Bayesian Variable Selection for Generalized Linear Models Using the Power-Conditional-Expected Prior

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

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

[1]Department of Mathematics, National Technical University of Athens, Athens, Greece
fouskakis@math.ntua.gr

[2]Athens University of Economics and Business, Department of Statistics, Athens, Greece
<ntzoufras,kperrakis>@aueb.gr

Abstract

The Zellner's g-prior, and its extensions for generalized linear models (GLMs), is a popular choice in the variable selection context. This prior can be expressed as a power-prior with fixed set of imaginary data. We assign an extra hierarchical level that introduces uncertainty for the imaginary data under the g-prior design, by borrowing ideas from the power expected posterior priors. For variable selection on normal regression models, the resulting power-conditional-expected-posterior (PCEP) prior is a conjugate normal-inverse gamma prior, which provides a consistent variable selection method and gives more weight to parsimonious models than the Zellner's g-prior. Moreover, we will try to extent this methodology for GLMs and examine possible connections with hyper-g priors.

Keywords: power expected-posterior priors; objective model selection methods; power prior; training sample; unit-information prior.

1 The role of imaginary data in g-priors

Let us consider a set of imaginary data $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_{n^*}^*)^T$ of size n^*. Then, for any model m_ℓ with parameter vector $\boldsymbol{\theta}_\ell$, likelihood $f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell)$ and baseline prior $\pi_\ell^N(\boldsymbol{\theta}_\ell)$, we can obtain a "sensible" prior for the model parameters from

$$\pi_\ell(\boldsymbol{\theta}_\ell|\boldsymbol{y}^*;\delta) \propto f(\boldsymbol{y}^*|\boldsymbol{\theta}_\ell, m_\ell)^{1/\delta}\pi_\ell^N(\boldsymbol{\theta}_\ell)\,. \tag{1}$$

This is the power-prior introduced by Ibrahim and Chen [3]. The parameter δ controls the weight that the imaginary data contribute to the "final" posterior distribution of $\boldsymbol{\theta}_\ell$. For $\delta = 1$, (1) is exactly equal to the posterior distribution of $\boldsymbol{\theta}_\ell$ after observing the imaginary data $\boldsymbol{y}^*$. For $\delta = 1/n^*$ the contribution of the imaginary data to the overall posterior is equal to one data point; i.e. a prior having a unit-information interpretation [4].

We focus on variable selection problems for GLM's, i.e. for any model m_ℓ with parameters $\boldsymbol{\theta}_\ell = (\beta_0, \boldsymbol{\beta}_\ell\,, \phi)$ and response data $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ with likelihood given

by

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\beta}_\ell\,,\phi) &= \exp\left(\sum_{i=1}^{n}\frac{y_i\vartheta_i - b(\vartheta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i,\phi)\right), \\ \vartheta_i &= g\circ b'^{-1}\Big(\beta_0 + \boldsymbol{X}_{(i)}\boldsymbol{\beta}_\ell\Big) \end{aligned}$$

where X_ℓ is a $n \times d_\ell$ design matrix and $g \circ b'^{-1}()$ is the inverse function of $g \circ b'(\vartheta) = g\big(b'(\vartheta)\big)$, ϑ_i and ϕ are the location and dispersion parameters of the exponential family, respectively, $a()$, $b()$, $c()$ are functions specifying the structure of the distribution, and $g()$ is the link function connecting the mean of Y_i with the linear predictor.

Under the power-prior approach for the regression coefficients $\boldsymbol{\beta}_\ell$ given β_0, ϕ, with $\pi_\ell^N(\boldsymbol{\beta}_\ell|\beta_0,\phi) \propto 1$;

$$\pi_\ell(\boldsymbol{\beta}_\ell|\beta_0,\phi,\boldsymbol{y}^*;\delta) \approx f_{N_{d_\ell}}\Big(\boldsymbol{\beta}_\ell\,;\,\widehat{\boldsymbol{\beta}}_\ell^*\,,\delta\big(\mathrm{X}_\ell^{*T}\boldsymbol{H}_\ell\mathrm{X}_\ell^*\big)^{-1}\Big),$$

where $\widehat{\boldsymbol{\beta}}_\ell^*$ is the MLE of $\boldsymbol{\beta}_\ell$ for data $\boldsymbol{y}^*$, $f_{N_d}(\boldsymbol{y}\,;\,\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the d-dimensional normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{H} = \mathrm{diag}(h_i)$ with $h_i^{-1} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^2 \alpha(\phi_i) b''(\vartheta_i)$.

Assuming that $\boldsymbol{y}^* = g^{-1}(0)\mathbf{1}_n$ and that $\alpha(\phi_i) = \phi/w_i$ (with w_i being a known fixed weight), we have

$$\pi_\ell(\boldsymbol{\beta}_\ell|\sigma^2,\boldsymbol{y}^*;\delta) \approx f_{N_{d_\ell}}\Big(\boldsymbol{\beta}_\ell\,;\,\mathbf{0}\,,\delta c\big(\mathrm{X}_\ell^{*T}\boldsymbol{W}_\ell\mathrm{X}_\ell^*\big)^{-1}\phi\Big),$$

where $\boldsymbol{W} = \mathrm{diag}(w_i)$ and $c = \left[\frac{\partial g(\mu_i)}{\partial \mu_i}\big|_{\mu_i=0}\right]^2 b''\big(g\circ b'^{-1}(0)\big)$. Thus, Zellner's g-prior can be interpreted as a power-prior on imaginary data.

2 Power-conditional-expected posterior (PCEP) priors

We implement the PCEP prior for the regression coefficients $\boldsymbol{\beta}_\ell$ conditionally on β_0 and ϕ; see for details in [1]. The conditional expected posterior (CEP) prior is

$$\pi_\ell^{CEP}(\boldsymbol{\beta}_\ell,\beta_0,\phi) = \pi_\ell^{CEP}(\boldsymbol{\beta}_\ell|\beta_0,\phi)\pi_\ell^N(\beta_0|\phi)\pi_\ell^N(\phi) \tag{2}$$

with

$$\pi_\ell^{CEP}(\boldsymbol{\beta}_\ell|\beta_0,\phi) = \int \pi_\ell^N(\boldsymbol{\beta}_\ell|\beta_0,\phi,\boldsymbol{y}^*)m_0^N(\boldsymbol{y}^*|\beta_0,\phi)d\boldsymbol{y}^*. \tag{3}$$

This prior is actually the same as the expected posterior prior of Perez and Berger [5] conditional on β_0 and ϕ. Similar to the power-expected posterior prior [2], we construct the PCEP prior by raising the likelihood, involved in (3), to a power $1/\delta$ that controls the effect of the training sample in the PCEP prior. Thus,

$$\begin{aligned} \pi_\ell^{PCEP}(\boldsymbol{\beta}_\ell,\beta_0,\phi;\delta) &= \pi_\ell^{PCEP}(\boldsymbol{\beta}_\ell|\beta_0,\phi;\delta)\pi_\ell^N(\beta_0,\phi) \\ &= \left[\int \pi_\ell^N(\boldsymbol{\beta}_\ell|\beta_0,\phi,\boldsymbol{y}^*;\delta)m_0^N(\boldsymbol{y}^*|\beta_0,\phi;\delta)d\boldsymbol{y}^*\right]\pi_\ell^N(\beta_0,\phi), \end{aligned} \tag{4}$$

where

$$\pi_\ell^N(\boldsymbol{\beta}_\ell|\beta_0,\phi,\boldsymbol{y}^*;\delta) = \frac{f(\boldsymbol{y}^*|\,\boldsymbol{\beta}_\ell\,,\beta_0,\phi;\delta)\pi_\ell^N(\boldsymbol{\beta}_\ell\,|\,\beta_0,\phi)}{m_\ell^N(\boldsymbol{y}^*|\,\beta_0,\phi;\delta)} \tag{5}$$

with $f(\boldsymbol{y}^*|\,\boldsymbol{\beta}_\ell\,,\beta_0,\phi;\delta) \propto f(\boldsymbol{y}^*|\boldsymbol{\beta}_\ell\,,\beta_0,\phi)^{1/\delta}$ being the density-normalized power likelihood. Moreover, $m_\ell^N(\boldsymbol{y}^*|\beta_0,\phi;\delta)$ is the prior predictive distribution, evaluated at $\boldsymbol{y}^*$, of model m_ℓ given β_0,ϕ with the power likelihood under the baseline prior $\pi_\ell^N(\boldsymbol{\beta}_\ell\,|\,\beta_0,\phi)$.

Hence, $\pi_\ell^{PCEP}(\boldsymbol{\beta}_\ell,\beta_0,\phi;\delta)$ in (4) can be considered as a mixture of g-priors where a hyper-prior is now placed on the random imaginary data $\boldsymbol{y}^*$ rather than the variance multiplicator g (which is here substituted by δ). The method is more parsimonious than the hyper-g and the g-prior and it is illustrated using a simple example.

3 Illustration: The crime data-set

Here, we provide results for a normal regression example where the PCEP results in a conjugate Normal-inverse-gamma set-up [1]. The data concern crime rates for 47 states and include 15 explanatory variables [6]. The response variable is the rate of crimes in a particular category per head of population. All variables, including the response and excluding the indicator covariate (X_2), have been initially log-transformed and then all variables have been centered.

With $p = 15$ covariates we were able to contact a full enumeration search. Posterior marginal inclusion probabilities, are presented in Table 1. All four methods (PCEP, BIC, g-prior, hyper-g) give approximately equal support to the most prominent covariates, while for the remaining ones the posterior inclusion probabilities are lower under the PCEP.

Finally, we have compared the split-half out-of-sample predictive performance of the MAP models and the full model using RMSE. All models achieved similar predictive performance with difference not large enough to infer in favor of the superiority of one model (and therefore of a method). Since the model indicated as MAP by PCEP was more parsimonious, we claim that for this example PCEP achieves similar predictive

Table 1: Posterior marginal inclusion probabilities for the Crime data

	Variables (log scale)	PCEP	BIC	Zellner's g-prior	Hyper-g prior ($\alpha = 3$)
X_1	Percentage of males aged 14-24	0.828	0.909	0.850	0.843
X_2	Indicator variable for a Southern state	0.193	0.229	0.231	0.295
X_3	Mean years of schooling	0.974	0.992	0.978	0.967
X_4	Police expenditure in 1960	0.664	0.687	0.665	0.662
X_5	Police expenditure in 1959	0.402	0.404	0.422	0.465
X_6	Labour force participation rate	0.120	0.161	0.157	0.226
X_7	Number of males per 1000 females	0.124	0.168	0.160	0.228
X_8	State population	0.287	0.359	0.330	0.385
X_9	Number of non-whites per 1000 people	0.632	0.776	0.679	0.686
X_{10}	Unemployment rate of urban males 14-24	0.165	0.226	0.208	0.272
X_{11}	Unemployment rate of urban males 35-39	0.558	0.696	0.600	0.608
X_{12}	Gross domestic product per head	0.256	0.363	0.312	0.377
X_{13}	Income inequality	0.997	0.999	0.997	0.995
X_{14}	Probability of imprisonment	0.872	0.946	0.896	0.889
X_{15}	Average time served in state prisons	0.278	0.409	0.333	0.382

performance using a lower number of covariates.

4 Discussion

The PCEP prior for normal regression models is a conjugate normal-inverse-gamma prior, resulting in a variable selection procedure with similar large sample properties to BIC, supporting more parsimonious models than the approach using g-prior or hyper-g prior. For the rest of the GLMs more advanced MCMC methods and/or Laplace based approximating techniques will be implemented. Further extension of the method using a hyperprior on δ will be also investigated in the future.

Acknowledgements/Funding details

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.

References

[1] Fouskakis, D., Ntzoufras, I. (2013). "Power-conditional-expected priors: Using g-priors with random imaginary data for variable selection." arXiv:1307.2449 [stat.CO] (submitted).

[2] Fouskakis, D., Ntzoufras, I., Draper, D. (2014). "Power-expected-posterior priors for variable selection in Gaussian linear models." *Bayesian Analysis*, forthcoming.

[3] Ibrahim, J., Chen, M. (2000). "Power prior distributions for regression models." *Statistical Science*, **15**, 46–60.

[4] Kass, R., Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, **90**, 928–934.

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

[6] Vandaele, W. (1978). "Participation in illegitimate activities: Ehrlich revisited." In *Bayesian Statistics*, pages 270–335. Washington, DC: U.S. National Academy of Sciences.