(Almost) All About PEP

O'Bayes 2019, Warwick, UK

Dimitris Fouskakis

fouskakis@math.ntua.gr

National Technical University of Athens

02 July, 2019

Main References

- Fouskakis, D. and Ntzoufras, I. (2013). Computation for intrinsic variable selection in normal regression models via expected-posterior prior. *Statistics and Computing*, **23**, 491-499.
- Fouskakis, D., Ntzoufras, I. and Draper, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis*, **10**, 75-107.
- Fouskakis, D. and Ntzoufras, I. (2016). Limiting behavior of the Jeffreys power-expected-posterior Bayes factor in Gaussian linear models. *Brazilian Journal of Probability and Statistics*, **30**, 299-320.
- Fouskakis, D. and Ntzoufras, I. (2016). Power-conditional-expected priors. Using g-priors with random imaginary data for variable selection. *Journal of Computational and Graphical Statistics*, **25**, 647-664.
- Fouskakis, D. and Ntzoufras I. (2017). Information consistency of the Jeffreys power-expected-posterior prior in Gaussian linear models. *Metron*, **75**, 371-380.
- Fouskakis, D., Ntzoufras I. and Perrakis K. (2018). Power-expected-posterior priors in generalized linear models. *Bayesian Analysis*, **13**, 721-748.
- Consonni, G., Fouskakis, D., Liseo, B. and Ntzoufras, I. (2018). Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis*, **13**, 627-679.
- Fouskakis, D., Ntzoufras I. and Perrakis K. (2019). Variations of power-expected-posterior priors in normal regressions models (under revision).
- Fouskakis, D. (2019). Priors via Imaginary Training Samples of Sufficient Statistics for Objective Bayesian Hypothesis Testing (submitted).
- Fouskakis, D., Innocent, J.K. and Pericchi, L. (2019). Power-Expected-Posterior Prior Bayes Factor Consistency for Nested Linear Models With Increasing Dimensions (submitted).
- Fouskakis, D. and Ntzoufras I. (2019). Power-Expected-Posterior Priors As Mixtures of G-Priors for Bayesian Model Averaging (submitted).
- Petrakis, N., Peluso, S., Fouskakis, D. and Consonni, G. (2019). Objective Methods for Graphical Structural Learning (submitted).

Outline

1. Objective Bayes model comparison

$$\text{model } M_0 : f(\boldsymbol{y}|\boldsymbol{\theta}_0, M_0), \quad \boldsymbol{\theta}_0 \in \Theta_0 \subseteq \mathbb{R}^{d_0}$$
$$\text{model } M_\ell : f(\boldsymbol{y}|\boldsymbol{\theta}_\ell, M_\ell), \quad \boldsymbol{\theta}_\ell \in \Theta_\ell \subseteq \mathbb{R}^{d_\ell},$$

- If M_0 is nested in M_ℓ, we can assume that $\boldsymbol{\theta}_\ell = (\boldsymbol{\theta}_0^T, \boldsymbol{\theta}_{\ell\backslash 0}^T)^T$, so that $\boldsymbol{\theta}_0$ is a parameter common between the two models, whereas $\boldsymbol{\theta}_{\ell\backslash 0}$ is model specific.
- The use of a common parameter $\boldsymbol{\theta}_0$ in nested model comparison is often made to justify the employment of the same, potentially improper, prior on $\boldsymbol{\theta}_0$ across models.

$$\pi(\boldsymbol{\theta}_0, \boldsymbol{\theta}_{\ell\backslash 0}|M_\ell) = \pi^N(\boldsymbol{\theta}_0|M_\ell)\pi(\boldsymbol{\theta}_{\ell\backslash 0}|\boldsymbol{\theta}_0, M_\ell).$$

Task: Construct an objective prior $\pi(\boldsymbol{\theta}_{\ell\backslash 0}|\boldsymbol{\theta}_0, M_\ell)$.

Example: Variable Selection in Normal Linear Regression Models

Each model M_ℓ is specified by

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

with parameters $\boldsymbol{\theta}_\ell = (\boldsymbol{\beta}_\ell, \sigma^2)$ of size $d_\ell = p_\ell + 2$.

- M_0: null model having the intercept only, with parameters $\boldsymbol{\theta}_0 = (\beta_0, \sigma^2)$,
- M_p: full model with all p covariates under consideration.
- For model M_ℓ we write $\boldsymbol{\beta}_\ell = (\beta_0, \boldsymbol{\beta}_{\ell\setminus 0}^T)^T$ and $\mathrm{X}_\ell = [\mathrm{X}_0, \mathrm{X}_{\ell\setminus 0}]$, where X_0 is the n-dimensional unit vector.

Task: Construct an objective prior for $\boldsymbol{\beta}_{\ell\setminus 0}$.

2. Posterior measures of evidence

Within the Bayesian framework the comparison between models M_0 and M_ℓ is evaluated via the **Posterior Odds** (PO)

$$PO_{M_0,M_\ell} \equiv \frac{\pi(M_0|\boldsymbol{y})}{\pi(M_\ell|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|M_0)}{f(\boldsymbol{y}|M_\ell)} \times \frac{\pi(M_0)}{\pi(M_\ell)} = BF_{M_0,M_\ell} \times O_{M_0,M_\ell}$$

which is a function of the **Bayes Factor** BF_{M_0,M_ℓ} and the **Prior Odds** O_{M_0,M_ℓ}.

In the above $f(\boldsymbol{y}|M)$ is the marginal likelihood under model M and $\pi(M)$ is the prior probability of model M. The marginal likelihood is given by:

$$f(\boldsymbol{y}|M) = \int f(\boldsymbol{y}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta},$$

where $f(\boldsymbol{y}|\boldsymbol{\theta}, M)$ is the likelihood under model M with parameters $\boldsymbol{\theta}$ and $\pi(\boldsymbol{\theta}|M)$ is the prior distribution of model parameters given model M.

Objective Prior : If $\pi(\boldsymbol{\theta}|M)$ is improper $\Rightarrow$ indeterminacy of BF.
If $\pi(\boldsymbol{\theta}|M)$ is proper with large variance $\Rightarrow$ Lindley's paradox.

3. Principles for objective model comparison

3.1. Criteria for objective Bayesian model choice (Bayarri et al. 2012, Annals)

- The *basic criterion (C1)*
- *Model selection consistency (C2)*
- *Information consistency (C3)*
- *Intrinsic consistency criterion (C4)*
- *Predictive matching (C5)*
- *Measurement invariance (C6)*
- The *group invariance criterion (C7)*

Robust Prior

Under the variable selection problem in normal linear models

$$\pi^R(\beta_{\ell\setminus 0}, \beta_0, \sigma | M_\ell) \propto \sigma^{-1} \int_0^{+\infty} N_{p_\ell - p_0}(\beta_{\ell\setminus 0} | \mathbf{0}, g\Sigma_{\ell\setminus 0}) \pi^R(g) dg,$$

where $\Sigma_{\ell\setminus 0} = \sigma^2 \left(\mathrm{V}_{\ell\setminus 0}^T \mathrm{V}_{\ell\setminus 0}\right)^{-1}$, $\mathrm{V}_{\ell\setminus 0} = \left(\mathrm{I}_n - \mathrm{X}_0 \left(\mathrm{X}_0^T \mathrm{X}_0\right)^{-1} \mathrm{X}_0^T\right) \mathrm{X}_{\ell\setminus 0}$, and

$$\pi^R(g) = a_r \left[\rho_{\ell,r}(b_r + n)\right]^{a_r} (g + b_r)^{-(a_r+1)} 1_{\{g > \rho_{\ell,r}(b_r+n) - b_r\}},$$

with $a_r, b_r > 0$ and $\rho_{\ell,r} \geq \frac{b_r}{b_r+n}$.

- While the result holds for a general matrix of common predictors X_0, note that, if $\mathrm{X}_0 = 1$ (i.e. when M_0 contains only the intercept), then $\mathrm{V}_{\ell\setminus 0} = \mathrm{Z}_{\ell\setminus 0}$, with $\mathrm{Z}_{\ell\setminus 0}$ denoting the column-wise centered version of $\boldsymbol{X}_{\ell\setminus 0}$. **Thus the robust prior is a mixture of g-prior.**
- $a_r = 1/2$, $b_r = 1$ and $\rho_{\ell,r}^{-1} = p_\ell + 1$.
- The hyper-g-prior and the hyper-g/n-prior are special cases of the robust prior.

3.2. Compatibility of priors

Informally this means that priors should be related across models (e.g. Consonni and Veronese, 2008). Compatibility is usually applied to nested models. One way to achieve compatibility is the prior under each model to be anchored to a common base measure (more later).

3.3. Validation of Bayesian approaches

An acceptable Bayesian procedure should correspond, at least asymptotically, to a prior which makes sense in the context where it is applied (see Berger and Pericchi, 1996).

3.4. Methods with good frequentist properties

Use prior distributions that lead to good frequentist performances, e.g. select priors based on the coverage of posterior intervals and false discovery rates (FDR) (for example Tansey et. al. 2018).

4. Methods for constructing objective priors

4.1. Unit information principle (Kass & Wasserman, 1995): A unit information prior (UIP) has an information content equivalent to a sample of size one.

4.2. Training samples : Use a **minimal subset of the data**, to convert an improper baseline prior to a proper posterior, and then use the remaining data to calculate the Bayes factor (**Intrinsic Bayes Factor - IBF**). **Intrinsic prior distributions** were originally introduced to provide a proper Bayesian interpretation for IBFs (Berger & Pericchi, 1996).

Alternatively, train the improper prior using a fraction of the full sample likelihood and calculate the marginal likelihood using the complementary fraction of the likelihood together with the newly trained prior (**Fractional Bayes Factor - FBF**); O'Hagan, 1995.

4.3. Imaginary observations (Good, 1950): Consider a thought experiment with an appropriate dataset $\boldsymbol{y}^*$ that will be used to specify the normalizing constants involved in the Bayes factors when using improper priors. In order to make the induced methods minimally informative, the notions of **minimal training sample** and the **UIP principles** are used in several occasions.

4.3.1. Fixed Imaginary Data

- **Power Prior** (Ibrahim & Chen, 2000):
  $\pi(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, a_0, M_\ell) \propto f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, M_\ell)^{a_0}\pi^N(\boldsymbol{\theta}_\ell|M_\ell)$. If n^* is the size of $\boldsymbol{y}^*$, $a_0 = 1/n^*$ makes the prior having a **unit information interpretation**.
- *g*-**Prior** (Zellner, 1986): For the variable selection problem in normal linear models

$$\boldsymbol{\beta}_{\ell\backslash 0}|\beta_0, \sigma^2, M_\ell \sim \mathrm{N}_{p_\ell}\left(\mathbf{0}, g\left(\mathrm{Z}_{\ell\backslash 0}^T\mathrm{Z}_{\ell\backslash 0}\right)^{-1}\sigma^2\right) \text{ and } \pi(\beta_0, \sigma^2|M_\ell) \propto 1/\sigma^2 ,$$

  with $\mathrm{Z}_{\ell\backslash 0}$ denoting the column-wise centered version of $\mathrm{X}_{\ell\backslash 0}$.
  - Usual choice: $g = n \rightarrow$ UIP.
  - Can be considered as a **power prior** with all imaginary data set equal to a pre-specified value.
  - **Information Paradox**.
- **Mixtures** *g*-**Prior** (Liang et al. 2008, JASA):
  - **Hyper** g: $\pi(g) = [(a_h - 2)/2](1+g)^{-a_h/2}$, $g > 0$. Default choice $a_h = 3$.
  - **Hyper** g/n: $\pi(g) = [(a_h - 2)/(2n)](1+g/n)^{-a_h/2}$, $g > 0$. Default choice $a_h = 3$.

  Both can be considered as a **power prior** with fixed imaginary data and a hyper-prior placed on a_0.

4.3.2. Random Imaginary Data

- **Expected-Posterior Priors (EPP)** (Pérez and Berger, 2002, Biometrika)
  The expected posterior prior (EPP) for the parameter under M_ℓ is the expectation of the posterior distribution given imaginary observations $\boldsymbol{y}^*$ of size n^*, where the expectation is taken with respect to a suitable probability measure $m^*(\boldsymbol{y}^*|M_*)$ under **a reference model** M_*.

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

  where $\pi^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, M_\ell) \propto f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, M_\ell)\pi^N(\boldsymbol{\theta}_\ell|M_\ell)$ is the posterior distribution of $\boldsymbol{\theta}_\ell$ under model M_ℓ conditionally on the imaginary data $\boldsymbol{y}^*$ for the given baseline (typically improper) prior $\pi^N(\boldsymbol{\theta}_\ell|M_\ell)$.

Properties

- **Nice Interpretation**
- In nested cases usually $M_* = M_0$ and $m^* = m_0^N$ is the marginal likelihood of M_0, evaluated at $\boldsymbol{y}^*$, under the baseline prior. In this case EPP = **Intrinsic Prior**.
- **Compatibility**.
- **Impropriety**: Impropriety of baseline priors causes no indeterminacy. Impropriety in m^* also does not cause indeterminacy, because m^* is common to the EPPs for all models.
- **Choice of** n^*: Usually we choose the smallest n^* for which the posterior is proper; this is the **minimal training sample size**.
- Main Issue: In variable selection problems specification of X_ℓ^*. Also the resulting prior can be influential when the sample size n is not much larger than the total number of parameters under the full model.

- **Power-Expected-Posterior (PEP) Priors.**
  - In EPP, set $n^* = n$ and therefore $\mathrm{X}_\ell^* = \mathrm{X}_\ell$ in variable selection problems.
  - Introduce an additional **power-parameter** δ and raise the likelihoods into this power, as in the **power prior approach**, in order to control the weight that the imaginary data contribute to the final prior.
  - In **normal models** you can substitute the power-likelihood terms with the density - normalized power - likelihoods$\rightarrow$ still normal with variance inflated by δ.
  - If $\delta = n^* \rightarrow$ UIP. Thus no excessive weight when the number of parameters is close to the number of data.

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

$$\pi_\ell^{PEP}(\boldsymbol{\theta}_\ell|\delta) = \int \underbrace{\pi_\ell^{N}(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*,\delta)}_{\Downarrow} \underbrace{m^*(\boldsymbol{y}^*,\delta)}_{\Downarrow} d\boldsymbol{y}^*$$

$$f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, M_\ell)^{1/\delta}$$

5. PEP - Variable selection in normal linear models

Each model M_ℓ is specified by

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

The baseline prior is $\pi_\ell^N(\boldsymbol{\beta}_\ell, \sigma_\ell^2) \propto \sigma_\ell^{-2}$ and the final PEP is:

$$\pi_\ell^{PEP}(\boldsymbol{\beta}_\ell, \sigma_\ell^2|\mathrm{X}_\ell^*, \delta) = \int f_{N_{d_\ell}}\left[\boldsymbol{\beta}_\ell; \widehat{\boldsymbol{\beta}}_\ell^*, \delta\,(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\sigma_\ell^2\right] \times f_{IG}\left(\sigma_\ell^2; \frac{n^*-d_\ell}{2}, \frac{RSS_\ell^*}{2\delta}\right) m_0^N(\boldsymbol{y}^*|\mathrm{X}_0^*, \delta)\, d\boldsymbol{y}^*,$$

where

$$m_0^N(\boldsymbol{y}^*\,|\,\mathrm{X}_0^*, \delta) \propto \pi^{\frac{1}{2}(d_0-n^*)}\,|\mathrm{X}_0^{*T}\mathrm{X}_0^*|^{-\frac{1}{2}}\,\Gamma\left(\frac{n^*-d_0}{2}\right) RSS_0^{*\,-\left(\frac{n^*-d_0}{2}\right)}.$$

Further Specifications - Assumptions & Properties

- **Default choices:** $\delta = n^* = n$ and $\mathrm{X}^*_\ell = \mathrm{X}_\ell$. Reference model is the null model.
- **Assumption:** $p < n$.
- Using **simple Monte Carlo techniques**, the marginal likelihood under the PEP prior can be estimated quite accurately.
- When comparing the full model M_p to a reduced model M_ℓ the resulting **Bayes factor** takes the simple form:

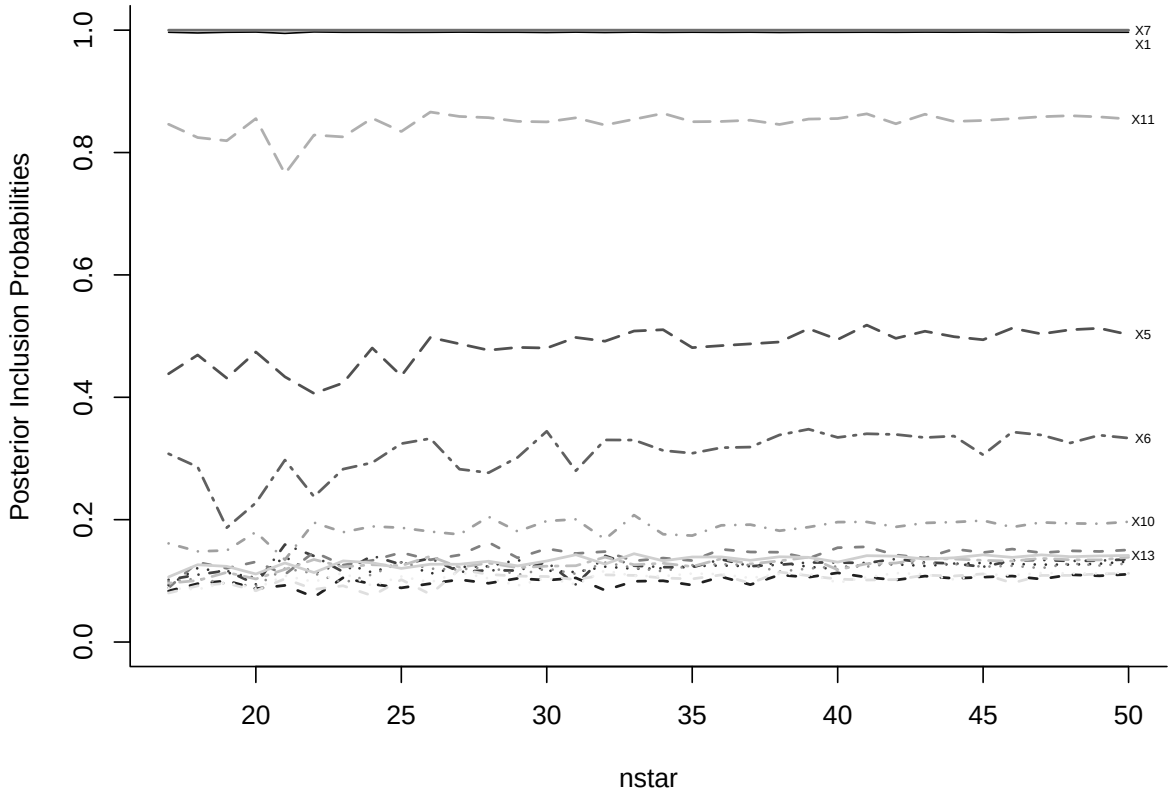
$$BF^{PEP}_{p\ell} = 2\frac{\Gamma(n-p)}{\Gamma^2\left(\frac{n-p}{2}\right)} \int_0^{\frac{\pi}{2}} \frac{(sin\varphi)^{n-d_\ell-1}(cos\varphi)^{n-p-1}\left(\delta + sin^2\varphi\right)^{\frac{n-p}{2}}}{\left(\delta\frac{RSS_p}{RSS_\ell} + sin^2\varphi\right)^{\frac{n-d_\ell}{2}}} d\varphi.$$

Properties (more)

- Desiderata? Yes! More later!

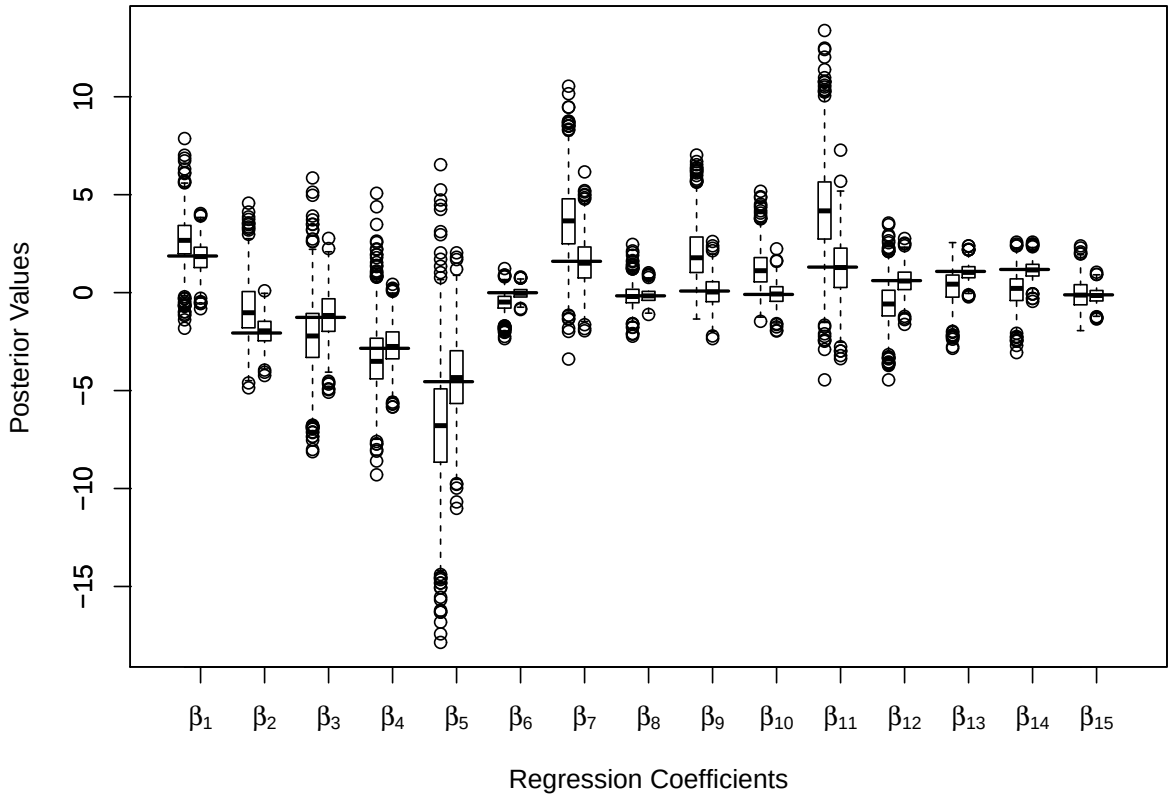
Sensitivity analysis on imaginary sample size

Figure 1: *Posterior marginal inclusion probabilities, for n^* values from 17 to $n = 50$, with the PEP prior methodology (simulated example for a variable selection problem in normal linear model).*



Sensitivity analysis on imaginary sample size (cont.)

Figure 2: *Boxplots of the posterior distributions of the regression coefficients. For each coefficient, the left-hand boxplot summarizes the EPP results and the right-hand boxplot displays the PEP posteriors; solid lines in both posteriors identify the MLEs. We used the first 20 observations from the simulated data-set and a randomly selected training sample of size* $n^* = 17$.



6. PCEP - Variable selection in normal linear models

Each model M_ℓ is now specified by

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

Since σ^2 appears in all models under comparison, we can assume a common prior distribution $\pi_\ell^N(\sigma^2) \propto \sigma^{-2}$ for all models $M_\ell \in \mathcal{M}$. By this way, we define the power-conditional-expected-posterior (PCEP) prior by

$$\pi_\ell^{PCEP}(\boldsymbol{\beta}_\ell, \sigma^2|\mathrm{X}_\ell^*, \delta) = \pi_\ell^{PCEP}(\boldsymbol{\beta}_\ell|\sigma^2, \mathrm{X}_\ell^*, \delta)\pi_\ell^N(\sigma^2),$$

with baseline prior

$$\pi_\ell^N(\boldsymbol{\beta}_\ell|\sigma^2) = f_{N_{d_\ell}}\left(\boldsymbol{\beta}_\ell; \mathbf{0}, g_0\left(\boldsymbol{X}_\ell^{*T}\boldsymbol{X}_\ell^*\right)^{-1}\sigma^2\right)$$

and $g_0 = \delta n^*$.

Further Specifications - Assumptions & Properties

- **Default choices:** $\delta = n^* = n$ and $\mathrm{X}_\ell^* = \mathrm{X}_\ell$. Reference model is the null model.
- **Assumption:** $p < n$
- **Closed Form Expression**:
  $\pi_\ell^{PCEP}(\beta_\ell|\sigma^2, \mathrm{X}_\ell^*, \delta) =$
  $f_{N_{d_\ell}}\left(\beta_\ell; \mathbf{0}, \delta\left\{\boldsymbol{X}_\ell^{*T}\left[w^{-1}\mathrm{I}_{n^*} - (\delta\Lambda_0^* + w\mathrm{H}_\ell^*)^{-1}\right]\boldsymbol{X}_\ell^*\right\}^{-1}\sigma^2\right)$, where
  $w = g_0/(g_0+\delta)$, $\Lambda_0^* = \left[\delta\mathrm{I}_{n^*} + g_0 n^{*-1}\mathbf{1}_{n^*}\mathbf{1}_{n^*}^T\right]^{-1}$ and
  $\mathrm{H}_\ell^* = \mathrm{X}_\ell^*(\mathrm{X}_\ell^{*T}\mathrm{X}_\ell^*)^{-1}\mathrm{X}_\ell^{*T}$.
- PCEP: A more dispersed $g = n$-prior with random imaginary data.
- **Basic criterion (*C1*)**
- **Consistent model selection procedure (*C2*)**
- **Information inconsistency (*C3*)**
- **Predictive matching (*C5*)**

7. PCEP (PEP) - Variable selection in GLM

- Response distribution: member of the exponential family (normal regression, binomial logistic regression, Poisson log-linear models).
- Linear predictor of the form $\mathrm{X}_\ell \boldsymbol{\beta}_\ell$.
- ϕ dispersion parameter (known in Logistic & Poisson regression).

Problem: Find the normalized power likelihood. For instance, for the binomial and Poisson regression models, the normalized power likelihoods are composed by products of discrete distributions that have no standard form.

Here we extend the definition of PCEP (PEP):

$$\pi_\ell^{\mathrm{PEP}}(\boldsymbol{\beta}_\ell, \phi|\boldsymbol{\delta}) = \pi_\ell^{\mathrm{PEP}}(\boldsymbol{\beta}_\ell|\phi, \boldsymbol{\delta})\pi_\ell^{\mathrm{N}}(\phi),$$

where $\boldsymbol{\delta} = (\delta_0, \delta_1)$ and

$$\begin{aligned}
\pi_\ell^{\mathrm{PEP}}(\boldsymbol{\beta}_\ell|\phi, \boldsymbol{\delta}) &= \int \pi_\ell^{N}(\boldsymbol{\beta}_\ell|\boldsymbol{y}^*, \phi, \delta_1) m_0^{\mathrm{N}}(\boldsymbol{y}^*|\phi, \delta_0)\mathrm{d}\boldsymbol{y}^*, \\
\pi_\ell^{N}(\boldsymbol{\beta}_\ell|\boldsymbol{y}^*, \phi, \delta_1) &= \frac{f_\ell(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \phi, \delta_1)\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell|\phi)}{m_\ell^{\mathrm{N}}(\boldsymbol{y}^*|\phi, \delta_1)}, \\
m_\ell^{\mathrm{N}}(\boldsymbol{y}^*|\phi, \delta_1) &= \int f_\ell(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \phi, \delta_1)\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell|\phi)\mathrm{d}\boldsymbol{\beta}_\ell, \\
m_0^{\mathrm{N}}(\boldsymbol{y}^*|\phi, \delta_0) &= \frac{\int f_0(\boldsymbol{y}^*|\beta_0, \phi, \delta_0)\pi_0^{\mathrm{N}}(\beta_0|\phi)\mathrm{d}\beta_0}{C_0}, \; C_0: \text{ normalizing constant} \\
f_\ell(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \phi, \delta_1) &= \frac{f_\ell(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \phi)^{1/\delta_1}}{k_\ell(\boldsymbol{\beta}_\ell, \phi, \delta_1)}, \quad f_0(\boldsymbol{y}^*|\beta_0, \phi, \delta_0) = \frac{f_\ell(\boldsymbol{y}^*|\beta_0, \phi)^{1/\delta_0}}{k_0(\beta_0, \phi, \delta_0)}.
\end{aligned}$$

In the original PCEP prior for normal regression models:

$$k_\ell \equiv k_\ell(\boldsymbol{\beta}_\ell, \phi, \delta_1) = \int f_\ell(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell, \phi, \delta_1)^{1/\delta_1} d\boldsymbol{y}^* \text{ and } \delta_1 = \delta$$

and

$$k_0 \equiv k_0(\beta_0, \phi, \delta_0) = \int f_0(\boldsymbol{y}^*|\beta_0, \phi, \delta_0)^{1/\delta_0} d\boldsymbol{y}^* \text{ and } \delta_0 = \delta$$

and C_0=1.

New variations of PEP depending on the selection of k_ℓ and k_0 and δ_1, δ_0:

- **DR-PEP**: $k_\ell = k_0 = 1$ (unnormalized likelihoods), $\delta_1 = \delta_0 = \delta$ and $C_0 = \int \int f_0(\boldsymbol{y}^*|\beta_0, \phi, \delta_0)\pi_0^{\mathrm{N}}(\beta_0|\phi)\mathrm{d}\beta_0\mathrm{d}\boldsymbol{y}^*$
- **CR-PEP**: δ_1 and $k_\ell = 1$ (unnormalized likelihood), $\delta_0 = 1 \Rightarrow k_0 = 1$ (original likelihood) and $C_0 = 1$.

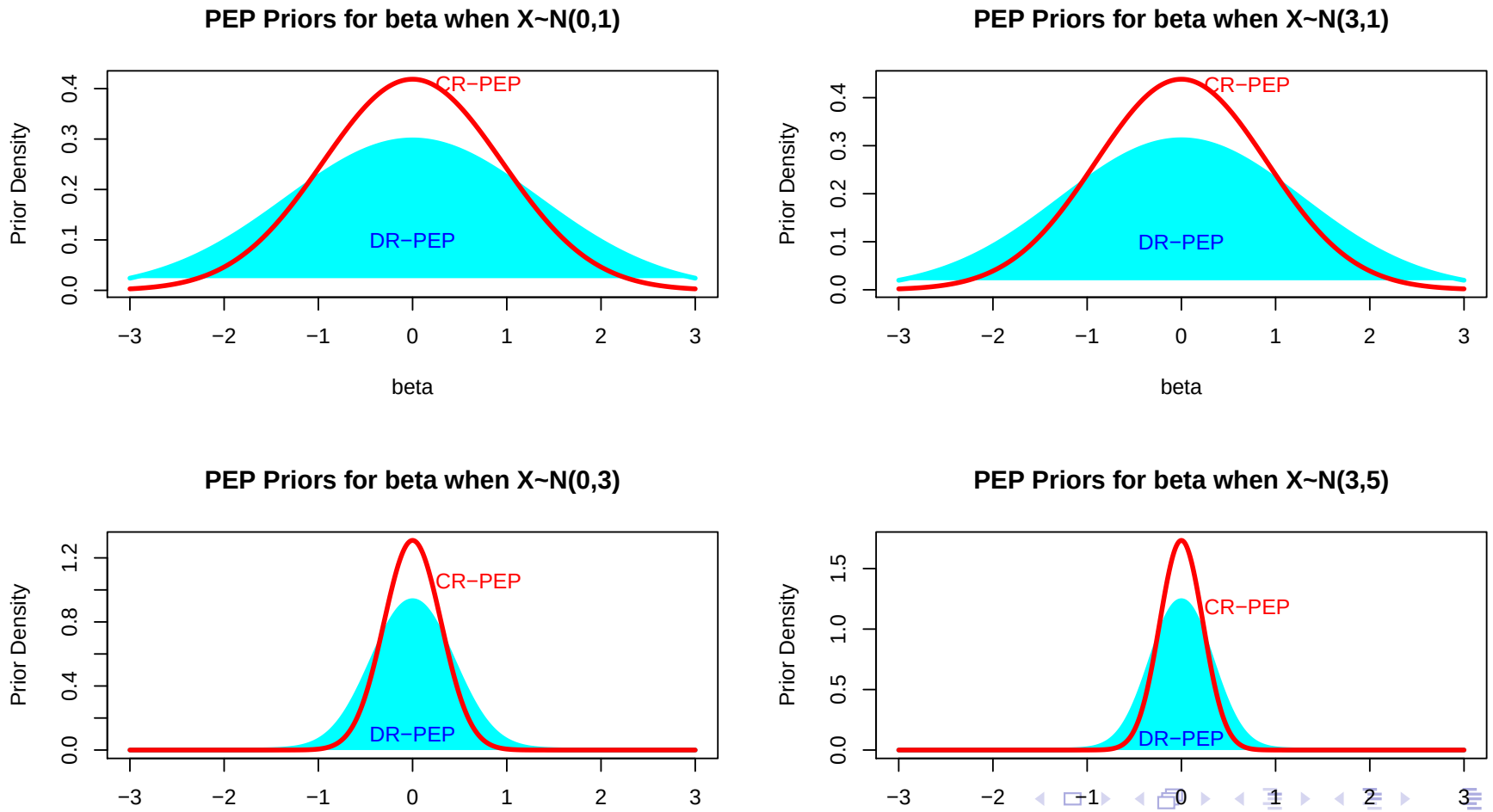
In normal regression models DR-PEP = PCEP.

Further Specifications - Assumptions

- **Default choices:** $n^* = n$ and thus $\mathrm{X}_\ell^* = \mathrm{X}_\ell$. Another advantage of setting $n^* = n$, which becomes more obvious in the GLM framework, is that one can now utilize large-sample approximations when needed for large n, for instance, for the baseline posterior in the PEP definition.
- Reference model is the null model.
- **Baseline prior:** Jeffreys prior for GLMs (Ibrahim and Laud, 1991, JASA) is used.
- **power parameters**: Under the DR-PEP $\delta_1 = \delta_0 = \delta = n^*$. Under the CR-PEP $\delta_1 = n^*$ and $\delta_0 = 1$.
- **Assumption:** $p < n$.

Simple Normal Linear Regression

Figure 3: *The marginal DR-PEP and CR-PEP priors, conditional on* $\sigma = 1$, *using a simple normal linear regression model.*



Hyper-δ Extensions

- Under the DR-PEP prior we place a hyper-prior for $\delta_1 = \delta_0 = \delta$.
- Under the CR-PEP prior we place a hyper-prior for $\delta_1 = \delta$, while $\delta_0 = 1$.
- Hyper-δ-prior: $\pi(\delta) = [(a-2)/2](1+\delta)^{-a/2}$, $\delta > 0$. Default choice $a = 3$.
- Hyper-δ/n-prior: $\pi(\delta) = [(a-2)/(2n)](1+\delta/n)^{-a/2}$, $\delta > 0$. Default choice $a = 3$.

Computation & Properties

- Under all six approaches (CR-PEP, DR-PEP, CR-PEP hyper-δ, CR-PEP hyper-δ/n, DR-PEP hyper-δ and DR-PEP hyper-δ/n), a Gibbs variable selection method has been used.
- Under all six approaches we have proved **predictive matching (*C5*).**
- Under all six approaches we have showed empirically **model selection consistency (*C2*).**
- For models with discrete responses and known dispersion parameter (such as Poisson and binomial models) **information inconsistency is not an issue** since the likelihood is bounded even for saturated models (***C3***); see for example Li & Clyde (2018).

Illustration (Poisson & Binomial Models)

- $n = 100$, $p = 5$ and $p = 3$ predictors for logistic and Poisson scenarios respectively.
- Each simulation is repeated 100 times.
- Each predictor is drawn from a standard normal distribution with pairwise correlation given by

$$\text{corr}(X_i, X_j) = r^{|i-j|}, \quad 1 \leq i < j \leq p.$$

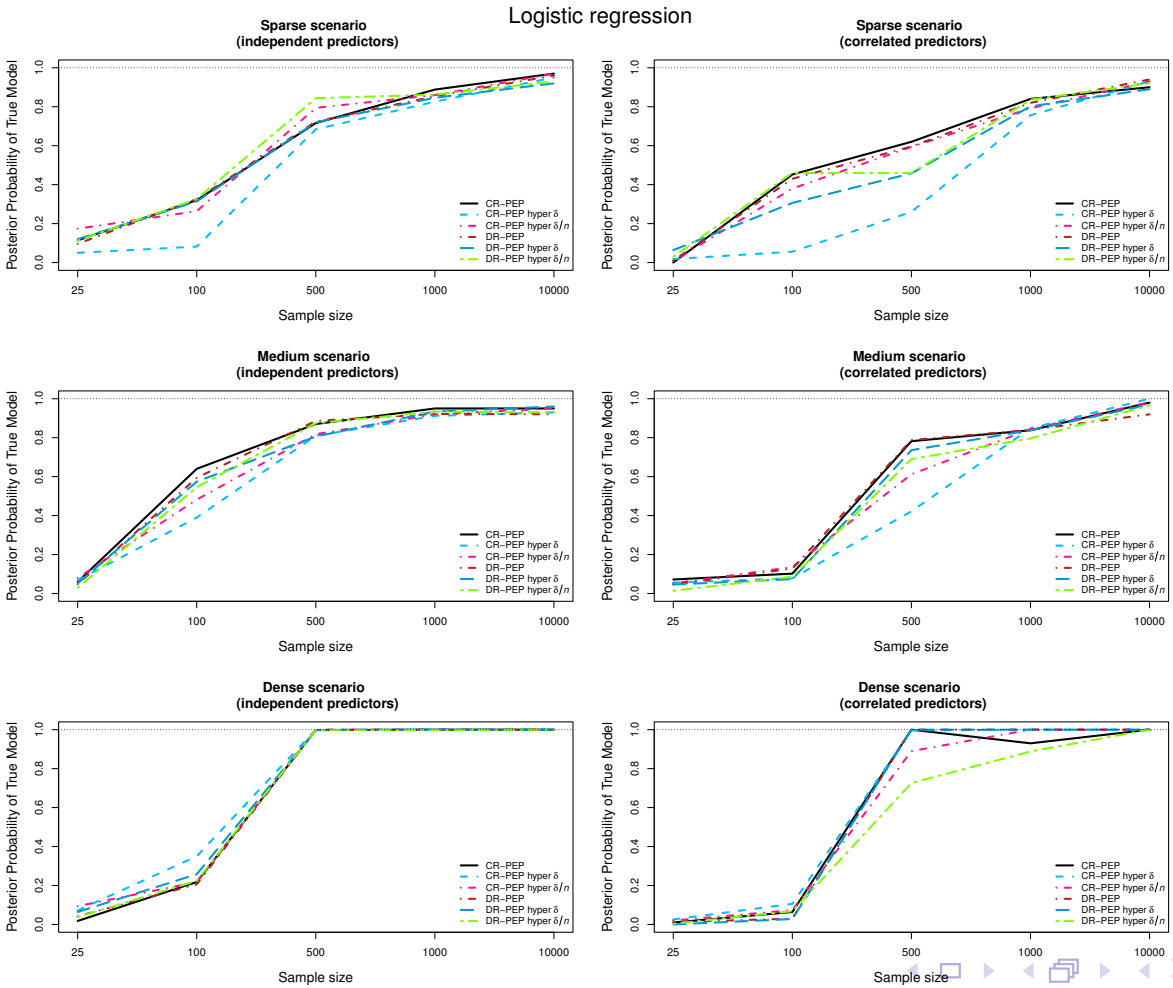
  with (i) independent predictors ($r = 0$) and (ii) correlated predictors ($r = 0.75$).
- $n \in \{25, 100, 500, 1000, 10000\}$.

Scenario	**Binomial Logistic** ($n = 100$)						**Poisson** ($n = 100$)			
	β_0	β_1	β_2	β_3	β_4	β_5	β_0	β_1	β_2	β_3
null	0.1	0	0	0	0	0	-0.3	0	0	0
sparse	0.1	0.7	0	0	0	0	-0.3	0.3	0	0
medium	0.1	1.6	0.8	-1.5	0	0	-0.3	0.3	0.2	0
full	0.1	1.75	1.5	-1.1	-1.4	0.5	-0.3	0.3	0.2	-0.15

Table 1: Simulation Binomial and Poisson regression scenarios.

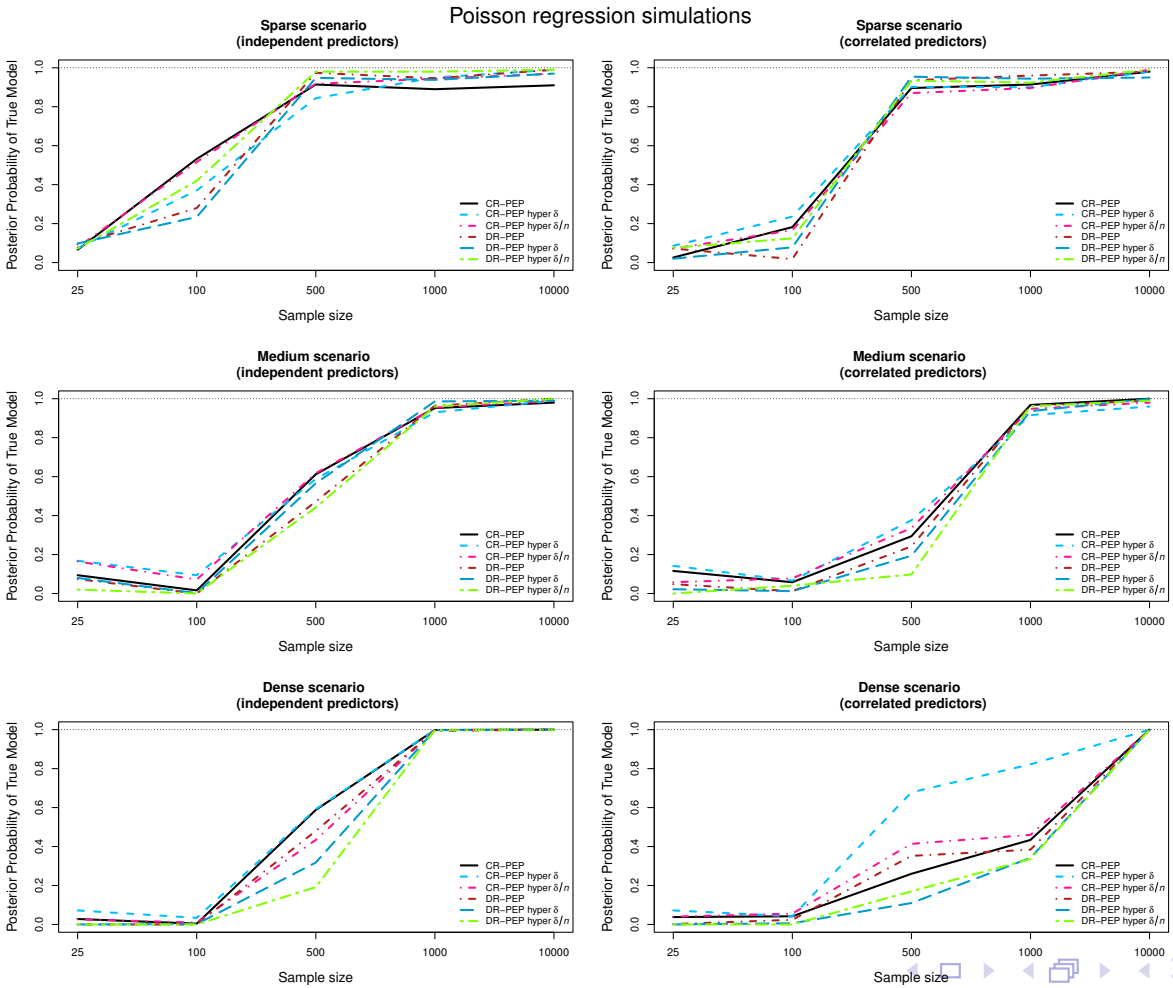
Binomial logistic regression results

Figure 4: *Posterior probabilities of the true model vs. sample size for the sparse, medium and dense logistic regression scenarios.*



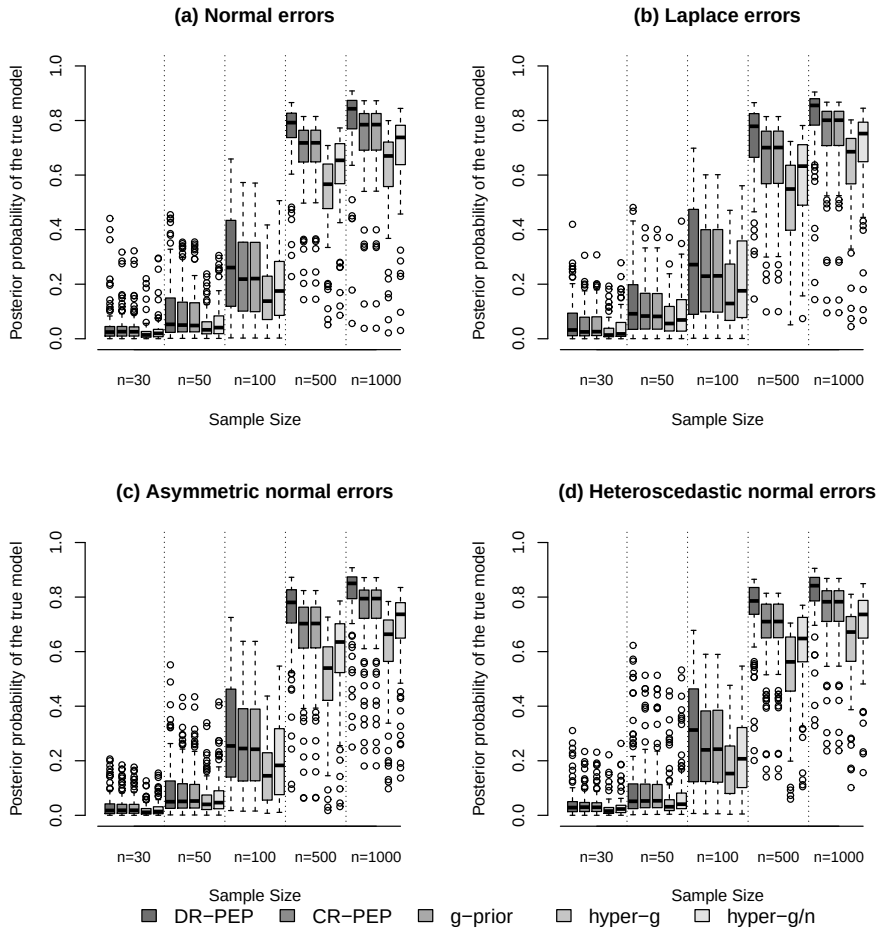
Poisson regression results

Figure 5: *Posterior probabilities of the true model vs. sample size for the sparse, medium and dense poisson regression scenarios.*



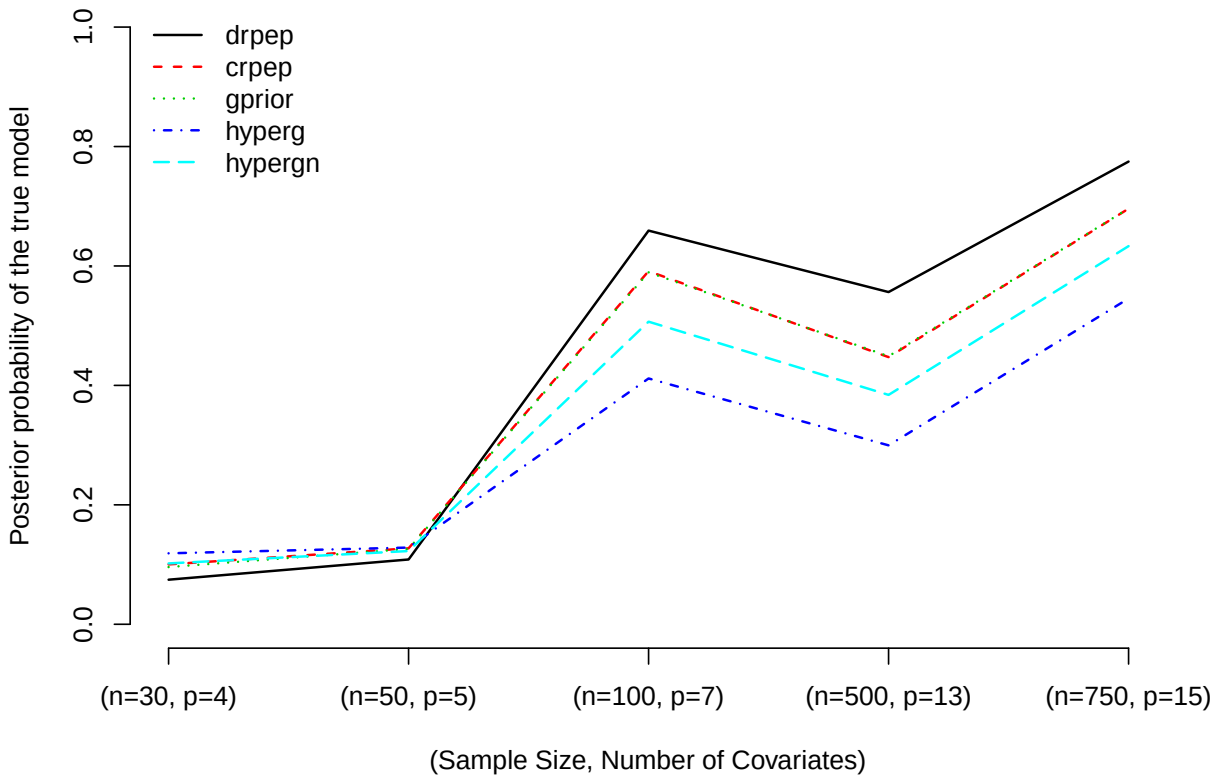
Normal Linear Regression: M-open Cases

Figure 6: *Posterior probabilities of the true model vs. sample size under the four model cases.*



Normal Linear Regression: Growing *n* and *p*

Figure 7: *Posterior probabilities of the true model vs. sample size and model dimensionality under independent covariates.*



8. PEP using sufficient statistics

- EPP and PEP prior can be expressed as an average over all possible sets of **sufficient statistics based on imaginary data coming from the baseline (simplest) model** instead of all possible sets of imaginary data.
- The above might result to a great reduction of the problem dimensionality. This can be beneficial especially in PEP where the dimension is n and also in cases where the prior and the posterior are not available in closed form expression.
- Calculations can be much easier.
- Assumptions:
  - $n > d_\ell$ for all ℓ.
  - $M_0 \subset M_\ell$ for all ℓ. Hence **the minimal sufficient statistic of each model under consideration is always a sufficient statistic of M_0.**

Advantages

- In PEP $n = n^*$. Therefore when we have large n we can use asymptotic properties of the MLE.
- We could use the fact that, under mild conditions and for some distributions, the Maximum Likelihood Estimator (MLE) is a sufficient statistic. Moreover, under some regularity conditions (and perhaps under an appropriate reparametrisation), the MLE is asymptotically normal. Thus, one could think of using the MLE as a sufficient statistic together with a normal approximation to its distribution, particularly for the large sample scenario.
- For instance, in contingency tables, we can work with the MLE of the logarithm of the odds ratio, which follows, asymptotically, the normal distribution (work in progress).

9. PEP - Consistency for nested linear models with increasing dimensions

Here we we examined the asymptotic behaviour of the PEP methodology when comparing **nested normal linear models**. Emphasis is given on the consistency of the Bayes factor of the full model M_p versus a generic submodel M_ℓ.

Table 2: Consistency of $BF_{p\ell}^{PEP}$ when model M_ℓ has dimension $dim(M_\ell) = i = O(1)$ and $\delta \in \{n, n-p\}$

	M_ℓ is correct	M_p is correct
$p = O(1)$	Consistent	Consistent
$p = O(n)$	Consistent	Not always Consistent

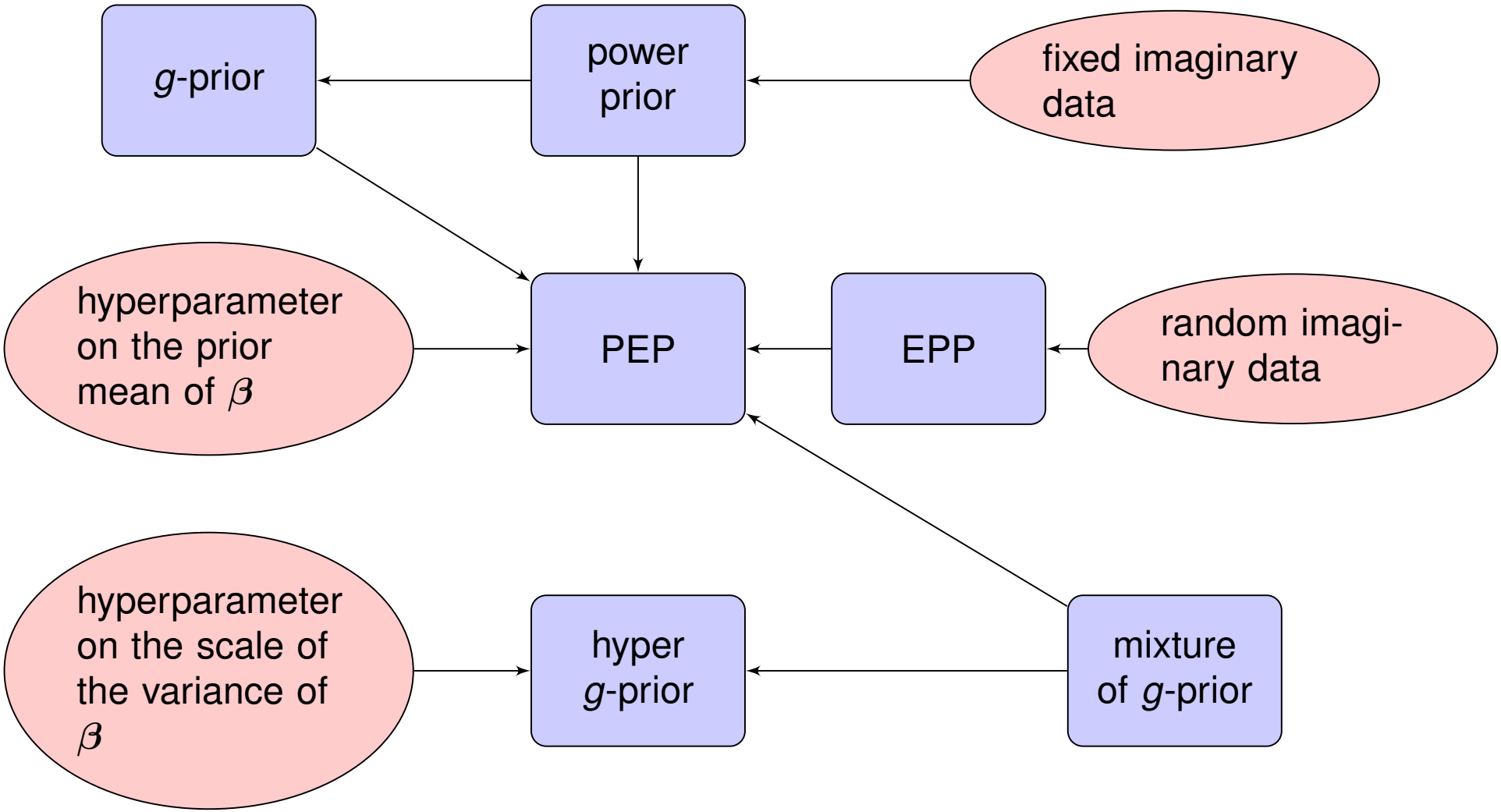
Table 3: Consistency of $BF_{p\ell}^{PEP}$ when model M_ℓ has dimension $dim(M_\ell) = i = O(1)$ and $\delta_\ell = p$

	M_ℓ is correct	M_p is correct
$p = O(1)$	Inconsistent	Consistent
$p = O(n)$	Consistent	Not always Consistent

Table 4: Consistency of $BF_{p\ell}^{PEP}$ when model M_ℓ has dimension $dim(M_\ell) = i = O(1)$ and $\delta_\ell = \delta > 0$

	M_ℓ is correct	M_p is correct
$p = O(1)$	Consistent if δ large	Consistent
$p = O(n)$	Not always Consistent	Not always Consistent

10. PEP priors as mixtures of g-priors for Bayesian model averaging



Model formulation

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a random sample. We would like to compare the nested models:

$$H_0 : \text{model } M_0 : \text{Normal}(\boldsymbol{y}|\mathrm{X}_0\boldsymbol{\beta}_0, \sigma_0^2), \quad \pi_0^N(\boldsymbol{\beta}_0, \sigma_0) \propto \sigma_0^{-(1+d_0)}$$

vs.

$$H_1 : \text{model } M_1 : \text{Normal}(\boldsymbol{y}|\mathrm{X}_1\boldsymbol{\beta}_1, \sigma_1^2), \quad \pi_1^N(\boldsymbol{\beta}_1, \sigma_1) \propto \sigma_1^{-(1+d_1)}$$

- X_0 is an $(n \times k_0)$ design matrix under model M_0
- X_1 is an $(n \times k_1)$ design matrix under model M_1
- $k_0 < k_1$ and M_0 is nested in M_1
- $\boldsymbol{\beta}_1 = \left(\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_e^T\right)^T$, $\mathrm{X}_1 = [\mathrm{X}_0|\mathrm{X}_e]$
- $\mathrm{P}_0 = \mathrm{X}_0 \left(\mathrm{X}_0{}^T\mathrm{X}_0\right)^{-1} \mathrm{X}_0{}^T$
- $\mathrm{P}_1 = \mathrm{X}_1 \left(\mathrm{X}_1{}^T\mathrm{X}_1\right)^{-1} \mathrm{X}_1{}^T$

Usual choices for d_0 and d_1 are $d_0 = d_1 = 0$ (resulting to the reference prior) or $d_0 = k_0$ and $d_1 = k_1$ (resulting to dependence Jeffreys prior).

PEP prior

$$\beta_e|t,\sigma_1,\beta_0 \sim N_{k_1-k_0}\left(\mathbf{0},\frac{\delta\sigma_1^2}{t}\mathrm{V}\right), t|\sigma_1 \sim \text{Beta}\left(\frac{n^*+d_0-k_1}{2},\frac{n^*+d_1-d_0-k_1}{2}\right),$$

$$(\beta_0,\sigma_1) \sim \pi_1^{PEP}(\beta_0,\sigma_1) \propto \sigma_1^{-(d_0+1)},$$

where $\mathrm{V}^{-1} = \mathrm{X}_e^{*T}(\mathrm{I}_{n^*} - \mathrm{P}_0^*)\mathrm{X}_e^*$.

- The EPP is directly available for $\delta = 1$.
- Under the usual case where the reference model M_0 is the null model, we have that $\mathrm{V} = (\mathrm{Z}_e^{*T}\mathrm{Z}_e^*)^{-1}$; where Z_e^* is the matrix of the centred (at the mean) imaginary covariates.
- In practice, when using the PEP prior with centred covariates and imaginary design matrices equal to actual ones ($n^* = n$), then the induced approach results in a **mixture of g-priors** with a different hyper-prior on $g = \delta/t$.

Comparison with other scale normal mixtures priors

A wide range of prior distributions for variable selection in normal linear models can be written as a normal scale mixture distribution:

$$f(\boldsymbol{\beta}_e, \boldsymbol{\beta}_{\mathbf{0}}, \sigma_1 | M_1) = \sigma_1^{-(d_0+1)} \int_0^{+\infty} f_{N_{k_1-k_0}}\left(\boldsymbol{\beta}_e;\, \mathbf{0},\, g\sigma_1^2 \Sigma_e\right) h(g|M_1) dg,$$

Under the PEP prior, the hyper-prior for $g = \delta/t$ is given by

$$g \sim SGBP\left(a = \tfrac{n^*+d_0-k_1}{2},\, b = \tfrac{n^*+d_1-d_0-k_1}{2},\, p = 1,\, q = \delta,\, s = \delta\right)$$

where *SGBP* stands for the *shifted generalized beta prime distribution*. In our case, since $q = s = \delta$ the hyper-prior simplifies to

$$\boxed{h(g|M_1) \propto (g-\delta)^{b-1} g^{-a-b},\ g \geq \delta.}$$

Moreover, most of the known priors used for variable selection assume that $\Sigma_e^{-1} = \mathrm{X}_e^T(\mathrm{I}_n - \mathrm{P}_0)\mathrm{X}_e$. This is also the case for the PEP prior if we consider $n^* = n$ and thus $\mathrm{X}_e^* = \mathrm{X}_e$.

Table 5: Mixing distributions of g under different prior setups ($d_1 = d_0 = 0$)

Prior	hyper-prior	$g \geq$	Parameters of the SGBP distribution a	b	p	q	s
PEP							
(General)	SGBP	δ	$\frac{n^* - k_1}{2}$	$\frac{n^* - k_1}{2}$	1	δ	δ
(Recommended)	SGBP	n	$\frac{n - k_1}{2}$	$\frac{n - k_1}{2}$	1	n	n
EPP							
(General)	SGBP	1	$\frac{n^* - k_1}{2}$	$\frac{n^* - k_1}{2}$	1	1	1
(Recommended)	SGBP	1	$\frac{1}{2}$	$\frac{1}{2}$	1	1	1
Robust							
(General)	SGBP	$\frac{b_r + n}{\rho_{1,r}^{-1}} - b_r$	a_r	1	1	$\frac{b_r + n}{\rho_{1,r}^{-1}}$	$\frac{b_r + n}{\rho_{1,r}^{-1}} - b_r$
(Recommended)	SGBP	$\frac{n+1}{k_0 + k_1} - 1$	$1/2$	1	1	$\frac{n+1}{k_0 + k_1}$	$\frac{n+1}{k_0 + k_1} - 1$
Hyper-g							
(General)	Beta′	0	$\frac{a_h}{2} - 1$	1	1	1	0
(Recommended)	Beta′	0	$1/2$	1	1	1	0
Hyper-g/n							
(General)	Beta′	0	$\frac{a_h}{2} - 1$	1	1	n	0
(Recommended)	Beta′	0	$1/2$	1	1	n	0

Prior distribution of the shrinkage parameter

- The prior distribution of $w = \delta/(\delta + t) = g/(g + 1)$ is given by

$$w \sim BTPD\left(a = \frac{n^* + d_0 - k_1}{2},\ b = \frac{n^* + d_1 - d_0 - k_1}{2},\ \theta = \frac{\delta}{\delta + 1}\ \lambda = 1,\ \kappa = 1\right)$$

where $BTPD(a, b, \theta, \lambda, \kappa)$ is the Beta truncated Pareto distribution with parameters a, b, θ, λ, κ and density function

$$f(w; a, b, \theta, \lambda, \kappa) = \frac{1}{B(a, b)} \frac{\kappa \theta^\kappa w^{-\kappa - 1}}{1 - \left(\frac{\theta}{\lambda}\right)^\kappa} \left[\frac{1 - \left(\frac{\theta}{w}\right)^\kappa}{1 - \left(\frac{\theta}{\lambda}\right)^\kappa}\right]^{a-1} \left[1 - \frac{1 - \left(\frac{\theta}{w}\right)^\kappa}{1 - \left(\frac{\theta}{\lambda}\right)^\kappa}\right]^{b-1}$$

for $\theta < w < \lambda$. The prior mean and the variance of w are now given by

$$\begin{aligned} E(w) &= {}_2F_1(1,\ a;\ a + b;\ -1/\delta) \text{ and} \\ Var(w) &= {}_2F_1(2,\ a;\ a + b;\ -1/\delta) - {}_2F_1(1,\ a;\ a + b;\ -1/\delta)^2, \end{aligned}$$

where ${}_2F_1(a, b; c; z)$ is the Gauss hyper-geometric function.

Prior distribution of the shrinkage parameter (cont.)

- When considering the usual EPP setup with the minimal training sample, then the prior mean of the shrinkage is far away from one for specific cases (e.g. for the reference prior or for the Jeffreys' dependence prior when $k_0 = 1$ and $k_1 = 2$). This is not the case for the PEP prior for which the prior mean of the shrinkage is close to one even for models of small dimension; for example, under the reference prior and for $k_0 = 1$ and $k_1 = 2$ we obtain a prior mean of the shrinkage equal to 0.86 and a prior standard deviation of the shrinkage equal to 0.071. Generally the global shrinkage w under the PEP prior is close to one implying that the prior is generally non-informative since most of the information is taken from the data.

Desiderata

- Obviously PEP prior satisfies the **basic criterion** (C1). Furthermore Fouskakis & Ntzoufras (2016) proved that the PEP prior leads to a **consistent model selection procedure** (criterion C2). Fouskakis & Ntzoufras (2017) showed that the PEP prior satisfies the **information consistency criterion** (C3). Additionally, as shown here, for $d_0 = 0$, PEP prior belongs to a more general class of conditional priors

$$\pi_1(\boldsymbol{\beta}_e, \boldsymbol{\beta}_0, \sigma_1) \propto \sigma_1^{-1-(k_1-k_0)} h_1\left(\frac{\boldsymbol{\beta}_e}{\sigma_1}\right), \tag{1}$$

where $h(\cdot|M_1)$ is a proper density with support $\mathbb{R}^{k_1-k_0}$. Bayarri et.al. (2012) prove that the **group invariance criterion** (C7) hold if and only if $\pi_1(\boldsymbol{\beta}_e, \boldsymbol{\beta}_0, \sigma_1)$ has the form of (1). Additionally, if $h(\cdot|M_\ell)$ is symmetric around zero, which is the case under the PEP prior, **predictive matching criterion** (C5) also holds. When, finally $\mathrm{X}_e^* = \mathrm{X}_e$, the conditional scale matrix has the form $\Sigma_e^{-1} = \mathrm{X}_e^T(\mathrm{I}_n - \mathrm{P}_0)\mathrm{X}_e$ and then **null predictive matching, dimensional predictive matching and the measurement invariance criterion** (C6) hold, according to Bayarri et.al. (2012).

Properties (more)

- All the **full conditional posterior distributions** can be obtained in closed form. Therefore we can easily implement a full Gibbs sampler to obtain the posterior estimates of interest for any given model or a Gibbs based variable selection sampler to obtain estimates of the posterior model weights.
- The marginal likelihood given g is given by

$$f(\boldsymbol{y}|g, M_1) = C_1 \times (g+1)^{\frac{n+d_0-k_1}{2}} \left(1 + g\,\frac{1-R_1^2}{1-R_0^2}\right)^{-\frac{n+d_0-k_0}{2}}$$

where C_1 is a constant for all models (assuming that the covariates of X_0 are included in all models) given by

$$C_1 = 2^{\frac{d_0}{2}-1}\pi^{\frac{k_0-n}{2}}|\mathrm{X}_0^T\mathrm{X}_0|^{-1/2}\Gamma\left(\frac{n+d_0-k_0}{2}\right)(1-R_0^2)^{-\frac{n+d_0-k_0}{2}}||\boldsymbol{y}-\bar{y}\,\mathbf{1}_n||^{-\frac{n+d_0-k_0}{2}}$$

Properties (more)

- The **full marginal likelihood** is given by

$$f(\boldsymbol{y}|M_1) = C_1 \times \frac{B\left(\frac{k_1-k_0}{2} + \frac{n^*+d_0-k_1}{2}, \frac{n^*+d_1-d_0-k_1}{2}\right)}{B\left(\frac{n^*+d_0-k_1}{2}, \frac{n^*+d_1-d_0-k_1}{2}\right)}$$

$$\times \quad (\delta+1)^{\frac{n+d_0-k_1}{2}} \left(1+\delta\frac{1-R_1^2}{1-R_0^2}\right)^{-\frac{n+d_0-k_0}{2}}$$

$$\times \quad F_1\left(b; \frac{n+d_0-k_0}{2}, -\frac{n+d_0-k_1}{2}, \frac{k_1-k_0}{2}+a+b; \frac{1-R_0^2}{1-R_0^2+\delta(1-R_1^2)}, \frac{1}{\delta+1}\right)$$

where $F_1(a, b_1, b_2, c; x, y)$ is the hypergeometric function of two variables or Appell hypergeometric function given by

$$F_1(a, b_1, b_2, c; x, y) = \frac{1}{B(a, c-a)} \int_0^1 t^{a-1}(1-t)^{c-a-1}(1-xt)^{-b_1}(1-yt)^{-b_2} dt.$$

Properties (more)

- **The posterior expectation of the shrinkage parameter** $w = \frac{g}{g+1}$ is given by

$$E(w|\boldsymbol{y}, M_1) \quad = \quad \frac{\delta}{\delta+1} \times \frac{\widetilde{a}+b}{\widetilde{a}} \times \frac{\widetilde{F}_1(1)}{\widetilde{F}_1(0)},$$

where $\widetilde{a} = \frac{k_e}{2} + a - 1$, $k_e = k_1 - k_0$ and

$$\widetilde{F}_1(\kappa) = F_1\left(b; \frac{n+d_0-k_0}{2}, -\frac{n+d_0-k_1}{2}+\kappa, \frac{k_e}{2}+a+b-\kappa; \frac{1-R_0^2}{1-R_0^2+\delta(1-R_1^2)}, \frac{1}{\delta+1}\right)$$

for $\kappa \in \{0, 1\}$.

BMA estimates

Let us now consider a set of models $M \in \mathcal{M}$ where the covariates of matrix X_0 are included in all models and we are interested in accounting the uncertainty about the additional columns/covariates of X_e. In the following, for computational simplicity, we considered the data/design matrices Z_0 and Z_e with the centered covariates of X_0 and X_e instead.

A BMA point estimate of $\widehat{\boldsymbol{y}}^{new}$ is given by

$$E(\widehat{\boldsymbol{y}}^{new}|\boldsymbol{y}) = \mathrm{Z}_0^{new}\widehat{\beta}_0^c + \sum_{M\in\mathcal{M}} E\left(\frac{g}{g+1}|\boldsymbol{y}, M\right)\mathrm{Z}_e^{new}\widehat{\beta}_e f(M|\boldsymbol{y}).$$

Prediction consistency can be easily proven following the arguments of Liang et. al. (2008, JASA) with the local empirical Bayes estimate for g to be given by

$$\widehat{g} = \max\left\{\delta,\ \frac{n + d_0 - k_1}{n - k_1}F_{10} - 1\right\}$$

where F_{10} is the usual F statistic comparing model M_1 with model M_0.

Thank You Warwick!