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

Ioannis Ntzoufras

Professor of Statistics

Department of Statistics, Athens University of Economics and Business

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

Università della Svizzera italiana

Lugano, 10 February 2016

Ioannis Ntzoufras: Preamble

1) Some details about myself.

2) Research interests.

3) Personal view on the field of ”data science”.

4) Personal contributions to data science.

5) The future of data science.

6) How the IDIDS can contribute to the future of data science.

Ioannis Ntzoufras: Basic Information

Studies

1. 1990–1994: B.Sc. in Statistics and Insurance Science, University of Piraeus, Greece.
2. 1994–1995: M.Sc. in Statistics with Applications in Medicine, Southampton University, UK.
3. 1995–1999: Ph.D. in Statistics (supervised by Professor Petros Dellaportas), Athens University of Economics and Business, Greece

Work Experience

1. 2004–today: Department of Statistics, Athens University of Economics and Business (Professor in Statistics from 22/12/2015).
2. 2000–2004: Lecturer in Quantitative Methods, Department of Business Administration, University of the Aegean.
3. February-June 2000: Temporary Lecturer in Statistics, Department of Statistics and Actuarial Science, University of the Aegean.

Ioannis Ntzoufras: Teaching Experience

List of taught courses (selection)

1. **Programming with** R: B.Sc. in Statistics, AUEB (2006–2016).
2. **Multivariate Statistics**: B.Sc. in Statistics, AUEB (2006–2016).
3. **Data Analysis**: B.Sc. in Statistics, AUEB (2006–2015).
4. **Biostatistics and Epidemiology**: B.Sc. in Statistics, AUEB (2006–2014).
5. **Advanced Data Analysis with** R: M.Sc. in Statistics; M.Sc. in Business Analytics, AUEB (2014–2016).
6. **Bayesian Modelling Using WinBUGS**: M.Sc. in Biostatistics, University of Athens (2002, 2004, 2006); M.Sc. in Statistics, AUEB (2010–2015); University of Pavia (2010, 2015; short courses); Msc Course in Statistics, La Sapienza University (2013); M.Sc. in Economics, Universita Cattolica del Sacro Cuore (2015); Ph.D. course in Statistics, University of Milano-Biccoca (2015).
7. **Short courses on Bayesian Variable Selection**: Herriot–Watt University (2009) University College Dublin (2011); AUEB (2015; 2nd Spring School on **R**).

Ioannis Ntzoufras: Areas of Expertise

- **Feature/Variable Selection** (model search algorithms, marginal likelihood estimates and objective priors) - 14 publications (in journals such that *Statistical Science, Bayesian Analysis, JRSSC, Annals of Applied Statistics, JCGS, Statistics and Computing* etc.).
- **Sport Analytics** (Predictive models for football, Measures of Competitive Balance, Performance Analysis in Voleyball) - 7 publications (in journals: *JRSSD, Journal of Quantitative Analysis in Sports, IMA Management Mathematics*).
- **Categorical data analysis and Graphical Models** – 6 publications (in journals: *Psychometrika, JCGS, CSDA, Soc.Methodology*).
- **Applied psychometric analysis** – 5 publications (in journals: *European Psychiatry, Psychiatry Research, Personality and Individual Differences, Schizophrenia research, Schizophrenia Bulletin*).
- **Bayesian methods for latent variable models** (model selection and estimation) – 3 publications (in journals: *Statistics and Computing, British J. of Math. and Stat. Psychology, J. Stat. Comp. Sim.*).
- **Other** – 8 publications (In journals: *Statistics in Medicine, J. of Stat. Software, Canadian J. of Statistics, North American Actuarial J.* etc.)

Personal view on the field of ”data science”

The fast development of computing facilities and of the web applications (mainly via e-shops and social media) emerged access to an increased amount of data leading to the problem of **BIG DATA**.

The need to make sense from massive data (sometimes instantly) has lead us to the necessity to redefine and unify sciences related to data to a new field called data science.

Data science includes

- Informatics and programming.
- Data mining and computer intelligence techniques.
- Statistics and Data Analysis.
- Data extraction, storage, handling and cleaning.
- Other Quantitative and Operational research techniques.
- Visualization and (written and oral) communication techniques.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data Science

- Data science is the study of extraction of knowledge from data with the goal of obtaining meaning from data and creating easy-to-access data products.
- Key word is science.

Why data science is considered to be only for big data?

- The access to web analytics data has lead to the need of scientists having skills from different areas of science. Being a statistician or programmer is not enough any more.
- We need new computational tools to extract information fast and apply meaningful statistical analysis.
- We need new statistical tools and methods to handle big-data problems.

But data science is not only for Big data!.

"There are a lot of small data problems that occur in big data. They dont disappear because youve got lots of the stuff. They get worse."

Professor David Spiegelhalter

Some other problems with Big-Data.

- Web-analytics data are usually biased since not all people (yet) use internet and mobile services.
- The problem of no insight/inference/understanding what is actually going on.
- False positives.
- Multiplicity and spurious correlations (big problem in feature selection in genetics)

Personal contributions to the field of Data Science

Feature Selection (main expertise): Prominent problem in the new era of Science. In Data Science it is called as the problem of *spurious correlations*.

Important Contributions in Feature Selection

- MCMC model search algorithms (Gibbs Variable Selection, RJMCMC)
- Objective Bayes techniques for selecting optimal number of predictors and models.
- Bayesian Lasso $\Rightarrow$ Specification of shrinkage parameter (without cross-validation).
- Avoiding the Bartlett-Lindley paradox by jointly defining priors on both model and parameter space.
- Variable selection accounting for data collection costs or cost restrictions.

Personal contributions to the field of Data Science

Current and Future Considerations for Feature Selection on Big-Data

- $n << p$ problems are also very important problem in this area.

  Current related research:

  - Adaptive model search algorithm for large model spaces (with C. Staerk and M. Kateri from Aachen University).
  - Extensions of PEP variable selection methodology by using LASSO type specifications (with K. Perrakis and D. Fouskakis).
  - Extending the EM variable selection method for general variable selection problems (future plans).

- **Large n small p** we are working to develop several techniques to handle such data which will directly applicable on sequential collection of data from web based applications.

Personal contributions to the field of Data Science

Contributions in Sport Analytics

- Building prediction models for football games using Poisson regression models.
- Analysing measures of competitive balance.
- Measuring the efficiency of each move in Voleyball (Performance analysis).

Contributions in Bayesian Latent Variable Models

- Studied the effect of data augmentation and latent variables on MCMC techniques.
- Developed methods for the efficient estimation of the marginal likelihood.

Postgraduate Teaching related with Data Science

- Teaching *Advanced Data Analysis with R* (36 hours course)
  - M.Sc. of Business Analytics, AUEB (2014–2015)
  - M.Sc. of Statistics, AUEB (2014–2015)
- *Short Introduction to Statistics and R* in the M.Sc. in Data Science, AUEB (2015).

The future of data science

1. Clear picture of the skills required by a Data Scientist.
2. New sources of data $\Rightarrow$ new problems.
3. New computing tools for analysing data.
4. New Statistical methods and models for big data which will help us interpret reality and causality and provide us answers replying to WHYs.
5. Sequential updating and incorporation of information coming from multiple sources.
6. More interaction between informatics and statistics leading to new areas of expertise and a unified data language.
7. Automation of procedures $\Rightarrow$ Making Data Science Useless?? (large discussion on the web)

How the IDIDS can contribute to the future of data Science

- Build solid statistical methods and theory for big data.
- Develop solid methodology for network analysis.
- New, modern, flexible M.Sc. in Data Science.
- Analysing large panel datasets of time series in Economics where dimensionality considerably grows.
- Combine global stock market data in network analysis.
- Develop pioneering methods for medical problems (with collaboration with the Center for Computational Medicine in Cardiology, CCMC, the Institute for Research in Biomedicine, IRB and the University Center for Statistics in the Biomedical Sciences, CUSSB, in Milan).
- Develop novel software for personalized services such as shopping or medicine.

How can I contribute to IDIDS

1. Develop powerful research team on Feature Selection methods in several fields.
2. Build a Sports Analytic research team with emphasis given in football data, performance analytics and on-line big data with ultimate aim the collaboration with professional teams and sport leagues.
3. To be involved in new areas of intriguing research such as network analysis.

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 — National Technical University of Athens

Department of Statistics — Athens University of Economics and Business

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

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

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:

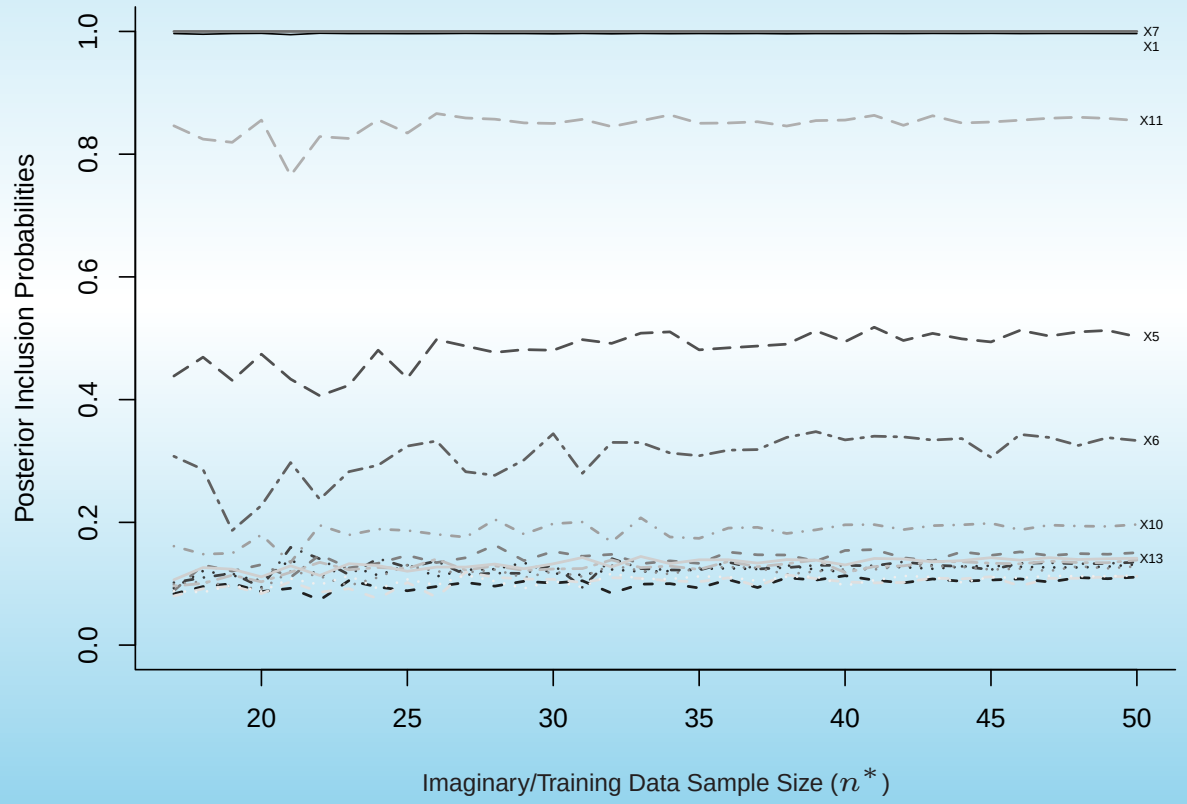
- Dependence of training sample size.
- Lack of robustness with respect to the sample irregularities.
- Excessive weight of the prior when the number of parameters is close to the number of data.

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, *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.

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

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

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

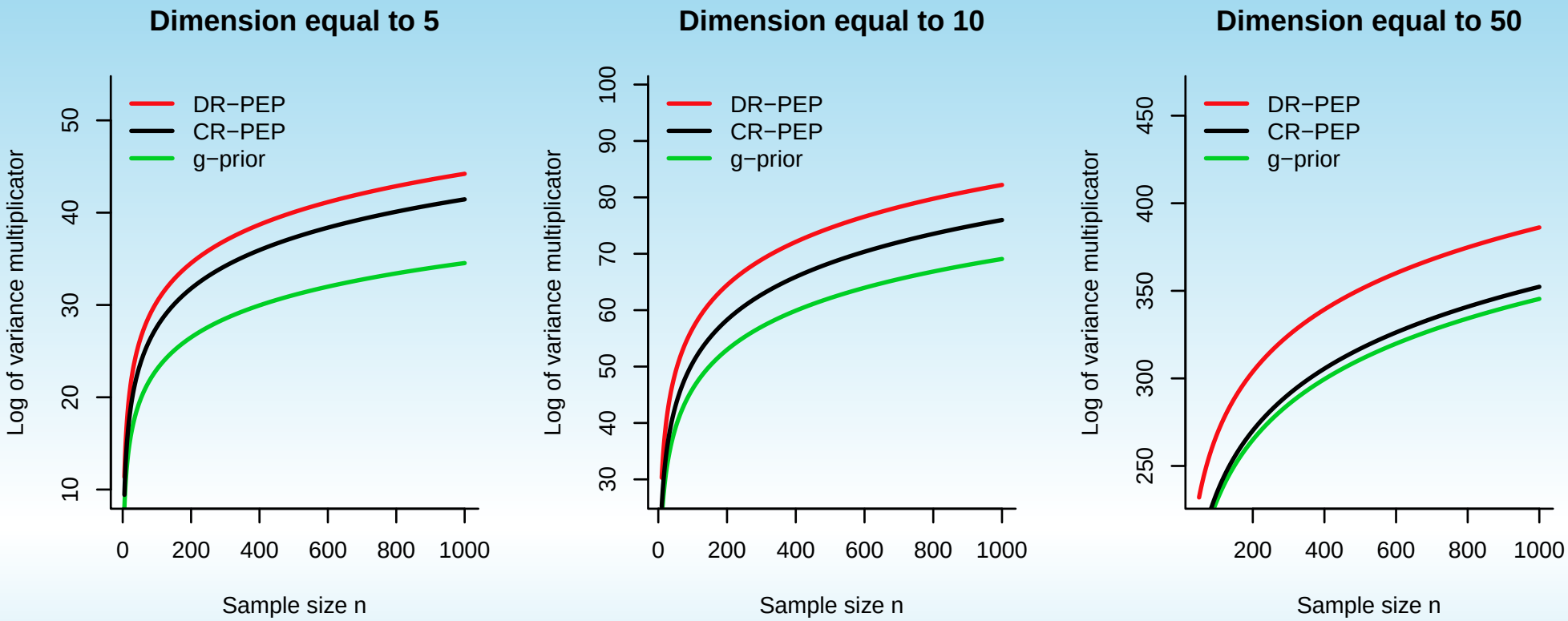


Figure 2: Log-variance multipliers of the DR-PEP, CR-PEP and g-priors versus sample size for $d_\ell = 5, 10, 50$.

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.

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

$$\pi_\gamma^{\mathrm{PEP}}(\boldsymbol{\beta}_\gamma) \approx \int\int f_{N_{d_\gamma}}\left(\boldsymbol{\beta}_\gamma; \widehat{\boldsymbol{\beta}^*_\gamma}, \delta\left(\mathbf{X}^{*T}_\gamma \mathbf{H}^*_\gamma \mathbf{X}^*_\gamma\right)^{-1}\right) m_0^{\mathrm{N}}(\boldsymbol{y}^*|\delta)\pi(\delta)\mathrm{d}\boldsymbol{y}^*\mathrm{d}\delta, \tag{3}$$

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

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

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

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 $\geq$ 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.


WHAT WE NEED TO BOOST SALES IS A VIRAL VIDEO. ON THE COUNT OF THREE, DO SOMETHING FUNNY.
TOM FISH BURNE
©marketoonist.com

Web-analytics Popularity Dataset

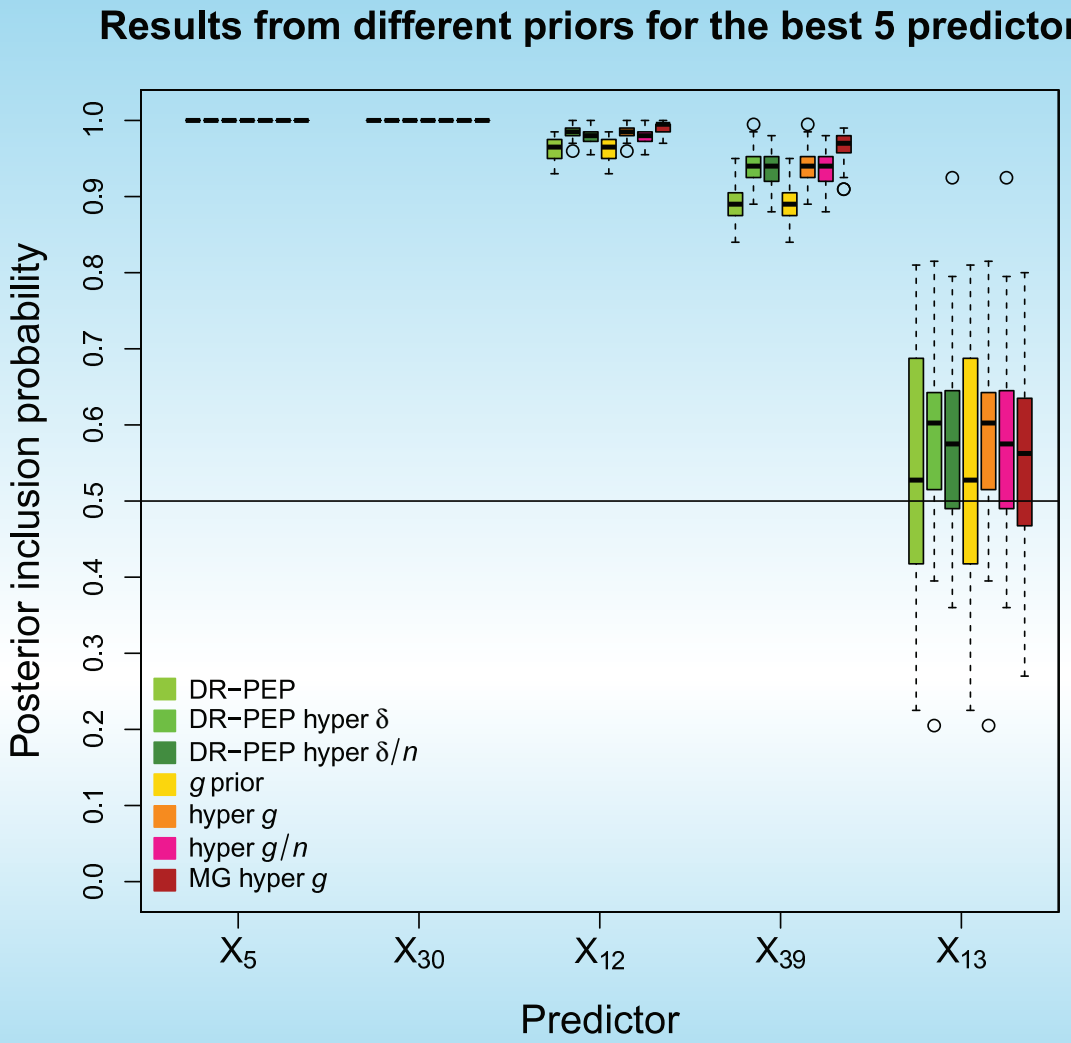


Figure 3: Comparison of Important Determinants of a Viral Post (40 batches).

Web-analytics Popularity Dataset

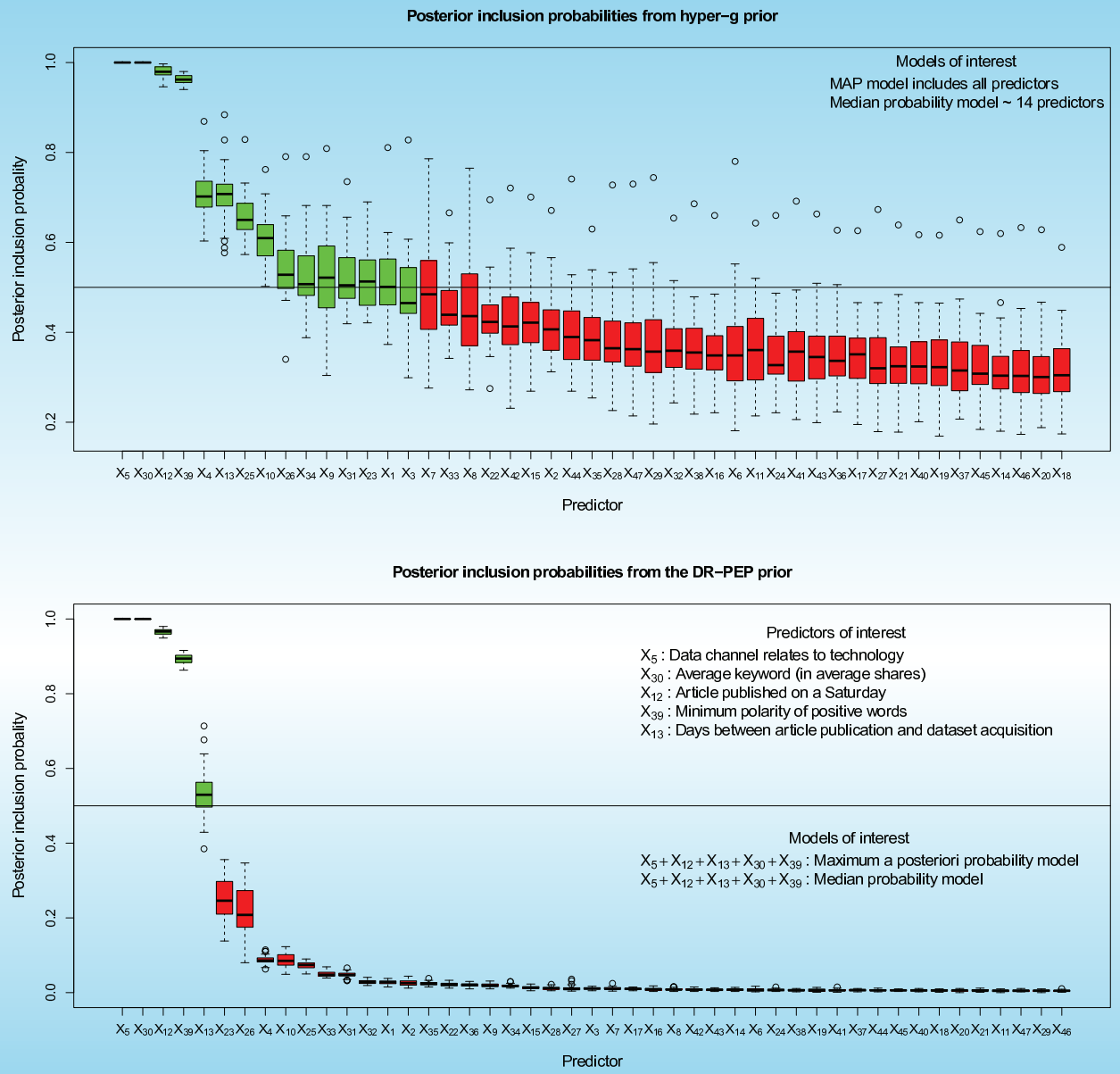


Figure 4: Posterior inclusion probabilities for hyper-g and DR-PEP (40 batches).

Web-analytics Popularity Dataset

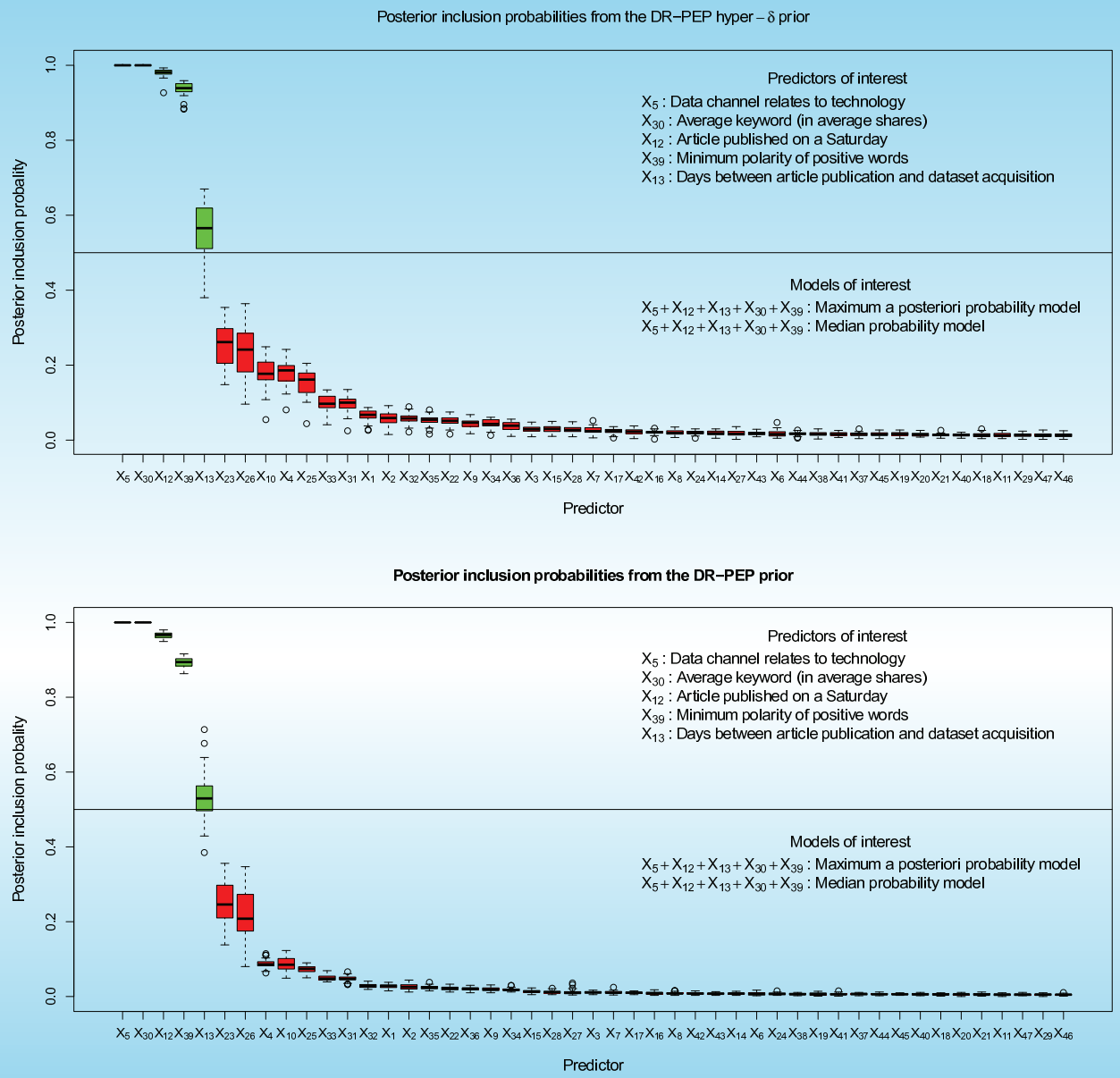


Figure 5: Posterior inclusion probabilities for hyper-g and DR-PEP (40 batches of size 200 iterations).

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

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

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.

	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 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.


THE END

Illustrative example 3: 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 3 (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 3: Potential predictors in the Pima Indians diabetes data set.

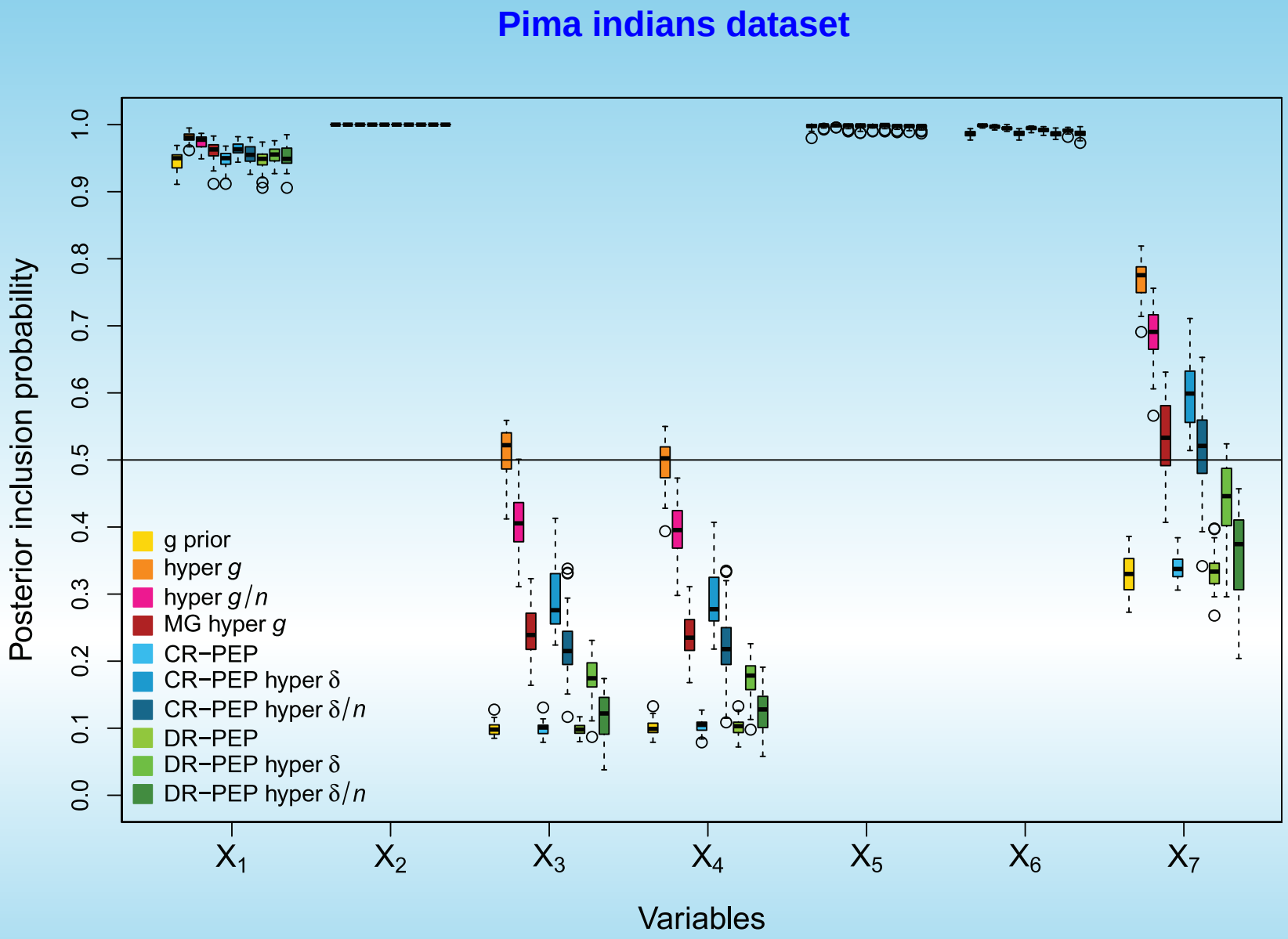


Figure 6: Boxplots of batched estimates of the posterior inclusion probabilities (40 batches of size 1000).