Power-Expected-Posterior Priors for Generalized Linear Models

Dimitris Fouskakis[1], **Ioannis Ntzoufras**[2], **Konstantinos Perrakis**[2]

[1] Department of Mathematics, National Technical University of Athens, Greece
[2] Department of Statistics, Athens University of Economics and Business, Greece
fouskakis@math.ntua.gr, {ntzoufras, kperrakis}@aueb.gr

1. Motivation

AIM: To develop an objective and fully automatic Bayesian variable selection procedure without the need of specifying any tuning parameters.

WHY:
(a) Information about the regression coefficients is usually not available;
(b) We wish to avoid Jeffreys-Lindley-Bartlett paradox.

Previous related work served as basis

- Intrinsic priors (Berger and Pericchi, 1996)
- Expected-posterior (EP) priors (Pérez and Berger, 2002)

Some characteristics

Are implemented in normal regression and probit models.
The implementation in GLMs is challenging.
Large sample approximations can not be applied due to the use of minimal training samples.

2. Expected Posterior Priors (EPP)

- **Expected-posterior prior** (EPP) is the posterior distribution of the parameter vector $\boldsymbol{\theta}_\ell$ for model M_ℓ, averaged over all possible imaginary samples $\boldsymbol{y}^* = (y_1^*, \ldots, y_{n^*}^*)^T$ coming from the predictive distribution $m_0(\boldsymbol{y}^*)$ of a reference model M_0.
- The EPP is given by

$$\pi_\ell^{EPP}(\boldsymbol{\theta}_\ell) = \int \pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*)\, m_0(\boldsymbol{y}^*)\, d\boldsymbol{y}^*\,, \qquad (1)$$

where

$$\pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*) = \frac{f_\ell(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell)\pi_\ell^N(\boldsymbol{\theta}_\ell)}{m_\ell(\boldsymbol{y}^*)}$$

$$m_0(\boldsymbol{y}^*) = \int f_0(\boldsymbol{y}^*|\boldsymbol{\theta}_0)\pi_0^N(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0\,. \qquad (2)$$

- M_0: reference model; M_ℓ: current model;
- $\boldsymbol{\theta}_\kappa$: parameter vector of model M_κ for $\kappa \in \{\ell, 0\}$;
- $\pi_\ell^N(\boldsymbol{\theta}_\ell)$: baseline prior of M_ℓ;
- $\pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*)$: posterior of $\boldsymbol{\theta}_\ell$ under M_ℓ with prior $\pi_\ell^N(\boldsymbol{\theta}_\ell)$;
- $m_\kappa(\boldsymbol{y}^*)$: marginal likelihood of M_κ; $\kappa \in \{\ell, 0\}$.

- M_0: set to the constant model.

3. Power-Expected-Posterior (PEP) Priors

Fouskakis et al. (2015):

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

$$\Downarrow \qquad \Downarrow \qquad \Downarrow$$

$$\pi_\ell^{PEP}(\boldsymbol{\theta}_\ell; \delta) = \int \pi_\ell^N(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*, \delta)\, m_0^N(\boldsymbol{y}^*|\delta)\, d\boldsymbol{y}^* \qquad (3)$$

we substitute the likelihood terms with powered-versions of the likelihoods (i.e. they are raised to the power of $1/\delta$).

PEP priors solve the following problems:

- Dependence of training sample size.
- Sensitivity to the selection of specific sub-samples.
- The prior is informative for models with $p \to n$.

Features of PEP

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

For Normal models

- The PEP prior (Fouskakis et al., 2015)
  - 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 and Pericchi, 2016).
- The Power-conditional-expected-posterior (PCEP) prior (Fouskakis and Ntzoufras, 2016) is similar to the g-prior with (i) more complicated variance structure, (ii) more dispersed and (iii) more parsimonious than the g-prior.
- PEP and PCEP $\Rightarrow$ consistent variable selection methods.

4. Extension to Generalized Linear Models

4.1 Definitions of the power-likelihood

Density-normalized power likelihood

For the **Normal case**, the definition of the power-likelihood seems quite clear via the normalization of the power likelihood:

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

since

$$f(y|\mu, \sigma^2) \sim N(\mu, \sigma^2) \Rightarrow f(y|\mu, \sigma^2, \delta) = N(\mu, \delta\,\sigma^2).$$

This is not the case for all distributions in the exponential family and hence for GLMs.

Alternative definitions of the power-likelihood

In the PEP representation (3), consider **the unormalized power-likelihood** and then normalize the posterior (which is also the approach in Ibrahim and Chen (2000). 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 $m_0^N(\boldsymbol{y}^*|\delta)$? $\Rightarrow$ **Two alternatives:**

(a) Diffuse Reference PEP (DR-PEP)

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

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

(b) Concentrated Reference PEP (CR-PEP)

Consider the **original likelihood** (without introducing any further uncertainty) with the marginal likelihood $m_0(\boldsymbol{y}^* \mid \delta)$ given by (2).

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

4.2 Comparisons in the normal case

- **DR-PEP prior** is the same with the original PEP prior.
- Both DR and CR-PEP priors are consistent.

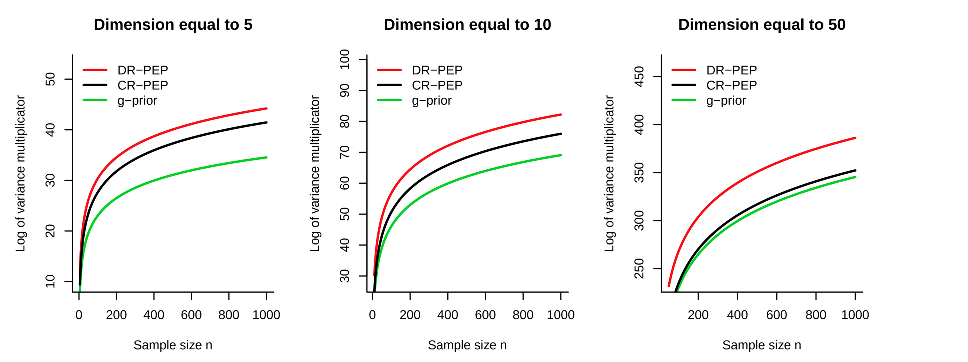

Figure 1: *Log-variance multipliers of the DR-PEP, CR-PEP and g-priors versus sample size for $d_\ell = 5, 10, 50$.*

4.3 A Gibbs based Variable Selection Sampler

We use an MCMC scheme with a full data augmentation:

- For each model $\boldsymbol{\gamma}$, we introduce a complement of $\boldsymbol{\beta}_\gamma$ denoted by $\boldsymbol{\beta}_{\setminus\gamma}$ for all coefficients not included in the model.
- A pseudoprior $\pi_\gamma(\boldsymbol{\beta}_{\setminus\gamma})$ is used as a proposal.
- Linear predictor: $\eta_i = \sum_{j=0}^{p} X_{ij}\gamma_j b_{\gamma,j}$;

  $b_{\gamma,j}$ is the element of $\boldsymbol{b}_\gamma = (\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{\setminus\gamma})$ for each X_j.
- A latent variable β_0 to represent the parameter(s) of model M_0.
- A latent vector of imaginary data $\boldsymbol{y}^*$.
- A Gibbs based variable selection algorithm is built on the augmented posterior

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

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

- We use Laplace approximation to evaluate the integral in the denominator.
- Jeffreys prior as a baseline: $\pi_\gamma^N(\boldsymbol{\beta}_\gamma) \propto |\mathbf{X}_\gamma^T W(\boldsymbol{\beta}_\gamma)\mathbf{X}_\gamma|^{1/2}$.

5. Hyper-delta PEP priors

Similarly to the hyper-g (Liang et al., 2008), the hyper-delta prior which introduces the prior

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

- 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 $n = n^*$.
- Another alternative option would be a hyper-δ/n prior.
- One additional step in the MCMC for δ. Use Metropolis-Hastings with proposal δ' from $q(\delta'|\delta) = \text{Gamma}(\delta, 1)$.

The hyper-δ DR-PEP prior can be approximated by

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

where $\widehat{\boldsymbol{\beta}}_\gamma^*$ is the MLE for $\boldsymbol{y}^*$ and $W_\gamma^* = W_\gamma(\widehat{\boldsymbol{\beta}}_\gamma^*)$.

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

6. Simulation study

Scenario	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: *Logistic and Poisson regression scenarios for Simulation Study 1 using independent ($r = 0$) and correlated predictors ($r = 0.75$).*

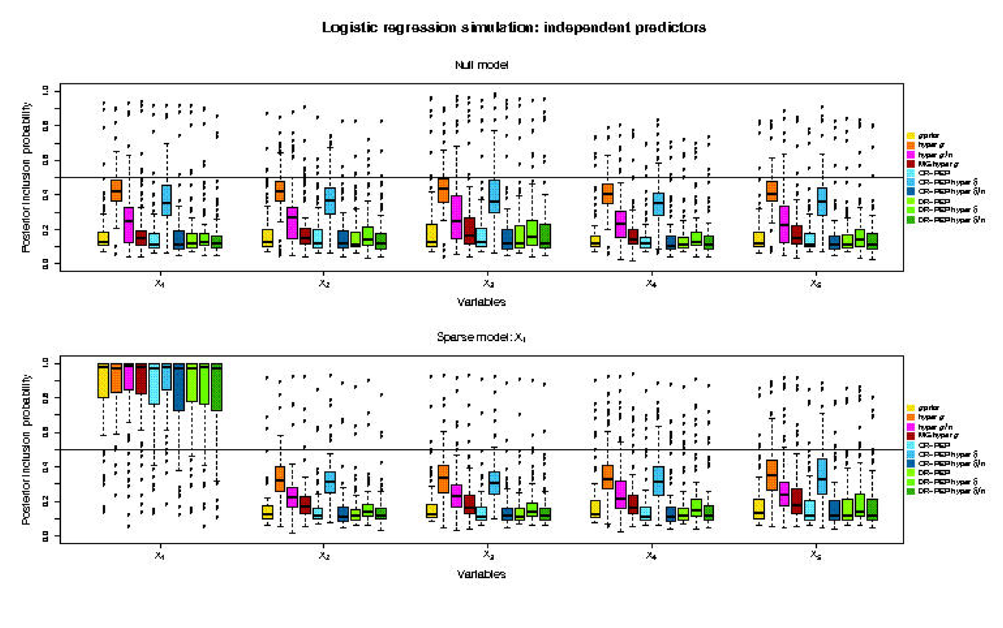

Figure 2: *Posterior inclusion probabilities for Simulation Study 1 from 100 replicated samples of the null, sparse, medium and full logistic regression model scenarios with correlated predictors ($r = 0.75$).*

Scenario	r	Prior distributions: g-prior	hyper g-prior	hyper g/n-prior	MG hyper g-prior	CR PEP	CR PEP hyper-δ	CR PEP hyper-δ/n	DR PEP	DR PEP hyper-δ	DR PEP hyper-δ/n
null	0.00	86	68	80	87	88	71	83	90	91	**94**
	0.75	91	68	90	94	95	75	91	95	**97**	95
sparse	0.00	75	74	74	75	76	68	**80**	73	68	69
	0.75	40	43	41	38	35	**44**	40	32	30	28
medium	0.00	29	43	37	36	27	**44**	30	28	25	20
	0.75	0	**5**	0	0	0	4	0	0	0	0
full	0.00	6	**23**	13	9	5	18	11	5	4	3
	0.75	0	0	1	0	0	**3**	0	0	0	0

Table 2: *Number of simulated samples (over 100 replications) that the MAP model coincides with the true model for the Poisson case in Simulation Study 1 (row-wise largest value in bold).*

7. Conclusions

- We extended the PEP prior formulation through the use of unnormalized power-likelihoods.
- They retain the features of the original PEP formulation.
- Hyper-δ and δ/n analogues of hyper-g priors do not suffer from the inflation of inclusion probabilities for non-important effects.
- DR-PEP seems that it is rather robust with respect to the fixed vs. random specification.

Acknowledgements – Funding Details

This work was supported by the Research Centre of the Athens University of Economics and Business (Funding program Action 2 for the support of basic research).

References

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

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 (forthcoming); arXiv:1307.2449 [stat.CO]* .

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., Ntzoufras, I. and Pericchi, L. R. (2016), 'On using sufficient statistics in (power) expected-posterior-priors for Bayesian model comparison', *(work in progress); early presentation is available at* http://www.math.ntua.gr/~fouskakis/Conferences/Cancun/Presentation_Cancun.pdf .

Fouskakis, D., Ntzoufras, I. and Perrakis, K. (2016), 'Power-Expected-Posterior Priors for Generalized Linear Models', *available at* http://stat-athens.aueb.gr/~jbn/papers/files/42_Fouskakis_Ntzoufras_Perrakis_2016_GLMPEP_v3.pdf .

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

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

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