Bayesian Variable Selection Using Power-Expected-Posterior Priors in GLMs

Ioannis Ntzoufras

Department of Statistics, Athens University of Economics and Business

Athens, Greece; e-mail: ntzoufras@aueb.gr.

Joint work with:

Dimitris Fouskakis & Konstantinos Perrakis

Department of Mathematics ***DZNE***

National Technical University of Athens ***(German Center for Neurodegenerative Diseases)***

8–12 May 2017, Thessaloniki
9th EMR-IBS and Italian Region conference

Synopsis

1. Introduction: Bayesian Model Selection and Power-Expected-Posterior (PEP) Priors
2. Alternative definitions of the power likelihood in PEP-priors
3. Implementing the method in GLMs (MCMC algorithm)
4. Additional results
5. Illustrations
6. Discussion

Introduction: Model Selection and Expected-Posterior Priors

Within the Bayesian framework the comparison between models M_0 and M_1 is evaluated via the **Posterior Odds** (PO)

$$PO_{01} \equiv \frac{\pi(M_0|\boldsymbol{y})}{\pi(M_1|\boldsymbol{y})} = \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} \times \frac{\pi(M_0)}{\pi(M_1)} = BF_{01} \times O_{01} \quad (1)$$

which is a function of the **Bayes Factor** (BF_{01}) and the **Prior Odds** (O_{01}).

In the above $m_\ell(\boldsymbol{y})$ is the marginal likelihood under model M_ℓ and $\pi(M_\ell)$ is the prior probability of model M_ℓ.

The marginal likelihood of model M_ℓ is given by

$$m_\ell(\boldsymbol{y}) = \int f_\ell(\boldsymbol{y}|\boldsymbol{\theta}_\ell)\pi_\ell(\boldsymbol{\theta}_\ell)d\boldsymbol{\theta}_\ell, \quad (2)$$

where $f_\ell(\boldsymbol{y}|\boldsymbol{\theta}_\ell)$ is the likelihood under model M_ℓ with parameters $\boldsymbol{\theta}_\ell$ and $\pi_\ell(\boldsymbol{\theta}_\ell)$ is the prior distribution of model parameters given model M_ℓ.

The Lindley-Bartlett-Jeffreys Paradox

For a single model inference $\Rightarrow$ a highly diffuse prior on the model parameters is often used (to represent ignorance).

$\Rightarrow$ Posterior density takes the shape of the likelihood and is insensitive to the exact value of the prior density function.

For multiple models inference $\Rightarrow$ BFs (and POs) are quite sensitive to the choice of the prior variance of model parameters.

$\Rightarrow$ For nested models, we support the simplest model with the evidence increasing as the variance of the parameters increase ending up to support of more parsimonious model no matter what data we have.

$\Rightarrow$ Under this approach, the procedure is quite informative since the data do not contribute to the inference.

$\Rightarrow$ Improper priors cannot be used since the BFs depend on the undefined normalizing constants of the priors.

Power-Expected-Posterior (PEP) Priors

Expected-Posterior priors (EPP; Perez and Berger, 2002, *Biometrika*)

$\Downarrow$

Power-Expected-Posterior Priors (PEP; Fouskakis, Ntzoufras and Draper, 2015, *Bayesian Analysis*).

$$\underbrace{\pi_\ell^{EPP}(\boldsymbol{\theta}_\ell)}_{\Downarrow} = \int \underbrace{\pi_\ell^{N}(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*)}_{\Downarrow} \underbrace{m_0^{N}(\boldsymbol{y}^*)}_{\Downarrow} d\boldsymbol{y}^*$$

$$\pi_\ell^{PEP}(\boldsymbol{\theta}_\ell;\delta) = \int \underbrace{\pi_\ell^{N}(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*,\delta)}_{\Downarrow} \underbrace{m_0^{N}(\boldsymbol{y}^*|\delta)}_{\Downarrow} d\boldsymbol{y}^*$$

we substitute the likelihood terms with powered-versions of the likelihoods

(i.e. they are raised to the power of $1/\delta$).

Features of PEP

PEP priors method amalgamates ideas from Intrinsic Priors, EPPs, Unit Information Priors and Power Priors, to unify ideas of Non-Data Objective Priors.

PEP priors solve the following problems:

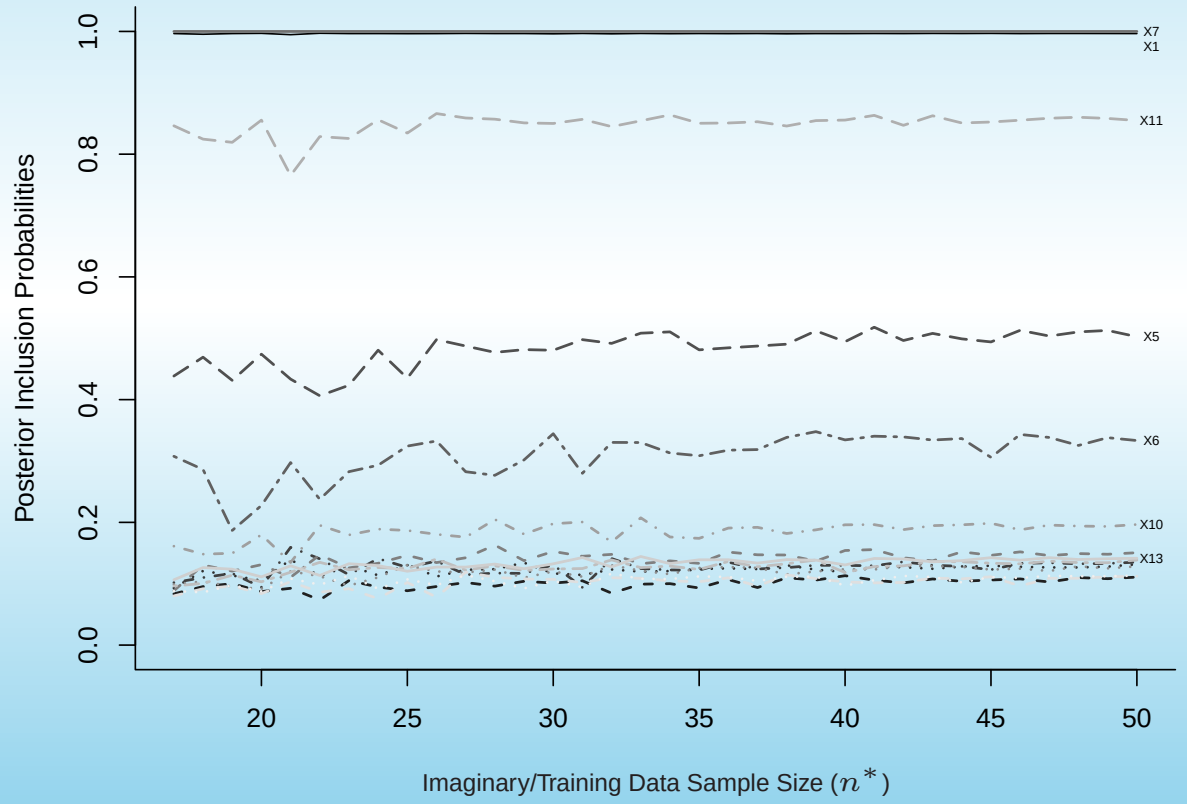
- Dependence of training sample size.
- Lack of robustness with respect to the sample irregularities.
- Excessive weight of the prior when the number of parameters is close to the number of data.

At the same time the PEP prior is a fully objective method and shares the advantages of Intrinsic Priors and EPPs.

- We choose $\delta = n^*$, $n^* = n$ and therefore $\mathrm{X}_\ell^* = \mathrm{X}_\ell$; by this way we dispense with the selection of the training samples.

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).*



Features of PEP (cont.)

For Normal models

- In Fouskakis, Ntzoufras & Draper, 2015 (*Bayesian Analysis*) we illustrated the the PEP prior approach
  - is robust with respect to the training sample size
  - is not informative when d_ℓ is close to n.
- The PEP prior can be expressed as a mixture of g-priors (Fouskakis, Ntzoufras & Pericchi, *submitted*).
- The Power-conditional-expected-posterior (PCEP) prior (Fouskakis & Ntzoufras, 2016, *JCGS*) is similar to the g-prior with (i) more complicated variance structure, (ii) more dispersed and (iii) more parsimonious than the g-prior
- Both PEP and PCEP are leading to consistent variable selection methods.

Extension to Generalized Linear Models

Definitions of the power-likelihood

Normal regression models: the definition of the power-likelihood seems quite clear.

We have worked with the density-normalized power likelihood since for any normal distribution with mean μ and variance σ^2 it holds that

$$f(y|\mu,\sigma^2,\delta) \quad = \quad \frac{f(y|\mu,\sigma^2)^{1/\delta}}{\int f(y|\mu,\sigma^2)^{1/\delta} dy} = N(\mu,\delta\,\sigma^2)$$

This is not the case for all distributions in the exponential family and hence for GLMs. May end up to a distribution which is not the same as the one in the original model formulation. For example

- In binary logistic regression $\Rightarrow$ it is still Bernoulli with success probability $\frac{\pi^{1/\delta}}{\pi^{1/\delta}+(1-\pi)^{1/\delta}}$.
- For the Binomial and the Poisson models, it results is some cumbersome distributions which increase computational complexity (without any obvious gain).

Alternative definitions of the power-likelihood

We consider the PEP representation

$$\pi_\ell^{PEP}(\boldsymbol{\theta}_\ell;\delta) \quad = \quad \int \pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*,\delta) m_0^N(\boldsymbol{y}^*|\delta) d\boldsymbol{y}^*$$

with δ controlling the amount of prior-information accounted in the final posterior (and the dispersion of the prior distribution).

We now consider **the unormalized power-likelihood** and then normalize the posterior (which is also the approach in Ibrahim and Chen, 2000, *Stat.Science*). Hence

$$\pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*,\delta) = \frac{f_\ell(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell)^{1/\delta}\pi_\ell^N(\boldsymbol{\theta}_\ell)}{\int f_\ell(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell)^{1/\delta}\pi_\ell^N(\boldsymbol{\theta}_\ell)d\boldsymbol{\theta}_\ell}$$

What about $m_0^N(\boldsymbol{y}^*|\delta)$?

Two alternatives for the marginal distribution

- Consider the **unormalized power-likelihood** and then normalize m_0^N:

$$m_0^N(\boldsymbol{y}^*, \delta) = \frac{\int f_0(\boldsymbol{y}^*|\boldsymbol{\theta}_0)^{1/\delta} \pi_0^N(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0}{\int\int f_0(\boldsymbol{y}^*|\boldsymbol{\theta}_0)^{1/\delta} \pi_0^N(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 d\boldsymbol{y}^*} \ .$$

  This will be noted as the *Diffuse Reference PEP (DR-PEP)*.

- Consider the **original likelihood** (without introducing any further uncertainty) i.e.

$$m_0^N(\boldsymbol{y}^*, \delta) = m_0(\boldsymbol{y}^*) = \int f_0(\boldsymbol{y}^*|\boldsymbol{\theta}_0) \pi_0^N(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 \ .$$

  This will be noted as the *Concentrated Reference PEP (CR-PEP)*.

In both cases the expected-posterior interpretation is retained with the first prior being more diffuse than the second.

Comparison of the two approaches in normal regression

Volume variance multipliers in normal regression models

The volume of the variance-covariance matrix in the g-prior and in the two PEP approaches is given by

$$\left|\text{Var}(\boldsymbol{\beta}_\ell|M_\ell)\right| = \varphi(n,d_\ell) \times |\boldsymbol{X}_\ell^T\boldsymbol{X}_\ell|^{-1}$$

- **G-prior with** $g = n \Rightarrow \varphi(n,d_\ell) = n^{d_\ell}$
- **DR-PEP prior** $\Rightarrow \varphi(n,d_\ell) = n^{2d_\ell}\left[\frac{2n+1}{(n+1)^2}\right]^{d_\ell-d_0}$
- **CR-PEP prior** $\Rightarrow \varphi(n,d_\ell) = n^{d_\ell}\left[\frac{n^2+2n}{n^2+2n+1}\right]^{d_\ell}\left[\frac{n^2+n+2}{n+2}\right]^{d_0}$.

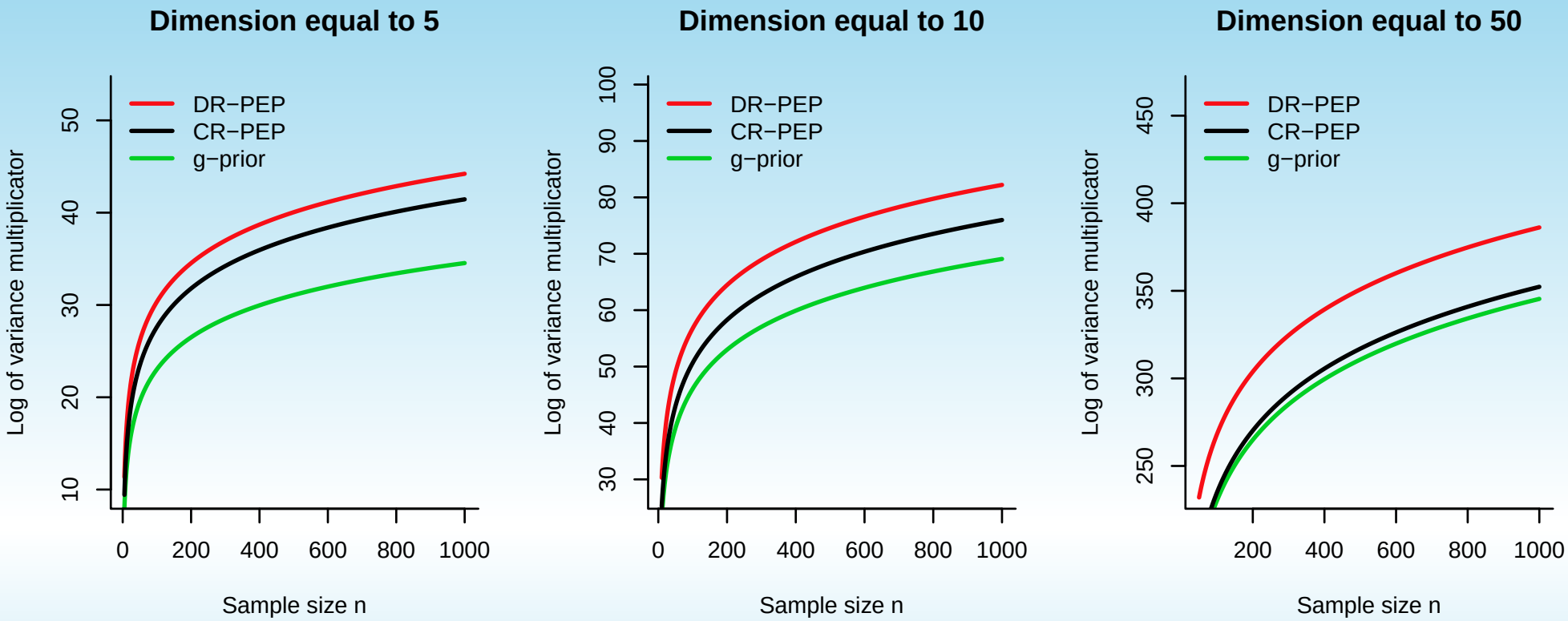


Figure 2: Log-variance multipliers of the DR-PEP, CR-PEP and g-priors versus sample size for $d_\ell = 5, 10, 50$.

The Final Formulation of the DR-PEP Prior

(Formulation and computation is similar for the CR-PEP prior)

The prior

$$\pi_{\boldsymbol{\gamma}}^{DRPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto \int\int \left\{ \frac{f_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\pi_{\boldsymbol{\gamma}}^{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{\int f_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\pi_{\boldsymbol{\gamma}}^{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})d\boldsymbol{\beta}_{\boldsymbol{\gamma}}} \right\} f_0(\boldsymbol{y}^*|\boldsymbol{\beta}_0)^{1/\delta}\pi_0^N(\boldsymbol{\beta}_0)d\boldsymbol{\beta}_0 d\boldsymbol{y}^*$$

Two possible approaches to simplify the above expression

- In GLMs, the posterior part can be well approximated by a normal distribution (Chen and Ibrahim, 2003, *Stat.Sinica*)
- Integral in the denominator can be well approximated using Laplace approximation

Constructing a Gibbs based Variable Selection Sampler

In order to estimate the posterior model probabilities, we use an MCMC scheme with full data augmentation by introducing

- For each model $\boldsymbol{\gamma}$, we introduce a complement of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denoted by $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ for all coefficients not included in the model.
- A pseudoprior $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})$ is defined to play the role of a proposal and the linear predictor can be rewritten as $\eta_i = \sum_{j=0}^{p} X_{ij}\gamma_j b_{\boldsymbol{\gamma},j}$ where $b_{\boldsymbol{\gamma},j}$ is the element of $\boldsymbol{b}_{\boldsymbol{\gamma}} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})$ which corresponds to covariate X_j.
- A latent parameter β_0 for the parameter of the reference model
- A latent vector of imaginary data $\boldsymbol{y}^*$

- We build a Gibbs based variable selection algorithm providing samples from the augmented posterior

$$\pi_{\boldsymbol{\gamma}}^{DRPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}\boldsymbol{\gamma}, \boldsymbol{y}^*, \beta_0|\boldsymbol{y})$$

$$\propto \frac{f_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}})\left[f_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})f_0(\boldsymbol{y}^*|\boldsymbol{\beta}_0)\right]^{1/\delta}}{\int f_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}})d\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})\pi_0^N(\boldsymbol{\beta}_0)\pi(\boldsymbol{\gamma})$$

- We use Laplace approximation to evaluate the integral in the denominator.
- In this work, we use the Jeffreys prior as a baseline prior.

Some additional results

- We have extended our approach by using hyper-priors for δ in a similar manner as hyper-g priors do.
- The hyper-δ extensions can be easily implemented by adding an additional step for δ in the Gibbs variable selection algorithm.
- The method is consistent for normal models. Simulations in Binomial and Poisson data indicate the same.
- Prediction matching holds for both PEP and hyper-pep versions.
- The DR-PEP coincides with the conditional version of PEP using density normalized likelihood for normal models.

Illustrative example 1: Pima Indians dataset

- Pima Indians diabetes data set (Ripley, 1996).
- $n = 532$ binary responses on diabetes presence (present=1, not present=0) according to the WHO criteria for signs of diabetes.
- $p = 7$ potential covariates which are listed in Table 1 (see next slide).
- The data also used by Holmes and Held (2006, *Bayesian Analysis*) and Bové and Held (2011, *Bayesian Analysis*).
- Beta-binomial prior on model space.

Covariate	Description
X_1	Number of pregnancies
X_2	Plasma glucose concentration (mg/dl)
X_3	Diastolic blood pressure (mm Hg)
X_4	Triceps skin fold thickness (mm)
X_5	Body mass index (kg/m^2)
X_6	Diabetes pedigree function
X_7	Age

Table 1: Potential predictors in the Pima Indians diabetes data set.

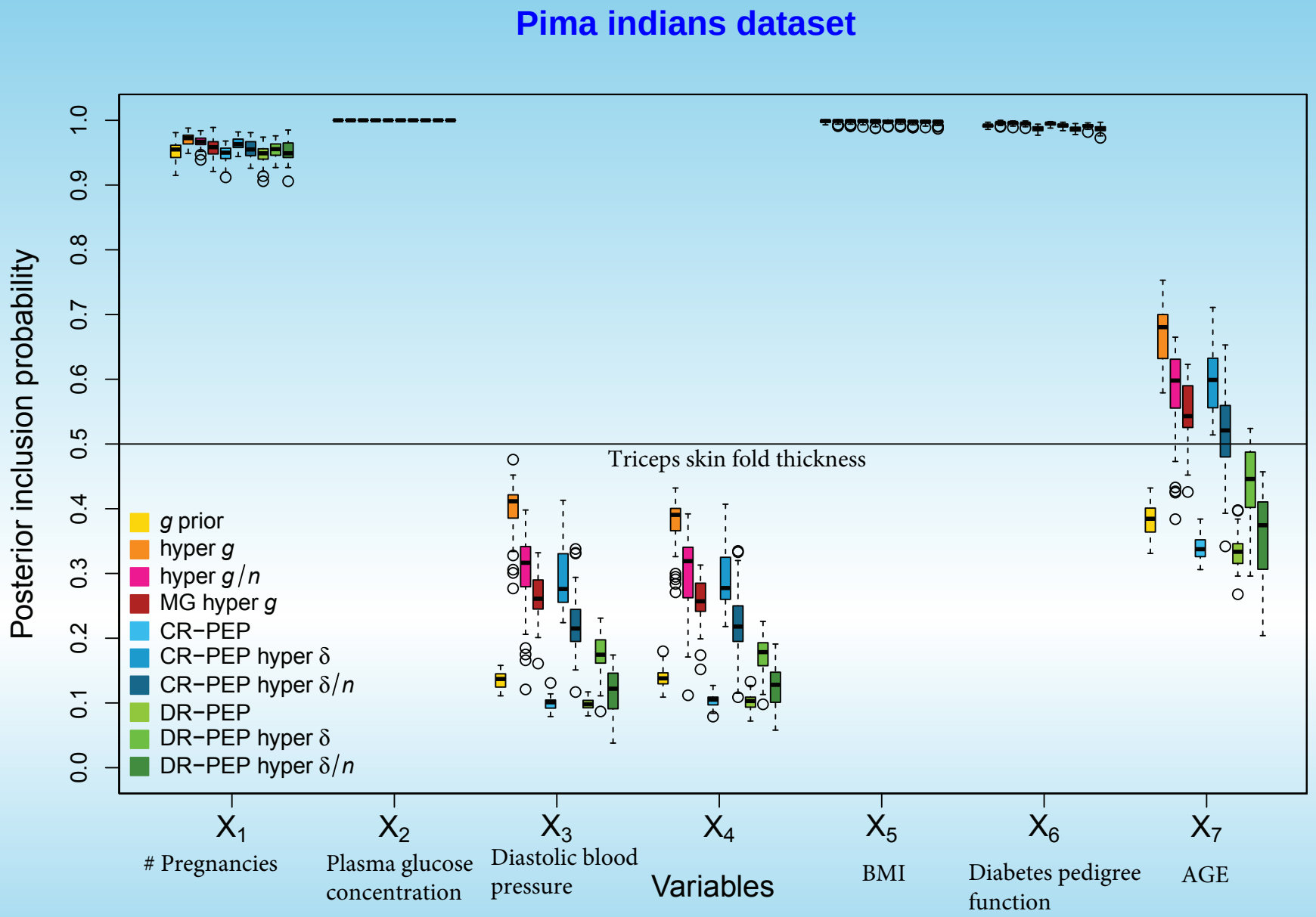


Figure 3: Boxplots of batched estimates of the posterior inclusion probabilities (40 batches of size 1000).

Illustrative example 2: Web-analytics Popularity Dataset

The dataset of this illustration refer to characteristics of the popular website of Mashable (`www.mashable.com`).

- The original content be publicly accessed and retrieved using the provided urls.
- All sites and related data were downloaded on January 8, 2015.
- Main variable: the number of shares which measures the popularity of the site/post.
- We are interested to identify the ingredients of a successful post and what it takes to for a post to become a viral.

Source UCI Machine Learning Repository `http://archive.ics.uci.edu/ml/`; see Fernandes et al. (2015) for more details.

Illustrative example 2: Web-analytics Popularity Dataset

Data used

Subset of 1000 observations.

47 Covariates (reduced from 60 after some initial exploratory analysis).

Response: Is a post a viral? (binary: one for posts with $\geq$ 1400 shares and zero otherwise).

Some covariates:

- Dummies for category type of the post (Lifestyle/Entertainment/Business/Social Media/Tech/World).
- Day of the week the post was published.
- Number of days on the air.
- Title and text length (in number of words).
- Rate of unique words in the content.
- Number of link, images, videos.
- Keyword statistics (number of shares).
- Positivity and negativity rates of words.
- Polarity indexes of text and title.
- Subjectivity of text and titles.

Web-analytics Popularity Dataset

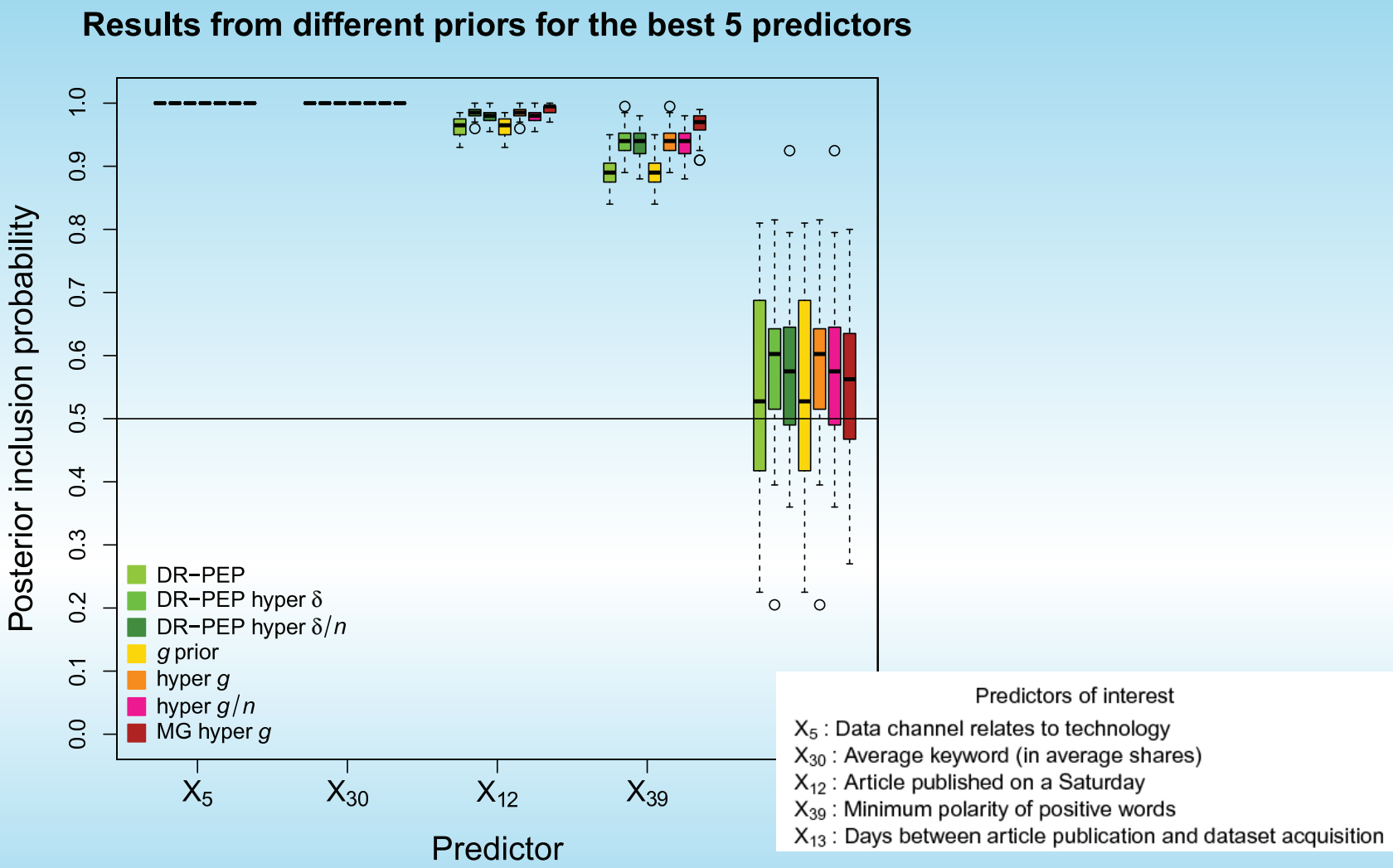
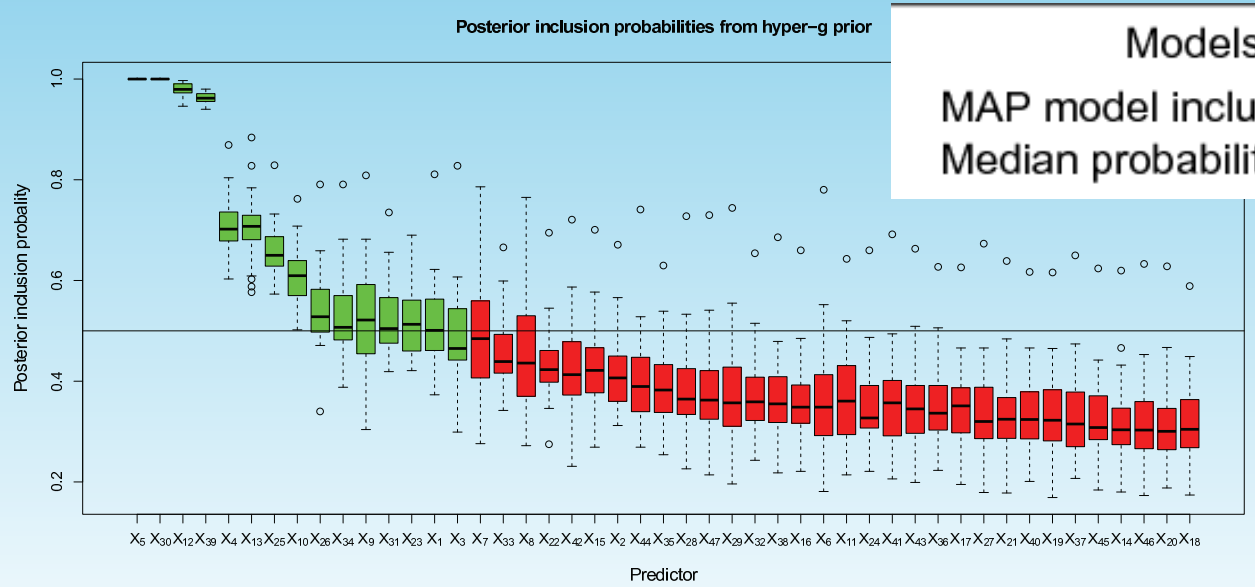


Figure 4: Comparison of Important Determinants of a Viral Post (40 batches).

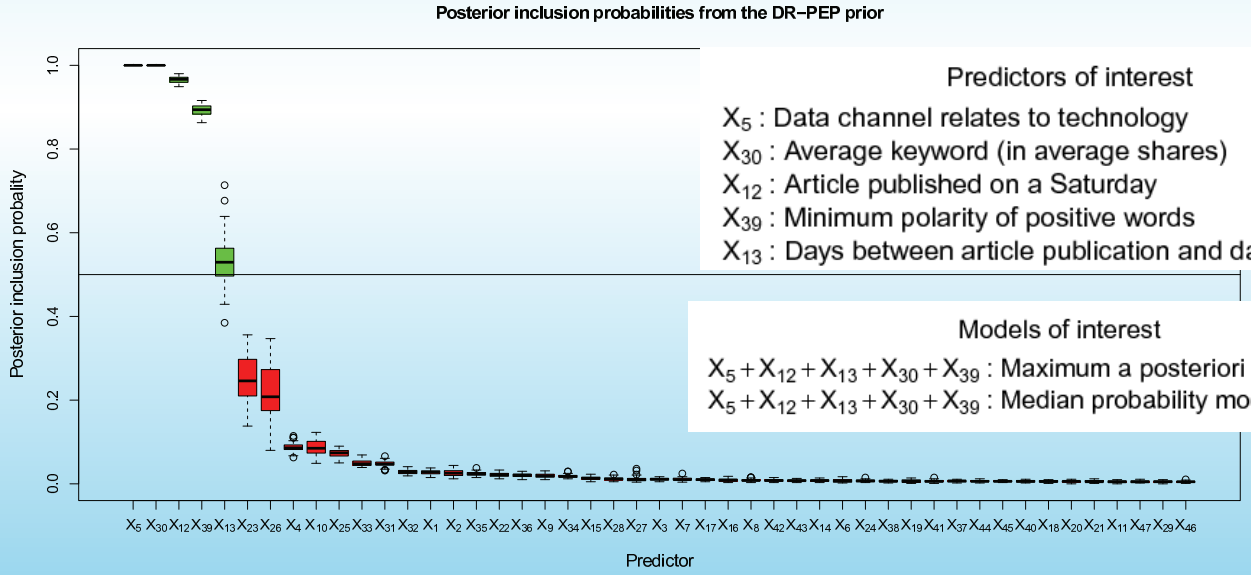
Web-analytics Popularity Dataset



Models of interest

MAP model includes all predictors

Median probability model ~ 14 predictors



Predictors of interest

X_5 : Data channel relates to technology

X_{30} : Average keyword (in average shares)

X_{12} : Article published on a Saturday

X_{39} : Minimum polarity of positive words

X_{13} : Days between article publication and dataset acquisition

Models of interest

$X_5 + X_{12} + X_{13} + X_{30} + X_{39}$: Maximum a posteriori probability model

$X_5 + X_{12} + X_{13} + X_{30} + X_{39}$: Median probability model

Figure 5: Posterior inclusion probabilities for hyper-g and DR-PEP (40 batches).

Web-analytics Popularity Dataset

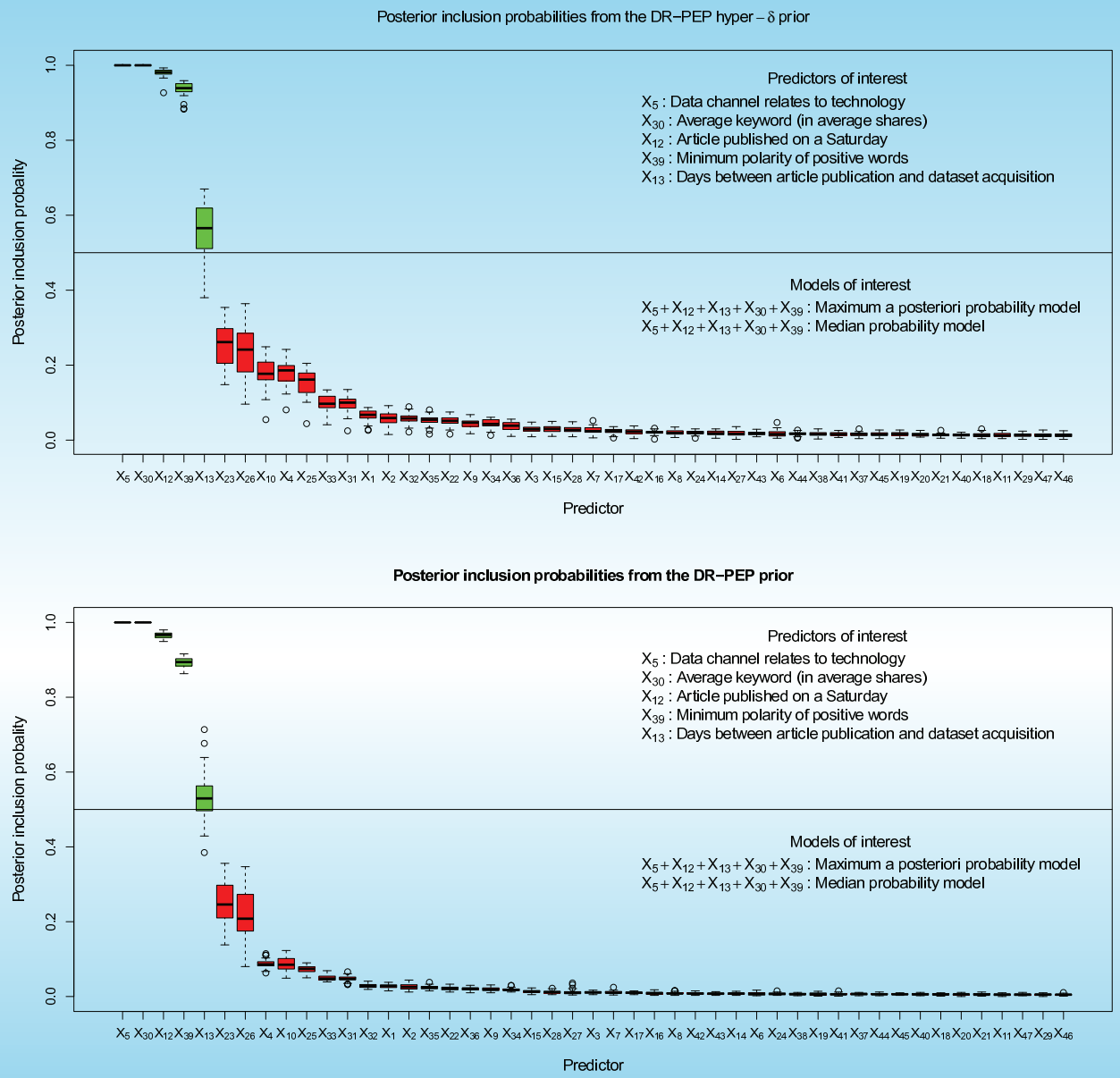


Figure 6: Posterior inclusion probabilities for hyper-g and DR-PEP (40 batches of size 200 iterations).

Illustrative example 3: Small Scale Data Simulation

- Also presented in Chen et al. (2008) and Li and Clyde (2015).
- $n = 100$, $p = 3$ predictors. Each simulation is repeated 100 times.
- Each predictor is drawn from a standard normal distribution with pairwise correlation given by

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

with (i) independent predictors ($r = 0$) and (ii) correlated predictors ($r = 0.75$).

Scenario	**Poisson** $(n = 100)$			
	β_0	β_1	β_2	β_3
null	-0.3	0	0	0
sparse	-0.3	0.3	0	0
medium	-0.3	0.3	0.2	0
full	-0.3	0.3	0.2	-0.15

Table 2: Four simulation scenarios for Poisson regression assuming independent and correlated predictors.

	Null		Sparse		Medium		Full	
Prior	0	0.75	0	0.75	0	0.75	0	0.75
g-prior	87	93	74	36	29	0	5	0
hyper g-prior	59	71	72	41	**45**	3	**21**	2
hyper g/n-prior	81	83	72	42	38	1	13	1
MG hyper g-prior*	84	90	72	37	32	0	10	0
CR PEP	88	95	76	35	27	0	5	0
CR PEP hyper-δ	71	75	68	**44**	44	**4**	18	**3**
CR PEP hyper-δ/n	83	91	**80**	40	30	0	11	0
DR PEP	90	95	73	32	28	0	5	0
DR PEP hyper-δ	91	**97**	68	30	25	0	4	0
DR PEP hyper-δ/n	**94**	95	69	28	20	0	3	0

Table 3: Number of times that the MAP model corresponds to the true model for 100 simulated datasets; column-wise largest value is in red.

Concluding remarks

- We have extended PEP-variable selection for GLMs
- Main problems
  - Definition of the power-likelihood - we have presented two alternatives
  - Computation - we have used an augmented Gibbs variable selection sampler
- CR-PEP and DR-PEP are more parsimonious than g-priors with similar properties.
- Work must be done to prove consistency in the general setup and extend methodology for *large* p*, small* n problems.
- Efficient computation for large scale data:

  EMVS (Rockova and George, 2014, *JASA*) or other fast alternatives should be explored.

Funding

This research has been co-financed in part by the European Union (European Social Fund-ESF) and by Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF)-Research Funding Program: Aristeia II/PEP-BVS.


THE END