}^* \\
&= m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*) \int \frac{m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*\,, \delta)}\, m_\ell^N(\boldsymbol{y}^*|\,\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)\, d\boldsymbol{y}^*\,.
\end{aligned}$$

(4) Generate $\boldsymbol{y}^{*(t)}$ $(t = 1, \ldots, T)$ from $m_\ell^N(\boldsymbol{y}^*|\,\boldsymbol{y}; \mathrm{X}_\ell\,, \mathrm{X}_\ell^*\,, \delta)$ and estimate the marginal likelihood by

$$\widehat{m}_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell\,, \mathrm{X}_\ell^*\,, \delta) = m_0^N(\boldsymbol{y}|\mathrm{X}_0, \mathrm{X}_0^*) \left[\frac{1}{T}\sum_{t=1}^{T} \frac{m_\ell^N(\boldsymbol{y}|\,\boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}{m_0^N(\boldsymbol{y}|\,\boldsymbol{y}^*; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)} \frac{m_0^N(\boldsymbol{y}^*|\,\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)}{m_\ell^N(\boldsymbol{y}^*|\,\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}\right], \tag{39}$$

since, according to Monte-Carlo scheme (3),

$$\begin{aligned}
m_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell\,, \mathrm{X}_\ell^*\,, \delta) &= \int \frac{m_\ell^N(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*)}{m_\ell^N(\boldsymbol{y}^*|\mathrm{X}_\ell^*\,, \delta)}\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,, \delta)\, m_\ell^N(\boldsymbol{y}^*|\,\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)\, d\boldsymbol{y}^* \\
&= m_0^N(\boldsymbol{y}|\mathrm{X}_0, \mathrm{X}_0^*) \int \frac{m_\ell^N(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}{m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)}\, \frac{m_0^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)}{m_0^N(\boldsymbol{y}|\boldsymbol{y}^*; \mathrm{X}_0, \mathrm{X}_0^*\,, \delta)} \\
&\qquad \times m_\ell^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*\,, \delta)\, d\boldsymbol{y}^*\,.
\end{aligned} \tag{40}$$

The first and second Monte-Carlo schemes are straightforward, since we only need to obtain a single sample of $\boldsymbol{y}^*$ from the prior and the posterior predictive distribution from model m_0, respectively; then we estimate the marginal likelihood of every model using those simulated values. Nevertheless we expect those estimates for the marginal likelihood of model m_ℓ to have large Monte-Carlo error, since the imaginary data are generated from importance functions that do not make full use of the data $\boldsymbol{y}$ (in the first Monte-Carlo scheme) or the stochastic structure of model m_ℓ (in both of these schemes).

The third and fourth Monte-Carlo schemes generate imaginary data from the posterior predictive distribution of the model under consideration, and thus we expect them to be more accurate. Moreover, in the fourth Monte-Carlo scheme, when we estimate Bayes factors we only need to evaluate posterior predictive distributions, which are available even in the case of baseline improper priors.

Using arguments similar to those in Section 2.3, it is clear that the marginal likelihoods $m_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell\,,\mathrm{X}_\ell^*\,,\delta)$ under the two baseline prior choices considered in this paper will end up with the same posterior odds and model probabilities for $g \to \infty$, $a_\ell = -\frac{d_\ell}{2}$ and $b_\ell = 0$. This is due to the fact that the posterior predictive densities involved in this expression become the same for the above-mentioned prior parameter values, while the corresponding prior predictive density will be the same up to some normalizing constants common across all models that cancel out in the calculation of posterior odds and model probabilities.

5 Model-search algorithm

For any number of variables p under consideration in our model-uncertainty problem, the number of models for which we need to evaluate the marginal likelihood is 2^p, which is enormous even when p is only moderately large. As a result, full enumeration (across all models) of the marginal likelihoods and the corresponding posterior model probabilities needed in Bayesian variable-selection problems becomes infeasible. For this reason, in such problems, advanced MCMC methods are typically used as model-search algorithms to identify the most important models and variables. Estimation of posterior model odds can then be performed efficiently within reduced model spaces in which unimportant variables have been excluded, according to the model search algorithm; see Fouskakis, Ntzoufras and Draper (2009) for an example of this approach in practice.

When the marginal likelihood is given in closed form, we may use the MCMC model composition (MC^3: Madigan and York (1995)) method, which is a simple Metropolis algorithm that can be employed to explore large model spaces. Variants of MC^3 have been used in Gaussian linear models by Hoeting, Madigan and Raftery (1996), Raftery, Madigan and Hoeting (1997) and Hoeting, Raftery and Madigan (2002). The MC^3 algorithm can be summarized as follows:

(1) For the current model $m \in \mathcal{M}$, propose a move to model $m' \in \mathcal{M}$ with probability $j(m, m')$.

(2) Calculate and store the marginal likelihood $f(\boldsymbol{y}|m')$ of model m'.

(3) Set $m = m'$ (i.e., accept the proposed model m') with probability $\alpha = \min\left[1, \frac{f(m'|\boldsymbol{y})}{f(m|\boldsymbol{y})}\frac{j(m',m)}{j(m,m')}\right]$.

(4) Store m as the current model.

(5) Repeat steps (1)–(4) until a target number of models is visited or a pre-specified CPU budget is exhausted.

Posterior model probabilities can be estimated in two ways: by considering the marginal likelihoods of the visited and proposed models stored in step (2), or by a frequency tabulation of the visited models in the output of the MCMC sampler.

When the marginal likelihood is not analytically tractable under the PEP prior, it can be obtained by one of the Monte-Carlo schemes described in Section 4. We can exploit the fact that, in order to estimate the marginal likelihood of any model $m_\ell \in \mathcal{M}$, in the first two Monte-Carlo schemes we need only to sample $\boldsymbol{y}^*$ from $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)$ and $m_0^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*, \delta)$, respectively.

Here (when marginal likelihoods are not available analytically) we propose to modify the standard MC^3 method by sampling a binary vector $\boldsymbol{\gamma}$ indicating the variables included in the model (see, e.g., George and McCulloch, 1993), using a Metropolis-within-Gibbs approach as follows.

(1) Generate $\boldsymbol{y}^{*(t)}$ $(t = 1, \ldots, T)$ from $m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)$ (this is the first Monte-Carlo marginal likelihood scheme) or $m_0^N(\boldsymbol{y}^*|\boldsymbol{y}; \mathrm{X}_0, \mathrm{X}_0^*, \delta)$ (this is the second scheme).

(2) For the current model m_ℓ, corresponding to the set of variable-inclusion indicators $\boldsymbol{\gamma}_\ell$, repeat the following:

For $j = 1, \ldots, p$ (selected in random order), repeat the following steps:

(a) Propose $\gamma_j' = 1 - \gamma_j$ with probability one.
(b) Keep the other covariates the same: $\gamma_l' = \gamma_l$ for all $l \neq j$.
(c) Identify $m_{\ell'}$ corresponding to the vector $\boldsymbol{\gamma}_{\ell'}$ with elements $\gamma_k', k = 1, \ldots, p$.
(d) If $m_{\ell'}$ is not previously visited, calculate and store its estimated marginal likelihood $f(\boldsymbol{y}|m_{\ell'}) = \widehat{m}_{\ell'}^{PE}(\boldsymbol{y}|\mathrm{X}_{\ell'}, \mathrm{X}_{\ell'}^*, \delta)$ given by (35, first scheme) or (36, second scheme).
(e) Set $m_\ell = m_{\ell'}$ (i.e., accept the proposed model $m_{\ell'}$) with probability

$$\alpha = \min\left[1, \frac{f(m_{\ell'}|\boldsymbol{y})}{f(m_\ell|\boldsymbol{y})}\right] = \min\left[1, \frac{\widehat{m}_{\ell'}^{PE}(\boldsymbol{y}|\mathrm{X}_{\ell'}, \mathrm{X}_{\ell'}^*, \delta)}{\widehat{m}_\ell^{PE}(\boldsymbol{y}|\mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)} \frac{f(m_{\ell'})}{f(m_\ell)}\right], \tag{41}$$

where $f(m_\ell)$ is the prior probability for model ℓ.

(3) Store m_ℓ as the current model.

(4) Repeat steps (2)–(3) until a target number of models is visited or a pre-specified CPU budget is exhausted.

For the third and fourth Monte-Carlo marginal-likelihood estimates, we start the above MC^3 algorithm from step (2) and in step (2)(d) we generate $\boldsymbol{y}^{*(t)}$ $(t = 1, \ldots, T)$ from $m_{\ell'}^N(\boldsymbol{y}^*|\,\boldsymbol{y}; \mathrm{X}_\ell, \mathrm{X}_\ell^*, \delta)$, which now depends on the proposed model, and estimate the marginal likelihood of that model using expressions (38) and (39), respectively.

6 Experimental results

In this section we illustrate the PEP-prior methodology on both simulated and real examples, and we perform sensitivity analyses to verify the stability of our findings. Results are presented for the two different baseline prior specifications described in Sections 2.1 and 2.2, i.e., the PEP prior with the g-prior as a baseline choice (we call this approach Z-PEP) and the PEP prior with the

independence Jeffreys prior as a baseline choice (we call this approach J-PEP). In both cases, the marginal likelihood (33) is not analytically tractable, and therefore initially we evaluate the four Monte-Carlo marginal-likelihood approaches presented in Section 4. Then we present results for $n^* = n$, followed by an extensive sensitivity analysis over different values of n^*. Our results are compared with those obtained by (a) the expected-posterior prior with minimal training sample, with power parameter $\delta = 1$ and the independence Jeffreys prior as baseline (we call this approach J-EPP) and (b) the expected intrinsic Bayes factor (EIBFs), i.e., the arithmetic mean of the IBFs over different minimal training samples (in Section 6.1.3 we also make some comparisons between the Z-PEP, J-PEP and IBF methods). Implementation details for J-EPP can be found in Fouskakis and Ntzoufras (2012), while computational details for the EIBF are provided in Appendix B. In all illustrations the design matrix X^* of the imaginary/training data is selected as a random subsample of size n^* of the rows of X.

Note that, since Pérez and Berger (2002) have shown that Bayes factors from the J-EPP approach become identical to those from the EIBF method as the sample size $n \to \infty$ (with the number of covariates p fixed), it is possible (for large n) to use EIBF as an approximation to J-EPP that is computationally much faster than the full J-EPP calculation. We take advantage of this fact below: for example, producing the results in Table 5 would have taken many days of CPU time with J-EPP; instead, essentially equivalent results were available in hours with EIBF. For this reason, one can regard the labels "J-EPP" and "EIBF" as more or less interchangeable in what follows.

6.1 A simulated example

Here we illustrate the PEP method by considering the simulated data set of Nott and Kohn (2005). This data set consists of $n = 50$ observations with $p = 15$ covariates. The first 10 covariates are generated from a multivariate Normal distribution with mean vector $\mathbf{0}$ and covariance matrix I_{10}, while

$$X_{ij} \sim N\big(0.3X_{i1} + 0.5X_{i2} + 0.7X_{i3} + 0.9X_{i4} + 1.1X_{i5}, 1\big) \text{ for } (j = 11, \ldots, 15; i = 1, \ldots, 50)\,, \quad (42)$$

and the response is generated from

$$Y_i \sim N\big(4 + 2X_{i1} - X_{i5} + 1.5X_{i7} + X_{i,11} + 0.5X_{i,13}, 2.5^2\big) \text{ for } i = 1, \ldots, 50\,. \quad (43)$$

With $p = 15$ covariates there are only 32,768 models to compare; we were able to conduct a full enumeration of the model space and MC^3 was not needed.

6.1.1 PEP prior results

To check the efficiency of the four Monte-Carlo marginal-likelihood estimates of Section 4, we initially performed a small experiment. For Z-PEP, we estimated the logarithm of the marginal likelihood for models $(X_1 + X_5 + X_7 + X_{11})$ and $(X_1 + X_7 + X_{11})$, by running each Monte-Carlo technique 100 times for 1,000 iterations and calculating the Monte-Carlo standard errors; Table 1 presents the results. It is clear that the third and fourth Monte-Carlo schemes produce much lower Monte-Carlo standard errors, as expected; what is interesting is how much better schemes (3) and (4) actually are. Therefore from now on we use the third scheme for Z-PEP and the fourth scheme for J-PEP, holding the number of iterations constant at 1,000.

Table 1: *Monte-Carlo standard errors (MCSEs) of the estimates of the logarithm of the marginal likelihoods for models* $(X_1 + X_5 + X_7 + X_{11})$ *and* $(X_1 + X_7 + X_{11})$*, for the simulated example of Section 6.1; the four Monte-Carlo approaches in Section 4 were each performed 100 times for 1,000 iterations, using the Z-PEP prior methodology.*

Monte-Carlo	MCSE for Model	
Scheme	$(X_1 + X_5 + X_7 + X_{11})$	$(X_1 + X_7 + X_{11})$
(1)	0.9563	0.8769
(2)	0.7061	0.5312
(3)	0.0339	0.0281
(4)	0.0353	0.0296

Table 2: *Posterior model probabilities for the best models, together with Bayes factors of the MAP model* (m_1) *against* $m_j, j = 2, \ldots, 7$*, for the Z-PEP and the J-PEP prior methodologies in the simulated example of Section 6.1.*

		Z-PEP		J-PEP		
		Posterior Model	Bayes		Posterior Model	Bayes
m_j	Predictors	Probability	Factor	Rank	Probability	Factor
1	$X_1 + X_5 + X_7 + X_{11}$	0.0783	1.00	(2)	0.0952	1.00
2	$X_1 + X_7 + X_{11}$	0.0636	1.23	(1)	0.1054	0.90
3	$X_1 + X_5 + X_6 + X_7 + X_{11}$	0.0595	1.32	(3)	0.0505	1.88
4	$X_1 + X_6 + X_7 + X_{11}$	0.0242	3.23	(4)	0.0308	3.09
5	$X_1 + X_7 + X_{10} + X_{11}$	0.0175	4.46	(5)	0.0227	4.19
6	$X_1 + X_5 + X_7 + X_{10} + X_{11}$	0.0170	4.60	(9)	0.0146	6.53
7	$X_1 + X_5 + X_7 + X_{11} + X_{13}$	0.0163	4.78	(10)	0.0139	6.87

Table 2 presents the posterior model probabilities (with a uniform prior on the model space) for the best models, together with Bayes factors for the Z-PEP and J-PEP prior methodologies. The MAP model for the Z-PEP prior includes four of the five true effects; the data-generating model is seventh in rank due to the small effect of X_{13}. Moreover, notice that when using the J-PEP prior the methodology becomes even more parsimonious; the MAP model is now $X_1 + X_7 + X_{11}$, which is the second-best model under the Z-PEP approach. When we focus on posterior inclusion probabilities (see Table 3) rather than posterior model probabilities and odds, although we observe again that J-PEP supports systematically more parsimonious models than Z-PEP, no noticeable differences between the inclusion probabilities using the two priors are observed (with the largest difference seen in the inclusion probabilities of X_5; these are about 0.5 for Z-PEP and about 0.4 for J-PEP).

6.1.2 Sensitivity analysis for the imaginary/training sample size n^*

To examine the sensitivity of the PEP approach to the sample size n^* of the imaginary/training data set, we present results for $n^* = 17, \ldots, 50$. Figures 1 and 2 display posterior marginal variable-inclusion probabilities and posterior model probabilities, respectively. As noted previously, to

Table 3: *Posterior inclusion probabilities for the Z-PEP and J-PEP prior methodologies for the simulated example of Section 6.1.*

	Covariate							
Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Z-PEP	0.997	0.110	0.129	0.133	0.503	0.337	1.000	0.150
J-PEP	0.993	0.088	0.108	0.121	0.395	0.253	1.000	0.117

	Covariate						
Method	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
Z-PEP	0.126	0.197	0.856	0.136	0.142	0.111	0.113
J-PEP	0.100	0.152	0.789	0.109	0.115	0.099	0.100

specify X* when $n^* < n$ we randomly selected a sub-sample of the rows of the original matrix X. Results are presented for Z-PEP; similar results for J-PEP have been omitted for brevity. The main conclusions from these figures can be summarized as follows:

- From Figure 1 it is evident that posterior inclusion probabilities are quite insensitive to a wide variety of values of n^*.
- From Figure 2 it is clear that posterior model probabilities of the best models are also quite robust for widely different values of n^*; also note that we obtain the same three best models for all values of n^*.
- In both figures we observe more variability for smaller values of n^*; this arises from the selection of the sub-samples used for the construction of X*.
- The principal conclusion is that, since the results are not sensitive to the choice of n^*, we can use $n^* = n$ and dispense with training samples altogether; this yields the advantages mentioned in the paper's Summary and in Section 1 (increased stability of the resulting Bayes factors, removal of the arbitrariness arising from individual training-sample selections, and substantial increases in computational speed, allowing many more models to be compared within a fixed CPU budget).

6.1.3 Comparisons with the intrinsic-Bayes-factor (IBF) and J-EPP approaches

Here we compare the PEP Bayes factor between the two best models ($(X_1 + X_5 + X_7 + X_{11})$ and $(X_1 + X_7 + X_{11})$) with the corresponding Bayes factors using J-EPP and IBF. For IBF and J-EPP we randomly selected 100 training samples of size $n^* = 6$ (the minimal training sample size for the estimation of these two models) and $n^* = 17$ (the minimal training sample size for the estimation of the full model with all $p = 15$ covariates), while for Z-PEP and J-PEP we randomly selected 100 training samples of sizes $n^* = 6, 17$ and $n^* = 5 \times k$ for $k = 5, \ldots, 10$. Each marginal-likelihood estimate in PEP was obtained with 1,000 iterations, using the third and fourth Monte-Carlo schemes for Z-PEP and J-PEP, respectively, and in J-EPP with 1,000 iterations, using the fourth Monte-Carlo scheme. Figure 3 presents the results as parallel boxplots, and motivates the following observations:

Figure 1: *Posterior marginal inclusion probabilities for different n^* with the Z-PEP prior methodology for the simulated example of Section 6.1.*

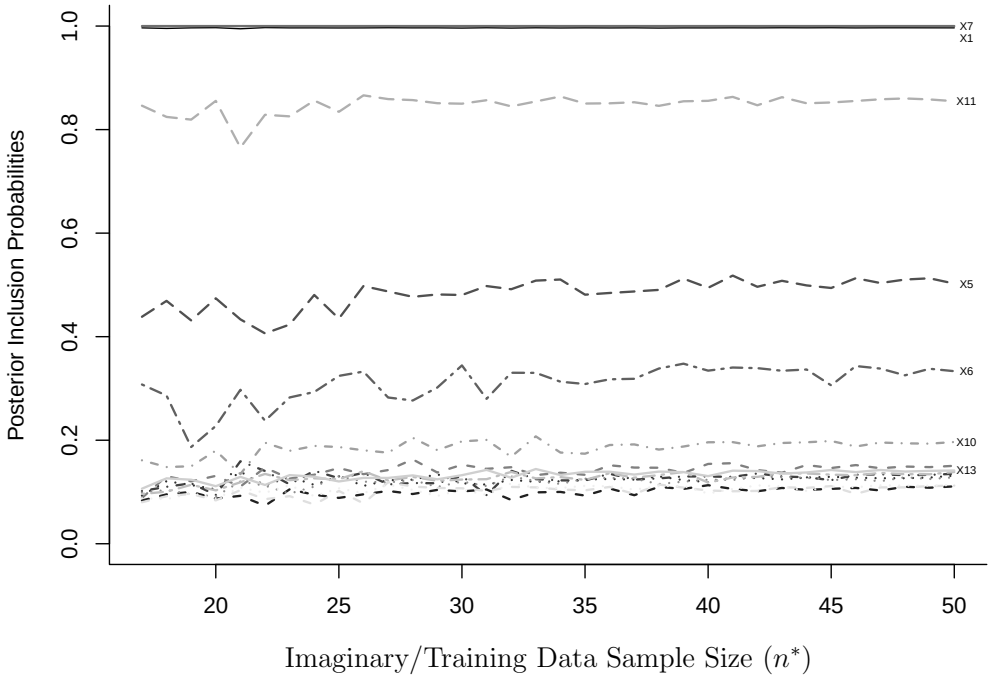


Figure 2: *Posterior model probabilities of the five best models obtained for each n^*, with the Z-PEP prior methodology for the simulated example of Section 6.1. The models shown are those that entered the top five models for at least one value of n^*; lines connect posterior model probabilities for models of the same rank over different values of n^*.*

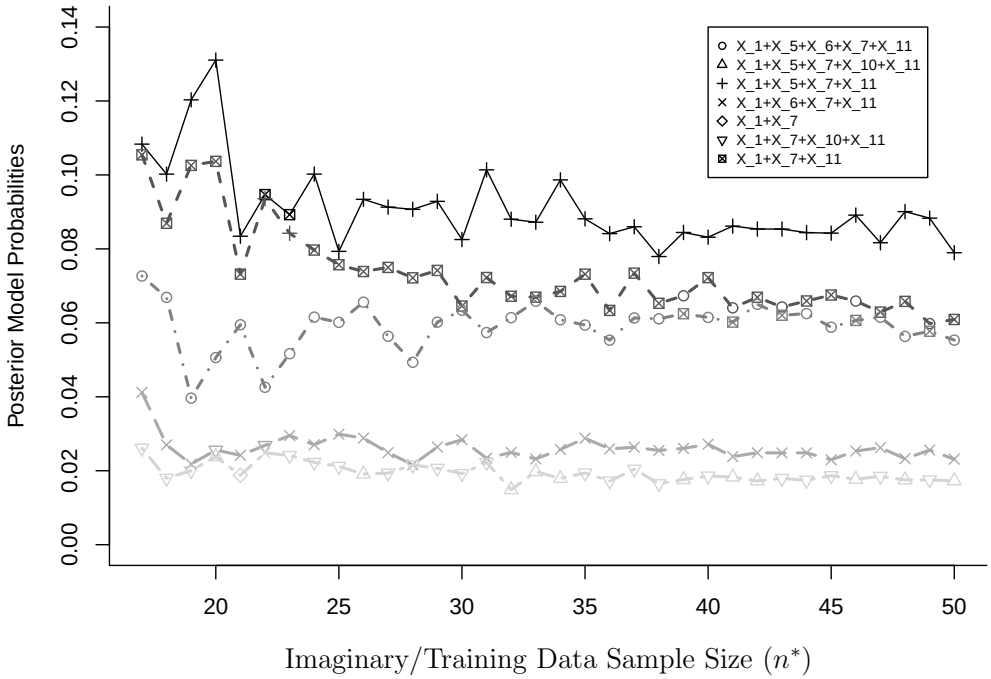
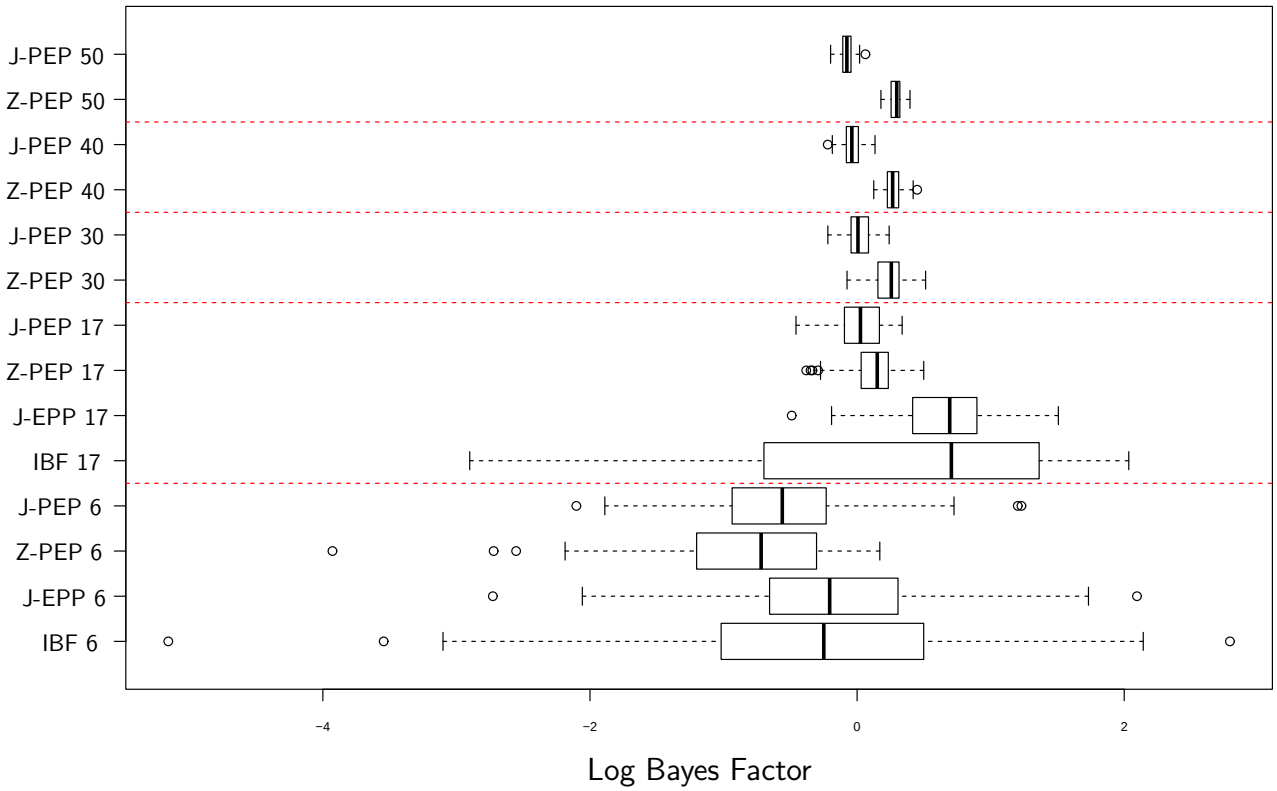


Figure 3: *Boxplots of the Intrinsic Bayes Factor (IBF) and Bayes factors using the J-EPP, J-PEP and Z-PEP approaches, on a logarithmic scale, in favor of model* $(X_1 + X_5 + X_7 + X_{11})$ *over model* $(X_1 + X_7 + X_{11})$ *for the simulated example of Section 6.1. For IBF and J-EPP, training samples of size* $n^* = 6$ *and 17 were used; for both PEP priors we used* $n^* = 6, 17$ *and* $n^* = 5 \times k$ *for* $k = 5, \ldots, 10$. *In the boxplot labels on the vertical axis, letters indicate methods and numbers signify training sample sizes.*



- For $n^* = 6$ and 17, although there are some differences between the median log Bayes factors across the four approaches, the variability across random training samples is so large as to make these differences modest by comparison; none of the methods finds a marked difference between the two models.

- With modest n^* values, which would tend to be favored by users for their advantage in computing speed, the IBF method exhibited an extraordinary amount of instability across the particular random training samples chosen: with $n^* = 6$ the observed variability of IBF estimated Bayes factors across the 100 samples was from $e^{-5.16} \doteq 0.005$ to $e^{+2.48} \doteq 11.89$, a multiplicative range of more than 2,300, and with $n^* = 17$ the corresponding span was from $e^{-2.90} \doteq 0.055$ to $e^{+2.03} \doteq 7.61$, a multiplicative variation of about 138. The instability of the J-EPP approach across training samples was smaller than with IBF but still large: for J-EPP the range of estimated Bayes factors for $n^* = 6$ was from $e^{-2.72} \doteq 0.065$ to $e^{+2.09} \doteq 8.08$ (a multiplicative span of about 125); the corresponding values for $n^* = 17$ were from 0.61 to 4.51, a multiplicative range of 7.4. The analogous multiplicative ranges for Z-PEP were considerably lower: 60.22, 2.41 and 1.24, respectively, for $n^* = 6, 17$ and 50; similarly for

J-PEP the corresponding multiplicative ranges were 28.01, 2.21 and 1.30.

- Figure 3 highlights the advantage of using $n^* = n$ with the PEP approach over the IBF and J-EPP methods with modest training samples: the Monte-Carlo uncertainty introduced in the IBF and J-EPP methods by the need to choose a random training sample creates a remarkable degree of sensitivity in those approaches to the particular samples chosen, and this undesirable behavior is entirely absent with the $n^* = n$ version of the PEP method. The observed variability for $n^* = n$ in the PEP approach is due solely to Monte-Carlo noise in the marginal-likelihood computation.

6.2 Variable selection in the Breiman-Friedman ozone data set

In this section we use a data set often examined in variable-selection studies — the ozone data of Breiman and Friedman (1985) — to implement the Z-PEP and J-PEP approaches and make comparisons with other methods. The scientific purpose of gathering these data was to study the relationship between ozone concentration and a number of meteorological variables, including temperature, wind speed, humidity and atmospheric pressure; data are from a variety of locations in the Los Angeles basin in 1976. The data set we used was slightly modified from its form in other studies, based on preliminary exploratory analyses we performed; our version of the data set has $n = 330$. As a response we used a standardized version of the logarithm of the ozone variable of the original data set. The standardized versions of 9 main effects, 9 quadratic terms, 2 cubic terms, and 36 two-way interactions (a total of 56 explanatory variables) were included as possible covariates. (Further details concerning the final data set used in this section are provided in Appendix C.)

6.2.1 Searching the model space

Full-enumeration search for the full space with 56 covariates was computationally infeasible, so we used the model search algorithm (MC^3) given in Section 5 for the Z-PEP prior methodology and the EIBF approach. For Z-PEP we used the third Monte-Carlo scheme with 1,000 iterations; for EIBF we used 30 randomly-selected minimal training samples ($n^* = 58$).

With such a large number of predictors, the model space in our problem is too large for the MC^3 approach to estimate posterior model probabilities with high accuracy in a reasonable amount of CPU time. For this reason, we implemented the following two-step method:

(1) First we used MC^3 to identify variables with high posterior marginal inclusion probabilities $P(\gamma_j = 1|\boldsymbol{y})$, and we created a reduced model space consisting only of those variables whose marginal probabilities were above a threshold value. According to Barbieri and Berger (2004), this method of selecting variables may lead to the identification of models with better predictive abilities than approaches based on maximizing posterior model probabilities. Although Barbieri and Berger proposed 0.5 as a threshold value for $P(\gamma_j = 1|\boldsymbol{y})$, we used the lower value of 0.3, since our aim was only to identify and eliminate variables not contributing to models with high posterior probabilities. The inclusion probabilities were based on the marginal-likelihood weights for the visited models.

(2) Then we used the same model search algorithm as in step (1) in the reduced space to estimate posterior model probabilities (and the corresponding odds).

Table 4: *Posterior inclusion probabilities using Z-PEP, J-PEP and EIBF for the reduced model space of the ozone data set.*

Index	Name	J-PEP	Z-PEP	EIBF
1	Day of year	1.000	1.000	1.000
2	Wind speed at LAX	0.985	0.992	0.976
5	Temperature at Sandburg	0.182	0.375	0.475
7	PG from LAX to Daggett	0.613	0.857	0.984
8	Inversion base temperature at LAX	1.000	1.000	1.000
9	Visibility at LAX	1.000	1.000	1.000
10	(Day of year)^2	1.000	1.000	1.000
12	(500 mb pressure height at VAFB)^2	0.618	0.840	0.980
13	(Humidity at LAX)^2	0.716	0.918	1.000
15	(Inversion base height at LAX)^2	0.983	0.988	1.000
16	(PG from LAX to Daggett)^2	1.000	1.000	1.000
18	(Visibility at LAX)^2	0.896	0.965	0.995
20	(Inversion base temperature at LAX)^3	0.401	0.641	0.923
23	(Day of year) $\times$ (Humidity at LAX)	0.006	0.011	0.027
26	(Day of year) $\times$ (PG from LAX to Daggett)	0.042	0.093	0.233
30	(Wind speed at LAX) $\times$ (Humidity at LAX)	0.011	0.027	0.073
36	(500 mb pressure height at VAFB) $\times$ (Humidity at LAX)	0.036	0.087	0.100
39	(500 mb pressure height at VAFB) $\times$ (PG from LAX to Daggett)	0.040	0.159	0.371
42	(Humidity at LAX) $\times$ (Temperature at Sandburg)	0.315	0.429	0.694
43	(Humidity at LAX) $\times$ (Inversion base height at LAX)	0.974	0.898	0.877
48	(Temperature at Sandburg) $\times$ (PG from LAX to Daggett)	0.017	0.026	0.028
51	(Inversion base height at LAX) $\times$ (PG from LAX to Daggett)	0.065	0.197	0.342

Notes: (1) Abbreviations used in this table: LAX = Los Angeles International Airport, mb = millibar, VAFB = Vandenberg Air Force Base, PG = pressure gradient (mm Hg). (2) Wind speed is measured in mph, temperature and inversion base temperature in °F, visibility in miles, inversion base height in feet, humidity in % and pressure height in m. (3) As mentioned in the text, all variables were standardized (mean 0, standard deviation 1) before exploring quadratic/cubic terms and interactions, to minimize collinearity.

Initially we ran MC^3 for 100,000 iterations for both the Z-PEP and EIBF approaches. The reduced model space was formed from those variables that in either run had posterior marginal inclusion probabilities above 0.3. With this approach we reduced the initial list of $p = 56$ available candidates down to 22 predictors. Table 4 (first two columns) presents those predictors.

In the reduced model space we then ran MC^3 for 220,000 iterations for the J-PEP, Z-PEP and EIBF approaches. For Z-PEP we used the fourth Monte-Carlo scheme with 1,000 iterations, for Z-PEP we used the third Monte-Carlo scheme with 1,000 iterations and for EIBF we used 30 randomly-selected minimal training samples ($n^* = 24$). The resulting posterior inclusion probabilities are presented in Table 4 (last three columns), and posterior model odds for the five best models under each approach are given in Table 5. Table 4 shows that the three methods give approximately equal support to the most prominent covariates; for the remaining predictors the

Table 5: *Posterior odds (PO_{1k}) of the five best models within each analysis versus the current model k, for the reduced model space of the ozone data set. Variables common in all three analyses were:* $X_1 + X_2 + X_8 + X_9 + X_{10} + X_{15} + X_{16} + X_{18} + X_{43}$.

J-PEP

Ranking			Additional Variables	Number of Covariates	Posterior Odds PO_{1k}
J-PEP	Z-PEP	EIBF			
1	(>5)	(>5)		9	1.00
2	(1)	(5)	$X_7 + X_{12} + X_{13} + X_{20}$	13	1.29
3	(>5)	(>5)	$X_7 + X_{13} + X_{20}$	12	1.46
4	(>5)	(>5)	$X_{12} + X_{20}$	11	1.87
5	(>5)	(>5)	X_{12}	10	2.08

Z-PEP

Ranking			Additional Variables	Number of Covariates	Posterior Odds PO_{1k}
Z-PEP	J-PEP	EIBF			
1	(2)	(5)	$X_7 + X_{12} + X_{13} + X_{20}$	13	1.00
2	(>5)	(>5)	$X_5 + X_7 + X_{12} + X_{13} + X_{20}$	14	1.19
3	(>5)	(3)	$X_5 + X_7 + X_{12} + X_{13} + X_{20} + X_{42}$	15	1.77
4	(>5)	(1)	$X_7 + X_{12} + X_{13} + X_{20} + X_{42}$	14	1.94
5	(>5)	(>5)	$X_7 + X_{12} + X_{13}$	12	2.30

EIBF

Ranking			Additional Variables	Number of Covariates	Posterior Odds PO_{1k}
EIBF	J-PEP	Z-PEP			
1	(>5)	(4)	$X_7 + X_{12} + X_{13} + X_{20} + X_{42}$	14	1.00
2	(>5)	(>5)	$X_5 + X_7 + X_{12} + X_{13} + X_{20} + X_{26} + X_{42}$	16	1.17
3	(>5)	(3)	$X_5 + X_7 + X_{12} + X_{13} + X_{20} + X_{42}$	15	1.30
4	(>5)	(>5)	$X_7 + X_{12} + X_{13} + X_{20} + X_{39} + X_{42}$	15	1.44
5	(2)	(1)	$X_7 + X_{12} + X_{13} + X_{20}$	13	1.58

posterior inclusion probabilities are lower for J-PEP, followed by the values for Z-PEP. This shows that the PEP methodology supports more parsimonious models than the EIBF approach.

When looking at the posterior model odds presented in Table 5, the MAP model under the Z-PEP approach is the only one that appears in the five most probable models in all approaches (with rank 2 in J-PEP and rank 5 in EIBF). Moreover, from this table it is clear that the J-PEP approach supports the most parsimonious models; at the other extreme, EIBF again gives the least support to the most parsimonious models.

6.2.2 Comparison of predictive performance

Here we examine the out-of-sample predictive performance of J-PEP, Z-PEP and J-EPP on the full model and the three MAP models found by each method implemented in the previous analysis. To do so, we randomly partitioned the data in half 50 times. For each partition, we generated

Table 6: *Comparison of the predictive performance of the PEP and J-EPP methods, using the full and MAP models in the reduced model space of the ozone data set.*

				$RMSE^*$			
Model	d_ℓ	R^2	R^2_{adj}	J-PEP	Z-PEP	J-EPP	Jeffreys Prior
Full	22	0.8500	0.8392	0.5988	0.5935	0.6194	0.5972
				(0.0087)	(0.0097)	(0.0169)	(0.0104)
J-PEP MAP	9	0.8070	0.8016	0.5975	0.6161	0.7524	0.6165
				(0.0063)	(0.0051)	(0.0626)	(0.0052)
Z-PEP MAP	13	0.8370	0.8303	0.5994	0.5999	0.6982	0.5994
				(0.0071)	(0.0060)	(0.0734)	(0.0049)
EIBF MAP	14	0.8398	0.8326	0.6182	0.5961	0.6726	0.5958
				(0.0066)	(0.0072)	(0.0800)	(0.0061)

Comparison with the full model (percentage changes)

				$RMSE$			
Model	d_ℓ	R^2	R^2_{adj}	J-PEP	Z-PEP	J-EPP	Jeffreys Prior
J-PEP MAP	−59%	−5.06%	−4.48%	−0.22%	+3.81%	+21.5%	+3.23%
Z-PEP MAP	−41%	−1.50%	−1.06%	+0.10%	+1.01%	+12.7%	+0.37%
EIBF MAP	−36%	−1.20%	−0.78%	+3.24%	+0.44%	+10.9%	−0.23%

Note: *Mean (standard deviation) over 50 different split-half out-of-sample evaluations.

an MCMC sample of $T = 1,000$ iterations from the model of interest m_ℓ and then computed the following measure of predictive accuracy:

$$RMSE_\ell = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_V}\sum_{i\in V}\left(y_i - \widehat{y}^{(t)}_{i|m_\ell}\right)^2}, \tag{44}$$

the root mean squared error for a validation dataset V of size $n_V = \lceil \frac{n}{2} \rceil$; here $\widehat{y}^{(t)}_{i|m_\ell} = \mathrm{X}_{\ell(i)}\boldsymbol{\beta}^{(t)}_\ell$ is the predicted value of y_i according to the assumed model ℓ for iteration t, $\boldsymbol{\beta}^{(t)}_\ell$ is the vector of model m_ℓ parameters for iteration t and $\mathrm{X}_{\ell(i)}$ is the ith row of the matrix X_ℓ of model m_ℓ.

Results for the full model and the MAP models are given in Table 6. For comparison purposes, we have also included the split-half $RMSE$ measures for these three models using predictions based on direct fitting of model (4) with the independence Jeffreys prior $f(\boldsymbol{\beta}_\ell\,, \sigma^2_\ell) \propto \frac{1}{\sigma^2_\ell}$, which can be viewed as a parametric bootstrap approach around the MLE for $\boldsymbol{\beta}_\ell$ and the unbiased estimate of σ^2_ℓ, allowing for variability based on their standard errors.

Table 6 shows that all $RMSE$ values for the PEP and the Jeffreys-prior approaches are similar, indicating that PEP provides predictive performance equivalent to that offered by the Jeffreys prior; also note that the PEP and the Jeffreys-prior $RMSE$s for the two PEP MAP models are close to the corresponding values for the full model, which has considerably higher dimension. (The point of this comparison is to demonstrate that the PEP approach, which can be used for variable selection, achieves a level of predictive accuracy comparable to that of the Jeffreys-prior approach, which cannot be used for variable selection, for reasons given in the first paragraph of Section 1.)

In contrast, with the J-EPP approach the $RMSE$ values of all four models are noticeably

higher than the corresponding values for the Jeffreys-prior and PEP approaches. Figure 4 provides the explanation, by showing the distribution of $RMSE$ values across the 50 random data splits, for each of the four implementations in each of the four models examined in Table 6. The J-EPP approach is predictively unstable as a function of its training samples, a behavior that is entirely absent in PEP's performance.

7 Discussion

The major contribution of the research presented here is to sharply diminish the effect of training samples on previously-studied expected-posterior-prior methodology. By combining ideas from the power-prior approach of Ibrahim and Chen (2000) and the unit-information prior of Kass and Wasserman (1995), we raise the likelihood involved in the expected-posterior prior to a power proportional to the inverse of the training sample size, resulting in prior information equivalent to one data point. In this way the effect of the training sample is minimal, regardless of its sample size, and we can choose training samples with size as big as the size of the original data; this choice promotes stability of the resulting Bayes factors, removes the arbitrariness arising from individual training-sample selections, and avoids the computational burden of averaging over many training samples. Additional advantages of our approach over methods that depend on training samples include the following.

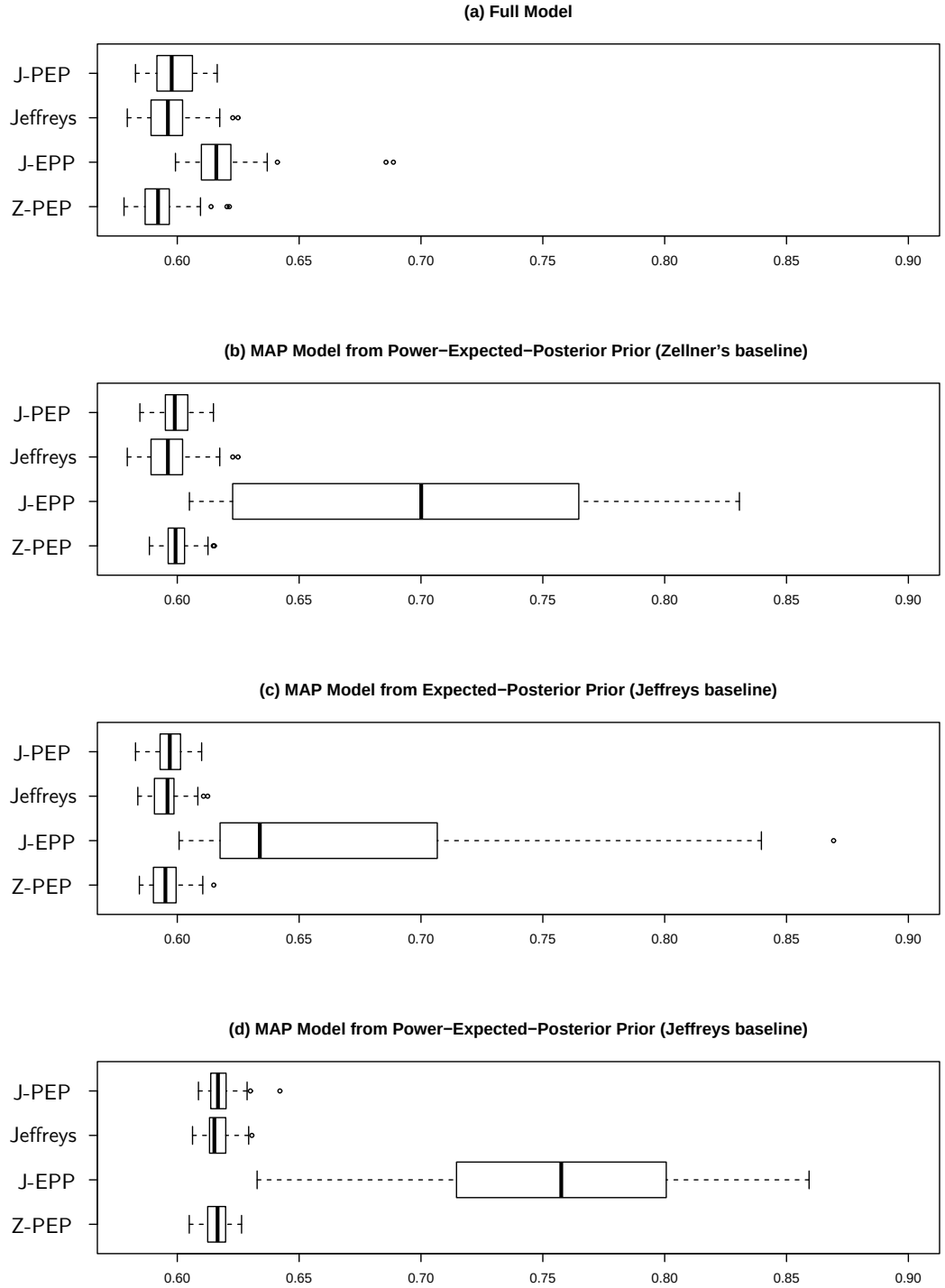
- In variable-selection problems in linear models, the training data refer to both y and X. Under the base-model approach, we can simulate training data y^* directly from the prior predictive distribution of a reference model, but we still need to consider a sub-sample X^* of the original design matrix X. The number of possible sub-samples of X can be enormous, inducing large variability, since some of those sub-samples can be highly influential for the posterior analysis. By using our approach, and working with training-sample sizes equal to the size of the full data set, we avoid the selection of such sub-samples by choosing $X^* = X$.

- The size of the full model is usually regarded as specifying the minimal training sample. This selection makes inference within the current data set coherent, but the size of the minimal training sample will change if additional covariates are added. This means that the expected-posterior prior distribution will depend on the number of covariates. Moreover, if the data derive from a highly structured situation (such as an analysis of covariance in a factorial design), any choice of a small part of the data to act as a training sample would be somewhat untypical. Finally, the effect of the minimal training sample will be large in settings where the actual sample size is small in comparison to the number of covariates.

It is worth noting that our method works in a totally different fashion than fractional Bayes factors. In the latter, the likelihood is partitioned based on two data subsets; one is used for building the prior within each model and the other is employed for model evaluation and comparison. In contrast, with our approach, the original likelihood is used only once, for simultaneous variable selection and posterior inference. Moreover, the fraction of the likelihood (power likelihood) — used in the expected-posterior expression of our prior distribution — refers solely to the imaginary data coming from a prior predictive distribution based on the reference model.

Our PEP approach can be implemented under any baseline prior choice; results using the g-prior and the independence Jeffreys prior as baseline prior choices are presented here. The

Figure 4: *Distribution of RMSE across 50 random partitions of the ozone data set, for the Jeffreys-prior, J-EPP, Z-PEP and J-PEP methods, in (a) the full model, (b) the Z-PEP MAP model, (c) the J-EPP MAP model, and (d) the J-PEP MAP model.*



conjugacy structure of the first, in Gaussian linear models, makes calculations easier and also offers flexibility in situations in which non-diffuse parametric prior information is available. When, on the other hand, the experimenter does not have strong prior information about the parameters of the competing models, the independence Jeffreys baseline prior can be viewed as a natural choice.

From our empirical results we conclude that our method

- is systematically more parsimonious (under either baseline prior choice) than the expected-posterior prior approach using the Jeffreys prior as a baseline prior and minimal training samples, while sacrificing no desirable performance characteristics to achieve this parsimony;
- is robust to the size of the training sample, thus supporting the use of the entire data set as a "training sample" and thereby promoting stability and fast computation; and
- identifies maximum a-posteriori models that achieve good out-of-sample predictive performance.

Our PEP approach could be applied to any prior distribution that is defined via imaginary training samples. Additional future extensions of our approach include implementation in generalized linear models, where computation is more demanding.

Appendix

A Prior mean and covariance matrix for model parameters

THEOREM 1: *Under the baseline prior setup (17), when* $(a_\ell > 1, a_0 > 1)$ *the PEP prior mean of* $\boldsymbol{\beta}_\ell$ *is* $\mathsf{E}(\boldsymbol{\beta}_\ell) = \mathbf{0}$*, and when* $(a_\ell > 2, a_0 > 1)$ *the PEP prior covariance matrix is*

$$\mathsf{V}(\boldsymbol{\beta}_\ell) = \left\{ \frac{\delta\, w}{a_\ell - 1 + \frac{n^*}{2}} \left[b_\ell + \frac{1}{2} \frac{b_0}{a_0 - 1} tr(\Lambda_\ell^* \Lambda_0^{*-1}) \right] \mathrm{I}_{d_\ell} + \frac{w^2 b_0}{a_0 - 1} (\mathrm{X}_\ell^{*^T} \mathrm{X}_\ell^*)^{-1} \mathrm{X}_\ell^{*^T} \Lambda_0^{*-1} \mathrm{X}_\ell^* \right\} (\mathrm{X}_\ell^{*^T} \mathrm{X}_\ell^*)^{-1}, \tag{45}$$

where $tr(\mathrm{A})$ *is the trace of the matrix* A*.*

Proof of Theorem 1. From (21), the prior mean is

$$\begin{aligned}
\mathsf{E}(\boldsymbol{\beta}_\ell) &= \int \boldsymbol{\beta}_\ell \, \pi_\ell^{PE}(\boldsymbol{\beta}_\ell \,|\, \mathrm{X}_\ell^*, \delta)\, d\boldsymbol{\beta}_\ell \\
&= \int \left\{ \int \boldsymbol{\beta}_\ell \, f_{St_{d_\ell}} \left[\boldsymbol{\beta}_\ell \,;\, 2\,a_\ell + n^*,\, w\,\widehat{\boldsymbol{\beta}}_\ell^*, \delta\, w\, (\mathrm{X}_\ell^{*^T} \mathrm{X}_\ell^*)^{-1} \frac{b_\ell + \frac{SS_\ell^*}{2}}{a + \frac{n^*}{2}} \right] d\boldsymbol{\beta}_\ell \right\} m_0^N(\boldsymbol{y}^* | \mathrm{X}_0^*, \delta)\, d\boldsymbol{y}^* \\
&= \int \mathsf{E}_{St_{d_\ell}} \left[\boldsymbol{\beta}_\ell \,;\, 2\,a_\ell + n^*,\, w\,\widehat{\boldsymbol{\beta}}_\ell^*, \delta\, w\, (\mathrm{X}_\ell^{*^T} \mathrm{X}_\ell^*)^{-1} \frac{b_\ell + \frac{SS_\ell^*}{2}}{a_\ell + \frac{n^*}{2}} \right] m_0^N(\boldsymbol{y}^* | \mathrm{X}_0^*, \delta)\, d\boldsymbol{y}^* \,;
\end{aligned} \tag{46}$$

here SS_ℓ^* is defined in Section 2.2 and $\mathsf{E}_{St_d}[\xi(\boldsymbol{z})\,;\, df,\, \boldsymbol{\mu},\, \Sigma]$ is the expectation of a function $\xi(\boldsymbol{z})$ of $\boldsymbol{z}$, where z follows a d-dimensional Student distribution with density $f_{St_d}(\boldsymbol{z}\,;\, df, \boldsymbol{\mu}, \Sigma)$ given by

$$f_{St_d}(\boldsymbol{y}\,;\, df, \boldsymbol{\mu}, \Sigma) = \frac{\Gamma\left(\frac{df+d}{2}\right)}{\Gamma\left(\frac{df}{2}\right)} (df\,\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \left[1 + \frac{1}{df} (\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right]^{-\frac{df+d}{2}}. \tag{47}$$

For $\xi(\boldsymbol{z}) = \boldsymbol{z}$, the expectation is $\boldsymbol{\mu}$, yielding

$$\begin{aligned}
\mathsf{E}(\boldsymbol{\beta}_\ell) &= \int w\,\widehat{\boldsymbol{\beta}}^*_\ell\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^* = w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*^T}\int \boldsymbol{y}^*\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^* \\
&= w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*^T}\int \boldsymbol{y}^* f_{St_{n^*}}\left(\boldsymbol{y}^*\,;\,2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)d\boldsymbol{y}^* \\
&= w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*^T}\,\mathsf{E}_{St_{n^*}}\left(\boldsymbol{y}^*\,;\,2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right) \\
&= w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*^T}\mathbf{0} = \mathbf{0}\,. \qquad (48)
\end{aligned}$$

Since the mean is zero, the covariance matrix is

$$\begin{aligned}
\mathsf{V}(\boldsymbol{\beta}_\ell) &= \mathsf{E}(\boldsymbol{\beta}_\ell\,\boldsymbol{\beta}_\ell^T) = \int \boldsymbol{\beta}_\ell\,\boldsymbol{\beta}_\ell^T\,\pi_\ell^{PE}(\boldsymbol{\beta}_\ell\,|\mathrm{X}_\ell^*\,,\delta)\,d\boldsymbol{\beta}_\ell \\
&= \int \mathsf{E}_{St_{d_\ell}}\left[\boldsymbol{\beta}_\ell\,\boldsymbol{\beta}_\ell^T\,;\,2\,a_\ell+n^*,\,w\,\widehat{\boldsymbol{\beta}}^*_\ell,\delta\,w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\frac{b_\ell+\frac{SS^*_\ell}{2}}{a_\ell+\frac{n^*}{2}}\right] m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\,. \qquad (49)
\end{aligned}$$

For $\boldsymbol{z}\sim St(df,\boldsymbol{\mu},\Sigma)$, $\mathsf{E}(\boldsymbol{z}\boldsymbol{z}^T)$ is given by

$$\mathsf{E}(\boldsymbol{z}\boldsymbol{z}^T) = \mathsf{V}(\boldsymbol{z}) + \mathsf{E}(\boldsymbol{z})\,[\mathsf{E}(\boldsymbol{z})]^T = \frac{df}{df-2}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^T\,, \qquad (50)$$

from which

$$\begin{aligned}
\mathsf{E}_{St_{d_\ell}}\left[\boldsymbol{\beta}_\ell\,\boldsymbol{\beta}_\ell^T\,;\,2\,a_\ell+n^*,\,w\,\widehat{\boldsymbol{\beta}}^*_\ell,\,\delta\,w\,(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\frac{b_\ell+\frac{SS^*_\ell}{2}}{a_\ell+\frac{n^*}{2}}\right] &= \delta\,w\left(\frac{b_\ell+\frac{SS^*_\ell}{2}}{a_\ell-1+\frac{n^*}{2}}\right)(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1} \\
&\quad + w^2\widehat{\boldsymbol{\beta}}^*_\ell\widehat{\boldsymbol{\beta}}^{*^T}_\ell\,. \qquad (51)
\end{aligned}$$

Substitution into (49) yields

$$\begin{aligned}
\mathsf{V}(\boldsymbol{\beta}_\ell) &= \int\left[\delta\,w\left(\frac{b_\ell+\frac{SS^*_\ell}{2}}{a_\ell-1+\frac{n^*}{2}}\right)(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1} + w^2\widehat{\boldsymbol{\beta}}^*_\ell\widehat{\boldsymbol{\beta}}^{*^T}_\ell\right] m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^* \\
&= \delta\,w\,\frac{1}{a_\ell-1+\frac{n^*}{2}}\left[b_\ell+\frac{1}{2}\int SS^*_\ell\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\right](\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1} \\
&\quad + w^2\int\widehat{\boldsymbol{\beta}}^*_\ell\widehat{\boldsymbol{\beta}}^{*^T}_\ell\, m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^* \\
&= \delta\,w\,\frac{1}{a_\ell-1+\frac{n^*}{2}}\left[b_\ell+\frac{1}{2}\mathsf{E}_{St_{n^*}}\left(\boldsymbol{y}^{*^T}\Lambda^*_\ell\boldsymbol{y}^*;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)\right](\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1} \\
&\quad + w^2(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*^T}\,\mathsf{E}_{St_{n^*}}\left(\boldsymbol{y}^*\boldsymbol{y}^{*^T};2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)\mathrm{X}_\ell^*(\mathrm{X}_\ell^{*^T}\mathrm{X}_\ell^*)^{-1}\,. \qquad (52)
\end{aligned}$$

Now (see, e.g., Scott, 1997, Theorem 9.18) for any symmetric matrix A and any random vector $\boldsymbol{z}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathsf{V}(\boldsymbol{z})$,

$$E(\boldsymbol{z}^T\mathrm{A}\boldsymbol{z}) = tr\Big[\mathrm{A}\,\mathsf{V}(\boldsymbol{z})\Big] + \boldsymbol{\mu}^T\mathrm{A}\boldsymbol{\mu}\,, \qquad (53)$$

if $E(\boldsymbol{z}\boldsymbol{z}^T)$ exists. Therefore for $\boldsymbol{z}\sim St(df,\boldsymbol{\mu},\Sigma)$,

$$E(\boldsymbol{z}^T\mathrm{A}\boldsymbol{z}) = \frac{df}{df-2}tr(A\Sigma) + \boldsymbol{\mu}^T\mathrm{A}\boldsymbol{\mu}\,, \qquad (54)$$

from which, in our case,

$$\mathsf{E}_{St_{n^*}}\left(\boldsymbol{y}^{*^T}\Lambda_\ell^*\,\boldsymbol{y}^*;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)=\frac{b_0}{a_0-1}tr(\Lambda_\ell^*\Lambda_0^{*-1})\,. \tag{55}$$

Moreover,

$$\mathsf{E}_{St_{n^*}}\left(\boldsymbol{y}^*\boldsymbol{y}^{*^T};2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)=\mathsf{V}_{St_{n^*}}\left(\boldsymbol{y}^*;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)+\mathbf{0}=\frac{b_0}{a_0-1}\Lambda_0^{*-1}\,. \tag{56}$$

Substituting (55) and (56) into (52), we obtain (45) as desired.

□

THEOREM 2: *Under the baseline prior setup (17), for* $(a_\ell>1,a_0>1)$ *the PEP prior mean of* σ_ℓ^2 *is*

$$\mathsf{E}\left(\sigma_\ell^2\right)=\frac{b_0}{a_0-1}\,\frac{\frac{1}{2}tr\left(\Lambda_\ell^*\Lambda_0^{*-1}\right)+\frac{(a_0-1)b_\ell}{b_0}}{\frac{n^*}{2}+a_\ell-1}\,, \tag{57}$$

and for $(a_\ell>2,a_0>2)$ *the PEP prior variance is*

$$\begin{aligned}\mathsf{V}\left(\sigma_\ell^2\right) &= \left[\left(\frac{n^*}{2}+a_\ell-1\right)\left(\frac{n^*}{2}+a_\ell-2\right)\right]^{-1}\left\{b_\ell^2+\frac{b_\ell\,b_0\,tr(\Lambda_\ell^*\Lambda_0^{*-1})}{a_0-1}\right.\\ &\quad\left.+\frac{b_0^2\left[2\,tr(\Lambda_\ell^*\Lambda_0^{*-1}\Lambda_\ell^*\Lambda_0^{*-1})+tr(\Lambda_\ell^*\Lambda_0^{*-1})^2\right]}{4(a_0-1)(a_0-2)}\right\}\\ &\quad-\left(\frac{b_0}{a_0-1}\right)^2\left[\frac{\frac{1}{2}tr(\Lambda_\ell^*\Lambda_0^{*-1})+\frac{(a_0-1)b_\ell}{b_0}}{\frac{n^*}{2}+a_\ell-1}\right]^2\,.\end{aligned} \tag{58}$$

Proof of Theorem 2. The prior mean of the variance parameter is

$$\begin{aligned}\mathsf{E}\left(\sigma_\ell^2\right) &= \int\sigma_\ell^2\,\pi_\ell^{PE}(\sigma_\ell^2|\mathrm{X}_\ell^*\,,\delta)\,d\sigma_\ell^2\\ &= \int\left[\int\sigma_\ell^2 f_{IG}\left(\sigma_\ell^2\,;\,a_\ell+\frac{n^*}{2},b_\ell+\frac{SS_\ell^*}{2}\right)d\sigma_\ell^2\right]m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\\ &= \int\mathsf{E}_{IG}\left(\sigma_\ell^2;\,a_\ell+\frac{n^*}{2},b_\ell+\frac{SS_\ell^*}{2}\right)m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\,.\end{aligned} \tag{59}$$

Here $\mathsf{E}_{IG}\left[\xi(Z);a,b\right]$ is the expectation of a function $\xi(Z)$ of Z, where Z follows an Inverse-Gamma distribution with parameters a and b; for $\xi(Z)=Z$ this expectation is $\frac{b}{a-1}$. Thus the prior mean of σ_ℓ^2 is

$$\begin{aligned}\mathsf{E}\left(\sigma_\ell^2\right) &= \int\frac{b_\ell+\frac{SS_\ell^*}{2}}{\frac{n^*}{2}+a_\ell-1}m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\\ &= \left(\frac{n^*}{2}+a_\ell-1\right)^{-1}\left[b_\ell+\frac{1}{2}\mathsf{E}_{St_{n^*}}\left(SS_\ell^*;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)\right]\\ &= \left(\frac{n^*}{2}+a_\ell-1\right)^{-1}\left[b_\ell+\frac{1}{2}\mathsf{E}_{St_{n^*}}\left(\boldsymbol{y}^*\Lambda_\ell^*\,\boldsymbol{y}^*;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)\right]\,.\end{aligned} \tag{60}$$

From (55),

$$\begin{aligned}\mathsf{E}\left(\sigma_\ell^2\right) &= \left(\frac{n^*}{2}+a_\ell-1\right)^{-1}\left[b_\ell+\frac{1}{2}\frac{b_0}{a_0-1}tr(\Lambda_\ell^*\Lambda_0^{*-1})\right] \\ &= \frac{b_0}{a_0-1}\,\frac{\frac{1}{2}tr(\Lambda_\ell^*\Lambda_0^{*-1})+\frac{(a_0-1)b_\ell}{b_0}}{\frac{n^*}{2}+a_\ell-1}\,. \end{aligned} \tag{61}$$

The prior variance of σ_ℓ^2 can be written as

$$\begin{aligned}\mathsf{V}\left(\sigma_\ell^2\right) &= \mathsf{E}\left(\sigma_\ell^4\right)-\left[\mathsf{E}\left(\sigma_\ell^2\right)\right]^2 \\ &= \mathsf{E}\left(\sigma_\ell^4\right)-\left(\frac{b_0}{a_0-1}\right)^2\left[\frac{\frac{1}{2}tr(\Lambda_\ell^*\Lambda_0^{*-1})+\frac{(a_0-1)b_\ell}{b_0}}{\frac{n^*}{2}+a_\ell-1}\right]^2\,, \end{aligned} \tag{62}$$

where

$$\begin{aligned}\mathsf{E}\left(\sigma_\ell^4\right) &= \int \sigma_\ell^4\,\pi_\ell^{PE}(\sigma_\ell^2|\mathrm{X}_\ell^*\,,\delta)\,d\sigma_\ell^2 \\ &= \int\left[\int \sigma_\ell^4 f_{IG}\left(\sigma_\ell^2\,;\,a_\ell+\frac{n^*}{2},b_\ell+\frac{SS_\ell^*}{2}\right)m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\sigma_\ell^2\right]d\boldsymbol{y}^* \\ &= \int \mathsf{E}_{IG}\left[(\sigma_\ell^2)^2\,;\,a_\ell+\frac{n^*}{2},b_\ell+\frac{SS_\ell^*}{2}\right]m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\,. \end{aligned} \tag{63}$$

For a random variate Z that follows an Inverse-Gamma distribution with parameters a and b,

$$\mathsf{E}(Z^2)=\mathsf{V}(Z)+\mathsf{E}(Z)^2=\frac{b^2}{(a-1)^2(a-2)}+\left(\frac{b}{a-1}\right)^2=\frac{b^2}{(a-1)(a-2)} \tag{64}$$

for $a>2$. Hence

$$\begin{aligned}\mathsf{E}\left(\sigma_\ell^4\right) &= \int\frac{\left(b_\ell+\frac{SS_\ell^*}{2}\right)^2}{\left(\frac{n^*}{2}+a_\ell-1\right)\left(\frac{n^*}{2}+a_\ell-2\right)}\,m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^* \\ &= \int\frac{\left(b_\ell^2+b_\ell SS_\ell^*+\frac{1}{4}SS_\ell^{*2}\right)}{\left(\frac{n^*}{2}+a_\ell-1\right)\left(\frac{n^*}{2}+a_\ell-2\right)}\,m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^* \\ &= \left[\left(\frac{n^*}{2}+a_\ell-1\right)\left(\frac{n^*}{2}+a_\ell-2\right)\right]^{-1}\left[b_\ell^2+b_\ell\int SS_\ell^*\,m_0^N(\boldsymbol{y}^*;\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\right. \\ &\qquad \left.+\frac{1}{4}\int SS_\ell^{*2}\,m_0^N(\boldsymbol{y}^*;\mathrm{X}_0^*\,,\delta)\,d\boldsymbol{y}^*\right] \\ &= \left[\left(\frac{n^*}{2}+a_\ell-1\right)\left(\frac{n^*}{2}+a_\ell-2\right)\right]^{-1}\left\{b_\ell^2+b_\ell\,\mathsf{E}_{St_{n^*}}\left(\boldsymbol{y}^{*^T}\Lambda_\ell^*\,\boldsymbol{y}^*;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)\right. \\ &\qquad \left.+\frac{1}{4}\,\mathsf{E}_{St_{n^*}}\left[(\boldsymbol{y}^{*^T}\Lambda_\ell^*\,\boldsymbol{y}^*)^2;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right]\right\}\,. \end{aligned} \tag{65}$$

The expectation $\mathsf{E}_{St_{n^*}}\left(\boldsymbol{y}^{*^T}\Lambda_\ell^*\boldsymbol{y}^*;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right)$ is given by (55). Moreover, if A is a symmetric matrix and $\boldsymbol{z}\sim N_d(\mathbf{0},\Sigma)$ then (see Scott, 1997, Theorem 9.21)

$$\mathsf{E}\left[(\boldsymbol{z}^T\mathrm{A}\boldsymbol{z})^2\right]=2\,tr(\mathrm{A}\Sigma\mathrm{A}\Sigma)+tr(\mathrm{A}\Sigma)^2\,. \tag{66}$$

By rewriting the multivariate Student distribution with density $f_{St_d}(\boldsymbol{y}\,;\,df,\boldsymbol{\mu},\Sigma)$ as a Normal-Inverse-Gamma scale mixture, we can calculate $\mathsf{E}_{St_{n^*}}\left[(\boldsymbol{y}^{*^T}\Lambda_\ell^*\,\boldsymbol{y}^*)^2;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right]$, as follows: if $\boldsymbol{z}\sim St_d(df,\mathbf{0},\Sigma)$ then

$$\begin{aligned}
\mathsf{E}_{St_d}\left[(\boldsymbol{z}^T\mathrm{A}\boldsymbol{z})^2\,;\,df,\mathbf{0},\Sigma\right] &= \int(\boldsymbol{z}^T\mathrm{A}\boldsymbol{z})^2 f_{St_d}(\boldsymbol{z}\,;\,df,\mathbf{0},\Sigma)\,d\boldsymbol{z}\\
&= \int(\boldsymbol{z}^T\mathrm{A}\boldsymbol{z})^2\left[\int f_{N_d}(\boldsymbol{z}\,;\,\mathbf{0},\Sigma\psi)\,f_{IG}(\psi\,;\,\frac{df}{2},\frac{df}{2})\,d\psi\right]d\boldsymbol{z}\\
&= \int\left[\int(\boldsymbol{z}^T\mathrm{A}\boldsymbol{z})^2 f_{N_d}(\boldsymbol{z}\,;\,\mathbf{0},\Sigma\psi)\,d\boldsymbol{z}\right]f_{IG}(\psi\,;\,\frac{df}{2},\frac{df}{2})\,d\psi \qquad (67)\\
&= \left[2\,tr(\mathrm{A}\Sigma\mathrm{A}\Sigma)+tr(\mathrm{A}\Sigma)^2\right]\int\psi^2 f_{IG}(\psi\,;\,\frac{df}{2},\frac{df}{2})d\psi\\
&= \left[2\,tr(\mathrm{A}\Sigma\mathrm{A}\Sigma)+tr(\mathrm{A}\Sigma)^2\right]\frac{\frac{df^2}{4}}{\left(\frac{df}{2}-1\right)\left(\frac{df}{2}-2\right)}\,. \qquad (68)
\end{aligned}$$

It now follows that

$$\mathsf{E}_{St_{n^*}}\left[(\boldsymbol{y}^{*^T}\Lambda_\ell^*\,\boldsymbol{y}^*)^2;2\,a_0,\mathbf{0},\frac{b_0}{a_0}\Lambda_0^{*-1}\right]=\frac{b_0^2}{(a_0-1)(a_0-2)}\left[2\,tr(\Lambda_\ell^*\Lambda_0^{*-1}\Lambda_\ell^*\Lambda_0^{*-1})+tr(\Lambda_\ell^*\Lambda_0^{*-1})^2\right]. \qquad (69)$$

By substituting (55) and (69) in (65), we obtain

$$\begin{aligned}
\mathsf{E}\left(\sigma_\ell^4\right) &= \left[\left(\frac{n^*}{2}+a_\ell-1\right)\left(\frac{n^*}{2}+a_\ell-2\right)\right]^{-1}\left\{b_\ell^2+\frac{b_\ell\,b_0\,tr(\Lambda_\ell^*\Lambda_0^{*-1})}{a_0-1}\right. \qquad (70)\\
&\quad+\left.\frac{b_0^2\left[2\,tr(\Lambda_\ell^*\Lambda_0^{*-1}\Lambda_\ell^*\Lambda_0^{*-1})+tr(\Lambda_\ell^*\Lambda_0^{*-1})^2\right]}{4(a_0-1)(a_0-2)}\right\}. \qquad (71)
\end{aligned}$$

Finally, the prior variance is then

$$\begin{aligned}
\mathsf{V}\left(\sigma_\ell^2\right) &= \mathsf{E}\left(\sigma_\ell^4\right)-\left[\mathsf{E}\left(\sigma_\ell^2\right)\right]^2 \qquad (72)\\
&= \left[\left(\frac{n^*}{2}+a_\ell-1\right)\left(\frac{n^*}{2}+a_\ell-2\right)\right]^{-1}\left\{b_\ell^2+\frac{b_\ell\,b_0\,tr(\Lambda_\ell^*\Lambda_0^{*-1})}{a_0-1}\right.\\
&\quad+\left.\frac{b_0^2\left[2\,tr(\Lambda_\ell^*\Lambda_0^{*-1}\Lambda_\ell^*\Lambda_0^{*-1})+tr(\Lambda_\ell^*\Lambda_0^{*-1})^2\right]}{4(a_0-1)(a_0-2)}\right\}-\left(\frac{b_0}{a_0-1}\right)^2\left[\frac{\frac{1}{2}tr(\Lambda_\ell^*\Lambda_0^{*-1})+\frac{(a_0-1)b_\ell}{b_0}}{\frac{n^*}{2}+a_\ell-1}\right]^2.
\end{aligned}$$

□

B Implementation of the EIBF

The results in this paper based on the Expected Intrinsic Bayes Factor (EIBF) were obtained using the following procedure:

(1) Generate S random splits.

(2) For every random split $s=1,2,\ldots,S$ we denote by $\boldsymbol{y}_{[s]}$ and $\mathrm{X}_{[s]}$ the data used for the evaluation of the posterior distribution and by $\boldsymbol{y}_{[\backslash s]}$ and $\mathrm{X}_{[\backslash s]}$ the remaining data used for the calculation of the Bayes factor.

(3) For every split we calculate the marginal likelihood of model m_ℓ by

$$\psi(\boldsymbol{y}|m_\ell, s) = \int f(\boldsymbol{y}_{[\setminus s]}|\boldsymbol{\beta}_\ell\,, \sigma^2_\ell, m_\ell; \mathrm{X}_{\ell[\setminus s]})\, f(\boldsymbol{\beta}_\ell\,, \sigma^2_\ell|\boldsymbol{y}_{[s]}, m_\ell; \mathrm{X}_{\ell[s]})\, d\boldsymbol{\beta}_\ell\, d\sigma^2_\ell\,, \tag{73}$$

where $\mathrm{X}_{\ell[s]}$ and $\mathrm{X}_{\ell[\setminus s]}$ are the submatrices of $\mathrm{X}_{[s]}$ and $\mathrm{X}_{[\setminus s]}$ corresponding to model m_ℓ, respectively.

(4) We then calculate the Bayes factor of model m_ℓ versus the reference model m_0 under the split s by

$$IBF^{[s]}_{\ell 0} = \frac{\psi(\boldsymbol{y}|m_\ell, s)}{\psi(\boldsymbol{y}|m_0, s)}\,. \tag{74}$$

(5) We calculate the EIBF by computing the arithmetic mean of $IBF^{[s]}_{\ell 0}$ over all splits $s = 1, \ldots, S$:

$$EIBF_{\ell 0} = \frac{1}{S}\sum_{s=1}^{S} IBF^{[s]}_{\ell 0}\,. \tag{75}$$

(6) All weights based on the EIBF are calculated by

$$W^{EIBF}_\ell = \frac{EIBF_{\ell 0}}{\sum_{m_\ell \in \mathcal{M}} EIBF_{\ell 0}}\,. \tag{76}$$

For the IBF approach we generate one random split and we calculate the Bayes factor as in step (4).

C Ozone data set details

Here we present a brief description of the data and the transformations used in the analyses of Section 6.2 based on the original ozone data.

- The **response variable** ozone (daily maximum of 24 hourly averages: midnight–1am, 1–2am, ..., 11pm–midnight) had substantial positive skew; within the Box-Cox power transformation family the optimal transformation was $ozone^{0.17}$. This is not far from the log transform, which is easier to interpret, so we standardized log(ozone) (subtracting off its mean and dividing by its standard deviation: all of the standardizations here and below are to stabilize the numerical work and reduce the correlations between main effects and their squares); the standardized variable is called `s_log_ozone`.

- **Month, day and weekday predictors**: Weekday had no effect on ozone and was omitted. We combined month and day to create a variable called `day_of_year` that ran through the consecutive integers 1–366 from 1 Jan to 31 Dec, and we then standardized this variable; the standardized variable is called `s_day_of_year` . We then created a variable called `s_day_of_year_2` by squaring `s_day_of_year`.

- **Temperature predictors**: Temperature at Sandburg and temperature at El Monte were very highly correlated, and the El Monte temperature variable had 139 missing values (versus only 2 missing values for the Sandburg temperature), so we omitted the El Monte temperature variable.

- The **remaining 8 covariates** were `pressure_500`, `humidity`, `temp_sandburg`, `inversion_height`, `wind`, `pressure_gradient`, `inversion_temp`, and `visibility`; these were standardized as above (the variable names of the resulting standardized versions are the same as those of the original variables preceded by `s_`). We also calculated squared versions of all standardized variables with a naming convention similar to that above (for example, the squared standardized wind variable was called `s_wind_2`).

- Omitting all rows of data for which one or more of the predictors were missing, our regression modeling was based on 330 days of data.
- Local-regression (`loess`) descriptive analyses of the relationships between the outcome `s_log_ozone` and each of our 9 predictor variables revealed cubic relationships between the outcome and the predictors `temp_sandburg` and `inversion_temp`, so we raised each of `s_temp_sandburg` and `s_inversion_temp` to the third power and included these two cubic terms among the total set of predictor variables.
- With 9 main effects there are $\frac{9 \cdot 8}{2} = 36$ pairwise interactions among the main effects; we also created these 36 variables by multiplying the standardized versions of the predictors in a pairwise manner. The resulting variables had names of the form `humidityXwind`.
- Our total set of predictor variables therefore had 9 main effects, 9 quadratic terms, 2 cubic terms, and 36 two-way interactions, for a total of 56 predictors.
- The final data set contains 330 rows and 57 columns; the first column is `s_log_ozone`, and the other 56 columns are the predictors.

Additional details on the ozone data analysis are available in a supplemental document provided on request; our version of the data set is also available from us.

References

Aitkin, M. (1991), 'Posterior Bayes factors', *Journal of the Royal Statistical Society Series B,* **53**, 111–142.

Armagan, A., Dunson, D. and Lee, J. (2012), 'Generalized double Pareto shrinkage', *arXiv:1104.0861v3,* available at `http://arxiv.org/abs/1104.0861` .

Balakrishnan, S. and Madigan, D. (2010), Priors on the variance in sparse Bayesian learning: the demi-Bayesian lasso, In *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, M.-H. Chen, P. Müller, D. Sun, and K. Ye (eds.), Springer Science and Business Media, pp. 346–359.

Barbieri, M. and Berger, J. (2004), 'Optimal predictive model selection', *Annals of Statistics,* **32**, 870–897.

Berger, J. and Pericchi, L. (1996*a*), The intrinsic Bayes factor for linear models, in J. Bernardo, J. Berger, A. Dawid, and A. Smith, eds., *Bayesian Statistics*, Vol. 5, Oxford University Press, pp. 25–44.

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

Berger, J. and Pericchi, L. (2004), 'Training samples in objective model selection', *Annals of Statistics,* **32**, 841–869.

Breiman, L. and Friedman, J. (1985), 'Estimating optimal transformations for multiple regression and correlation', *Journal of the American Statistical Association,* **80**, 580–598.

Carvalho, C., Polson, N. and Scott, J. (2010), 'The horseshoe estimator for sparse signal', *Biometrika,* **97**, 465–480.

Casella, G., Girón, F., Martínez, M. and Moreno, E. (2009), 'Consistency of Bayesian procedures for variable selection', *Annals of Statistics,* **37**, 1207–1228.

Casella, G. and Moreno, E. (2006), 'Objective Bayesian variable selection', *Journal of the American Statistical Association,* **101**, 157–167.

Fahrmeir, L., Kneib, T. and Konrath, S. (2010), 'Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection', *Statistics and Computing,* **20**, 203–219.

Fernandez, C., Ley, E. and Steel, M. (2001), 'Benchmark priors for Bayesian model averaging', *Journal of Econometrics,* **100**, 381–427.

Fouskakis, D. and Ntzoufras, I. (2012), 'Computation for intrinsic variable selection in normal regression models via expected-posterior prior', *Statistics and Computing, forthcoming* .

Fouskakis, D., Ntzoufras, I. and Draper, D. (2009), 'Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care', *Annals of Applied Statistics,* **3**, 663–690.

George, E. and Foster, D. (2000), 'Calibration and empirical Bayes variable selection', *Biometrika,* **87**, 731–748.

George, E. and McCulloch, R. (1993), 'Variable selection via Gibbs sampling', *Journal of the American Statistical Association,* **88**, 881–889.

George, E. and McCulloch, R. (1997), 'Approaches for Bayesian variable selection', *Statistica Sinica,* **7**, 339–373.

Good, I. (2004), *Probability and the Weighting of Evidence*, Haffner, New York, USA.

Griffin, J. E. and Brown, P. J. (2010), 'Inference with normal-gamma prior distributions in regression problems', *Bayesian Analysis,* **5**, 171–188.

Hans, C. (2009), 'Bayesian lasso regression', *Biometrika,* **96**, 835–845.

Hans, C. (2010), 'Model uncertainty and variable selection in Bayesian lasso regression', *Statistics and Computing,* **20**, 221–229.

Hoeting, J., Madigan, D. and Raftery, A. (1996), 'A method for simultaneous variable selection and outlier identification in linear regression', *Computational Statistics and Data Analysis,* **22**, 251–270.

Hoeting, J., Raftery, A. and Madigan, D. (2002), 'A method for simultaneous variable and transformation selection in linear regression', *Journal of Computational and Graphical Statistics,* **11**, 485–507.

Ibrahim, J. and Chen, M. (2000), "Power prior distributions for regression models", *Statistical Science,* **15**, 46–60.

Ishwaran, H. and Rao, J. (2005), 'Spike and slab variable selection: Frequentist and Bayesian strategies', *The Annals of Statistics,* **33**, 730–773.

Iwaki, K. (1997), 'Posterior expected marginal likelihood for testing hypotheses', *Journal of Economics, Asia University,* **21**, 105–134.

Kass, R. and Wasserman, L. (1995), 'A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion', *Journal of the American Statistical Association,* **90**, 928–934.

Li, Q. and Lin, N. (2010), 'The Bayesian elastic net', *Bayesian Analysis,* **5**, 847–866.

Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. (2008), 'Mixtures of g priors for Bayesian variable selection', *Journal of the American Statistical Association,* **103**, 410–423.

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

Mitchell, T. and Beauchamp, J. (1988), 'Bayesian variable selection in linear regression', *Journal of the American Statistical Association,* **83**, 1023–1036.

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

Nott, D. and Kohn, R. (2005), 'Adaptive sampling for Bayesian variable selection', *Biometrika,* **92**, 747–763.

O'Hagan, A. (1995), 'Fractional Bayes factors for model comparison', *Journal of the Royal Statistical Society Series B,* **57**, 99–138.

Park, T. and Casella, G. (2008), 'The Bayesian lasso', *Journal of the American Statistical Association,* **103**, 681–686.

Pérez, J. (1998), *Development of Expected Posterior Prior Distribution for Model Comparisons*, PhD thesis, Department of Statistics, Purdue University, USA.

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

Raftery, A., Madigan, D. and Hoeting, J. (1997), 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association,* **92**, 179–191.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics,* **6**, 461–464.

Scott, J. (1997), *Matrix Analysis for Statistics*, Wiley Series in Probability and Statistics, New York, USA.

Spiegelhalter, D., Abrams, K. and Myles, J. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Statistics in Practice, Wiley, Chichester, UK.

Spiegelhalter, D. and Smith, A. (1988), 'Bayes factors for linear and log-linear models with vague prior information', *Journal of the Royal Statistical Society Series B,* **44**, 377–387.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B,* **58**, 267–288.

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

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

Zou, J. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society Series B,* **67**, 301–320.