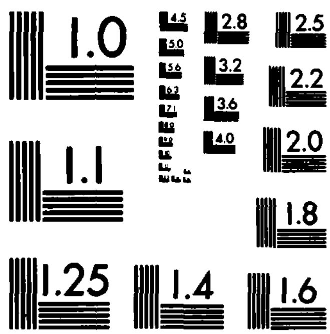
AD-A120 969 ON MINIMUM INFORMATION PRIOR DISTRIBUTIONS(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER H AKAIKE AUG 82 MRC-TSR-2410 DAAG29-80-C-0041 1/1

UNCLASSIFIED F/G 12/1 NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

MRC Technical Summary Report #2410

2

ON MINIMUM INFORMATION PRIOR DISTRIBUTIONS

Hirotugu Akaike

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

August 1982

(Received May 19, 1982)

Approved for public release
Distribution unlimited

Sponsored by

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

ON MINIMUM INFORMATION PRIOR DISTRIBUTIONS

Hirotugu Akaike*

Technical Summary Report #2410
August 1982

ABSTRACT

An explicit formulation of the concept of non-informative prior distribution over a finite number of possibilities is given. Numerical examples show that the formulation 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 discussed.

AMS(MOS) Subject Classifications: 62A15, 62F15

Key Words: Prior distribution; Bayes procedure; Entropy; Information; Information inequality; Prediction.

Work Unit Number - 4 - Statistics and Probability

* The Institute of Statistical Mathematics, Tokyo, Japan.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

The concept of non-informative prior distribution has been useful in developing Bayesian procedures for practical applications. However, rigorous analysis of the concept in the case of finite number of possible alternatives has never been sufficiently developed. In this paper a new definition, the minimum information prior distribution, is introduced based on the predictive point of view. The characteristic of the minimum information prior distribution is analyzed numerically and non-trivial examples of determination of prior probability distribution over a finite number of possibilities are reported.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

ON MINIMUM INFORMATION PRIOR DISTRIBUTIONS

Hirotugu Akaike*

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 (1946) is well-known. However, its definition is based on the concept of invariance of the distribution by the transformation of the parameter and its application is limited to the case where the family of possible data distributions is continuously parametrized. Lindley (1956) applied the Shannon entropy to develop an information theoretic analysis of the structure of Bayesian modeling. This work prompted the works by Zellner (1977) and Bernardo (1979) on the definition of the least informative prior distribution based on some deinitions of the amount of information. For an extensive reference on the

* The Institute of Statistical Mathematics, Tokyo, Japan.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

literature on non-informative prior distributions readers are referred to Bernardo (1979).

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. Coventionally 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 (1974, p. 376). The analysis of Bernardo (1979) also leads to this prior distribution. In our approach we define the mimimum information prior distribution as the prior distribution which "let the data speak most" in predicting the behavior of a future observation which is similar in nature to the present data. Such a prior distribution is obtained by keeping the 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.

Our analysis shows that the uniform 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. 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.

Much remains to be done on the theoretical analysis of the minimum information prior distribution. However, the numerical examples clearly show that the concept may find direct applications in practical problems where the data distributions can be lumped into finite number of possibilities. Comparison of the present definition with other similar definitions is briefly discussed in the final section.

2. DEFINITION OF THE MINIMUM INFORMATION PRIOR DISTRIBUTION

Consider a set of data distributions $\{f_k(\cdot)\}$ $(k = 1,2,\ldots,K)$. 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 k^{th} 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, 1951)

$$I(w) = \iint p(y,x) \log\left(\frac{p(y,x)}{p(y)p(x)}\right) dy\, dx$$

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 distribtion. 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 $-\sum w_k \log w_k$, 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 $I(w)$ 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 $(0 \leq \alpha \leq 1)$. Denote $I(w)$ by $I(\alpha)$. The concavity of $I(w)$ for general w holds if it holds that

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

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

$$\iint p_u(y,x)\log\left[\frac{p_u(y,x)}{p_u(y)p_u(x)}\right] dy\ dx \leq \iint p_u(y,x)\log\left[\frac{p_v(y,x)}{p_v(y)p_v(x)}\right] dy\ dx$$

which is equivalent to

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

where $I(q,p) = \iint q(y,x)\log(q(y,x)/p(y,x))dy\ dx$ 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 $I(qq,pp) = \iint q(y,x)\log\{q(y)q(x)/(p(y)p(x))\}\ dy\ dx$ 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 repect to the measure $dy\ dx$ and $q(\cdot)$ and $p(\cdot)$ denote corresponding marginal distributions. The minimization leads to the variational analysis of

$$R(q) = I(qq,pp) + \lambda(I(q,p) - \Theta) + \mu(\iint q(y,x)dy\ dx - 1),$$

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 $\iint r(y,x)[\log\{q(y)q(x)/(p(y)p(x))\} + \lambda\ \log(q(y,x)/p(y,x))]dy\ dx = 0$. 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 $I(q,p) = C\, I(qq,pp)$, 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

$$\iint q(y,x)dy\, dx = \iint \left(\frac{q(y)q(x)}{p(y)p(x)}\right)^{c} p(y,x)dy\, dx$$

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 $I(q,p) \leq I(qq,pp)$ 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 $0,1,2,\ldots,I$. The quantities useful for the numerical maximaization 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 $s(y,x) = \log\{p_w(y,x)/(p_w(y)p_w(x))\}$, $Dff(k,y,x) = f_k(y)f_k(x) - f_K(y)f_K(x)$ 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 $w_1, w_2, \ldots, w_{K-1}$; whereas w_K is given by $w_K = 1 - w_1 - \ldots - w_{K-1}$.

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

$$f_k(x) = {}_NC_x\, p_k^{\,x}(1-p_k)^{N-x},$$

where N and p_k $(k=1,2...,K)$ 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 $w_k \geq 0$. 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

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.

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

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.

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.

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 result of Table 3 clearly suggests that some clustering of data distributions is required when there is significant overlap among the distributions.

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 prababilities 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 introduced in this paper 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 number as such criterion, see, for example, Akaike (1982). Once the above two principles are accepted the definition of the minimum information prior distribution follow quite naturally.

Contrary to the usual preconception 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. At least in principle the present analysis can be extended to more complex situations, if only the necessary numerical procedure is properly developed.

If we followed Lindley (1956) we could have defined the minimum information prior distribution as that w_k which maximizes

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

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

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

where $p(k|x) = f_k(x)w_k/p(x)$, the prior distribution that maximizes $I_o(w)$ 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 present definition leads to a prior distribution which is closer to the uniform distribution than that by our definition. This shows that the predictive point of view demands more adaptive choice of the prior distribution.

Table 5. Prior Distributions Maximizing $I_o(w)$

	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 (1977) is based on a modification of $I_o(w)$ to avoid the analytical difficulty in handling $I_o(w)$. The criterion is based on somewhat 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 (1979) 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.

Since statistics is developed to handle problems in the real world, no procedure can claim its superiority to others unless it is tested with real applications. In that sense much remains to be done to clarify the practical implication of the minimum information prior distribution. Nevertheless, the clarity of its technical meaning and the reasonable behavior of the numerical examples suggest the potential of the minimum information prior as a conceptual resort in terminating the notorious indefinite digression in Bayesian modeling.

ACKNOWLEDGEMENT

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

REFERENCES

Akaike, Hirotugu (1982), "Statistical Inference and the Measurement of Entropy," in Proceedings of the Conference on Scientific Inference, Data Analysis and Robustness, eds. G.E.P. Box and T. Leonard, New York: Academic Press. To appear.

Bernardo, Jose M. (1979), "Reference Posterior Distributions for Bayesian Inference," (with discussion), Journal of the Royal Statistical Society, Ser. B, 41, 113-147.

Cox, D.R. and Hinkley, D.V. (1974), Theoretical Statistics, London: Chapman & Hall.

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.

Kullback, S. and Leibler, R.A. (1951), "On Information and Sufficiency," Annals of Mathematical Statistics, 22, 79-86.

Lindley, D.V. (1956), "On a Measure of the Information Provided by an Experiment," Annals of Mathematical Statistics, 27, 986-1005.

Zellner, Arnold (1977), "Maximal Data Information Prior Distributions," in New Developments in the Applications of Bayesian Methods, eds. A. Aykac and C. Brumat, Amsterdan: North-Holland, 211-232.

HA/jk

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 2410	2. GOVT ACCESSION NO.: AD-A120 969
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): ON MINIMUM INFORMATION PRIOR DISTRIBUTIONS	5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Hirotugu Akaike	8. CONTRACT OR GRANT NUMBER(s): DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709	12. REPORT DATE: August 1982
	13. NUMBER OF PAGES: 15
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Prior distribution; Bayes procedure; Entropy; Information; Information inequality; Prediction.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An explicit formulation of the concept of non-informative prior distribution over a finite number of possibilities is given. Numerical examples show that the formulation 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 discussed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)