Power-Expected-Posterior Priors in Generalized Linear Models

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

Available at `http://stat-athens.aueb.gr/~jbn/papers/obayes15.htm`.

Synopsis

1. From the expected-posterior prior (EPP) to the power-expected-posterior (PEP) prior
2. Alternative definitions of the power likelihood in PEP-priors
3. Implementing the method in GLMs (MCMC algorithm)
4. Illustrations
5. Using Mixtures of PEP priors
6. Illustrations (continued)
7. Discussion

1 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.

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 $\boldsymbol{y}^* = (y_1^*, \ldots, y_{n^*}^*)^T$ coming from a "suitable" predictive distribution $m^*(\boldsymbol{y}^*)$.

  Hence the EPP for the parameters of any model M_ℓ is

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

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

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. $m^*(\boldsymbol{y}^*) = m_0^N(\boldsymbol{y}^*)$.
- 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*).

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

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

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:

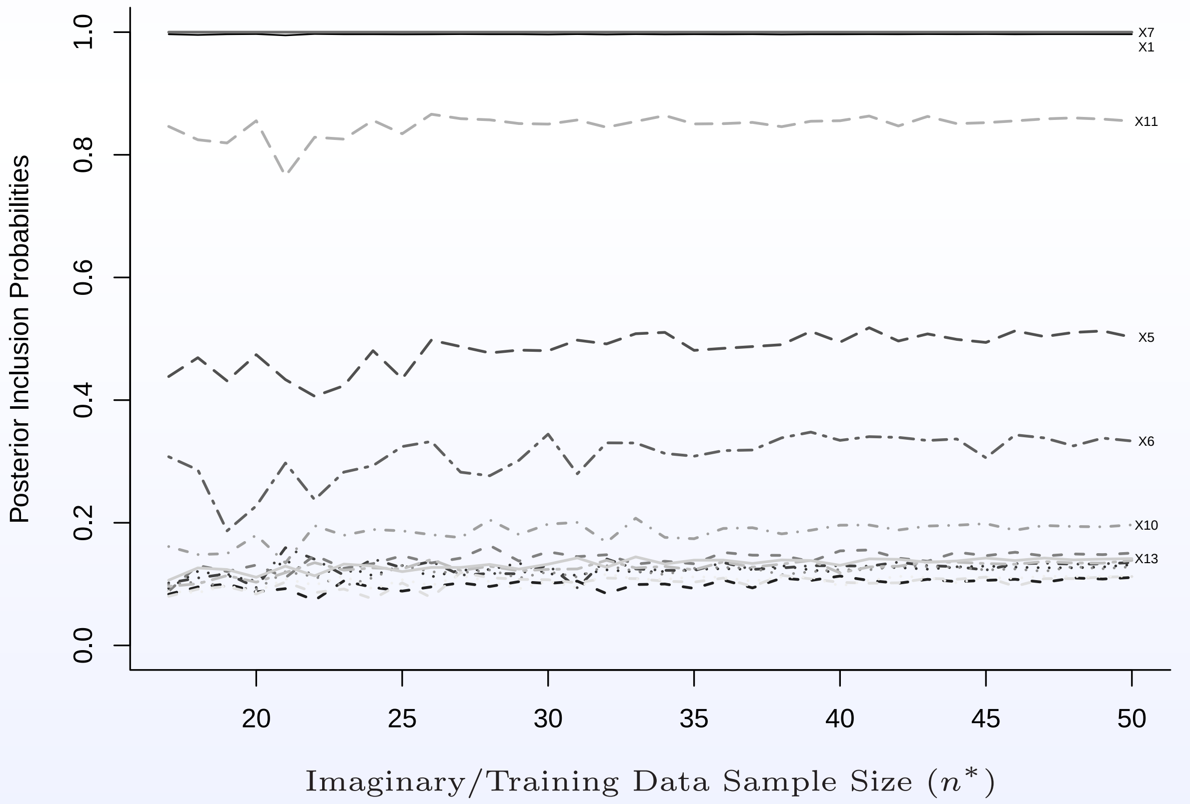
- Dependence of training sample size.
- Lack of robustness with respect to the sample irregularities.
- Excessive weight of the prior when the number of parameters is close to the number of data.

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 $X^*_\ell = X_\ell$; 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, *unpublished work, presented in ISBA2014*).
- The Power-conditional-expected-posterior (PCEP) prior (Fouskakis & Ntzoufras, 2015, *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.

2 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.

Definitions of the power-likelihood

Density-normalized power likelihoods in GLMs: May end up to a distribution which is not the same as the one in the original model formulation.

- In binary logistic regression $\Rightarrow$ power likelihood is still Bernoulli with success probability

$$\frac{\pi^{1/\delta}}{\pi^{1/\delta} + (1-\pi)^{1/\delta}}.$$

- This is not the case for the Binomial and the Poisson models resulting is some cumbersome distributions which increase computational complexity (without any obvious gain).

Alternative definitions of the power-likelihood

We consider the PEP representation

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

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.

Features of the diffuse-reference PEP

- Still has the interpretation of a posterior density given some imaginary data $\boldsymbol{y}^*$ "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 $\boldsymbol{y}^*$ "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

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

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

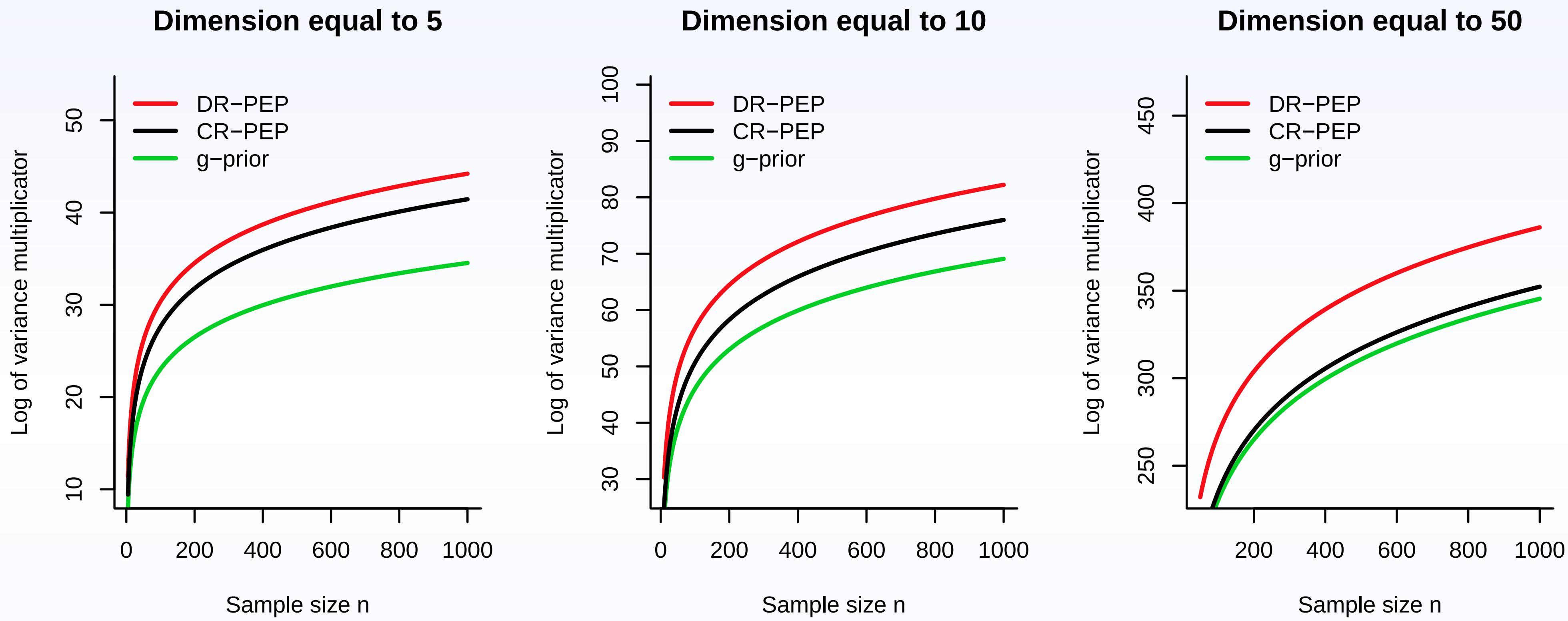


Figure 2: Log-variance multipliers of the DR-PEP, CR-PEP and g-priors versus sample size for $d_\ell = 5, 10, 50$.

Formulation in GLMs

- $M_\ell \rightarrow \boldsymbol{\gamma}$: Binary variable inclusion indicators ($\boldsymbol{\gamma}$) 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.
- $\mathbf{X}_{\boldsymbol{\gamma}}$ is the $n \times d_{\boldsymbol{\gamma}}$ data matrix for model $\boldsymbol{\gamma}$ with $d_{\boldsymbol{\gamma}} = \sum_{j=0}^{p} \gamma_j$ covariates.
- $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the parameter vector of length $d_{\boldsymbol{\gamma}}$ with the effects of each covariate
- The linear predictor vector is given by $\boldsymbol{\eta}_{\boldsymbol{\gamma}} = \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}$

We focus our presentation on the DR-PEP (computation is similar for the CR-PEP)

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

- 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

The posterior

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

The marginal likelihood

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

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

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

$$\propto \frac{f_{\gamma}(\boldsymbol{y}|\boldsymbol{\beta}_{\gamma})\Big[f_{\gamma}(\boldsymbol{y}^*|\boldsymbol{\beta}_{\gamma})f_0(\boldsymbol{y}^*|\boldsymbol{\beta}_0)\Big]^{1/\delta}}{\int f_{\gamma}(\boldsymbol{y}^*|\boldsymbol{\beta}_{\gamma})^{1/\delta}\pi_{\gamma}^{N}(\boldsymbol{\beta}_{\gamma})d\boldsymbol{\beta}_{\gamma}}\pi_{\gamma}^{N}(\boldsymbol{\beta}_{\gamma})\pi_{\gamma}^{N}(\boldsymbol{\beta}_{\setminus\gamma})\pi_{0}^{N}(\boldsymbol{\beta}_{0})\pi(\boldsymbol{\gamma})$$

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

The MCMC algorithm - Gibbs variable selection for PEP

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

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

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

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

Step 5: Sample $\boldsymbol{y}^*$ from the full conditional posterior (given $\boldsymbol{\beta}_\gamma$, 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 $m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\delta)$ and $m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^{*'}|\delta)$ which are computed by using Laplace approximation.

Details about the proposal for y_i^*

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

Binomial response: A product binomial proposal distribution is used with probability of success equal to

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

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

Poisson regression:

CR-PEP: A product Poisson proposal distribution is used with mean equal to $\lambda_i = \lambda_{0,i}\lambda_{\boldsymbol{\gamma},i}^{1/\delta}$, with $\lambda_{0,i} = \exp(\beta_0)$ and $\lambda_{\boldsymbol{\gamma},i} = \exp(\eta_{\boldsymbol{\gamma},i})$.

DR-PEP: The same strategy for DR-PEP failed and we have used a simple Poisson proposal with mean $\lambda_i = y_i^*$.

Further Remarks about the MCMC

- Metropolis-Hastings steps are not needed for $\boldsymbol{\beta}_{\setminus\gamma}$ and $\boldsymbol{\gamma}$
  - $\boldsymbol{\beta}_{\setminus\gamma}$ is sampled directly from the pseudo-prior distribution.
  - $\boldsymbol{\gamma}$ is sampled directly from the full conditional Bernoulli distribution.
- The pseudo-prior of $\boldsymbol{\beta}_{\setminus\gamma}$ 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 $\boldsymbol{\beta}_{\gamma}$ and β_0 (normal proposals based on MLEs).

Illustrative example 1: Simulated Binomial data

$n = 200$ binary responses with $p = 10$ potential covariates.

For $i = 1, \ldots, n$

$$\begin{aligned} X_{ij} &\sim N(0,1) \text{ for } j = 1, \ldots, 5 \\ X_{ij} &\sim N(0.3X_{i1} + 0.5X_{i2} + 0.7X_{i3} + 0.9X_{i4} + 1.1X_{i5}, 1) \text{ for } j = 6, \ldots, 10 \\ Y_i &\sim Bernoulli(p_i) \end{aligned}$$

Three scenarios:

Null: $logit(p_i) = 0.1$

Sparse: $logit(p_i) = 0.1 - 0.9X_{i3} + 1.2X_{i7} + 0.4X_{i10}$

Dense: $logit(p_i) = 0.1 + 0.6X_{i1} - 0.9X_{i3} + X_{i5} + 0.9X_{i6} + 1.2X_{i7} - 1.2X_{i8} - 0.5X_{i9}$

Scenario 1 (Null): True model = null

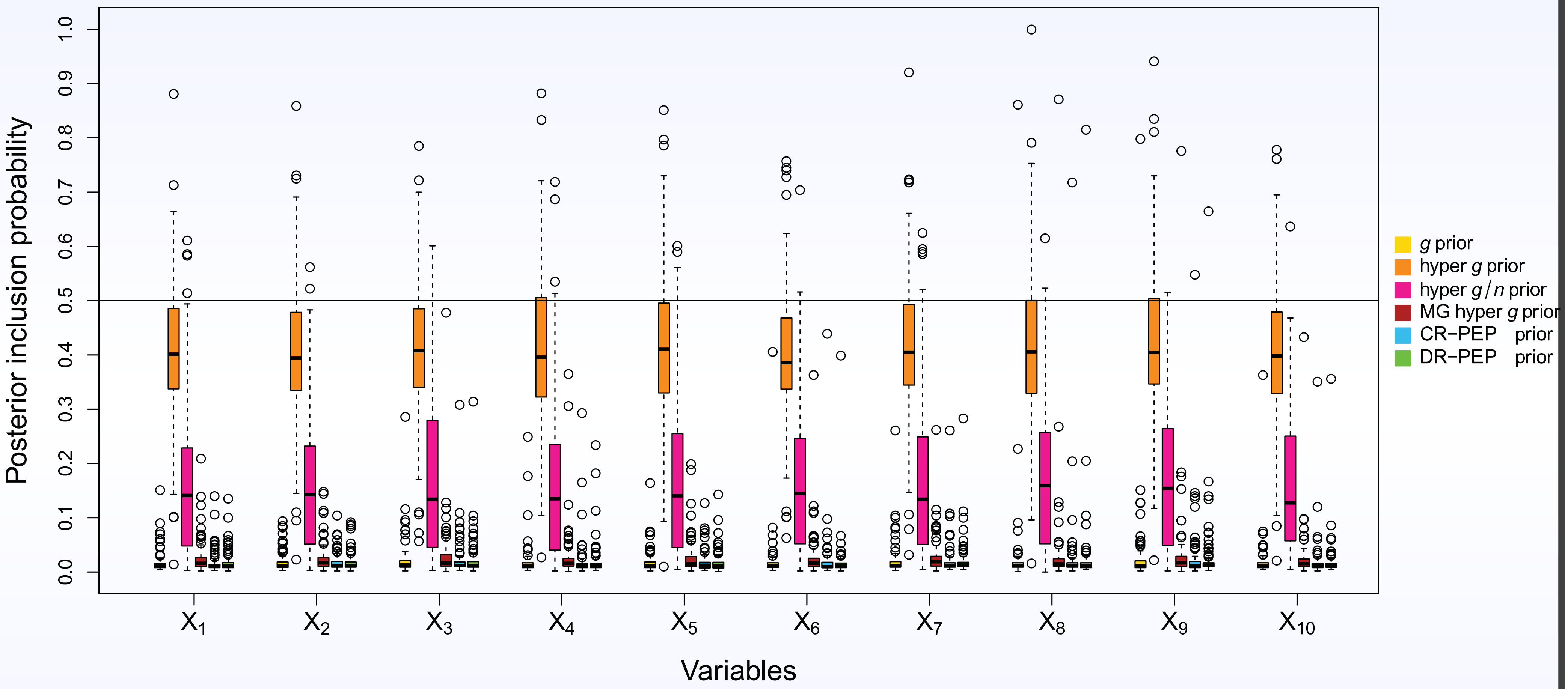


Figure 3: Posterior inclusion probabilities from 100 samples.

MG hyper-g prior: Maruyama & George (2011, Annals of Statistics) prior.

Scenario 2 (Sparse): True model = $0.1 - 0.9X_3 + 1.2X_7 + 0.4X_{10}$

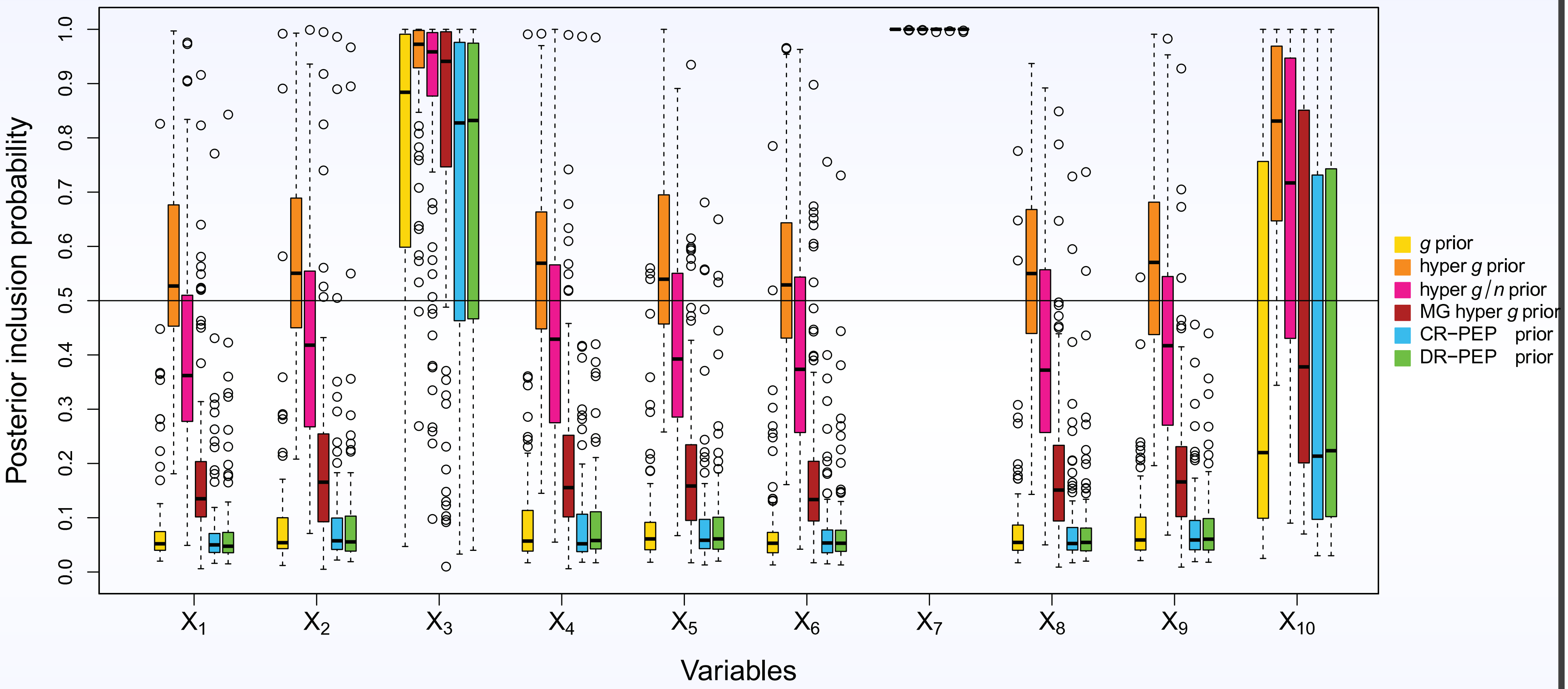


Figure 4: Posterior inclusion probabilities from 100 samples.

MG hyper-g prior: Maruyama & George (2011, Annals of Statistics) prior.

Scenario 3 (Dense)

True model $= 0.1 + 0.6X_1 - 0.9X_3 + X_5 + 0.9X_6 + 1.2X_7 - 1.2X_8 - 0.5X_9$

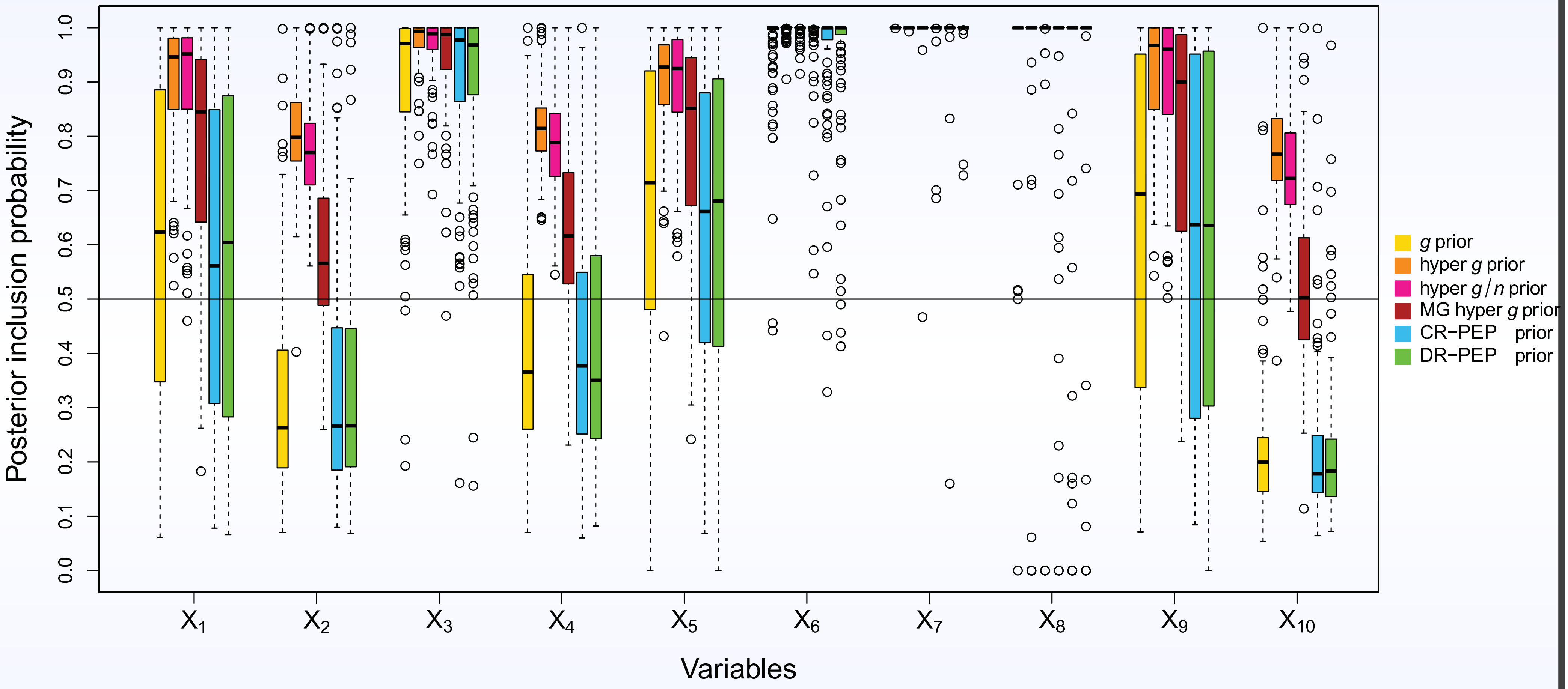


Figure 5: Posterior inclusion probabilities from 100 samples.

MG hyper-g prior: Maruyama & George (2011, Annals of Statistics) prior.

3 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

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

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 $\delta/(1+\delta)$

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

- 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 $\boldsymbol{\beta}_\gamma$, β_0, $\boldsymbol{y}^*$ and γ).

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

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

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

$$A_\delta = \Big\{ f_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\boldsymbol{\beta}_\gamma) f_0(\boldsymbol{y}^*|\beta_0) \Big\}^{\Delta\delta} \times \frac{\pi(\delta')}{\pi(\delta)} \times \frac{\widehat{m}^{\text{N}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\delta)}{\widehat{m}^{\text{N}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\delta')} \times \frac{q(\delta|\delta')}{q(\delta'|\delta)}.$$

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

4 Illustrative examples

- A real life example
- A Poisson simulated study

Illustrative example 2: 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.

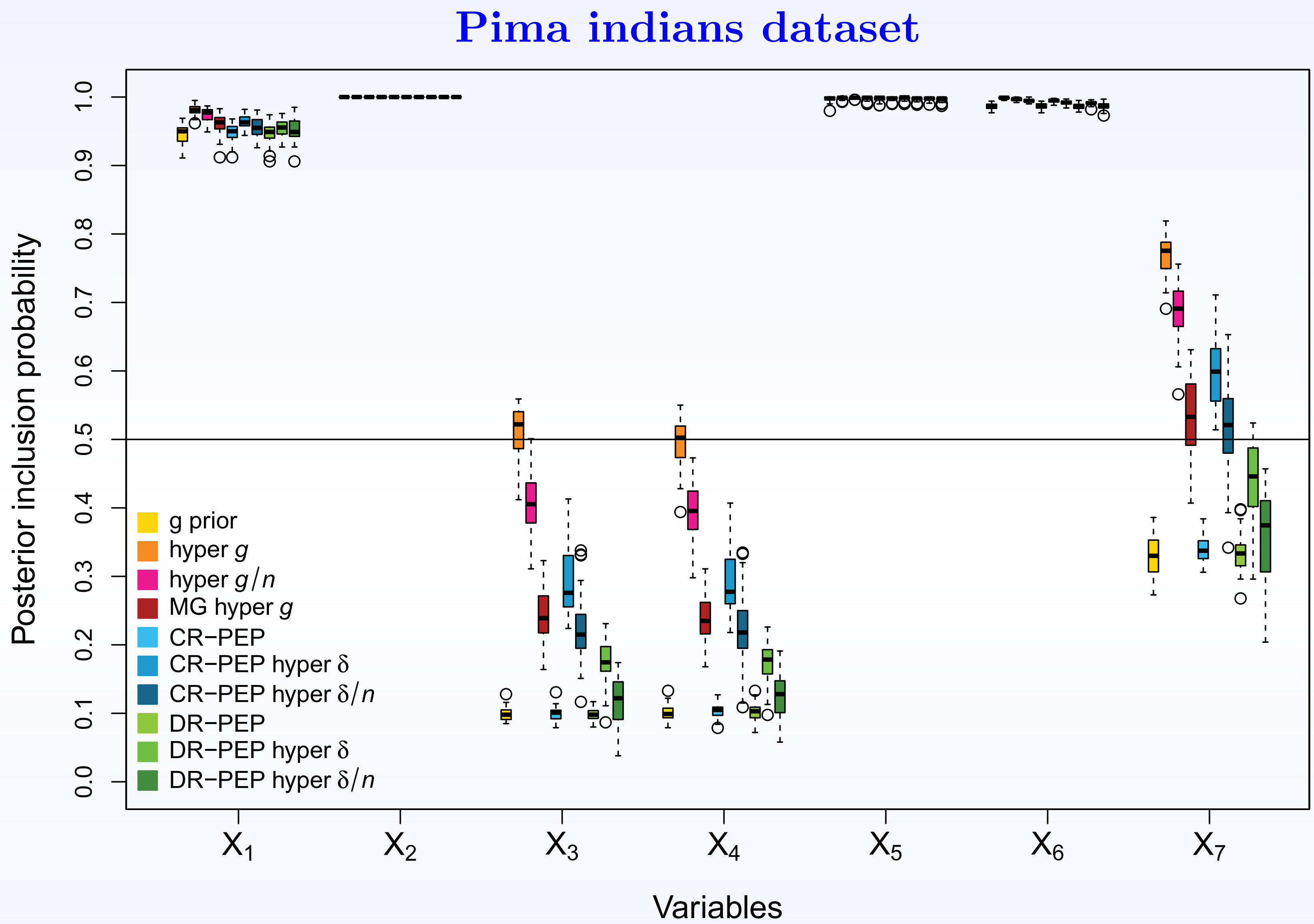


Figure 6: Boxplots of batched estimates of the posterior inclusion probabilities (40 batches of size 1000).

Pima indians dataset

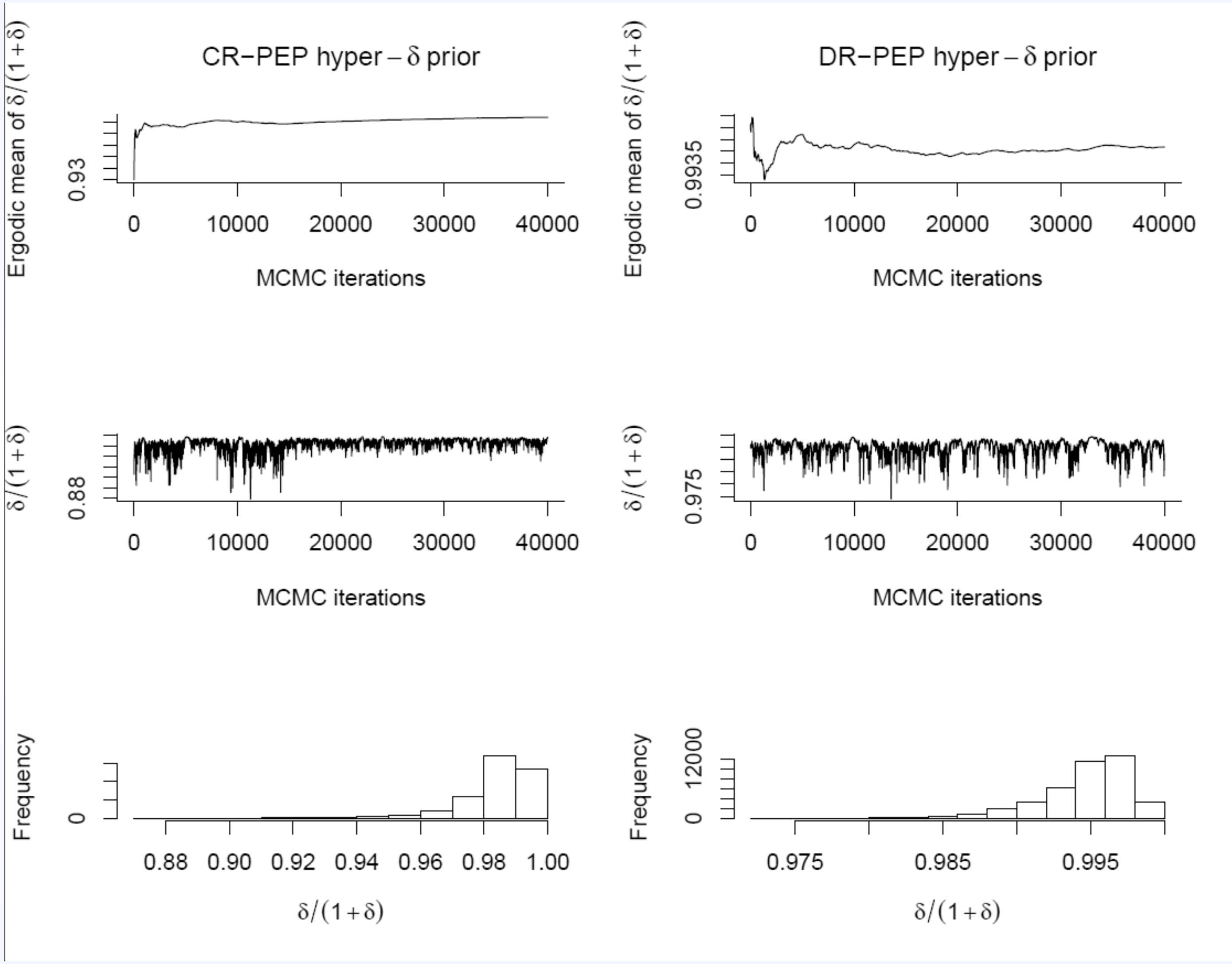


Figure 7: MCMC plots for the shrinkage factor $\delta/(1+\delta)$ for the CR-PEP and DR-PEP hyper-δ priors (40000 iterations).

Illustrative example 3: Poisson Simulated data

- 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

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

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.

Comments on the rates of identifying the true model

i) Variable selection methods using PEP priors perform well; 6 out of the 8 best MAP success patterns are observed in one of the PEP priors.

ii) Variable selection methods using PEP priors support more parsimonious models than the competing methods.

iii) For the null and the sparse scenarios, PEP priors perform overall better than the competing methods.

iv) For the medium model scenario, the PEP priors perform more or less equally well to the other methods.

v) When the true model is the full model

- All methods generally fail in the correlated scenario
- The CR-PEP hyper-δ and δ/n priors are performing generally well in comparison the the competing methods.
- The rest of the PEP priors have lower MAP success rates than the competing methods using hyper-g priors.

Current directions of research

- We are working to extend the consistency results for the GLMs setup
- **Main direction**: To extend the methodology in *large p, small n* problems.
- Use double exponential as baseline prior
  - g-prior type of behaviour when d_γ is smaller than n
  - LASSO type of shrinkage and behaviour when d_γ is bigger than n or we have extreme collinearities
- What about computation?

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

Concuding 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.


THE
END

Appendix A: Detailed description of the MCMC algorithm

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

$$O_j = \frac{f_{\boldsymbol{\gamma}_j^1}(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}_j^1})}{f_{\boldsymbol{\gamma}_j^0}(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}_j^0})} \times \left[\frac{f_{\boldsymbol{\gamma}_j^1}(\boldsymbol{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}_j^1})}{f_{\boldsymbol{\gamma}_j^0}(\boldsymbol{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}_j^0})}\right]^{1/\delta} \times \frac{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}_j^1}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_j^1})}{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}_j^0}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_j^0})} \times \frac{\pi_{\boldsymbol{\gamma}_j^1}(\boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}_j^1})}{\pi_{\boldsymbol{\gamma}_j^0}(\boldsymbol{\beta}_{\setminus \boldsymbol{\gamma}_j^0})} \times \frac{\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}_j^0}(\boldsymbol{y}^*|\delta)}{\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}_j^1}(\boldsymbol{y}^*|\delta)} \times \frac{\pi(\boldsymbol{\gamma}_j^1)}{\pi(\boldsymbol{\gamma}_j^0)}.$$

where

- $\boldsymbol{\gamma}_j^1 = (\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})$
- $\boldsymbol{\gamma}_j^0 = (\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})$
- $\boldsymbol{\gamma}_{\setminus j}$ is $\boldsymbol{\gamma}$ without element j

Step 2: Update $\boldsymbol{\beta}_\gamma$ from the full conditional posterior (given the current values of $\boldsymbol{\gamma}$ and $\boldsymbol{y}^*$) using a Metropolis step.

(a) Propose $\boldsymbol{\beta}_\gamma^{'}$ from the proposal distribution $q(\boldsymbol{\beta}_\gamma) = \mathrm{N}_{d_\gamma}(\widetilde{\boldsymbol{\beta}}_\gamma, \widetilde{\Sigma}_{\boldsymbol{\beta}_\gamma});$

– $\widetilde{\boldsymbol{\beta}}_\gamma$ is the MLE with response $(\boldsymbol{y}, \boldsymbol{y}^*)$ and weights $\mathbf{w} = (\mathbf{1}_n, \delta^{-1}\mathbf{1}_{n^*})$

– $\widetilde{\Sigma}_{\boldsymbol{\beta}_\gamma}$ the corresponding estimated variance-covariance matrix of $\widetilde{\boldsymbol{\beta}}_\gamma$.

(b) Accept the proposed values with probability $\alpha_{\boldsymbol{\beta}_\gamma} = \min\{1, A_{\boldsymbol{\beta}_\gamma}\}$; where $A_{\boldsymbol{\beta}_\gamma}$ is given by

$$A_{\boldsymbol{\beta}_\gamma} = \frac{f_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{\beta}_\gamma^{'})}{f_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{\beta}_\gamma)} \times \left[\frac{f_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\boldsymbol{\beta}_\gamma^{'})}{f_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\boldsymbol{\beta}_\gamma)}\right]^{1/\delta} \times \frac{\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_\gamma^{'})}{\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_\gamma)} \times \frac{q(\boldsymbol{\beta}_\gamma)}{q(\boldsymbol{\beta}_\gamma^{'})}.$$

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

Step 4: Sample β_0 from the full conditional posterior (given the current values of $\boldsymbol{y}^*$) using a Metropolis step.

(a) Propose $\beta_0^{'}$ from $q(\beta_0) = \mathrm{N}(\widehat{\beta}_0^*, \delta\widehat{\sigma}^2_{\widehat{\beta}_0^*});$

– $\widehat{\beta}_0^*$ is the MLE with response $\boldsymbol{y}^*$

– $\widehat{\sigma}_{\widehat{\beta}_0^*}$ is corrsponding standard error of $\widehat{\beta}_0^*$.

(b) Accept the proposed move with probability

$$\alpha_{\beta_0} = \min\left\{1,\ \left[\frac{f_0\big(\boldsymbol{y}^*|\beta_0'\big)}{f_0\big(\boldsymbol{y}^*|\beta_0\big)}\right]^{1/\delta} \times \frac{\pi_0^{\mathrm{N}}\big(\beta_0'\big)}{\pi_0^{\mathrm{N}}\big(\beta_0\big)} \times \frac{q\big(\beta_0\big)}{q\big(\beta_0'\big)}\right\}.$$

Step 5: Sample $\boldsymbol{y}^*$ from the full conditional posterior (given $\boldsymbol{\beta}_\gamma$, β_0 and γ) using a Metropolis step.

(a) Propose $\boldsymbol{y}^{*'}$ from $q(\boldsymbol{y}^*)$; see Slide 25 for details.

(b) Compute $\widehat{m}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{y}^*|\delta)$ and $\widehat{m}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{y}^{*'}|\delta)$ using Laplace approximation.

(c) Accept the proposed values with probability $\alpha_{\boldsymbol{y}^*} = \min\{1, A_{\boldsymbol{y}^*}\}$; where $A_{\boldsymbol{y}^*}$ is given by

$$A_{\boldsymbol{y}^*} = \left[\frac{f_{\boldsymbol{\gamma}}(\boldsymbol{y}^{*'}|\boldsymbol{\beta}_\gamma)}{f_{\boldsymbol{\gamma}}(\boldsymbol{y}^*|\boldsymbol{\beta}_\gamma)} \times \frac{f_0(\boldsymbol{y}^{*'}|\beta_0)}{f_0(\boldsymbol{y}^*|\beta_0)}\right]^{1/\delta} \times \frac{\widehat{m}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{y}^*|\delta)}{\widehat{m}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{y}^{*'}|\delta)} \times \frac{q(\boldsymbol{y}^*)}{q(\boldsymbol{y}^{*'})}.$$

Illustrative example 1: Simulated Binomial data (continued)

The following plots also include comparisons with hyper-delta PEP priors for example 1.

Scenario 1 (Null): True model = null

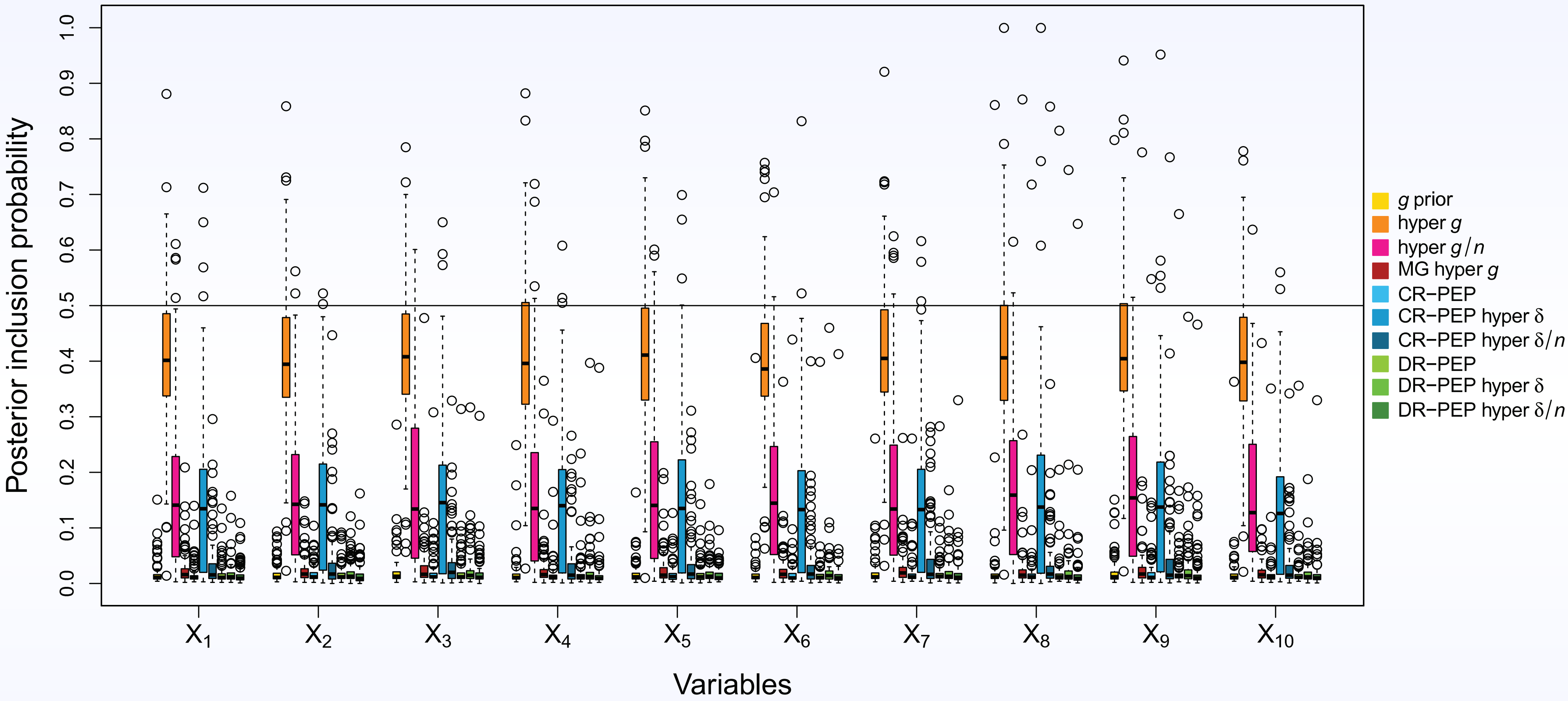


Figure 8: Posterior inclusion probabilities from 100 samples.

MG hyper-g prior: Maruyama & George (2011, Annals of Statistics) prior.

Scenario 2 (Sparse): True model $= 0.1 - 0.9X_3 + 1.2X_7 + 0.4X_{10}$

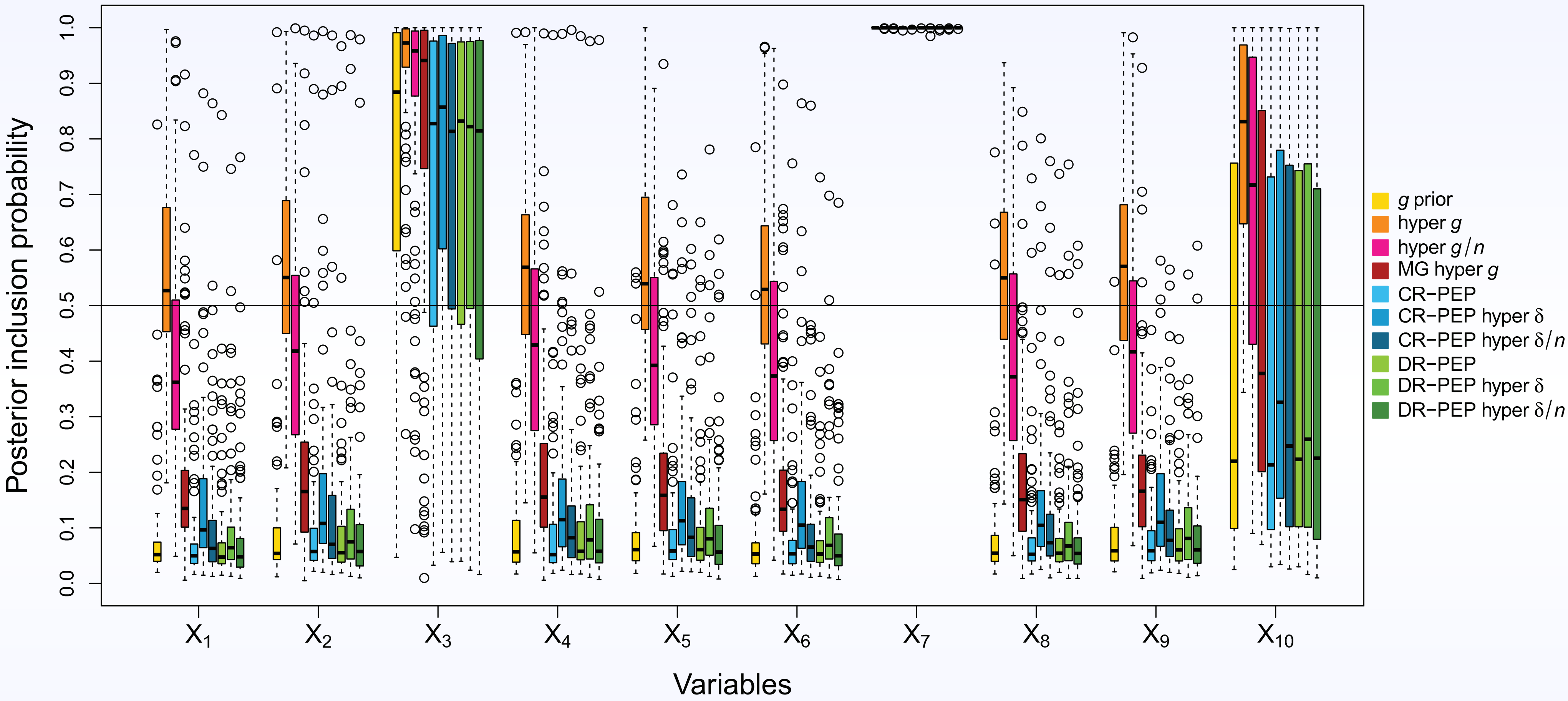


Figure 9: Posterior inclusion probabilities from 100 samples.

MG hyper-g prior: Maruyama & George (2011, Annals of Statistics) prior.

Scenario 3 (Dense)

True model $= 0.1 + 0.6X_1 - 0.9X_3 + X_5 + 0.9X_6 + 1.2X_7 - 1.2X_8 - 0.5X_9$

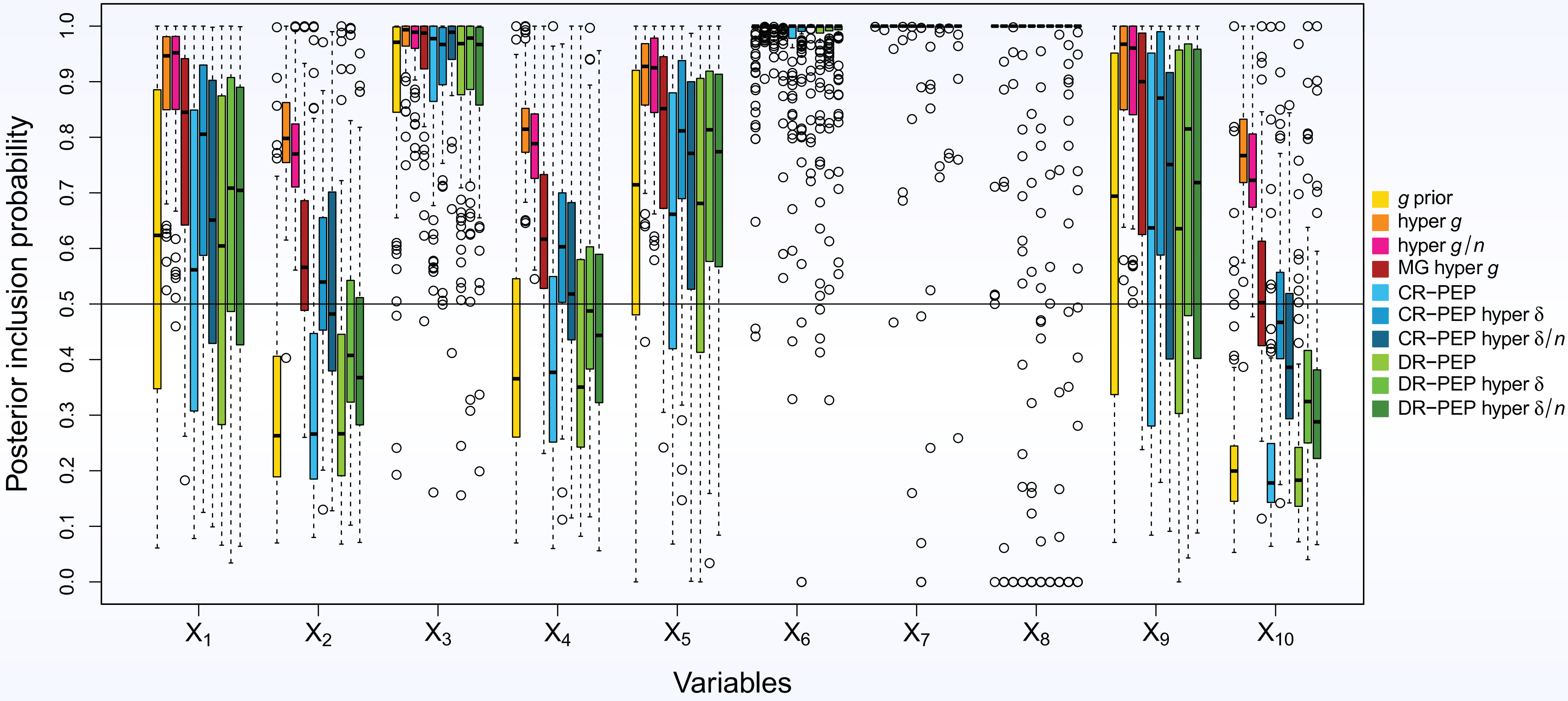


Figure 10: Posterior inclusion probabilities from 100 samples.

MG hyper-g prior: Maruyama & George (2011, Annals of Statistics) prior.

Illustrative example 3: Poisson Simulated data (continued)

Here, we will also find plots of posterior inclusion probabilities for each scenario of simulated example 3.

Independent Covariates

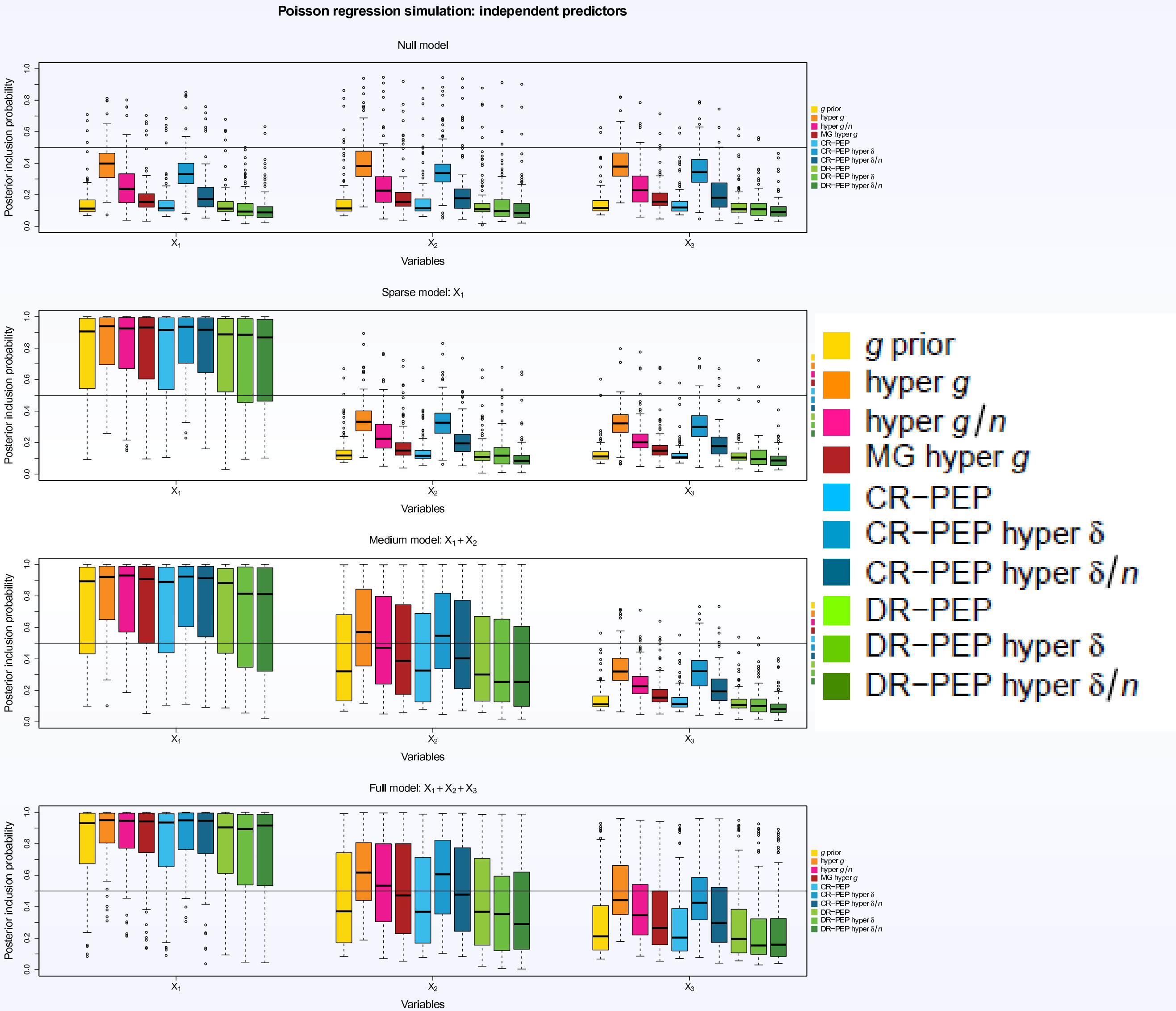


Figure 11: Posterior inclusion probabilities from 100 samples.

Correlated Covariates

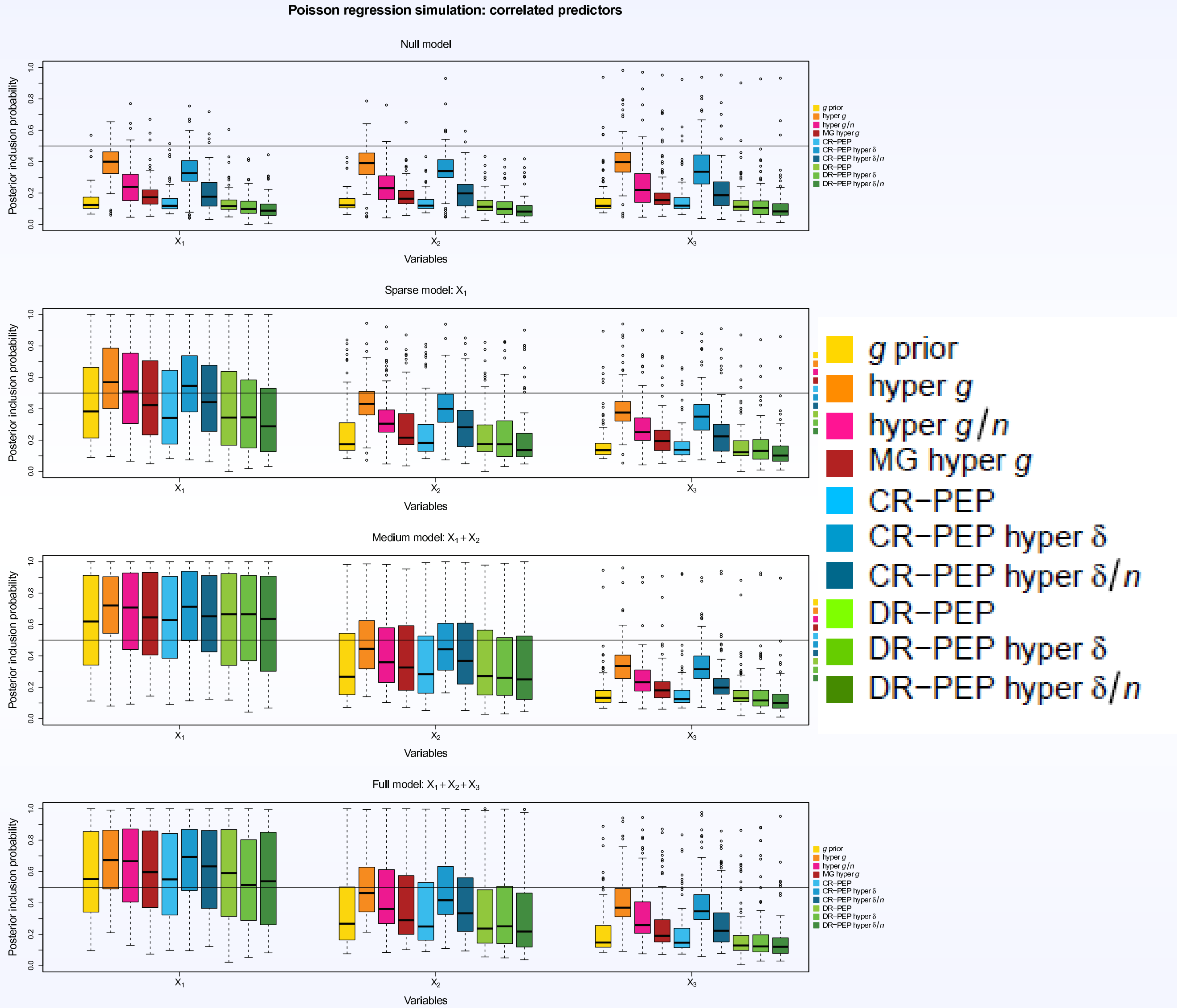


Figure 12: Posterior inclusion probabilities from 100 samples.