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

Dimitris Fouskakis,

Department of Mathematics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Athens, Greece; e-mail: fouskakis@math.ntua.gr.

Joint work with:

Ioannis Ntzoufras & David Draper

Department of Statistics
Athens University of Economics and Business
Athens, Greece; e-mail: ntzoufras@aueb.gr

Department of Applied Mathematics and Statistics
University of California
Santa Cruz, USA; e-mail: draper@ams.ucsc.edu

Presentation is available at:
www.math.ntua.gr/~fouskakis/Presentations/Irvine/Presentation_Irvine.pdf.
Synopsis

1. Bayesian Variable Selection
2. Prior Specification
3. Expected-Posterior Priors
4. Motivation
5. Power-Expected Posterior (PEP) Methodology
6. MCMC for Sampling from the Posterior
7. Marginal Likelihood Computation
8. Simulated Example
9. Real Life Example
1 Bayesian Variable Selection

Within the Bayesian framework the identification of the “best set of predictors” between the \( \mathcal{M} = \{0, 1\}^p \) competitors is equivalent (assuming a zero-one loss function) to find the model \( m \) with the highest posterior model probability, defined as

\[
f(m|y) = \frac{f(y|m)f(m)}{\sum_{m_\ell \in \mathcal{M}} f(y|m_\ell)f(m_\ell)},
\]

where \( f(y|m) \) is the marginal likelihood under model \( m \) and \( f(m) \) is the prior probability of model \( m \). The marginal likelihood function in the above calculation can be further expanded to include the effect of the model parameters:

\[
f(y|m) = \int f(y|\theta_m, m)f(\theta_m|m)d\theta_m,
\]

where \( f(y|\theta_m, m) \) is the likelihood under model \( m \) with parameters \( \theta_m \) and \( f(\theta_m|m) \) is the prior distribution of model parameters given model \( m \).
Posterior odds and Bayes factors

Based on the above posterior model probabilities, pairwise comparisons of any two models, $m_k$ and $m_\ell$, is given by the **Posterior Odds (PO)**

$$PO_{m_k,m_\ell} \equiv \frac{f(m_k|y)}{f(m_\ell|y)} = \frac{f(y|m_k)}{f(y|m_\ell)} \times \frac{f(m_k)}{f(m_\ell)} = B_{m_k,m_\ell} \times O_{m_k,m_\ell}$$

which is a function of the **Bayes Factor** $B_{m_k,m_\ell}$ and the **Prior Odds** $O_{m_k,m_\ell}$.

The posterior model probability can be then expressed entirely in terms of Bayes Factors and Prior Odds as

$$f(m|y) = \left( \sum_{m_\ell \in \mathcal{M}} PO_{m_\ell,m} \right)^{-1} = \left[ \sum_{m_\ell \in \mathcal{M}} B_{m_\ell,m} \times O_{m_\ell,m} \right]^{-1}.$$
2 Prior Specification

Prior on the model space

- Uniform prior on the model space
  \[ f(m) \propto I[m \in \mathcal{M}] . \]
  The above prior is equivalent of assuming that each covariate has prior probability 0.5 of entering the model.

- Beta-Binomial hierarchical prior on the model size \( W \)
  \[ W \sim Bin(p, \theta) \text{ and } \theta \sim Beta(\alpha, \beta). \]
  If \( \alpha = \beta = 1 \) then we have a uniform prior on the model size.
Prior on model parameters

- Proper prior distributions (conjugate if available).
  - For example in the case of the Gaussian regression models a popular choice is the **Zellner’s g-prior** (Zellner, 1986).
  - Main issue: Specification of hyperparameter $g$ that controls the prior variance.
  - Large values of $g \rightarrow$ Bartlett’s paradox (e.g., Bartlett, 1957 Biometrika).
  - For $g = n \Rightarrow$ **unit information prior** (Kass & Wasserman, 1995, JASA).
  - Beta prior on $\frac{g}{g+1} \rightarrow$ Mixtures of $g$-priors (e.g. Liang *et al.*, 2008, JASA).
- Non-local priors (e.g. Johnson and Rossell, 2010, RSSS B).
  - They have zero mass for values of the parameter under the null hypothesis.
  - Products of independent normal moment priors.
Prior on model parameters (cont.)

- Shrinkage priors.
  - E.g. Bayesian Lasso (Park and Casella, 2008, JASA, horseshoe prior (Carvalho et al., 2010, Biometrika), etc.

- Improper (reference) priors (defined up to arbitrary constants).
  - Objectivity.
  - Jeffreys prior.
  - Bayes factors cannot be determined.

- Priors defined via imaginary data.
  - Expected-Posterior prior (Pérez & Berger, 2002, Biometrika).
3 Expected-Posterior Priors

Expected-posterior priors (EPPs) are defined as the posterior distribution of a parameter vector for the model under consideration, averaged over all possible imaginary samples $y^*$ coming from a “suitable” predictive distribution $m^*(y^*)$. Hence the EPP for the parameters of any model $m_\ell \in M$, with $M$ denoting the model space, is

$$
\pi^E_\ell (\theta_\ell) = \int f(\theta_\ell|y^*, m_\ell) m^*(y^*) dy^*
$$

$$
= \int \pi^N_\ell (\theta_\ell|y^*) m^*(y^*) dy^*
$$

$$
= \int \frac{f(y^*|\theta_\ell, m_\ell)\pi^N_\ell (\theta_\ell)}{\int f(y^*|\theta_\ell, m_\ell)\pi^N_\ell (\theta_\ell)d\theta_\ell} m^*(y^*) dy^*. \quad (1)
$$
• $\pi_N^\ell(\theta_\ell)$ can be proper or improper. If improper, Bayes factors can be defined.

• Different choices of $m^*$ include:
  
  – **Base-model approach.** Select a “reference” model $m_0$ for the training sample and define $m^*(y^*) = m_0^N(y^*) \equiv f(y^*|m_0)$ to be the prior predictive distribution, evaluated at $y^*$, for the reference model $m_0$ under the baseline prior $\pi_0^N(\theta_0)$. 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 the constant model (with no predictors).

  – Use of empirical distribution for $m^*$.

  – Subjective specification of $m^*$.

• Under the base-model approach the resulting expected-posterior Bayes factors becomes essentially identical to the arithmetic **Intrinsic Bayes Factors** (IBF), Berger and Pericchi (1996, JASA).
We need to specify the imaginary responses $y^*$ (of size $n^*$) and the imaginary design matrix $X^*$ (of size $n^* \times d$). Then the EPP, will depend on $X^*$ but not on $y^*$, since the latter is integrated out.

- How large should such training samples be, i.e how large should $n^*$ be?
  - Minimal training sample. With $p$ covariates and unknown error variance, we set $n^* = (p + 1) + 1 = d + 1$. Then to specify $X^*$ we randomly select a sub-sample of the rows of the original matrix $X$ ($n^*$ should be $< n$).
  - 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.

- How should the rows of $X$ be chosen?
  - With $(n, p) = (100, 50)$ there are about $10^{29}$ possible choices.
  - Arithmetic mean of the Bayes factors over all possible $X^*$.
  - Take a random sample from the set of all possible training samples, but this adds an extraneous layer of Monte-Carlo noise to the model-comparison process.

- 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”.
4 Motivation

1. Produce a less influential expected-posterior prior.

2. Diminish the effect of training samples on the expected-posterior prior methodology.

We combine ideas from the power prior approach and unit information prior approach → power-expected-posterior prior:

- We raise the likelihood involved in the expected-posterior prior distribution to the power of $1/\delta$.

- For $\delta = n^*$ this produces a prior information content equivalent to one data point.

- The method is sufficiently insensitive to the size of $n^*$; 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.
Specification

• We focus on variable selection problems for **Gaussian linear models**. We consider two models $m_\ell$ ($\ell = 0, 1$) with parameters $\theta_\ell = (\beta_\ell, \sigma_\ell^2)$ and likelihood specified by

$$Y | X_\ell, \beta_\ell, \sigma_\ell^2, m_\ell \sim N_n(X_\ell \beta_\ell, \sigma_\ell^2 I_n),$$

where $Y = (Y_1, \ldots, Y_n)$ is a vector containing the responses for all subjects, $X_\ell$ 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, $\beta_\ell$ is a vector of length $d_\ell$ summarizing the effects of the covariates on the response $Y$ and $\sigma_\ell^2$ is the error variance.

• Two different Baseline Prior Choice:
  – **Independence Jeffrey’s prior**, i.e. $p(\beta_\ell, \sigma_\ell^2 | m_\ell) \propto \sigma_\ell^{-2}$.
  – Zellner’s g-prior.

• Base-model approach; the null model is the reference model.
5 Power-Expected-Posterior (PEP) Methodology

- The PEP prior is:

\[ \pi_{\ell}^{PE}(\beta_\ell, \sigma_\ell^2 | X^*_\ell, \delta) = \pi_{\ell}^N(\beta_\ell, \sigma_\ell^2 | X^*_\ell) \times \int \frac{m_0^N(y_\ell^* | X_\ell^* \beta_\ell, \delta \sigma_\ell^2 I_{n^*})}{m_\ell^N(y_\ell^* | X^*_\ell, \delta)} f_{N_{n^*}}(y_\ell^* ; X_\ell^* \beta_\ell, \delta \sigma_\ell^2 I_{n^*}) \, dy_\ell^*, \]

- The posterior under the PEP prior becomes:

\[ \pi_{\ell}^{PE}(\beta_\ell, \sigma_\ell^2 | y; X_\ell, X^*_\ell, \delta) \propto \int f(\beta_\ell, \sigma_\ell^2 | y, y^*_\ell, m_\ell ; X_\ell, X^*_\ell, \delta) m_\ell^N(y | y^*_\ell ; X_\ell, X^*_\ell, \delta) m_0^N(y^*_\ell | X_0^* \beta_\ell, \delta \sigma_\ell^2 I_{n^*}) \, dy^*_\ell, \]

- \( f(\beta_\ell, \sigma_\ell^2 | y, y^*_\ell, m_\ell ; X_\ell, X^*_\ell, \delta) \): posterior using data \( y \), design matrix \( X_\ell \) under prior \( f(\beta_\ell, \sigma_\ell^2 | y^*_\ell, m_\ell ; X^*_\ell, \delta) \) (i.e. the posterior under power normal likelihood and baseline prior).

- \( m_\ell^N(y | y^*_\ell ; X_\ell, X^*_\ell, \delta) \): marginal likelihood of model \( m_\ell \), using data \( y \), design matrix \( X_\ell \) and prior \( f(\beta_\ell, \sigma_\ell^2 | y^*_\ell, m_\ell ; X^*_\ell, \delta) \).
PEP under the Jeffreys baseline

Baseline Prior:

\[
\pi^N_\ell (\beta, \sigma^2 | X^*_\ell) = \frac{c_\ell}{\sigma^2_\ell},
\]  
(2)

where \(c_\ell\) is an unknown normalizing constant.

J-PEP prior:

\[
\pi^{PE}_\ell (\beta, \sigma^2 | X^*_\ell, \delta) = \int f_{\mathcal{N}_d_\ell} \left[ \beta; \hat{\beta}^*_\ell, \delta \left( X^*_\ell X^*_\ell \right)^{-1} \sigma^2_\ell \right] f_{\mathcal{IG}} \left( \sigma^2_\ell, \frac{n^* - d_\ell}{2}, \frac{RSS^*_\ell}{2} \right)
\]

\[
\times m^N_0 (y^* | X^*_0, \delta) \, dy^* ,
\]

with \(\hat{\beta}^*_\ell = (X^*_\ell X^*_\ell)^{-1} X^*_\ell y^*\), \(RSS^*_\ell = y^T (I_n^* - X^*_\ell (X^*_\ell X^*_\ell)^{-1} X^*_\ell y^*) y^* \) and

\[
m^N_\ell (y^* | X^*_\ell, \delta) = c_\ell \pi^{\frac{d_\ell - n^*}{2}} |X^*_\ell X^*_\ell|^{-\frac{1}{2}} \Gamma \left( \frac{n^* - d_\ell}{2} \right) \left( \frac{RSS^*_\ell}{2} \right)^{-\frac{n^* - d_\ell}{2}}.
\]
The posterior under the J-PEP prior:

\[
\pi_{PE}^{PE}(\beta_\ell, \sigma_\ell^2 | y; X_\ell, X_*^\ell, \delta) \propto \int f_{Nd_\ell}(\beta_\ell; \tilde{\beta}^N, \tilde{\Sigma}^N \sigma_\ell^2) f_{IG}(\sigma_\ell^2; \tilde{a}_\ell^N, \tilde{b}_\ell^N) \\
\times m_N^N(y|y^*; X_\ell, X_*^\ell, \delta) m_0^N(y^*|X_0^*, \delta) \, dy^*,
\]

with

\[
\tilde{\beta}^N = \tilde{\Sigma}^N (X^T_\ell y + \delta^{-1} X_*^T \ell y^*), \quad \tilde{\Sigma}^N = \left[ X^T_\ell X_\ell + \delta^{-1} X^T_\ell \ell X_*^\ell \right]^{-1},
\]

\[
\tilde{a}_\ell^N = \frac{n + n^* - d_\ell}{2}, \quad \tilde{b}_\ell^N = \frac{SS^N_\ell + \delta^{-1} RSS^*_\ell}{2} + b_\ell;
\]

where

\[
SS^N_\ell = (y - X_\ell \hat{\beta}_\ell^*)^T \left[ I_n + \delta X_\ell (X_*^T \ell X_*^\ell)^{-1} X^T_\ell \right]^{-1} (y - X_\ell \hat{\beta}_\ell^*),
\]

while,

\[
m_N^N(y|y^*; X_\ell, X_*^\ell, \delta) = f_{Stn} \left\{ y; n^* - d_\ell, X_\ell \hat{\beta}_\ell^*, \frac{RSS^*_\ell}{\delta(n^* - d_\ell)} \left[ I_n + \delta X_\ell (X_*^T \ell X_*^\ell)^{-1} X^T_\ell \right] \right\}.
\]
PEP under the Zellner’s g-prior baseline

Baseline Prior:

$$\pi^N_\ell (\beta_\ell | \sigma^2_\ell ; X^*_\ell) = f_{N_{d_\ell}} \left[ \beta_\ell ; 0, g (X^*_\ell T X^*_\ell)^{-1} \sigma^2_\ell \right] \quad \text{and} \quad \pi^N_\ell (\sigma^2_\ell) = f_{IG} \left( \sigma^2_\ell ; a_\ell, b_\ell \right). \quad (3)$$

Z-PEP Prior:

$$\pi^{PE}_\ell (\beta_\ell, \sigma^2_\ell | X^*_\ell, \delta) = \int f_{N_{d_\ell}} \left[ \beta_\ell ; w \widehat{\beta}^*_\ell, w \delta (X^*_\ell T X^*_\ell)^{-1} \sigma^2_\ell \right] f_{IG} \left( \sigma^2_\ell ; a_\ell + \frac{n^*}{2}, b_\ell + \frac{SS^*_\ell}{2} \right) \times m^N_0 \left( y^* | X^*_0, \delta \right) dy^*. \quad (4)$$

Here $$w = \frac{g}{g+\delta}$$, $$\widehat{\beta}^*_\ell = (X^*_\ell T X^*_\ell)^{-1} X^*_\ell T y^*$$, $$SS^*_\ell = y^* T \Lambda^*_\ell y^*$$ and

$$m^N_0 \left( y^* | X^*_0, \delta \right) = f_{St_{n^*}} \left( y^* ; 2 a_\ell, 0, \frac{b_\ell \Lambda^*_{\ell^{-1}}}{a_\ell} \right),$$

with

$$\Lambda^*_{\ell^{-1}} = \delta \left[ I_{n^*} - \frac{g}{g+\delta} X^*_\ell (X^*_\ell T X^*_\ell)^{-1} X^*_\ell T \right]^{-1} = \delta I_{n^*} + g X^*_\ell (X^*_\ell T X^*_\ell)^{-1} X^*_\ell T.$$
PEP under the Zellner’s g-prior baseline (cont.)

**Theorem 1.** Under the baseline prior setup (3), the power-expected-posterior prior mean of $\beta_\ell$ is equal to zero (i.e. $E[\beta_\ell] = 0$) while the power-expected-posterior prior variance is given by

$$V[\beta_\ell] = \left\{ \frac{\delta w}{a_\ell - 1 + n^*/2} \left[ b_\ell + \frac{1}{2} \frac{b_0}{a_0 - 1} tr(\Lambda_\ell^* \Lambda_0^{-1}) \right] I_{d_\ell} + \frac{w^2 b_0}{a_0 - 1} (X_\ell^* X_\ell^*)^{-1} X_\ell^* \Lambda_0^{-1} X_\ell^* \right\} (X_\ell^* X_\ell^*)^{-1}.$$  

where $tr(A)$ is the trace of matrix $A$.

**Theorem 2.** Under the baseline prior setup (3), the power-expected-posterior prior mean of $\sigma_\ell^2$ is given by

$$E[\sigma_\ell^2] = \frac{b_0}{a_0 - 1} \left\{ \frac{1}{2} tr(\Lambda_\ell^* \Lambda_0^{-1}) + (a_0 - 1) b_\ell / b_0 \right\} \frac{n^*/2 + a_\ell - 1}{n^*/2}$$

while the power-expected-posterior prior variance is given by

$$V[\sigma_\ell^2] = E[\sigma_\ell^4] - E[\sigma_\ell^2]^2$$

$$= \left\{ \left( \frac{n^*}{2} + a_\ell - 1 \right) \frac{n^*}{2} + a_\ell - 2 \right\}^{-1} \left( \frac{b_\ell^2 b_0 tr(\Lambda_\ell^* \Lambda_0^{-1})}{a_0 - 1} + \frac{b_0^2 \left\{ 2 tr(\Lambda_\ell^* \Lambda_0^{-1} \Lambda_\ell^* \Lambda_0^{-1}) + tr(\Lambda_\ell^* \Lambda_0^{-1})^2 \right\}}{4(a_0 - 1)(a_0 - 2)} \right)$$

$$- \left( \frac{b_0}{a_0 - 1} \right)^2 \left( \frac{1}{2} tr(\Lambda_\ell^* \Lambda_0^{-1}) + (a_0 - 1) b_\ell / b_0 \right)^2 \frac{n^*/2 + a_\ell - 1}{n^*/2}.$$
The posterior under the Z-PEP prior:

\[
\pi_{\ell}^{PE}(\beta_\ell, \sigma_\ell^2 | y; X_\ell, X^*_\ell, \delta) \propto \int f_{N_{d_\ell}}(\beta_\ell; \tilde{\beta}_N^\ell, \tilde{\Sigma}^N_{\sigma_\ell^2}) f_{IG}(\sigma_\ell^2; \tilde{a}_N^\ell, \tilde{b}_N^\ell) \\
\times m_N^n(y|y^*; X_\ell, X^*_\ell, \delta) m_0^N(y^*|X^*_0, \delta) dy^*,
\]

with

\[
\tilde{\beta}^N = \tilde{\Sigma}^N (X^T_\ell y + \delta^{-1}X^{*T}_\ell y^*), \quad \tilde{\Sigma}^N = \left[ X^T_\ell X_\ell + (w \delta)^{-1}X^{*T}_\ell X^*_\ell \right]^{-1},
\]

\[
\tilde{a}_N^\ell = \frac{n + n^*}{2} + a_\ell, \quad \tilde{b}_N^\ell = \frac{SS^N_\ell + SS^*_\ell}{2} + b_\ell;
\]

where

\[
SS^N_\ell = (y - w X_\ell \hat{\beta}^*_\ell)^T \left[ I_n + \delta w X_\ell (X^{*T}_\ell X^*_\ell)^{-1} X^T_\ell \right]^{-1} (y - w X_\ell \hat{\beta}^*_\ell),
\]

while

\[
m_N^n(y|y^*; X_\ell, X^*_\ell, \delta) = f_{St_n} \left\{ y; 2a_\ell + n^*, w X_\ell \hat{\beta}^*_\ell, \frac{2b_\ell + SS^*_\ell}{2a_\ell + n^*} \left[ I_n + w \delta X_\ell (X^{*T}_\ell X^*_\ell)^{-1} X^T_\ell \right] \right\}
\]
Hyper-parameters

- 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 $f(\beta_\ell, \sigma^2_\ell | y^*, m_\ell ; X^*_\ell, \delta)$. In this manner, the entire PEP prior accounts for information equal to $(1 + \frac{1}{\delta})$ 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 can be used to guide the choice of $a$ and $b$.)

- By comparing the posterior distributions under the two different baseline schemes described before, it is straightforward to prove that they coincide for large $g$ (and therefore for $w \approx 1$), $a_\ell = -d_\ell/2$ and $b_\ell = 0$. 
6 MCMC for Sampling from the Posterior

We consider the following augmented conditional distribution:

\[
f(\beta_\ell, \sigma^2_\ell, y^* | y; X_\ell, X^*_\ell, \delta) \propto f_{N_{d_\ell}}(\beta_\ell; \tilde{\beta}^N_\ell, \tilde{\Sigma}^N_\ell \sigma^2_\ell) \times f_{IG}(\sigma^2_\ell; \tilde{a}^N_\ell, \tilde{b}^N_\ell) \times f(y^* | y; X_\ell, X^*_\ell, \delta),
\]

with

\[
f(y^* | y; X_\ell, X^*_\ell, \delta) \propto m^N_\ell (y^* | y, X_\ell, X^*_\ell, \delta) \frac{m^N_0 (y^* | X^*_0, \delta)}{m^N_\ell (y^* | X^*_\ell, \delta)},
\]

with

\[
m^N_\ell (y^* | y, X_\ell, X^*_\ell, \delta) = \int \int f(y^* | \beta_\ell, \sigma^2_\ell, m_\ell ; X^*_\ell, \delta) f(\beta_\ell, \sigma^2_\ell | y, m_\ell ; X_\ell) \, d\beta_\ell \, d\sigma^2_\ell.
\]

For the Zellner’s baseline prior, (4) becomes equal to

\[
f_{St_{n^*}} \left\{ y^* ; 2 a_\ell + n, \frac{g}{g+1} X_\ell \hat{\beta}_\ell, \frac{2b_\ell + SS_\ell}{2 a_\ell + n} \left[ \delta I_{n^*} + \frac{g}{g+1} X^* (X^T X_\ell)^{-1} X^*_\ell \right] \right\},
\]

where \( SS_\ell = y^T \left[ I_n - \frac{g}{g+1} X_\ell (X^T X_\ell)^{-1} X^T_\ell \right] y \), while, for the Jeffreys baseline prior, (4) becomes

\[
f_{St_{n^*}} \left\{ y^* ; n - d_\ell, X_\ell \hat{\beta}_\ell, \frac{SS_\ell}{n - d_\ell} \left[ \delta I_{n^*} + X^* (X^T X_\ell)^{-1} X^*_\ell \right] \right\},
\]

with the posterior sum of squares now given by \( SS_\ell = y^T \left[ I_n - X_\ell (X^T X_\ell)^{-1} X^T_\ell \right] y \).
Using the previous expressions, we can specify the following MCMC scheme:

1. Generate $y^*$ from $f(y^* | y; X_\ell, X_\ell^*, \delta)$.
2. Generate $\sigma^2_\ell$ from $IG(\tilde{a}^N_\ell, \tilde{b}^N_\ell)$.
3. Generate $\beta_\ell$ from $N_d(\bar{\beta}^N_\ell, \bar{\Sigma}^N_\sigma^2_\ell)$.

In Step 1, we can generate the imaginary data $y^*$ by using a Metropolis-Hastings algorithm with proposal $q(y^*) = m^N_\ell(y^* \mid y, X_\ell, X_\ell^*, \delta)$ given in (4) and acceptance probability

$$\alpha = \min \left[ 1, \frac{m^N_0(y^* \mid X_0^*, \delta) m^N_\ell(y^* \mid X_\ell^*, \delta)}{m^N_\ell(y^* \mid X_\ell^*, \delta) m^N_0(y^* \mid X_0^*, \delta)} \right].$$
7 Marginal Likelihood Computation

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

$$m_{\ell}^{PE}(y|X_\ell, X_\ell^*, \delta) = \int \int f(y|\beta_\ell, \sigma_{\ell}^2, m_\ell; X_\ell) \pi_{\ell}^{PE}(\beta_\ell, \sigma_{\ell}^2|X_\ell^*, \delta) \, d\beta_\ell \, d\sigma_{\ell}^2$$

$$= m_{\ell}^N(y|X_\ell, X_\ell^*) \int \frac{m_{\ell}^N(y^*|y, X_\ell, X_\ell^*, \delta)}{m_{\ell}^N(y^*|X_\ell^*, \delta)} \, m_{0}^N(y^*|X_0^*, \delta) \, dy^*.$$ 

In the above expression $m_{\ell}^N(y|X_\ell, X_\ell^*)$ is the marginal likelihood of model $m_\ell$ for the actual data under the baseline prior. Therefore, under the Zellner’s baseline prior is given by

$$m_{\ell}^N(y|X_\ell, X_\ell^*) = \frac{2}{a_{\ell}} \left[ I_n + g X_\ell \left( X_\ell^T X_\ell^* \right)^{-1} X_\ell^T \right]$$

while under the Jeffreys baseline prior is given by

$$m_{\ell}^N(y|X_\ell, X_\ell^*) = c_{\ell} \pi \frac{d_{\ell}-n}{2} |X_\ell^T X_\ell|^{-\frac{1}{2}} \Gamma \left( \frac{n-d_{\ell}}{2} \right) R S S_{\ell}^{-\frac{n-d_{\ell}}{2}},$$

with $R S S_{\ell} = y^T (I_{n^*} - X_\ell (X_\ell^T X_\ell)^{-1} X_\ell^T) y.$
Marginal likelihood computation (cont.)

1. Generate $y^*(t) (t = 1, \ldots, T)$ from $m^N_0(y^*|X_0^*, \delta)$ and estimate the marginal likelihood by

$$
\hat{m}^{PE}_\ell(y|X_\ell, X^*_\ell, \delta) = m^N_\ell(y|X_\ell, X^*_\ell) \left[ \frac{1}{T} \sum_{t=1}^{T} \frac{m^N_\ell(y^*(t)|y, X_\ell, X^*_\ell, \delta)}{m^N_\ell(y^*(t)|X_\ell^*, \delta)} \right].
$$

2. Generate $y^*(t) (t = 1, \ldots, T)$ from $m^N_0(y^*|y, X_0, X^*_0, \delta)$ and estimate the marginal likelihood by

$$
\hat{m}^{PE}_\ell(y|X_\ell, X^*_\ell, \delta) = m^N_0(y|X_0, X^*_0) \left[ \frac{1}{T} \sum_{t=1}^{T} \frac{m^N_0(y^*(t)|y^*, X_\ell, X^*_\ell, \delta)}{m^N_0(y^*(t)|X_0^*, \delta)} \right].
$$

3. Generate $y^*(t) (t = 1, \ldots, T)$ from $m^N_\ell(y^*|y, X_\ell, X^*_\ell, \delta)$ and estimate the marginal likelihood by

$$
\hat{m}^{PE}_\ell(y|X_\ell, X^*_\ell, \delta) = m^N_\ell(y|X_\ell, X^*_\ell) \left[ \frac{1}{T} \sum_{t=1}^{T} \frac{m^N_\ell(y^*(t)|X_\ell^*, \delta)}{m^N_\ell(y^*(t)|X^*_\ell, \delta)} \right].
$$

4. Generate $y^*(t) (t = 1, \ldots, T)$ from $m^N_\ell(y^*|y; X_\ell, X^*_\ell, \delta)$ and estimate the marginal likelihood by

$$
\hat{m}^{PE}_\ell(y|X_\ell, X^*_\ell, \delta) = m^N_0(y|X_0, X^*_0) \left[ \frac{1}{T} \sum_{t=1}^{T} \frac{m^N_\ell(y^*(t)|y^*, X_\ell, X^*_\ell, \delta)}{m^N_\ell(y^*(t)|X_0^*, \delta)} \frac{m^N_0(y^*|y; X_0, X^*_0, \delta)}{m^N_\ell(y^*|y; X_\ell, X^*_\ell, \delta)} \right].
$$


We propose to modify the standard $MC^3$ method by sampling a binary vector $\gamma$ indicating the variables included in the model (see, e.g., George & McCulloch, JASA, 1993), using a Metropolis-within-Gibbs approach as follows.

1. Generate $y^{*(t)} (t = 1, \ldots, T)$ from $m_N^0(y^*|X^*_0, \delta)$ (this is the first Monte-Carlo marginal likelihood scheme) or $m_N^0(y^*|y; X_0, X^*_0, \delta)$ (this is the second scheme).

2. For the current model $m_\ell$, corresponding to the set of variable-inclusion indicators $\gamma_\ell$, 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 remaining covariates the same: $\gamma'_l = \gamma_l$ for all $l \neq j$.
     - (c) Identify $m_\ell'$, corresponding to the vector $\gamma_\ell'$, with elements $\gamma'_{k}, k = 1, \ldots, p$.
     - (d) If $m_\ell'$ is not previously visited, calculate and store its estimated marginal
likelihood \( f(m_ℓ', y) = \hat{m}_ℓ^{PE}(y|X_ℓ, X^*_ℓ, δ) \) given by (5, first scheme) or (6, second scheme).

(e) Set \( m_ℓ = m_ℓ' \) (i.e., accept the proposed model \( m_ℓ' \)) with probability

\[
\alpha = \min \left[ 1, \frac{f(m_ℓ', y)}{f(m_ℓ|y)} \right] = \min \left[ 1, \frac{\hat{m}_ℓ^{PE}(y|X_ℓ', X^*_ℓ', δ)}{\hat{m}_ℓ^{PE}(y|X_ℓ, X^*_ℓ, δ)} \frac{f(m_ℓ')}{f(m_ℓ)} \right]. \tag{9}
\]

(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 \( y^*(t) \) \((t = 1, \ldots, T)\) from \( m_ℓ^N(y^*|y; X_ℓ, X^*_ℓ, δ) \), which now depends on the proposed model, and estimate the marginal likelihood of that model using expressions (7) and (8), respectively.
8 Simulated Example

We illustrate the proposed methods by considering the simulated data-set Nott & Kohn (2005, Biometrika). This data-set consists of $n = 50$ observations and $p = 15$ covariates. The first 10 covariates are generated from a standardized Normal distribution while

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

and the response from

$$Y_i \sim N \left( 4 + 2X_{i1} - X_{i5} + 1.5X_{i7} + X_{i11} + 0.5X_{i13}, \ 2.5^2 \right), \ \text{ for } \ i = 1, \ldots, 50.$$
In order to check the efficiency of the four Monte-Carlo estimates 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 1000 iterations and we calculated the Monte-Carlo standard errors.

Table 1: Monte-Carlo standard errors of the estimates of the logarithm of the marginal likelihoods for models $X_1 + X_5 + X_7 + X_{11}$ and $X_1 + X_7 + X_{11}$ when running the four Monte-Carlo techniques 100 times for 1000 iterations, using the PI methodology

<table>
<thead>
<tr>
<th>Monte-Carlo Scheme</th>
<th>Model</th>
<th>$X_1 + X_5 + X_7 + X_{11}$</th>
<th>$X_1 + X_7 + X_{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.9563</td>
<td>0.8769</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.7061</td>
<td>0.5312</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.0339</td>
<td>0.0281</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.0353</td>
<td>0.0296</td>
</tr>
</tbody>
</table>

The third and fourth Monte-Carlo scheme produces much lower Monte-Carlo standard errors as expected. The third/fourth scheme is the one that will be used, for the Z-PEP/J-PEP methodology, in all the following runs, keeping the number of iterations equal to 1000.
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.

<table>
<thead>
<tr>
<th>$m_j$</th>
<th>Predictors</th>
<th>Z-PEP</th>
<th></th>
<th>J-PEP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posterior Model Probability</td>
<td>Bayes Factor</td>
<td>Rank</td>
<td>Posterior Model Probability</td>
<td>Bayes Factor</td>
</tr>
<tr>
<td>1</td>
<td>$X_1 + X_5 + X_7 + X_{11}$</td>
<td>0.0783</td>
<td>1.00</td>
<td>(2)</td>
<td>0.0952</td>
</tr>
<tr>
<td>2</td>
<td>$X_1 + X_7 + X_{11}$</td>
<td>0.0636</td>
<td>1.23</td>
<td>(1)</td>
<td>0.1054</td>
</tr>
<tr>
<td>3</td>
<td>$X_1 + X_5 + X_6 + X_7 + X_{11}$</td>
<td>0.0595</td>
<td>1.32</td>
<td>(3)</td>
<td>0.0505</td>
</tr>
<tr>
<td>4</td>
<td>$X_1 + X_6 + X_7 + X_{11}$</td>
<td>0.0242</td>
<td>3.23</td>
<td>(4)</td>
<td>0.0308</td>
</tr>
<tr>
<td>5</td>
<td>$X_1 + X_7 + X_{10} + X_{11}$</td>
<td>0.0175</td>
<td>4.46</td>
<td>(5)</td>
<td>0.0227</td>
</tr>
<tr>
<td>6</td>
<td>$X_1 + X_5 + X_7 + X_{10} + X_{11}$</td>
<td>0.0170</td>
<td>4.60</td>
<td>(9)</td>
<td>0.0146</td>
</tr>
<tr>
<td>7</td>
<td>$X_1 + X_5 + X_7 + X_{11} + X_{13}$</td>
<td>0.0163</td>
<td>4.78</td>
<td>(10)</td>
<td>0.0139</td>
</tr>
</tbody>
</table>
Sensitivity analysis on $n^*$ for Z-PEP

Figure 1: Posterior marginal inclusion probabilities for different $n^*$ for the Z-PEP methodology
Sensitivity analysis on $n^*$ for Z-PEP (cont.)

Figure 2: Posterior model probabilities of the six best models obtained for each $n^*$, for the Z-PEP methodology; bullets indicate different models entered the top six models for at least one value of $n^*$ and lines connect posterior model probabilities for models of the same rank over different values of $n^*$.
Comparisons

- In the following we compare the Bayes factor between the two best models $(X_1 + X_5 + X_7 + X_{11}$ versus $X_1 + X_5 + X_7)$ according to Z-PEP and J-PEP with the corresponding Bayes factors using J-EP (EPP with no power and with Jeffreys as baseline) and IBF.

- For IBF and EPP we randomly select 100 training samples of size $n^* = 6$ (minimal training samples for the estimation of these two models) and $n^* = 17$ (minimal training sample for the estimation of the full model when we consider all $p = 15$ covariates), while for Z-PEP & J-PEP we randomly select 100 training samples of sizes $n^* = 6$, 17 and $n^* = 20 + k \times 5$ for $k = 1, \ldots, 5$. Each marginal likelihood estimate in Z-PEP is obtained with 1000 iterations, using the third Monte-Carlo scheme and in J-PEP and J-EPP with 1000 iterations, using the fourth Monte-Carlo scheme.
Comparisons (cont.)

Log Bayes Factor
9 Real Life Example

We use the ozone data (Breinman & Friedman 1985, JASA) to implement our approaches. The data we used were slightly changed based on some initial preliminary exploratory analysis. As a response we use a standardized version of the logarithm of the ozone concentration variable of the original data set. The standardized versions of nine (9) main effects, 9 quadratic terms, 2 cubic terms, and 36 two-way interactions (a total of 56 variables) where used as possible covariates. The main effects are:

\[
\begin{array}{l|l}
X_1 & \text{Day of Year} \\
X_2 & \text{Wind speed (mph) at LAX} \\
X_3 & 500 \text{ mb pressure height (m) at VAFB} \\
X_4 & \text{Humidity (\%) at LAX} \\
X_5 & \text{Temperature (°F) at Sandburg} \\
X_6 & \text{Inversion base height (feet) at LAX} \\
X_7 & \text{Pressure gradient (mm Hg) from LAX to Daggett} \\
X_8 & \text{Inversion base temperature (°F) at LAX} \\
X_9 & \text{Visibility (miles) at LAX} \\
\end{array}
\]
(1) First we used \( MC^3 \) to identify variables with high posterior marginal inclusion probabilities, and we created a reduced model space consisting only of those variables whose marginal probabilities were above 0.3.

(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).

- We ran \( MC^3 \) for 100,000 iterations for both the Z-PEP and the EIBF methods. For EIBF we used 30 randomly-selected minimal training samples \((n^* = 58)\). 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.

- We then ran \( MC^3 \) for 220,000 iterations for the Z-PEP, J-PEP and the EIBF approaches. For EIBF we used 30 randomly-selected minimal training samples \((n^* = 24)\).
Table 3: Posterior odds ($PO_{1k}$) of the three best models within each analysis versus the current model $k$ for the reduced model space of 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

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Number of Covariates</th>
<th>Posterior Odds $PO_{1k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>J-PEP</td>
<td>Z-PEP</td>
<td>EIBF</td>
</tr>
<tr>
<td>1</td>
<td>(&gt;3)</td>
<td>(&gt;3)</td>
</tr>
<tr>
<td>2</td>
<td>(1)</td>
<td>(&gt;3)</td>
</tr>
<tr>
<td>3</td>
<td>(&gt;3)</td>
<td>(&gt;3)</td>
</tr>
</tbody>
</table>

### Z-PEP

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Number of Covariates</th>
<th>Posterior Odds $PO_{1k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-PEP</td>
<td>J-PEP</td>
<td>EIBF</td>
</tr>
<tr>
<td>1</td>
<td>(2)</td>
<td>(&gt;3)</td>
</tr>
<tr>
<td>2</td>
<td>(&gt;3)</td>
<td>(&gt;3)</td>
</tr>
<tr>
<td>3</td>
<td>(&gt;3)</td>
<td>(3)</td>
</tr>
</tbody>
</table>

### EIBF

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Number of Covariates</th>
<th>Posterior Odds $PO_{1k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIBF</td>
<td>J-PEP</td>
<td>Z-PEP</td>
</tr>
<tr>
<td>1</td>
<td>(&gt;3)</td>
<td>(&gt;3)</td>
</tr>
<tr>
<td>2</td>
<td>(&gt;3)</td>
<td>(&gt;3)</td>
</tr>
<tr>
<td>3</td>
<td>(&gt;3)</td>
<td>(3)</td>
</tr>
</tbody>
</table>
We evaluate and compare the out-of-sample predictive performance of the three highest a-posteriori model indicated by the above illustrated methods. To do so, we divide the data in half by considering 50 randomly selected splits. For each split, we generate an MCMC sample of $T$ iterations from the model of interest $m_{\ell}$ and then calculate the average root mean square error by

$$ARMSE_{\ell} = \frac{1}{T} \sum_{t=1}^{T} RMSE_{\ell}^{(t)}$$

with

$$RMSE_{\ell}^{(t)} = \sqrt{\frac{1}{n_{V}} \sum_{i \in V} (y_i - \hat{y}_{i|m_{\ell}}^{(t)})^2}$$

being the root mean square error for the validation dataset $V$ of size $n_{V}$ calculated for the $t$-iteration of the MCMC, where $\hat{y}_{i|m_{\ell}}^{(t)} = X_{\ell(i)} \beta_{\ell}^{(t)}$ are the expected values of $y_i$ according to the assumed model for iteration $t$, $\beta_{\ell}^{(t)}$ are the model parameters for iteration $t$ and $X_{\ell(i)}$ is the $i$-th row of matrix $X_{\ell}$ of model $m_{\ell}$. Detailed results for the two highest a-posteriori models are provided on the next Table; results for the full model were also added for comparison reasons. For comparison purposes, we have also included the split-half $RMSE$ measures for these three models using predictions based on direct fitting of the model with the independence Jeffreys prior.
Comparison of the predictive performance (cont.)

Table 4: Comparison of the predictive performance of the PEP and J-EP methods using the full and the MAP models in the reduced model space of the ozone data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>$d_\ell$</th>
<th>$R^2$</th>
<th>$R^2_{adj}$</th>
<th>J-PEP</th>
<th>Z-PEP</th>
<th>J-EP</th>
<th>Jeffreys Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>22</td>
<td>0.8500</td>
<td>0.8392</td>
<td>0.5988 (0.0087)</td>
<td>0.5935 (0.0097)</td>
<td>0.6194 (0.0169)</td>
<td>0.5972 (0.0104)</td>
</tr>
<tr>
<td>J-PEP MAP</td>
<td>9</td>
<td>0.8070</td>
<td>0.8016</td>
<td>0.5975 (0.0063)</td>
<td>0.6161 (0.0051)</td>
<td>0.7524 (0.0626)</td>
<td>0.6165 (0.0052)</td>
</tr>
<tr>
<td>Z-PEP MAP</td>
<td>13</td>
<td>0.8370</td>
<td>0.8303</td>
<td>0.5994 (0.0071)</td>
<td>0.5999 (0.0060)</td>
<td>0.6982 (0.0734)</td>
<td>0.5994 (0.0049)</td>
</tr>
<tr>
<td>EIBF MAP</td>
<td>14</td>
<td>0.8398</td>
<td>0.8326</td>
<td>0.6182 (0.0066)</td>
<td>0.5961 (0.0072)</td>
<td>0.6726 (0.0800)</td>
<td>0.5958 (0.0061)</td>
</tr>
</tbody>
</table>

Comparison with the full model (percentage changes)

<table>
<thead>
<tr>
<th>Model</th>
<th>$d_\ell$</th>
<th>$R^2$</th>
<th>$R^2_{adj}$</th>
<th>J-PEP</th>
<th>Z-PEP</th>
<th>J-EP</th>
<th>Jeffreys Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>J-PEP MAP</td>
<td>-59%</td>
<td>-5.06%</td>
<td>-4.48%</td>
<td>-0.22%</td>
<td>+3.81%</td>
<td>+21.5%</td>
<td>+3.23%</td>
</tr>
<tr>
<td>Z-PEP MAP</td>
<td>-41%</td>
<td>-1.50%</td>
<td>-1.06%</td>
<td>+0.10%</td>
<td>+1.01%</td>
<td>+12.7%</td>
<td>+0.37%</td>
</tr>
<tr>
<td>EIBF MAP</td>
<td>-36%</td>
<td>-1.20%</td>
<td>-0.78%</td>
<td>+3.24%</td>
<td>+0.44%</td>
<td>+10.9%</td>
<td>-0.23%</td>
</tr>
</tbody>
</table>

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

(b) MAP Model using the Z–PEP prior
(c) MAP Model using the J–EP prior

(d) MAP Model using the J–PEP prior
Thank You Irvine!