

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

Department of Statistics

National Technical University of Athens

Athens University of Economics and Business



National Technical University of Athens, 13 May 2016

Synopsis

1. Introduction: Model Selection and Expected-Posterior Priors
2. Alternative definitions of the power likelihood in PEP-priors
3. Implementing the method in GLMs (MCMC algorithm)
4. Using Mixtures of PEP priors
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)**

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

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

In the above $m_\ell(\mathbf{y})$ 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

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

where $f_\ell(\mathbf{y}|\boldsymbol{\theta}_\ell)$ 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.

Expected-Posterior Priors (EPP)

- Pérez & Berger (2002, Biometrika) developed priors for use in model comparison, through utilization of the device of **imaginary training samples**.
- They defined the **expected-posterior prior** (EPP) as the posterior distribution of a parameter vector for the model under consideration, averaged over all possible imaginary samples $\mathbf{y}^* = (y_1^*, \dots, y_{n^*}^*)^T$ coming from a “suitable” predictive distribution $m^*(\mathbf{y}^*)$.

Hence the EPP for the parameters of any model M_ℓ is

$$\pi_\ell^{EPP}(\boldsymbol{\theta}_\ell) = \int \pi_\ell^N(\boldsymbol{\theta}_\ell | \mathbf{y}^*) m^*(\mathbf{y}^*) d\mathbf{y}^*, \quad (3)$$

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

Features of EPP

- **Impropriety** of baseline priors causes no indeterminacy. **Impropriety** in m^* also does not cause indeterminacy, because it is common to the EPPs for all models.
- It makes priors **compatible** across models, through their dependence on a common data distribution.
- Usually we consider as m^* the marginal likelihood of a reference model.
 - Usual choices in regression models are the **null** and full **model**.
 - Here we consider the null, i.e. $m^*(\mathbf{y}^*) = m_0^N(\mathbf{y}^*)$.
- In nested cases usually the reference model is the simplest model. In this case EPP is the same as the **Intrinsic Prior**.
- We choose the smallest n^* for which the posterior is proper: **minimal training sample size**.
- **Main Issue**: In variable selection problems specification of X_ℓ^* .

Power-Expected-Posterior (PEP) Priors

Fouskakis, Ntzoufras and Draper (2015, *Bayesian Analysis*).

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

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 $X_\ell^* = X_\ell$; by this way we dispense with the selection of the training samples.

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, *unpublished work, presented in ISBA2014*).
- The Power-conditional-expected-posterior (PCEP) prior (Fouskakis & Ntzoufras, 2016, *to appear in 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.

1 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

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

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

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

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

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

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

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

Two alternatives for the marginal distribution

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

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

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

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

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

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.

Features of the diffuse-reference PEP

- Still has the interpretation of a posterior density given some imaginary data \mathbf{y}^* “weighted” by n^*/δ data-points and averaged over a data distribution.
- The same type of uncertainty is introduced both in the “posterior” and the predictive (averaged) part.
- In normal regression models
 - Equivalent to using the the density-normalized power likelihood.
 - It is equivalent to PEP and PCEP.
 - It leads to a consistent model selection method.
 - It is more dispersed (and parsimonious) than the g-prior.

Features of the concentrated-reference PEP

- Still has the interpretation of a posterior density given some imaginary data \mathbf{y}^* “weighted” by n^*/δ data-points averaged over the predictive distribution of **the actual reference model**.
- Different type of uncertainty is introduced both in the “posterior” and the predictive (averaged) part.
- Less dispersed than the diffuse version of PEP.
- In normal regression models
 - It is less dispersed (and parsimonious) than PEP (and DR-PEP) and more dispersed (and parsimonious) than the g-prior.
 - It leads to a consistent model selection method.

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

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

- **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** $\Rightarrow \varphi(n, d_\ell) = n^{d_\ell} \left[\frac{n^2+2n}{n^2+2n+1} \right]^{d_\ell} \left[\frac{n^2+n+2}{n+2} \right]^{d_0}$.

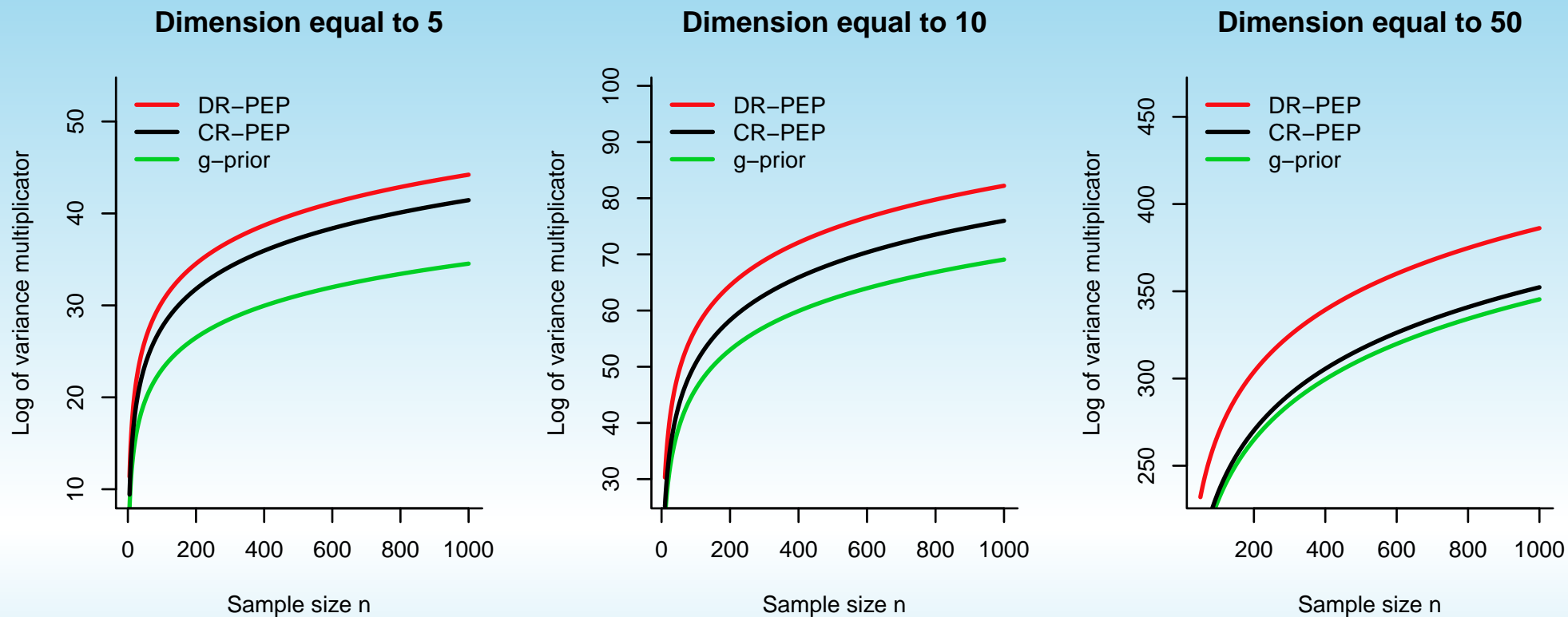


Figure 2: Log-variance multipliers of the DR-PEP, CR-PEP and g -priors versus sample size for $d_\ell = 5, 10, 50$.

Formulation in GLMs

- $M_\ell \rightarrow \gamma$: Binary variable inclusion indicators (γ) in order to search the model space using Gibbs sampling (George and McCulloch, 1993, *JASA*)
- $Y_i \sim$ a distribution member of the exponential family.
The parameters of the distribution are associated with the linear predictor via a link function.
- p covariates.
- \mathbf{X} is the $n \times (p + 1)$ data matrix with the first column to be the constant and rest containing the data of each covariate.
- \mathbf{X}_γ is the $n \times d_\gamma$ data matrix for model γ with $d_\gamma = \sum_{j=0}^p \gamma_j$ covariates.
- β_γ is the parameter vector of length d_γ with the effects of each covariate
- The linear predictor vector is given by $\eta_\gamma = \mathbf{X}_\gamma \beta_\gamma$

The Final Formulation of the DR-PEP Prior

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

The prior

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

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

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

$$\pi_{\gamma}^{DRPEP}(\boldsymbol{\beta}_{\gamma}, \boldsymbol{\beta}_{\setminus\gamma}, \gamma, \mathbf{y}^*, \boldsymbol{\beta}_0 | \mathbf{y})$$

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

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

The MCMC algorithm - Gibbs variable selection for PEP

For each iteration t ($t = 1, 2, \dots, N$),

Step 1: For $j = 1, \dots, p$, we update $\gamma_j \sim \text{Bernoulli} \left(\frac{O_j}{1+O_j} \right)$, with

Step 2: We update β_γ from the full conditional posterior (given the current values of γ and \mathbf{y}^*) using a Metropolis step and proposals build using MLEs from a model with response $(\mathbf{y}, \mathbf{y}^*)$ and weights $\mathbf{w} = (\mathbf{1}_n, \delta^{-1} \mathbf{1}_{n^*})$

Step 3: Update $\beta_{\setminus\gamma}$ from the pseudo-prior $\pi_\gamma(\beta_{\setminus\gamma}) = N_{d_{\setminus\gamma}} \left(\hat{\beta}_{\setminus\gamma}, \mathbf{I}_{d_{\setminus\gamma}} \hat{\sigma}_{\beta_{\setminus\gamma}}^2 \right)$.

Step 4: Sample β_0 from the full conditional posterior (given \mathbf{y}^*) using a Metropolis step with a normal proposal with mean the MLE with response \mathbf{y}^* and variance equal to $\delta \hat{\sigma}_{\hat{\beta}_0^*}^2$ with the latter being the corresponding variance of the MLE.

Step 5: Sample \mathbf{y}^* from the full conditional posterior (given β_γ , β_0 and γ) using a Metropolis step.

- The proposal depends on the model likelihood i.e. the stochastic part of the model; for details see next slide.
- In the acceptance probabilities we need to evaluate the marginal likelihoods $m_\gamma^N(\mathbf{y}^* | \delta)$ and $m_\gamma^N(\mathbf{y}^{*'} | \delta)$ which are computed by using Laplace approximation.

Details about the proposal for y_i^*

For \mathbf{y}^* we construct proposals depending on the likelihood of the model.

Binomial response: Propose $y_i^{*'} \sim \text{Bernoulli}(\pi_i)$ with

$$\pi_i = \frac{(\pi_{0,i}^D \pi_{\gamma,i})^{1/\delta}}{(\pi_{0,i}^D \pi_{\gamma,i})^{1/\delta} + [(1 - \pi_{0,i})^D (1 - \pi_{\gamma,i})]^{1/\delta}}$$

with $D = 1$ for DR-PEP and $D = \delta$ for CR-PEP, $\pi_{0,i} = [1 + \exp(-\beta_0)]^{-1}$ and $\pi_{\gamma,i} = [1 + \exp(-\eta_{\gamma,i})]^{-1}$; where $\eta_{\gamma,i}$ is the i -th element of $\boldsymbol{\eta}_{\gamma} = \mathbf{X}_{\gamma} \boldsymbol{\beta}_{\gamma}$.

Poisson regression:

CR-PEP: Propose $y_i^{*'} \sim \text{Poisson}(\lambda_i)$ with $\log \lambda_i = \beta_0 + \eta_{\gamma,i}/\delta$.

DR-PEP: Propose $y_i^{*'} \sim \text{Poisson}(y_i^*)$.

Further Remarks about the MCMC

- Exact distributions are used for $\beta_{\setminus\gamma}$ and γ
 - $\beta_{\setminus\gamma}$ is sampled directly from the pseudo-prior distribution.
 - γ is sampled directly from the full conditional Bernoulli distribution.
- The pseudo-prior of $\beta_{\setminus\gamma}$ serves the role of the proposal and it does not influence the posterior but it does influence the efficiency of the MCMC algorithm.
- No specific fine tuning is required for the proposal distributions of β_{γ} and β_0 (normal proposals based on MLEs).

2 Hyper-delta PEP priors

PEP priors with fixed δ are similar in notion and behaviour as the g-priors.

We extend our approach by using hyper-priors for δ in a similar manner as hyper-g priors do.

Under this setting, the hyper- δ PEP prior can be approximated by

$$\pi_{\gamma}^{\text{PEP}}(\boldsymbol{\beta}_{\gamma}) \approx \int \int f_{N_{d_{\gamma}}}(\boldsymbol{\beta}_{\gamma}; \widehat{\boldsymbol{\beta}}_{\gamma}^*, \delta (\mathbf{X}_{\gamma}^{*T} \mathbf{H}_{\gamma}^* \mathbf{X}_{\gamma}^*)^{-1}) m_0^N(\mathbf{y}^* | \delta) \pi(\delta) d\mathbf{y}^* d\delta, \quad (4)$$

where $\widehat{\boldsymbol{\beta}}_{\gamma}^*$ is the MLE given the imaginary data.

This approximation cannot be applied when using EPPs with minimal training samples.

Similarly to the hyper-g (Liang *et al.*, 2008, *JASA*), the hyper-delta prior is given by

$$\pi(\delta) = \frac{a-2}{2} (1+\delta)^{-a/2},$$

which introduces the following prior for $\delta/(1+\delta)$

$$\frac{\delta}{1+\delta} \sim \text{Beta}\left(1, \frac{a}{2} - 1\right)$$

- We use $a = 3$ as suggested by Liang *et al.* (2008, *JASA*).
- $\frac{\delta}{1+\delta}$ has an interpretation similar to a shrinkage parameter since it accounts for the proportion of information (in data-points) coming from the actual data when $n = n^*$ — in the general case this will be given by $n/(n + n^*/\delta)$.
- Another alternative option would be a hyper- δ/n prior of the form

$$\pi(\delta) = \frac{a-2}{2n} \left(1 + \frac{\delta}{n}\right)^{-a/2}.$$

Additional MCMC step for δ

Step 6: Sample of δ from the full conditional posterior (given the current values of β_γ , β_0 , \mathbf{y}^* and γ).

(a) Propose δ' from $q(\delta'|\delta) = \text{Gamma}(\delta, 1)$.

(b) Compute the Laplace approximations $\hat{m}_\gamma^N(\mathbf{y}^*|\delta)$ and $\hat{m}_\gamma^N(\mathbf{y}^*|\delta')$.

(c) Accept the proposed move with probability $\alpha_\delta = \min\{1, A_\delta\}$, where A_δ is given by

$$A_\delta = \left\{ f_\gamma(\mathbf{y}^*|\beta_\gamma) f_0(\mathbf{y}^*|\beta_0) \right\}^{\Delta\delta} \times \frac{\pi(\delta')}{\pi(\delta)} \times \frac{\hat{m}_\gamma^N(\mathbf{y}^*|\delta)}{\hat{m}_\gamma^N(\mathbf{y}^*|\delta')} \times \frac{q(\delta|\delta')}{q(\delta'|\delta)}.$$

where $\Delta\delta = 1/\delta' - 1/\delta$

Illustrative example 1: 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 1: 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 ≥ 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

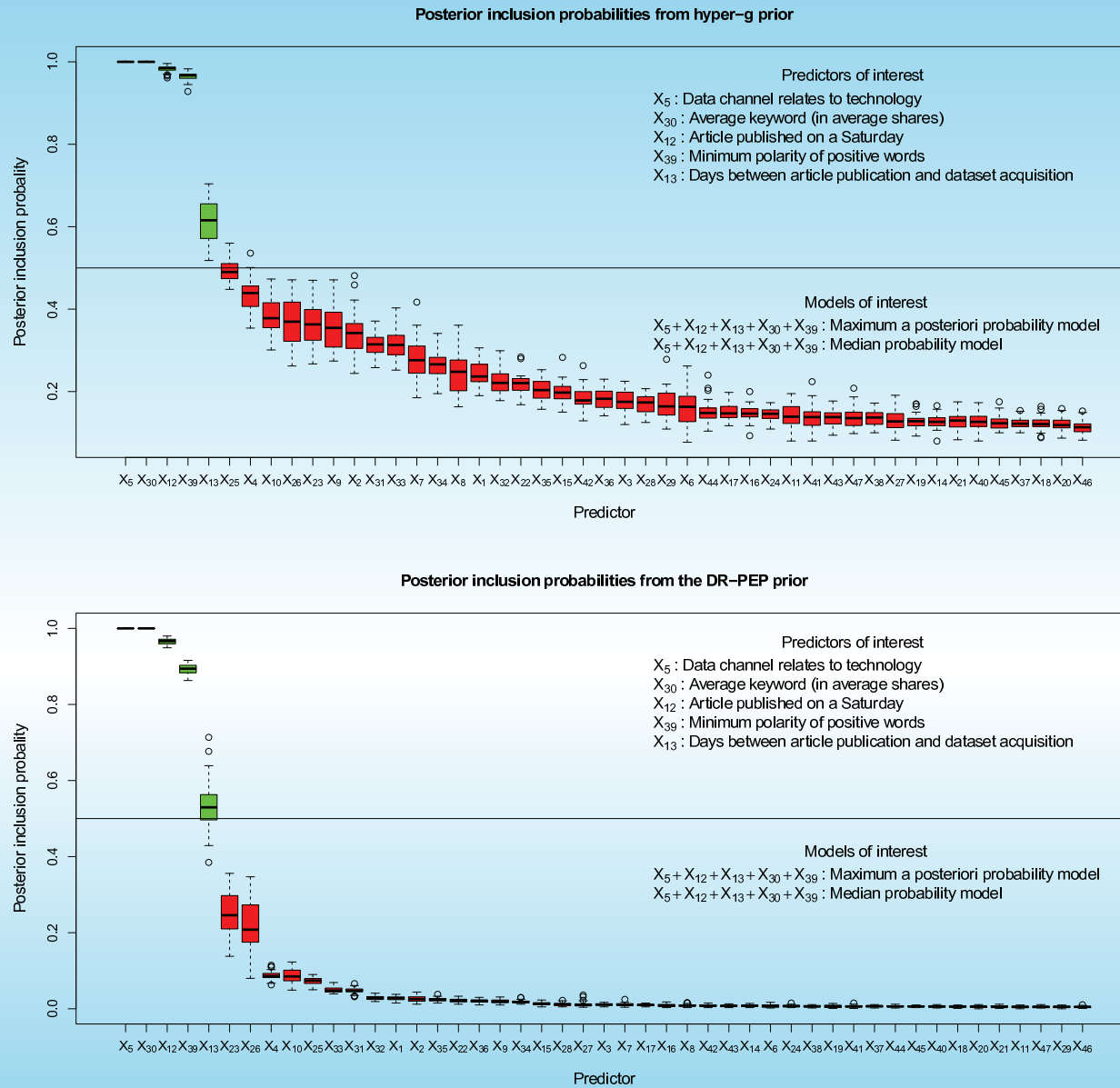


Figure 3: Posterior inclusion probabilities for hyper-g and DR-PEP (40 batches).

Web-analytics Popularity Dataset

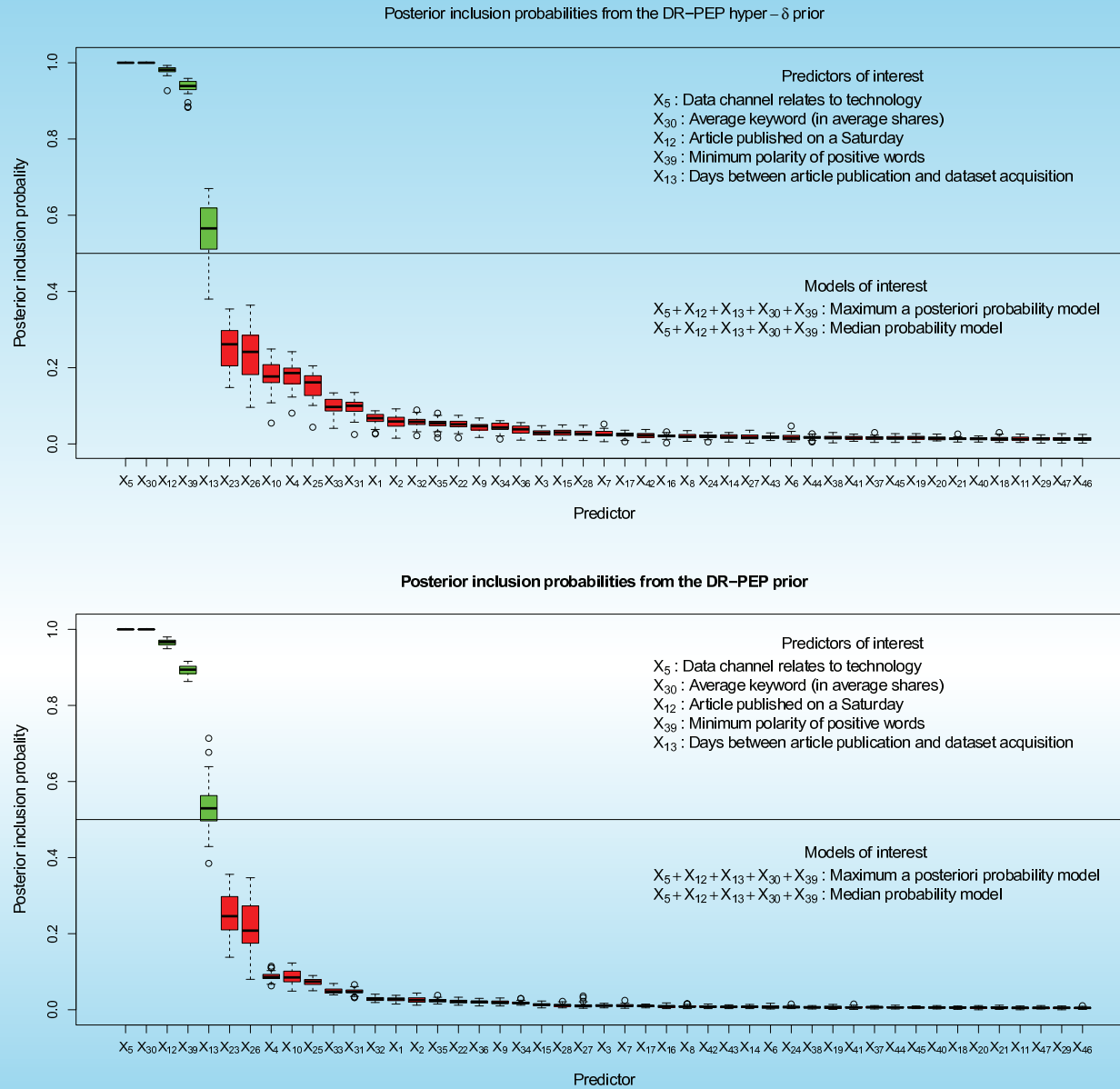


Figure 4: Posterior inclusion probabilities for hyper-g and DR-PEP (40 batches of size 200 iterations).

Illustrative example 2: 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

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

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

Prior	Null		Sparse		Medium		Full	
	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 2: 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.
- Computation should be improved to be implemented in big data:
EMVS (Rockova and George, 2014, *JASA*) or other fast alternatives should be explored.

Future plans: Moving towards really Big datasets

Ideas for Large p – small n problems

- In PEP, use LASSO baseline to result in a combination of LASSO-g-prior properties.
- Implement EM variable selection for PEP and general type of priors.
- Use population based MCMCs ideas and approaches partitioning the model space.

Ideas for Large n – reasonable p problems

- Use sequential updating of variable selection techniques to enable the application in sequentially collected data (common in web-analytics).
- Borrow ideas from meta-analysis for splitting the analysis in sub-samples and then combining information.

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.



