Ann. Inst. Statist. Math.
35 (1983), Part A, 167-174

ON A NONINFORMATIVE PRIOR DISTRIBUTION FOR BAYESIAN INFERENCE OF MULTINOMIAL DISTRIBUTION'S PARAMETERS

SHINTARO SONO

(Received Apr. 27, 1982; revised Nov. 2, 1982)

Summary

Noninformative prior distributions for Bayesian inference, for example, Jeffreys' priors, are much useful for the so-called "objective Bayesian inference" and make it possible to develop a method more powerful and flexible than traditional methods. In this paper a noninformative prior distribution, which is different from usual Jeffreys' priors, is introduced for Bayesian inference of multinomial distribution's parameters, using the assumption of the prior independence of the transformed parameters and the approximate data-translated likelihood function, and a short theoretical consideration for the inference based on the prior is attempted.

1. Introduction

Consider the sample $(N_i)_{i=1}^{K-1}$ distributed by multinomial distribution, $M(N; (\theta_i)_{i=1}^{K-1})$, i.e.,

$$p((N_i)_{i=1}^{K-1}|(\theta_i)_{i=1}^{K-1}) = N!\left(\prod_{i=1}^{K} N_i!\right)^{-1} \prod_{i=1}^{K} \theta_i^{N_i},$$

where N and N's are positive and nonnegative integers, respectively, and satisfy $N=\sum_{i=1}^{K} N_i$, and $(\theta_i)_{i=1}^{K-1} \in \Delta^{(K-1)} := \left\{(\theta_i)_{i=1}^{K-1} \in]0, 1[^{K-1}; \sum_{i=1}^{K-1} \theta_i < 1\right\}$ and $1=\sum_{i=1}^{K} \theta_i$; $(\theta_i)_{i=1}^{K-1}$ is unknown but fixed and K is greater than 1. For Bayesian inference of the unknown but fixed parameters, θ's, based on the data, N's, a noninformative prior, for example, Jeffreys' prior, i.e., $P((\theta_i)_{i=1}^{K-1}) = c \prod_{i=1}^{K} \theta_i^{-1/2}$, must be introduced. But the objective Bayesian inference like Box and Tiao [1], is based on H.P.D. (Highest Posterior Density) regions, especially, standardized H.P.D. regions, i.e., H.P.D. regions on the transformed parameters defining the approximate

data-translated likelihood function. Therefore, for the objective Bayesian inference, it is reasonable to consider the transformation of θ's defining the approximate data-translated likelihood function in some sense. The inference of θ's should be accomplished based on the transformed parameters. In the one-dimensional case, i.e., $K=2$, the transformation of θ is directly obtained from the Jeffreys' prior, i.e., the transformation is $t=\frac{2}{\pi}\operatorname{Arcsin}\sqrt{\theta}$. The prior induced from this transformation, i.e., the prior induced by $p(t)=1$, is identical with the Jeffreys' prior. But in the multidimensional case the relation between the two kinds of priors is not clear because the transformed parameters on which the locally uniform prior is introduced are not uniquely determined. In the following sections the transformed parameters having the locally uniform prior are defined in some reasonable manner and the Bayesian inference based on the transformation is considered.

2. Likelihood and prior

From the definition of the model in Section 1 the likelihood function of $(\theta_i)_{i=1}^{K-1} \in \Delta^{(K-1)}$ is

$$l((\theta_i)_{i=1}^{K-1}|(N_i)_{i=1}^{K-1}) \propto \prod_{i=1}^{K} \theta_i^{N_i} . \tag{2.1}$$

Consider the transformations of $\theta=(\theta_i)_{i=1}^{K-1} \in \Delta^{(K-1)}$ to $s=(s_i)_{i=1}^{K-1} \in I^{(K-1)}:=]0,\ 1[^{K-1}$ and $t=(t_i)_{i=1}^{K-1} \in I^{(K-1)}$,

$$\prod_{i=j}^{K-1} s_i = \sum_{i=1}^{j} \theta_i , \qquad j=1, \cdots, K-1 , \tag{2.2}$$

$$t_i=\frac{2}{\pi}\operatorname{Arcsin}(\sqrt{s_i}) , \qquad i=1, \cdots, K-1 . \tag{2.3}$$

The transformation, (2.2), is equivalent to $\theta_i=(1-s_{i-1})\prod_{j=i}^{K-1} s_j=\prod_{j=i}^{K-1} s_j-\sum_{j=1}^{i-1}\theta_j$, $s_0:=0$, $i=1, \cdots, K-1$, and $s_i=\left(\sum_{j=1}^{i}\theta_j\right)\Big/\left(\sum_{j=1}^{i+1}\theta_j\right)$, $i=1, \cdots, K-1$, and the Jacobians of (2.2) and (2.3) are easily obtained as $\det\frac{\partial(\theta)}{\partial(s)}=\prod_{j=2}^{K-1} s_j^{j-1}=\prod_{j=2}^{K-1}\left(\sum_{i=1}^{j}\theta_i\right)$, and $\det\frac{\partial(t)}{\partial(s)}=\pi^{-(K-1)}\left(\prod_{j=1}^{K-1}\sqrt{s_j(1-s_j)}\right)^{-1}$ with the conventional rule, $\sum_{a}^{b}:=0$ and $\prod_{a}^{b}:=1$, for $a>b$. Therefore, if the locally uniform prior is introduced on $t \in I^{(K-1)}$, i.e., if $p(t)=1$, $t \in I^{(K-1)}$, then the following priors are readily obtained:

$$p(s)=\pi^{-(K-1)}\prod_{j=1}^{K-1}(s_j(1-s_j))^{-1/2} , \tag{2.4}$$

$$(2.5)\qquad p(\theta)=\pi^{-(K-1)}\left(\prod_{i=1}^{K}\theta_i\right)^{-1/2}\left(\prod_{j=2}^{K-1}\left(\sum_{i=1}^{j}\theta_i\right)\right)^{-1/2}$$

$$=\left(\pi^{K-1}\sqrt{\theta_1\theta_2\cdots\theta_{K-1}\left(1-\sum_{i=1}^{K-1}\theta_i\right)(\theta_1+\theta_2)(\theta_1+\theta_2+\theta_3)\cdots(\theta_1+\theta_2+\cdots+\theta_{K-1})}\right)^{-1},$$

where $s\in I^{(K-1)}$ and $\theta\in\Delta^{(K-1)}$.

The locally uniform prior on t is derived from the assumption of prior independence of the parameters, s_i, $i=1,\cdots,K-1$, i.e.,

$$(2.6)\qquad p(s)=\prod_{i=1}^{K-1}p(s_i)\,,\qquad s\in I^{(K-1)}\,,$$

and this assumption, (2.6), have the reasonable Bayesian interpretation such that the prior knowledge of $(s_i)_{i=1}^{K-1}$ is vague enough in comparison with the information included in the likelihood function of $(s_i)_{i=1}^{K-1}$ and the prior knowledge of each s_i is almost invariant even if all other s's are specified. Because, if (2.6) is employed, then the likelihood of each s_i, given other s's, is transformed into the approximate data-translated likelihood on t_i by (2.3), therefore, it is reasonable to introduce the locally uniform prior on each t_i, and, from (2.3) and (2.6), the locally uniform prior on t is obtained.

3. Posterior distributions

For the evaluation of the posteriors of t etc. consider the integral,

$$(*)\qquad S((\alpha_i)_{i=1}^{K-1};\alpha_K;(\gamma_i)_{i=2}^{K-1}):=$$
$$\int_{\Delta^{(K-1)}}d\theta\prod_{i=1}^{K-1}\theta_i^{\alpha_i-1}\cdot\left(1-\sum_{i=1}^{K-1}\theta_i\right)^{\alpha_K-1}\cdot\prod_{j=2}^{K-1}\left(\sum_{i=1}^{j}\theta_i\right)^{\gamma_j-1},$$

where all α's and γ's are positive real numbers. Using the transformation, (2.2), the value of $(*)$ is obtained as

$$(3.1)\qquad S((\alpha_i)_{i=1}^{K-1};\alpha_K;(\gamma_i)_{i=2}^{K-1})=\prod_{j=1}^{K-1}B\left(\sum_{i=1}^{j}\alpha_i+\sum_{i=2}^{j}\gamma_i-j+1,\alpha_{j+1}\right)$$

where $B(a,b)$ is a beta function, represented by gamma functions as $B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$. From $p(\theta)=\left(\det\frac{\partial(\theta)}{\partial(t)}\right)^{-1}$ the posterior of t is readily obtained because $p(t|N\text{'s})=p(\theta|N\text{'s})\cdot\det\frac{\partial(\theta)}{\partial(t)}\propto l(\theta|N\text{'s})\ p(\theta)\cdot\det\frac{\partial(\theta)}{\partial(t)}=l(\theta|N\text{'s})$, and the normalizing constant is evaluated by the formula, (3.1), and $\det\frac{\partial(\theta)}{\partial(t)}=\pi^{K-1}\left(\prod_{i=1}^{K}\theta_i\cdot\prod_{j=2}^{K-1}\left(\sum_{i=1}^{j}\theta_i\right)\right)^{1/2}$, therefore

$$p(t|N's)=\pi^{K-1}\cdot C(N's)\cdot\prod_{i=1}^{K}\theta_i^{N_i}, \qquad t\in I^{(K-1)}, \tag{3.2}$$

where

$$C(N's)^{-1}:=S\left(\left(N_i+\frac{1}{2}\right)_{i=1}^{K-1}; N_K+\frac{1}{2}; \left(\frac{1}{2}\right)_{i=2}^{K-1}\right)$$
$$=\prod_{j=1}^{K-1}B\left(\sum_{i=1}^{j}N_i+\frac{1}{2}, N_{j+1}+\frac{1}{2}\right).$$

Similarly the posteriors of θ and s are obtained:

$$p(\theta|N's)=C(N's)\prod_{i=1}^{K}\theta_i^{N_i-1/2}\cdot\prod_{j=2}^{K-1}\left(\sum_{i=1}^{j}\theta_i\right)^{-1/2}, \qquad \theta\in\Delta^{(K-1)} \tag{3.3}$$

$$p(s|N's)=C(N's)\prod_{i=1}^{K}\theta_i^{N_i-1/2}\cdot\prod_{j=2}^{K-1}\left(\sum_{i=1}^{j}\theta_i\right)^{1/2}, \qquad s\in I^{(K-1)}. \tag{3.4}$$

And remark:

$$p(t|N's)=\prod_{i=1}^{K-1}p(t_i|N's)$$
$$=\pi^{K-1}\cdot C(N's)\cdot\prod_{j=1}^{K-1}s_j^{\sum_{i=1}^{j}N_i}\cdot(1-s_j)^{N_{j+1}}, \quad t\in I^{(K-1)}. \tag{3.5}$$

4. H.P.D. region, inference, and test

For the construction of the standardized H.P.D. region, i.e., the H.P.D. region on $t\in I^{(K-1)}$, which is represented by θ's terms under the transformations, (2.2) and (2.3), consider the modal value and the ratios of the posterior of t. The modal value of $p(t|N's)$, $\hat{t}$, is readily obtained from (3.2) because $\hat{t}$ is the transformed value of $\hat{\theta}$ which is maximizing the likelihood of θ, i.e., $\hat{\theta}_i=N_i/N$, $i=1,\cdots,K$, and the $(1-\alpha)$ H.P.D. region on $t=\{t\in I^{(K-1)};\ p(t|N's)/p(\hat{t}|N's)>r(\alpha)\}$, where $P(p(t|N's)/p(\hat{t}|N's)>r(\alpha)|N's)=1-\alpha$. Under the transformations, (2.2) and (2.3) (assuming $N_i>0$, $i=1,\cdots,K$),

$$R:=R(\theta)=\prod_{i=1}^{K}\theta_i^{N_i}\Big/\prod_{i=1}^{K}\hat{\theta}_i^{N_i}$$
$$=p(t|N's)/p(\hat{t}|N's), \qquad \theta\in\Delta^{(K-1)}, \tag{4.1}$$

therefore the value, $r(\alpha)$, satisfies $P(R>r(\alpha)|N's)=1-\alpha$. The moments of R, $E(R^{\nu}|N's)$, where ν is any real number greater than -1, are evaluated to approximate $r(\alpha)$; using (3.1), (3.3), and (4.1),

$$E(R^{\nu}|N's) \tag{4.2}$$

$$= \int_{\Delta^{(K-1)}} d\theta\, p(\theta | N\text{'s}) \cdot R^{\nu}$$

$$= \left(\prod_{i=1}^{K} \hat{\theta}_i^{N_i}\right)^{-\nu} \cdot \frac{S\left(\left((\nu+1)N_i+\frac{1}{2}\right)_{i=1}^{K-1}; (\nu+1)N_K+\frac{1}{2}; \left(\frac{1}{2}\right)_{i=2}^{K-1}\right)}{S\left(\left(N_i+\frac{1}{2}\right)_{i=1}^{K-1}; N_K+\frac{1}{2}; \left(\frac{1}{2}\right)_{i=2}^{K-1}\right)}$$

$$= \left(\prod_{i=1}^{K} \hat{\theta}_i^{N_i}\right)^{-\nu} \cdot \prod_{j=1}^{K-1} \frac{B\left((\nu+1)\left(\sum_{i=1}^{j} N_i\right)+\frac{1}{2}, (\nu+1)N_{j+1}+\frac{1}{2}\right)}{B\left(\sum_{i=1}^{j} N_i+\frac{1}{2}, N_{j+1}+\frac{1}{2}\right)}, \qquad \nu > -1,$$

and, using duplication formula of gamma function, $\Gamma(2z) = 2^{2z-1}\pi^{-1/2}\Gamma(z) \cdot \Gamma\left(z+\frac{1}{2}\right)$,

$$(4.3) \quad B\left((\nu+1)a+\frac{1}{2}, (\nu+1)b+\frac{1}{2}\right) = \frac{\pi\Gamma(2(\nu+1)a)\Gamma(2(\nu+1)b)}{2^{2(\nu+1)(a+b)-2}\Gamma((\nu+1)a)\Gamma((\nu+1)b)\Gamma((\nu+1)(a+b)+1)}, \qquad a, b > 0.$$

From (4.2), (4.3), and Stirling's formula with remainder term (see Whittaker and Watson [2], pp. 251–252) the following formulas are obtained:

$$(4.4) \quad \mathrm{E}(R^{\nu} | N\text{'s}) = (1+\nu)^{-(K-1)/2} \times \exp\left(\sum_{m=1}^{M} ((1+\nu)^{-(2m-1)}-1) \cdot C_m + R(M; N, \hat{\theta}; \nu)\right), \qquad \nu > -1,$$

where $R(M; N, \hat{\theta}; \nu)$ is the remainder term such that if $(N_i)_{i=1}^{K-1} \sim M(N; \theta^{(*)})$ for some fixed $\theta^{(*)} \in \Delta^{(K-1)}$, then there exists a constant, $C(M; \theta^{(*)})$ (depending only on M and $\theta^{(*)}$), satisfying

$$(\#) \quad \limsup_{N\to\infty} N^{2M+1} \sup_{|\nu| \leqq 1-\delta} |R(M; N, \hat{\theta}; \nu)| \leqq C(M; \theta^{(*)}), \quad \text{almost surely,}$$

for any fixed positive integer, M, and any $\delta \in]0, 1[$, and the rth derivative of $R(M; N, \hat{\theta}; \nu)$ with respect to ν has the same property, (#), as $R(M; N, \hat{\theta}; \nu)$ has for any positive integer, r. (In (4.4) put

$$C_m := \frac{(-1)^{m-1}B_m}{2m(2m-1)} \cdot \left((2^{-(2m-1)}-1)\sum_{i=1}^{K} \hat{\theta}_i^{-(2m-1)} + (2^{-(2m-1)}-2)\sum_{j=2}^{K-1}\left(\sum_{i=1}^{j} \hat{\theta}_i\right)^{-(2m-1)} - 1\right) \cdot N^{-(2m-1)}, \qquad m = 1, \cdots, M,$$

where B's are Bernoulli numbers defined as

$$(x/(e^x-1))+x/2-1=\sum_{m=1}^{\infty}(-1)^{m-1}\cdot\frac{B_m}{(2m)!}x^{2m}\,.\,)$$

And, using (4.4), the cumulant generating function of $S:=S(\theta):=-2\cdot\log R$ is obtained:

$$\begin{aligned}(4.5)\quad K(\xi;S):&=\log \mathrm{E}\,(\exp(\xi\cdot S)|N\text{'s})=\log \mathrm{E}\,(R^{-2\xi}|N\text{'s})\\&=\sum_{r=1}^{\infty}\frac{\xi^r}{r!}\cdot 2^{r-1}\cdot(r-1)!\cdot(K-1)\\&\quad\times\left(1+\frac{2}{(r-1)!(K-1)}\cdot\sum_{m=1}^{M}C_m\cdot\prod_{j=0}^{r-1}(2m-1+j)\right)\\&\quad+R(M;N,\hat\theta;-2\xi),\ \xi\in\left]-\frac{1}{2},\frac{1}{2}\right[\,.\end{aligned}$$

Therefore the rth cumulant of $S=-2\log R$, $K_r(S)$, is given as

$$\begin{aligned}(4.6)\quad K_r(S)&=2^{r-1}\cdot(r-1)!\cdot(K-1)\\&\quad\times\left(1+2((r-1)!\cdot(K-1))^{-1}\sum_{m=1}^{M}C_m\cdot\prod_{j=0}^{r-1}(2m-1+j)\right)\\&\quad+\overset{(r)}{\mathrm{E}}(M;N,\hat\theta)\,,\qquad r=\text{any positive integer}\,,\end{aligned}$$

where

$$\overset{(r)}{\mathrm{E}}(M;N,\hat\theta):=\frac{d^r}{d\xi^r}R(M;N,\hat\theta;-2\cdot\xi)\Big|_{\xi=0}\,.$$

Especially, when $M=1$,

$$(4.7)\qquad K_r(S)=2^{r-1}\cdot(r-1)!\cdot(K-1)\cdot(1-r\cdot d(N;\hat\theta))+\overset{(r)}{\mathrm{E}}(1;N,\hat\theta)\,,$$

where

$$d(N;\hat\theta):=(12\cdot(K-1)\cdot N)^{-1}\cdot\left(\sum_{i=1}^{K}\hat\theta_i^{-1}+3\cdot\sum_{j=2}^{K-1}\left(\sum_{i=1}^{j}\hat\theta_i\right)^{-1}+2\right).$$

Using $K_r((1-d(N;\hat\theta))\cdot\chi^2_{K-1})=2^{r-1}\cdot(r-1)!\cdot(K-1)\cdot(1-d(N;\hat\theta))^r$ and the property of the remainder term in (4.4), we have

$$(4.8)\qquad K_r(S)=K_r((1-d(N;\hat\theta))\chi^2_{K-1})+F(r;N,\hat\theta)\,,$$

where $F(r;N,\hat\theta)$ is the remainder term such that if $(N_i)_{i=1}^{K-1}\sim M(N;\theta^{(*)})$ for some fixed $\theta^{(*)}\in\Delta^{(K-1)}$, then there exists a constant, $C(r;\theta^{(*)})$ (depending only on r and $\theta^{(*)}$), satisfying

$$\limsup_{N\to\infty}N^2\cdot F(r;N,\hat\theta)\leqq C(r;\theta^{(*)})\,,$$

almost surely, for any positive integer r. (From the definition of $\hat{\theta}_i := N_i/N$, $d(N; \hat{\theta}) \in]0, 1[$ is clear, therefore the right hand of (4.8) is well-defined.)

From (4.5) the distribution of $S(\theta) = -2 \cdot \log R(\theta)$ (of course, "posterior") converges χ^2_{K-1}, almost surely. Therefore $\chi^2_{K-1}(\alpha)$ satisfying $P(\chi^2_{K-1} < \chi^2_{K-1}(\alpha)) = 1-\alpha$ gives the approximate value $-2 \cdot \log r(\alpha; N, \hat{\theta})$, where $P(S < -2 \cdot \log r(\alpha; N, \hat{\theta}) | N\text{'s}) = 1-\alpha$. But more precise approximation is given by (4.8), i.e.,

$$-2 \cdot \log r(\alpha; N, \hat{\theta}) \doteqdot (1-d(N; \hat{\theta}))\chi^2_{K-1}(\alpha) . \tag{4.9}$$

Therefore, using (4.9), the approximate $(1-\alpha)$ H.P.D. region on t transformed into θ is obtained as

$$\{\theta \in \Delta^{(K-1)}; S(\theta) < (1-d(N; \hat{\theta}))\chi^2_{K-1}(\alpha)\} , \tag{4.10}$$

and, consequently, the Bayesian inference of θ is accomplished by the approximate $(1-\alpha)$ Bayesian region induced from the approximate $(1-\alpha)$ H.P.D. region on t, (4.10). And, from (4.9) and (4.10), the Bayes test of the hypothesis, $\theta = \theta^{(0)}$, for a specified $\theta^{(0)} \in \Delta^{(K-1)}$ is readily obtained, i.e., "If $(-2 \cdot \log R(\theta^{(0)}))/(1-d(N; \hat{\theta})) < \chi^2_{K-1}(\alpha)$, then '$\theta = \theta^{(0)}$' is accepted, and if $(-2 \cdot \log R(\theta^{(0)}))/(1-d(N; \hat{\theta})) \geqq \chi^2_{K-1}(\alpha)$, then '$\theta = \theta^{(0)}$' is rejected," because $P(S(\theta) \leqq S(\theta^{(0)}) | N\text{'s}) < (\geqq) 1-\alpha$ is equivalent to $S(\theta^{(0)}) < (\geqq) -2 \cdot \log r(\alpha; N, \hat{\theta})$, respectively.

The inference of marginal parameters of θ based on the marginal posterior distributions given by (3.5) is accomplished by the similar argument and so is omitted.

5. Concluding remarks

It should be remarked that the prior, (2.5), is dependent on the ordering of θ's. Hence, for practice, the ordering must be specified from the statistical sense of θ's in each problem. But the Bayes test for θ based on the standardized H.P.D. regions is precisely numerically equivalent to the usual likelihood ratio test for any ordering of θ's (See Section 4).

The generalizations of the prior, (2.5), and the posterior, (3.3), are readily obtained by the application of the integral, $(*)$. This generalized prior is inevitably induced from the beta distribution on each s_i, i.e., the natural conjugate prior of s_i, given other s's.

Acknowledgement

The writer wishes to thank Professor Y. Suzuki and the referee for critical and valuable suggestions.

GRADUATE SCHOOL OF TOKYO UNIVERSITY

REFERENCES

[1] Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Massachusetts.

[2] Whittaker, E. T. and Watson, G. N. (1927). *A Course of Modern Analysis*, 4th ed., Cambridge University Press, London.