Ann. Inst. Statist. Math.
32 (1980), Part A, 171-179

IGNORANCE PRIOR DISTRIBUTION OF A HYPERPARAMETER AND STEIN'S ESTIMATOR

HIROTUGU AKAIKE

(Received Nov. 13, 1979; revised Feb. 8, 1980)

Summary

The concept of ignorance prior distribution is extended to the case of a hyperparameter. This leads to a procedure of formulating the partial ignorance of the original parameter. Its application to the estimation of the mean of a multivariate normal distribution with a particular hyperparameterized prior distribution of the mean leads to an improper prior distribution with the corresponding posterior mean very close to the James-Stein estimate.

1. Introduction

Stein ([5], p. 280) discussed a strange inconsistency between the posterior distribution of the mean ξ of a multivariate normal distribution with covariance matrix equal to the identity, obtained by assuming the uniform improper prior distribution, and the distribution of the observation x when ξ is known. Efron [3] expounded the phenomenon as one of the controversies in the foundation of statistics. This is typically represented by the inconsistency between the posterior inequality $\text{Prob}\{||\xi||>||x|| \mid x\}>0.50$ and the data distributional inequality $\text{Prob}\{||x||>||\xi|| \mid \xi\}>0.50$, where $||x||$ denotes the Euclidean norm of x. Efron argues that, being emphasized by the introduction of the James-Stein estimate, this inconsistency suggests that a theory of objective Bayesian inference, if exists, must be a great deal more subtle than previously expected.

Strawderman [7] obtained proper Bayes minimax estimators of ξ by assuming the prior distribution of ξ to be normal with zero mean and covariance matrix $\sigma^2 I$ with σ^2 distributed according to the distribution $(1-a)^{-1}\lambda^{-a}d\lambda$, where $\lambda=1/(1+\sigma^2)$ and $0\leqq a<1$. He acknowledges the suggestion by Stein for the choice of the prior distribution for which the calculations are tractable. This convenience for the calculation was also the main reason of the choice of the improper prior

density $\pi(\xi)=||\xi||^{-q+3}$ by Stein [5], where q is the dimension of ξ, to avoid the inconsistency mentioned in the preceding paragraph. Later, Stein [6] developed an argument for the use of a nearly harmonic prior density in the absence of firm prior information. He developed a comparison between the James-Stein positive part estimate and the related formal Bayes estimate based on the improper harmonic prior density $\pi(\xi)=||\xi||^{-q+2}$ and concluded that the difference would not be significant.

The purpose of the present paper is to provide an answer to the question raised by Efron about the so-called objective Bayesian inference. It is shown that, in the Strawderman's Bayesian model, the uniform prior distribution over $(0, \infty)$ of the transformed parameter $\tau=\log(1+\sigma^2)$ provides a reasonable representation of ignorance of σ^2. This is based on the extension of our interpretation of Jeffreys' ignorance prior distribution as an impartial prior distribution (Akaike [1]) to the case of an ignorance prior distribution of a hyperparameter. The ignorance prior distribution of σ^2 thus obtained represents a partial ignorance of θ and corresponds to the limiting case $a=1$ of the Strawderman's prior distribution. It is confirmed analytically, and also experimentally, that the corresponding prior distribution of ξ is close to the harmonic prior suggested by Stein.

2. Representation of partial ignorance

When we adopt a distribution $p(x|\theta)$ as an approximation to the distribution $p(x|\theta')$, we measure the degree of approximation by the entropy of $p(x|\theta')$ with respect to $p(x|\theta)$ which is defined by

$$B\{p(\cdot|\theta'), p(\cdot|\theta)\}=-\int \frac{p(x|\theta')}{p(x|\theta)} \log \left\{\frac{p(x|\theta')}{p(x|\theta)}\right\} p(x|\theta)dx\,. \tag{2.1}$$

Jeffreys' ignorance prior distribution can now be interpreted as the locally, or sometimes globally, impartial prior distribution of the parameter θ, defined as the uniform distribution in the space of some transform ϕ of the parameter θ where a small deviation $\Delta\phi=\phi'-\phi$ causes equal decrease of entropy for every ϕ, where ϕ and ϕ' correspond to θ and θ' of (2.1), respectively (Akaike [1]).

Now we want to incorporate our prior information by specifying a particular prior distribution of θ. To represent our uncertainty in the choice of the prior distribution, we consider the use of a family of prior distributions $p(\theta|\rho)$ parameterized by a hyperparameter ρ. If we specify a proper prior distribution $p(\rho)$ of ρ and the data x is observed the Bayes procedure transforms the prior distribution $p(\theta|\rho)p(\rho)$ into the posterior distribution $p(\theta|x|\rho)p(\rho|x)$, where the two factors are

respectively given by

$$p(\theta|x|\rho)=\frac{p(x|\theta)p(\theta|\rho)}{p(x|\rho)} \quad \text{and} \quad p(\rho|x)=\frac{p(x|\rho)p(\rho)}{p(x)}\,, \tag{2.2}$$

defined with $p(x|\rho)=\int p(x|\theta)p(\theta|\rho)d\theta$ and $p(x)=\int p(x|\rho)p(\rho)d\rho$. Since we are certain that under the assumption of $p(\theta|\rho)$ we will adopt $p(\theta|x|\rho)$ as our posterior distribution the choice of $p(\rho)$ affects our inference only through the definition of $p(\rho|x)$. Thus we can extend the definition of ignorance prior distribution to the present situation by replacing $p(x|\theta)$ of the preceding paragraph by $p(x|\rho)$. The resulting prior distribution $p(\theta)=\int p(\theta|\rho)p(\rho)d\rho$, which may be improper, would then be a representation of our partial ignorance of θ.

3. Partial ignorance of the mean of a multivariate normal distribution

To treat the problem discussed in Introduction we put $\theta=\xi$ and assume that $p(x|\theta)$ represents a q-dimensional normal distribution with mean θ and covariance matrix equal to the identity I. We also assume that $p(\theta|\rho)$ is a q-dimensional normal distribution with mean 0 and covariance matrix ρI. By convolving the above two distributions we get $p(x|\rho)$ which is a q-dimensional normal distribution with mean 0 and covariance matrix $(1+\rho)I$. By (2.1) we have

$$B\{p(\cdot|\rho'),\, p(\cdot|\rho)\}=\left(\frac{q}{2}\right)\left\{\log\left(\frac{1+\rho'}{1+\rho}\right)-\frac{1+\rho'}{1+\rho}+1\right\},$$

and we can see that the entropy is a function of $\log(1+\rho')-\log(1+\rho)$ only. This shows that by the transformation $\tau=\log(1+\rho)$ the same difference in τ produces the same decrease of entropy everywhere. Thus the ignorance prior distribution of ρ, which is globally impartial to the choice of ρ in this case, is defined by the uniform distribution with respect to $d\tau$.

The (improper) prior distribution of θ is

$$\begin{aligned} p(\theta)&=\int_0^\infty\left(\frac{1}{2\pi\rho}\right)^{q/2}\exp\left(-\frac{1}{2\rho}||\theta||^2\right)\frac{d\rho}{1+\rho} \\ &=\left(\frac{1}{2\pi}\right)^{q/2}s^{-q/2}\int_0^\infty v^{q/2-1}\left(\frac{s}{s+v}\right)\exp(-v)dv\,, \end{aligned} \tag{3.1}$$

where $s=||\theta||^2/2$ and $v=s/\rho$. Obviously $p(\theta)$ is approximated by $Cs^{-q/2}$ when s is large and by $Cs^{-q/2+1}$ when s is close to 0. This suggests that our prior distribution shows a good agreement with the Stein's

suggestion of nearly harmonic prior density. It puts more weight on smaller values of $||\theta||$ than the harmonic prior $||\theta||^{-q+2}$ which is obtained by replacing $(1+\rho)^{-1}d\rho$ by $d\rho$ in (3.1).

4. Estimation of the mean of a multivariate normal distribution

For the present $p(x|\theta)$ the entropy $B\{p(\cdot|\theta'), p(\cdot|\theta)\}$ is equal to $-(1/2)||\theta-\theta'||^2$. Thus we can see that the quadratic loss coincides with our entropy criterion. When ρ is known the Bayes estimate of θ is given by $\{\rho/(1+\rho)\}x$. The mean of the posterior distribution of θ is then given by $\theta^*=\mathrm{E}_{\rho|x}\{\rho/(1+\rho)\}x$, where $\mathrm{E}_{\rho|x}$ denotes the expectation with respect to the posterior distribution of ρ when x is observed. This posterior distribution $p(\rho|x)$ is given by

$$p(\rho|x)=C(1+\rho)^{-q/2-1}\exp\left(-\frac{1}{2(1+\rho)}||x||^2\right).$$

Thus we have

$$\mathrm{E}_{\rho|x}\{1/(1+\rho)\}=T_{q+2}(s)/T_q(s)\,,$$

where $s=||x||^2/2$ and

$$T_q(s)=\int_1^\infty u^{-q/2-1}\exp\left(-\frac{s}{u}\right)du\,.$$

With the change of the variable u into $v=s/u$ we get $T_q(s)=s^{-q/2}\gamma(q/2, s)$, where $\gamma(a, x)$ denotes a normalized incomplete gamma function

$$\gamma(a, x)=\int_0^x v^{a-1}\exp(-v)dv\,.$$

From the equality $\gamma(q/2, s)=(q/2-1)\gamma(q/2-1, s)-s^{q/2-1}\exp(-s)$ for $q>2$, we get the recurrence formula

$$T_{q+2}(s)=\frac{q}{2s}T_q(s)-\frac{1}{s}\exp(-s)$$

with $T_2(s)=s^{-1}\{1-\exp(-s)\}$ and $T_1(s)=s^{-1/2}2\pi^{1/2}\{\phi((2s)^{1/2})-0.5\}$, where $\phi(x)$ denotes a cumulative distribution function of a unit normal variate.

Since we have $||x-\theta^*||=(2s)^{1/2}\{T_{q+2}(s)/T_q(s)\}$ we get $||x-\theta^*||\leqq 2^{1/2}$ for $s\leqq 1$. For $s\geqq 1$ we have $||x-\theta^*||=(2s)^{1/2}\{s^{-1}\gamma(q/2+1, s)/\gamma(q/2, s)\}\leqq 2^{1/2}\gamma(q/2+1, \infty)/\gamma(q/2, 1)$. Thus we can see that the difference of the losses $||\theta^*-\theta||^2$ and $||x-\theta||^2$ is bounded. Brown ([2], p. 898) has shown that such an estimate θ^*, which is a generalized posterior mean defined by an improper prior distribution $p(\theta)$, is admissible if

$$\int_1^\infty (r^{q-1}f(r))^{-1}dr=\infty\,,$$

where $f(r)$ is defined by $f(r)=\int p(x|\theta)p(\theta)d\theta$ for $r=||x||$. In the present case we have

$$f(r)=\int_0^\infty \left(\frac{1}{2\pi}\right)^{q/2}(1+\rho)^{-q/2}\exp\left(-\frac{1}{2(1+\rho)}r^2\right)\frac{d\rho}{1+\rho}\,.$$

With the change of the variable ρ into $t=r^2/(2(1+\rho))$ we can show that $(r^{q-1}f(r))^{-1}>r\pi^{q/2}/\Gamma(q/2)$. Thus the present θ^* satisfies the Brown's condition of admissibility.

Although Strawderman [7] limited his prior distribution of $\lambda=1/(1+\rho)$ to $\lambda^{-a}/(1-a)$ $(0\leqq a<1)$, it is easy to see that his proof of minimaxity holds for $q\geqq 4$ with the improper prior density λ^{-1} which corresponds to the case $a=1$. Since $\lambda^{-1}d\lambda=(1+\rho)^{-1}d\rho$ is identical to our ignorance prior distribution of ρ, this shows that our θ^* dominates x as an estimate of θ and is minimax. In particular, we have $x-\theta^*=\{T_{q+2}(s)/T_q(s)\}x$, where

$$\frac{T_{q+2}(s)}{T_q(s)}=\left(\frac{q-2}{||x||^2}\right)\left(\frac{q}{q-2}-\frac{2}{q-2}\,\frac{\exp(-s)}{T_q(s)}\right).$$

Taking into account the equality

$$\frac{T_q(s)}{\exp(-s)}=\int_1^\infty u^{-q/2-1}\exp\left\{s\left(1-\frac{1}{u}\right)\right\}du\,,$$

we can see that the quantity within the above second parentheses grows from 0 to $q/(q-2)$ as s is increased from 0 to infinity. Since the positive part James-Stein estimator is defined by $\{1-(q-2)/||x||^2\}^+x$, where $+$ denotes the positive part, this suggests that the present θ^* is more shrinking than the James-Stein estimator for large values of $||x||^2$, while it is less shrinking for small values of $||x||^2$.

5. Numerical results

The comparison of the performance of the present estimate θ^* with that of James-Stein estimate has been performed by a Monte Carlo experiment. The experiment was performed by first sampling u from $N(0, I)$ to define $\theta=\sigma u$ and then defining the observation by $x=\theta+z$ with z independently sampled from $N(0, I)$. The trial was repeated 1000 times for a set of σ's. The positive part James-Stein estimate, θ^* and the estimate corresponding to the harmonic prior $||\theta||^{-q+2}$, computed by the formula (24) of Stein ([6], p. 361), were obtained together with the least squares estimate and the Bayes estimate when σ was known. These estimates are denoted by J-S$^+$, θ^*, harmonic, LS and BAYES, respectively. For an estimate $\theta(x)$ the loss is defined by $||\theta(x)-\theta||^2$.

The sample averages of the loss are shown in the Table. The square root of (sample variance of $||\theta^*-\theta||^2$)/1000 is denoted by SD and its values are shown within the parentheses in the Table.

Table

Comparison of average losses. SD denotes the square root of the sample variance divided by the sample size.

σ	LS	BAYES	J-S+	θ^* (SD)	harmonic
(q=2)					
0.125	1.951	0.029	1.951	0.825 (.035)	1.951
0.25	1.951	0.113	1.951	0.851 (.036)	1.951
0.5	1.951	0.385	1.951	0.946 (.036)	1.951
0.75	1.951	0.697	1.951	1.078 (.038)	1.951
1.0	1.951	0.972	1.951	1.226 (.041)	1.951
1.5	1.951	1.352	1.951	1.496 (.048)	1.951
2.0	1.951	1.565	1.951	1.689 (.054)	1.951
3.0	1.951	1.762	1.951	1.891 (.060)	1.951
5.0	1.951	1.881	1.951	1.979 (.062)	1.951
20.0	1.951	1.948	1.951	1.955 (.062)	1.951
100.0	1.951	1.951	1.951	1.948 (.061)	1.951
(q=4)					
0.125	3.895	0.061	1.406	1.052 (.037)	1.929
0.25	3.895	0.231	1.518	1.130 (.037)	1.984
0.5	3.895	0.783	1.892	1.414 (.039)	2.179
0.75	3.895	1.405	2.332	1.808 (.044)	2.438
1.0	3.895	1.948	2.708	2.231 (.051)	2.707
1.5	3.895	2.692	3.207	2.940 (.064)	3.142
2.0	3.895	3.109	3.458	3.368 (.072)	3.410
3.0	3.895	3.497	3.675	3.713 (.080)	3.660
5.0	3.895	3.738	3.810	3.852 (.083)	3.807
20.0	3.895	3.883	3.887	3.888 (.083)	3.887
100.0	3.895	3.894	3.894	3.893 (.083)	3.894
(q=8)					
0.125	7.989	0.122	1.347	1.342 (.044)	2.028
0.25	7.989	0.467	1.640	1.547 (.044)	2.213
0.5	7.989	1.598	2.634	2.297 (.047)	2.879
0.75	7.989	2.888	3.786	3.323 (.056)	3.768
1.0	7.989	4.020	4.814	4.382 (.069)	4.666
1.5	7.989	5.573	6.148	6.000 (.092)	6.033
2.0	7.989	6.439	6.844	6.863 (.106)	6.796
3.0	7.989	7.235	7.445	7.518 (.116)	7.441
5.0	7.989	7.715	7.797	7.835 (.119)	7.797
20.0	7.989	7.979	7.984	7.989 (.121)	7.984
100.0	7.989	7.991	7.991	7.991 (.121)	7.991
(q=20)					
0.125	20.152	0.306	1.422	1.718 (.047)	2.195
0.25	20.152	1.168	2.254	2.360 (.048)	2.821
0.5	20.152	3.969	5.007	4.676 (.058)	5.061
0.75	20.152	7.151	8.149	7.689 (.080)	7.941
1.0	20.152	9.946	10.843	10.541 (.106)	10.655
1.5	20.152	13.810	14.407	14.404 (.145)	14.387
2.0	20.152	15.993	16.394	16.415 (.162)	16.393
3.0	20.152	18.040	18.247	18.256 (.176)	18.247
5.0	20.152	19.316	19.397	19.399 (.184)	19.397
20.0	20.152	20.087	20.092	20.091 (.189)	20.092
100.0	20.152	20.147	20.148	20.147 (.190)	20.148

It is remarkable that the performance of the positive part James-Stein estimator J-S$^+$ becomes quite similar to that of θ^* even for $q=8$. For $q=20$, practically there is no difference. Although the numerical result is not included in the Table, the comparative pattern of shrinking of the estimators was just as predicted in the preceding sections. Incidentally, the harmonic estimator is not performing well compared with θ^* when σ is small. This is due to the difference of the prior densities $d\rho$ and $(1+\rho)^{-1}d\rho$. It is interesting to see that θ^* is performing satisfactorily even when $q=2$, where θ^* is not minimax.

6. Concluding remarks

Although the analysis of statistical characteristics of Stein type estimates has been well developed, there remained, as was mentioned by Stein ([6], p. 346), the problem of choice among admissible estimates.

Recently Takada [8] developed an interesting characterization of the positive part James-Stein estimator as the posterior mode corresponding to a particular choice of the improper prior distribution. However, by his paper, it is not clear how we can motivate the choice of such a prior distribution. The result of the present paper shows that the concept of partial ignorance combined with a particular hyperparameterized prior distribution provides a reasonable solution to the problem contemplated by Stein. Once this observation is confirmed we can quickly recognize that the same idea will be applicable to other Bayesian models with hyperparameters.

It is often mentioned that the practical application of the James-Stein estimate is rather limited. The present analysis shows that this is natural as the estimate can be viewed as an approximation to a generalized Bayes estimate with a rather limited structure of the prior distribution. It is usually necessary to incorporate further prior information on the preference of parameter values through the specification of the hyperparameterized prior distribution $p(\theta|\rho)$ to make the estimate useful for a particular application.

Acknowledgements

The author wishes to thank Professor T. Hida for drawing his attention to the Efron's article on controversies in the foundation of statistics. Thanks are due to Ms. E. Arahata for her help in preparing the numerical results reported in this paper.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. (1978). A new look at the Bayes procedure, *Biometrika*, **65**, 53-59.
[2] Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problem, *Ann. Math. Statist.*, **42**, 855-903.
[3] Efron, B. (1978). Controversies in the foundation of statistics, *Amer. Math. Monthly*, **85**, 231-246.
[4] Jeffreys, H. (1961). *Theory of Probability*, 3rd ed., Oxford University Press, London.
[5] Stein, C. (1962). Confidence sets for the mean of a multivariate normal distribution, *J. R. Statist, Soc.*, B, **24**, 265-296.
[6] Stein, C. (1973). Estimation of the mean of a multivariate normal distribution, *Proceeding of the Prague Symposium on Asymptotic Statistics*, 345-381.
[7] Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean, *Ann. Math. Statist.*, **42**, 385-388.
[8] Takada, Y. (1979). Stein's positive part estimator and Bayes estimator, *Ann. Inst. Statist. Math.*, **31**, A, 177-183.