Please cite this article as:
Paweł Kopciuszewski, Mixture methods of priori distribution choice in a binary response case, Scientific Research of the Institute of Mathematics and Computer Science, 2005, Volume 4, Issue 1, pages 66-70.
The website: http://www.amcm.pcz.pl/

MIXTURE METHODS OF PRIOR DISTRIBUTION CHOICE IN A BINARY RESPONSE CASE

Paweł Kopciuszewski

Institute of Mathematics and Computer Science, Czestochowa University of Technology

Abstract. The prior choice methods are strongly explored in Bayesian analysis because of its important applications. The use of any method to the prior selection does not guarantee the choice of the exactly one proper distribution. That is why the mixture of ones is proposed here. In the paper the mixtures of ML-II, conjugate and Jeffrey's priors are considered. The whole Bayesian analysis is executed to the binary response case.

1. Mixture of ML-II and conjugate prior

The application of any prior choice method is not enough to the unique prior identification. That is why the mixture of some prior choice methods is used. The most known is the mixture of ML-II method and the method which chooses the prior from the ε-contaminated distributions [1]. In the paper the prior for the variable with binary responses is proposed. There are described two mixtures of the ML-II method and the Jeffrey's method with the conjugate one.

The class of Beta distributions is the conjugate family to a binomial distribution. There is no formula to choose any Beta distribution from the whole class of Beta distributions [2]. We apply other method of prior selection to unique identification of prior distribution. Jointly, the conjugate and ML-II methods are considered to the prior selection. ML-II method is strictly connected with Bayes Factor which tests any Bayesian models quality [1]. It gives us the prior under the knowledge of the observed vector. That is why it is an Empirical Bayes prior [2]. If some additional prior information is given then it is an informative prior otherwise it is called the noninformative prior. In short, this method let us to choose the prior such that the marginal distribution of the observed vector in the observed point achieves its maximum.

Let $f(y|\theta)$ be the likelihood of the observable random variable Y with the response $y \in R^n$ under the conditional parameter θ. Let $\pi(\theta)$ be the prior of the parameter $\theta \in R^p$. In the analysis we limit the all coordinates of the vector $y = (y_1,\ldots,y_n)$ to the zero-one distribution. This is the most interesting distribution from the binomial class from the practical point of view, especially in the cases where the probability of the given event is searched for. Then the appearance of

the event or not appearance of this are considered. Assume that the considered parameter θ is the probability of any event.
From this the likelihood is given by the formula:

$$\mathrm{f}(\mathrm{y} \mid \theta) = \theta^k (1-\theta)^{n-k} \tag{1}$$

where

$$\mathrm{f}(y_i \mid \theta = \theta 1_{\{y_i=1\}}(y_i) + (1-\theta)1_{\{y_i=0\}}(y_i)$$

where k is the number of the event occurrence in the sample of all y coordinates.

The prior $\pi(\theta)$ conjugate to the binomial distribution is a Beta distribution which conditionally depends on two parameters α and β. This prior is the hyper-prior of the first stage in the prior hierarchy. Hence the considered model is the first stage hyperprior model [3]. The prior conditional density is written as follows:

$$\pi(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \tag{2}$$

Then the posterior $\pi(\theta|y)$ is proportional to the following function of the parameter θ

$$f(y|\theta) \cdot \pi(\theta|\alpha,\beta) = \theta^k (1-\theta)^{n-k} \frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \tag{3}$$

Hence the marginal $f(y)$ equals

$$f(y) = \int_0^1 f(y \mid \theta) \cdot \pi(\theta \mid \alpha,\beta) d\theta \tag{4}$$

If we find the hyperparameters α, β in which the marginal $f(y)$ has its maximum value then the unique prior will be found.

Theorem 1.
If parameters α and β tend to zero than the prior tends to the ML-II prior.

Proof.
The marginal $f(y)$ can be written as follows:

$$f(y) = \int_0^1 \theta^k (1-\theta)^{y-k} \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta =$$

$$\int_0^1 \theta^{k+\alpha-1}(1-\theta)^{y-k+\beta-1} \frac{1}{B(\alpha,\beta)} d\theta \tag{5}$$

which gives us the following quotient:

$$\frac{B(k+\alpha, y-k+\beta)}{B(\alpha,\beta)} = \frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)\Gamma(\alpha+\beta)}{\Gamma(n+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)} =$$

$$\frac{(k+\alpha-1)\cdot(k+\alpha-2)\ldots\cdot\alpha\cdot\Gamma(\alpha)\cdot(\beta+n-k-1)\cdot(\beta+n-k-2)\cdot\ldots\cdot\beta\cdot\Gamma(\beta)\cdot\Gamma(\alpha+\beta)}{(\alpha+\beta+n-1)\cdot(\alpha+\beta+n-2)\cdot\ldots\cdot(\alpha+\beta)\cdot\Gamma(\alpha+\beta)\cdot\Gamma(\alpha)\cdot\Gamma(\beta)} =$$

$$= \frac{(k+\alpha-1)\cdot(k+\alpha-2)\ldots\cdot\alpha\cdot(\beta+n-k-1)\cdot(\beta+n-k-2)\cdot\ldots\cdot\beta}{(\alpha+\beta+n-1)\cdot(\alpha+\beta+n-2)\cdot\ldots\cdot(\alpha+\beta)} \tag{6}$$

Let us mark the above quotient by $g(\alpha,\beta)$.
Hence we have:

$$h(\alpha,\beta) = \ln(g(\alpha,\beta)) =$$

$$\sum_{i=1}^{k}\ln(k+\alpha-i) + \sum_{i=1}^{n-k}\ln(n-k+\beta-i) - \sum_{i=1}^{n}\ln(n+\alpha+\beta-i) \tag{7}$$

To find the all extreme values of the function h the following derivatives have to equal to zero:

$$\frac{dh}{d\alpha} = \sum_{i=1}^{k}\frac{1}{k+\alpha-i} - \sum_{i=1}^{n-k}\frac{1}{n+\alpha+\beta-i} \tag{8}$$

$$\frac{dh}{d\beta} = \sum_{i=1}^{n-k}\frac{1}{n-k+\beta-i} - \sum_{i=1}^{n}\frac{1}{n+\alpha+\beta-i} \tag{9}$$

Certainly, these derivatives are not positive $\frac{dh}{d\alpha} < 0$ and $\frac{dh}{d\beta} < 0$ and then the above system of equations has not a unique solution. That is why the h has not an extreme values. More precise analysis of the h monotonicity implies that if parameters α and β are closer to zero than the h achieves the higher values and consequently the prior tends to the ML-II prior.

In the paper [4] the approximation of the ML-II prior has been given. This approximation gives us another prior distribution which in generally has not to be the same as the ML-II prior. The theorem suggests to minimize the following expression $A(\alpha,\beta)$ to find the approximation of ML-II prior

$$A(\alpha,\beta) = (E(p) - \hat{p})^2 + D^2(p) \tag{10}$$

where:

$$E(p)=\frac{\alpha}{\alpha+\beta},\quad D^2(p)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

In this case the maximum likelihood estimator equals $\hat{p}=\frac{k}{n}$.

If the uniform prior is considered for the parameter θ then the support of this distribution is considered to be the interval [0, 1.5·θ]. Certainly the upper bound of this interval can not to be higher than one. Uniform distribution with support [0,1] is the special case of the distribution from the beta distribution class. Certainly, then the posterior mode $\frac{k}{n}$ is the same as ML estimator.

From the theorem (2) in the paper [4] it results that $V(\theta)=(E(p)-\hat{p})^2+D^2(p)$ should be such small as it is possible to identify the approximation of ML-II prior

$$V(\theta)=\left(\frac{\alpha}{\alpha+\beta}-\frac{k}{n}\right)^2+\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{11}$$

It is clear that $V(\theta)$ tends to the infinity if α and β tends to zero.
That is why if we decide to choose the approximation of ML-II prior the parameters α and β should be near zero.
Assume additionally that $\alpha \in [\alpha_{\min},\alpha_{\max}]$ and $\beta \in [\beta_{\min}, \beta_{\max}]$, where $\alpha_{\min}$, $\alpha_{\max}$, $\beta_{\min}$, $\beta_{\max} > 0$.
Then the ML-II prior approximation exists and it equals

$$\frac{1}{B(\alpha_{\min},\beta_{\min})}\cdot\theta^{\alpha_{\min}-1}\cdot(1-\theta)^{\beta_{\min}-1} \tag{12}$$

Similar conditions could be obtained when the symmetry of the prior has been assumed.

2. Mixture of Jeffrey's and conjugate priors

Theorem 2.
Posterior distribution under the mixture Jeffrey's and conjugate priors equals:

$$\theta^k(1-\theta)^{n-k}\sqrt{\frac{1}{\pi}\left(\frac{1}{\theta}+\frac{1}{1-\theta}\right)} \text{ and its posterior mode is } \frac{k-1}{n-2}. \tag{13}$$

Proof:
The Jeffrey's prior $\pi(\theta)$ is proportional to the square root of the Fisher information determinant, that is

$$\pi(\theta) \propto \sqrt{E\left[\frac{\partial \ln f(y \mid \theta)}{\partial \theta}\right]^2} = \sqrt{n \cdot E\left[\frac{\partial \ln f(y_1 \mid \theta)}{\partial \theta}\right]^2} \tag{14}$$

In the considered case the Jeffrey's prior equals

$$\sqrt{n \cdot E\left[\frac{1_{\{y_1=1\}}(y_1) - 1_{\{y_i=0\}}(y_1)}{\theta 1_{\{y_1=1\}}(y_1) + (1-\theta) 1_{\{y_i=0\}}(y_1)}\right]^2} \tag{15}$$

After some further calculations we obtain:

$$\pi(\theta) \propto \sqrt{n\left(\frac{1}{p} + \frac{1}{1-p}\right)} \tag{16}$$

Integrate the density over the parameter support we obtain he following proper prior density

$$\pi(\theta) = \sqrt{\frac{1}{\pi}\left(\frac{1}{p} + \frac{1}{1-p}\right)} \tag{17}$$

Then the posterior equals

$$\pi(\theta \mid y) = \theta^k (1-\theta)^{n-k} \sqrt{\frac{1}{\pi}\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right)} \tag{18}$$

Some simple calculations implies that posterior maximum is achieved in (13).

The jointly use of two prior selection methods leads to the unique prior choice. As it has been showed in a binary response case both the empirical Bayes and classical prior can be obtained with use of ML-II, conjugate and Jeffrey's methods.

References

[1] Berger J., Berliner L.M., Robust Bayes analysis with ε-contaminated priors, The Annals of Statistics 1986, 14, 461-468.

[2] Gelman A.B., Carlin J.S., Stern H.S., Rubin D.B., Bayesian Data Analysis, Chapman & Hall, New York 2000.

[3] Robert C.P., The Bayesian Choice, A Decision-Theoretic Motivation, Springer, New York 2004.

[4] Kopciuszewski P., Aproksymacja rozkładów ML-II apriori, 30 Konferencja Zastosowań Matematyki, Instytut Matematyki PAN, Warszawa 2001.