Ann. Inst. Statist. Math.
35 (1983), Part A, 139-149

ON MINIMUM INFORMATION PRIOR DISTRIBUTIONS*

HIROTUGU AKAIKE

(Received Jan. 12, 1983)

Summary

The formulation of the concept of non-informative prior distribution over a finite number of possibilities is considered and the minimum information prior distribution is defined as the prior distribution that adds minimum expected amount of information to the posterior distribution. Numerical examples show that the definition leads to non-trivial results. An information inequality is established to assure the validity of numerical results. The relation of the present work to other works on the same subject is briefly reviewed and finally a minimax type prior distribution is introduced that exhibits the impartial property which is lacking in the minimum information prior distribution.

1. Introduction

In a practical application of the Bayes procedure the available prior information is not usually sufficient to completely specify the prior distribution. This often leads to the consideration of another prior distribution, the hyperprior distribution, over a set of possible prior distributions. The process may then be repeated indefinitely by considering a prior distribution over a set of possible prior distributions, until we come to the point where no more information is available to continue the process. The concept of non-informative or ignorance prior distribution has been developed to serve in this type of situation.

The ignorance prior distribution developed by Jeffreys [5] is well-known. However, its definition is based on the concept of invariance of the distribution by the transformation of the parameter and is limited to the case where the family of possible data distributions is continuously parametrized. Lindley [7] applied the Shannon entropy to develop an information theoretic analysis of the structure of Bayesian modeling. This work prompted the works by Zellner [8] and Bernardo [3] on the

* This work was partly supported by the United States Army Contract No. DAAG 29 -80-C-0041 in Mathematics Research Center, University of Wisconsin-Madison.

definition of the least informative prior distribution based on some definitions of the amount of information. For an extensive reference on the literature on non-informative prior distributions readers are referred to Bernardo [3].

In the present paper we consider the basic problem of specifying a prior distribution over a finite number of data distributions when no further prior information is available. Conventionally the uniform distribution which allocates equal probability to each data distribution is considered to be a reasonable choice in such a situation; see, for example, Cox and Hinkley ([4], p. 376). The analysis of Bernardo [3] also leads to this prior distribution. Here we define the minimum information prior distribution as the prior distribution which "let the data speak most" in predicting the behavior of a future observation which is independent of, but identically distributed as, the present data. This seems to provide a description of the common objective of statistical data analysis, the identification of the probability distribution that generated the present data.

A natural characterization of such a prior distribution is obtained by keeping the corresponding simultaneous distribution of the present and future observations as far away as possible from the state of independence. The deviation from the independence is measured by the Kullback-Leibler information number. By this definition the uniform prior distribution is a reasonable choice only when the possible data distributions do not show significant overlap. This is the situation where the likelihoods can clearly discriminate the hypotheses, a situation where the Bayesian modeling is practically unnecessary.

Numerical results show that when the overlap of the data distributions becomes significant the optimal choice of the prior distribution depends critically on the mutual relation of the data distributions. In particular, it is observed that some of the prior probabilities go down to zero when the overlap becomes extremely significant. These numerical examples constitute the first example of determination of non-trivial non-informative prior distributions over finite possibilities. A newly obtained information inequality assures the validity of numerically obtained minimum information prior distributions.

Comparison of the present definition with other similar definitions is briefly discussed in the final section. The appearance of zero prior probabilities by these definitions is baffling and the minimax information prior distribution is defined that minimizes the maximum expected deviation of the true distribution from the posterior distribution. Numerical results are included to show the potential of this definition for practical applications.

2. Definition of the minimum information prior distribution

Consider a set of data distributions $\{f_k(\cdot)\}$ $(k=1, 2, \cdots, K)$. The simultaneous distribution of the present and future observations x and y is defined by

$$p(y, x)=\sum_{k=1}^{K} f_k(y)f_k(x)w_k ,$$

where w_k denotes the prior probability of the kth distribution $f_k(\cdot)$. The deviation of this simultaneous distribution from the state of independence is measured by the Kullback-Leibler information (Kullback and Leibler [6])

$$I(w)=\int\int p(y, x) \log \left(\frac{p(y, x)}{p(y)p(x)}\right)dydx$$

where $p(\cdot)=\sum f_k(\cdot)w_k$.

The quantity $I(w)$ is non-negative and becomes zero when $p(y, x)=p(y)p(x)$. In this case we have $p(y|x)=p(y)$, where $p(y|x)$ denotes the probability density of y conditional on x, and the structure defined by $\{f_k(y)f_k(x)w_k\}$ does not allow any transmission of information from the present observation x to the expected behavior of the future observation y. This represents the situation where all the relevant information about y is represented by $\{f_k(y)\}$ and $\{w_k\}$. Since the specification of the prior distribution $w=\{w_k\}$ has to be done before the observation of x the above specification of w is acceptable only when we have complete information on the behavior of y.

When we are not confident in uniquely specifying a prior distribution we may consider a set of possible w's. However, this necessitates the introduction of a prior distribution over the possible prior distributions and eventually leads to the infinite digression of searching for prior distributions of prior distributions. One strategy to stop this digression is to introduce a prior distribution which is least prejudiced against every possibility. The prior distribution discussed in the preceding paragraph for which $p(y, x)=p(y)p(x)$ holds can be considered as maximally prejudiced, or informative, in the sense that no further observation of x can influence on the inference of y. If this interpretation is accepted then it is natural to consider the prior distribution with the corresponding probability distribution $p(y, x)$ furthest away from $p(y)p(x)$ as the least informative. This observation leads to the definition of the minimum information prior distribution: we call a prior distribution $\{w_k\}$ the minimum information prior distribution, with respect to $\{f_k(\cdot)\}$, when it gives the maximum of $I(w)$. In the rest

of the paper, unless stated otherwise, it is tacitly assumed that the data distributions $f_k(x)$ are mutually absolutely continuous.

3. Some analysis of $I(w)$

The basic criterion $I(w)$ can be represented as

$$I(w)=\text{Shannon entropy of } p_w(y)p_w(x) \\ -\text{Shannon entropy of } p_w(y, x)\ ,$$

where $p_w(x)$ and $p_w(y, x)$ respectively denote $p(x)$ and $p(y, x)$ defined by the prior distribution w and the Shannon entropy of a probability distribution $p(z)$ is defined by $-\int p(z)\log p(z)dz$. For the purpose of comparison of distributions the Shannon entropy may be considered as a measure of deviation from the uniform distribution. Thus the above representation of $I(w)$ shows that the minimum information prior distribution that maximizes $I(w)$ will maximize the dependence between x and y, keeping the marginal distribution $p_w(x)$ as close to the uniform distribution as possible.

In the exceptional situation where the data distributions are completely separated, i.e., $f_k(x)f_j(x)=0$ for $k\neq j$, $I(w)$ reduces to $-\sum \mathrm{w}_k \log w_k$, the Shannon entropy of the prior distribution w. This is maximized at $w_k=1/K$. This shows that when the data distributions are well separated the uniform prior distribution will provide a good approximation to the minimum information prior distribution.

When some of the data distributions show significant overlap we can expect that the solution will no longer be close to the uniform distribution. Since no single w_k can come close to 1, as this will minimize $I(w)$, we can further expect that some w_k's will be forced to go down to zero and a distribution in a lower dimensional space of w will appear as the solution. The numerical examples of the next section show the validity of these expectations.

If the concavity of $I(\mathrm{w})$ is shown that will assure the validity of the minimum information prior distribution obtained by a numerical procedure based on a local search for the maximum of $I(w)$. Consider a prior distribution $w=\alpha u+(1-\alpha)v$ defined by a pair of prior distributions u and v and α $(0\leqq\alpha\leqq 1)$. Denote $I(w)$ by $I(\alpha)$. The concavity of $I(w)$ for general w holds if it holds that

$$I(0)+\left(\frac{dI(\alpha)}{d\alpha}\right)_{\alpha=0}\geqq I(1)$$

for any pair of u and v. This inequality reduces to

$$\int\int p_u(y, x) \log \left[\frac{p_u(y, x)}{p_u(y)p_u(x)}\right] dydx \leqq \int\int p_u(y, x) \log \left[\frac{p_v(y, x)}{p_v(y)p_v(x)}\right] dydx$$

which is equivalent to

$$I(p_u, p_v) \leqq I(p_u p_u, p_v p_v) ,$$

where $I(q, p)=\int\int q(y, x) \log (q(y, x)/p(y, x))dydx$ and $p_u p_u(y, x)$ denotes $p_u(y)p_u(x)$.

This last inequality is an information inequality that shows that $p_v(y)p_v(x)$ is more sensitive to the variation of v than $p_v(y, x)$, i.e., an observation from $p_v(y)p_v(x)$ is more informative about v than that from $p_v(y, x)$. To prove the inequality we consider the minimum of

$$I(qq, pp)=\int\int q(y, x) \log \{q(y)q(x)/(p(y)p(x))\} dydx$$

for a given $p(y, x)$, under the condition $I(q, p)=\theta$, a positive constant. Here $q(y, x)$ and $p(y, x)$ denote arbitrary symmetric probability density functions with respect to the measure $dydx$ and $q(\cdot)$ and $p(\cdot)$ denote corresponding marginal distributions. The minimization leads to the variational analysis of

$$R(q)=I(qq, pp)+\lambda(I(q, p)-\theta)+\mu\left(\int\int q(y, x)dydx-1\right) ,$$

where λ and μ are Lagrange multipliers. By considering a small perturbation $r(y, x)$ $(=r(x, y))$ of $q(y, x)$ it can be seen that the stationary solution must satisfy the relation

$$\int\int r(y, x)[\log \{q(y)q(x)/(p(y)p(x))\}+\lambda \log (q(y, x)/p(y, x))]dydx=0 .$$

This shows that we have an equality

$$\log (q(y, x)/p(y, x))=c \log \{q(y)q(x)/(p(y)p(x))\}$$

and accordingly

$$I(q, p)=cI(qq, pp) ,$$

where $c=-\lambda^{-1}>0$. Due to the convexity of $I(qq, pp)$ with respect to q the stationary solution gives the minimum of $I(qq, pp)$ under the given constraints.

Since we have

$$\int\int q(y, x)dydx=\int\int \left(\frac{q(y)q(x)}{p(y)p(x)}\right)^c p(y, x)dydx$$

c must be equal to or less than 1, if $q(y)/p(y)$ and $q(x)/p(x)$ are positively correlated under $p(y, x)$. In this case $I(q, p) \leqq I(qq, pp)$ holds for any q. For the particular choice $p(y, x)=p_v(y, x)$ it can easily be seen that the positivity of the correlation holds for any symmetric $q(y, x)$. This completes the proof of the information inequality.

4. Numerical investigation

For the simplicity of numerical analysis we consider the case where the variables x and y take only integral values $0, 1, 2, \cdots, I$. The quantities useful for the numerical maximization of $I(w)$ are

$$I(w)=\sum_y \sum_x p_w(y, x)s(y, x)$$

$$\frac{\partial I(w)}{\partial w_k}=\sum_y \sum_x Dff(k, y, x)s(y, x)$$

$$\frac{\partial^2 I(w)}{\partial w_j \partial w_k}=\sum_y \sum_x \frac{Dff(j, y, x)Dff(k, y, x)}{p_w(y, x)}-2\sum_x \frac{Df(j, x)Df(k, x)}{p_w(x)} ,$$

where

$$s(y, x)=\log \{p_w(y, x)/(p_w(y)p_w(x))\} ,$$

$$Dff(k, y, x)=f_k(y)f_k(x)-f_K(y)f_K(x) \qquad \text{and}$$

$$Df(k, x)=f_k(x)-f_K(x) \ (=\sum_y Dff(k, y, x)) .$$

To apply the ordinary optimization procedure $I(w)$ is maximized with respect to $w_1, w_2, \cdots, w_{K-1}$; whereas w_K is given by $w_K=1-w_1-\cdots-w_{K-1}$.

As a typical set of data distributions $\{f_k(\cdot)\}$ we adopted a set of binomial distributions

$$f_k(x)={}_N\mathrm{C}_x\, p_k^x(1-p_k)^{N-x} ,$$

where N and p_k $(k=1, 2, \cdots, K)$ were properly chosen for each particular example. The uniform distribution $w_k=1/K$ was used as the initial guess to start the numerical optimization. An ordinary unconstrained numerical optimization procedure was applied with a minor modification to satisfy the non-negativity constraint $w_k \geqq 0$. For the examples to be discussed in the following the absolute values of the gradients at the solutions were at most of the order 10^{-6}, except for those w_k's which were zero where the gradients took significant negative values.

The first example was designed to see the effect of relative location of the data distributions on the determination of the minimum information prior distribution. Three sets of data distributions were considered, each composed of three data distributions, i.e., $K=3$. These

were defined respectively by ($p_1=0.1$, $p_2=0.5$, $p_3=0.9$), ($p_1=0.2$, $p_2=0.5$, $p_3=0.8$) and ($p_1=0.3$, $p_2=0.5$, $p_3=0.7$). The parameter N of the binomial distribution was put equal to 20. The minimum information prior distributions obtained numerically are given in Table 1 along with the corresponding p_k's. The numbers were rounded at the fourth decimal point.

Table 1. Effect of relative location ($N=20$)

k	w_k	p_k	w_k	p_k	w_k	p_k
1	.347	.1	.409	.2	.500	.3
2	.307	.5	.182	.5	.000	.5
3	.347	.9	.409	.8	.500	.7

The result of Table 1 shows that as the three data distributions come closer to each other the distribution at the center loses its prior probability. One might expect that if the data distributions are brought further closer then eventually the prior probability will concentrate on the distribution at the center. This does not happen for this example with $K=3$. However that type of behavior is observed locally in the example to be discussed after the next where $K=5$.

The second example was designed to check the effect of increased dispersions of the data distributions. With $K=3$ the p_k's used to define the binomial distributions were $p_1=0.25$, $p_2=0.5$ and $p_3=0.75$. To get distributions with successively increasing dispersions N was put equal to 80, 40, 30 and 20. The corresponding minimum information prior distributions are given in Table 2 along with the p_k's. It can be seen that as N is decreased, i.e., as the overlap of the data distributions is increased, the minimum information prior distribution deviates from the uniform distribution over the three data distributions to the one over the two end distributions, just as in the case of the first example.

Table 2. Effect of increased dispersions ($K=3$)

	N				
	80	40	30	20	p_k
w_1	.340	.373	.410	.500	0.25
w_2	.321	.255	.179	.000	0.5
w_3	.340	.373	.410	.500	0.75

The third example was chosen to illustrate further the complexity of the possible shape of the minimum information prior distribution for an increased K, the number of possible data distributions. In this example K was put equal to 5 and the p_k's were $p_1=0.1$, $p_2=0.325$, $p_3=$

0.5, $p_4=0.675$, $p_5=0.9$. The value of N was successively put equal to 70, 60, 50, 40, 30, 25, 20, 10 and 5. The corresponding minimum information prior distributions are given in Table 3 along with the p_k's. The result of Table 3 clearly suggests that some clustering of data distributions is required when there is significant overlap among the distributions.

Table 3. Effect of increased dispersions ($K=5$)

	N										
	70	60	50	40	30	25	20	15	10	5	p_k
w_1	.245	.253	.256	.262	.276	.289	.347	.361	.402	.500	.1
w_2	.196	.200	.244	.238	.224	.211	.000	.000	.000	.000	.325
w_3	.117	.094	.000	.000	.000	.000	.307	.278	.195	.000	.5
w_4	.196	.200	.244	.238	.224	.211	.000	.000	.000	.000	.675
w_5	.245	.253	.256	.262	.276	.289	.347	.361	.402	.500	.9

The fourth and the last example was designed to see the effect of the difference of dispersions among the data distributions. Only two data distributions were considered. The result is given in Table 4. It can be seen that the data distributions defined with $p_k=.5$ which have larger variances than those defined with $p_k=.9$ are receiving lower prior probabilities. Due to the relatively good separations of the data distributions the differences of the prior probabilities are rather small.

Table 4. Effect of the difference of dispersions

	N					
	20	15	10	5	2	p_k
w_1	.497	.494	.488	.471	.439	.5
w_2	.503	.506	.512	.529	.561	.9

5. Discussion

The definition of the minimum information prior distribution is based on two principles. The first is to specify the purpose of the inference based on the present data as the prediction of another similar future observation. The second is to evaluate the deviation of $p(y, x)$ from $p(y)p(x)$ by the Kullback-Leibler information $I(w)$. For the discussion of the adequacy of the Kullback-Leibler information as such criterion, see, for example, Akaike [2]. Contrary to the usual conception of the uniform distribution as the non-informative prior distribution for a finite set of possible data distributions, the numerical result has shown the

necessity of careful analysis of the mutual relation among the data distributions.

If we followed Lindley [7] we could have defined the minimum information prior distribution as that w which maximizes

$$I_0(w)=\sum_k w_k \int p_k(x) \log \left[\frac{p_k(x)}{p(x)}\right] dx \ .$$

Such a prior distribution may be characterized as the one that keeps the probability distribution $p_k(x)w_k$ over (x, k) as far away as possible from the state of independence defined by $p(x)w_k$. Since we have the relation

$$I_0(w)=\int p(x)\left\{\sum_k p(k|x) \log \left[\frac{p(k|x)}{w_k}\right]\right\} dx \ ,$$

where $p(k|x)=f_k(x)w_k/p(x)$, the prior distribution that maximizes $I_0(w)$ may also be characterized as the one that produces maximum expected change in the transition from $\{w_k\}$ to $\{p(k|x)\}$.

This definition leads to a numerical optimization problem which is simpler than that of our definition. The result corresponding to Table 3 is given in Table 5 for this definition. The computations for the cases $N=40$ and 30 were omitted. By comparing Table 5 with Table 3 we can see that the definition leads to a prior distribution which is closer to the uniform distribution than that by our definition.

Table 5. Prior distributions maximizing $I_0(w)$

	N										
	70	60	50	40	30	25	20	15	10	5	p_k
w_1	.226	.232	.239			.270	.284	.310	.363	.424	.1
w_2	.194	.193	.192			.178	.162	.117	.000	.000	.325
w_3	.158	.149	.138			.103	.108	.147	.275	.151	.5
w_4	.194	.193	.192			.178	.162	.117	.000	.000	.675
w_5	.226	.232	.239			.270	.284	.310	.363	.424	.9

The maximal data information prior distribution introduced by Zellner [8] is based on a modification of $I_0(w)$ to avoid the analytical difficulty in handling $I_0(w)$. The criterion is based on a formal use of the Shannon entropy and its technical meaning is rather unclear, unless we accept the Shannon entropy literally as a representation of the amount of information. The reference prior distribution introduced by Bernardo [3] is somewhat similar to our minimum information prior distribution. However, it is based on the concept of infinitely repeated observation of x, instead of the one single observation in our definition,

and inevitably leads to the uniform prior distribution when the number of possible data distributions is finite.

The appearance of zero prior probabilities in the foregoing numerical examples is baffling and suggests the necessity of considering other definitions of the non-informative prior distribution. The minimum information prior distribution was defined so as to minimize the expected deviation of the predictive distribution $p(y|x)$ from the original distribution $p(y)$ as measured by the Kullback-Leibler information. The concept of impartiality suggests the minimization of

$$\operatorname*{Max}_{k} \int f_k(x) \int f_k(y) \log \left[\frac{f_k(y)}{p(y|x)} \right] dydx \ .$$

We will call a posterior distribution that minimizes the above quantity the minimax information prior distribution.

Using the same notations as in preceeding sections, Table 6 shows a pair of numerically obtained minimax prior distributions. It can be seen that in the example on the left-hand side the distribution of the prior probabilities among the data distributions within a cluster is nearly

Table 6. Examples of minimax information prior distributions

	N				
p_k	25		25		p_k
.2	w_1	.086			
.225	w_2	.085	w_1	.26	.225
.25	w_3	.085			
.45	w_4	.098			
.475	w_5	.097			
.5	w_6	.097	w_2	.48	.5
.525	w_7	.097			
.55	w_8	.098			
.75	w_9	.086			
.775	w_{10}	.086	w_3	.26	.775
.8	w_{11}	.087			

uniform. The sums of the prior probabilities for the three clusters are almost equal to the corresponding prior probabilities of the minimax information prior distribution concentrated on the "cores" of the clusters. This seems to be in good conformity with what we expect of an ignorance prior as locally uniform distribution.

It has been observed (Akaike [1], pp. 29–30) that in the inferential use of the negentropy, or the Kullback-Leibler information $I(p, q)=$

$E_y \log [p(y)/q(y)]$, where E_y denotes the expectation with respect to $p(y)$, usually $p(\cdot)$ is factual and $q(\cdot)$ is hypothetical. It can be seen that in the definition of the minimum information prior distribution the roles are interchanged, while in the minimax information prior distribution the normal ordering is restored. This suggests that the latter is based on a more natural use of the Kullback-Leibler information.

Much remains to be done to confirm the practical utility of the concept of the ignorance prior distribution over finite alternatives. Nevertheless the result presented in this paper suggests that a proper combination of the predictive point of view and the concept of negentropy or the Kullback-Leibler information will lead to a useful definition.

Acknowledgements

The author is grateful to D. M. Titterington and C. F. Wu for helpful discussions on the information inequality and to E. Arahata for her help in computing.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. (1977). On entropy maximization principle, *Applications of Statistics* (ed. P. R. Krishnaiah), North-Holland, Amsterdam, 27-41.

[2] Akaike, H. (1983). Statistical inference and the measurement of entropy, *Scientific Inference, Data Analysis and Robustness* (eds. G. E. P. Box, T. Leonard and C. F. Wu), Academic Press, New York, 165-189.

[3] Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion), *J. R. Statist. Soc.*, B, **41**, 113-147.

[4] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman & Hall, London.

[5] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society of London*, Ser. A, **186**, 453-461.

[6] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statist.*, **22**, 79-86.

[7] Lindley, D. V. (1956). On a measure of the information provided by an experiment, *Ann. Math. Statist.*, **27**, 986-1005.

[8] Zellner, A. (1977). Maximal data information prior distributions, *New Developments in the Applications of Bayesian Methods* (eds. A. Aykac and C. Brumat), North-Holland, Amsterdam, 211-232.