THE PROBABILITY OF CONVERGENCE OF AN ITERATIVE PROCESS OF INVERTING A MATRIX

By Joseph Ullman

Columbia University

Introduction. The inversion of a matrix is a computational problem of wide application. This is a further study of an efficient iterative method of matrix inversion described by Harold Hotelling [1], with an examination of the probability of convergence in relation to the accuracy of the initial approximation. The lines of investigation were suggested both by his article and by helpful comments made during the course of the research.

The inverse of a matrix can be obtained to any desired degree of accuracy by using a variation of the Doolittle method, and starting with a sufficient number of accurate decimal places in the matrix being inverted. This procedure becomes inefficient if the order of the matrix is large, or if the desired degree of accuracy is very great. In either case the efficiency can be greatly increased by first obtaining an approximation to a small number of decimal places and then applying a method of iteration until the desired accuracy is achieved.

1. Iterative methods. Hotelling's method of iteration is as follows. Let A be the matrix to be inverted and let C_0 be the approximation to the inverse. Calculate in turn $C_1, C_2, \cdots$ where,

$$C_{m+1} = C_m(2 - AC_m). \tag{1.1}$$

This sequence of matrices will converge to the inverse of A if the roots of

$$D = 1 - AC_0, \tag{1.2}$$

are all less than unity in absolute value.

The iterative method (1.1) will be generalized to yield a class of iterative methods, one element of which will be shown to be more efficient, in certain cases, than method (1.1). The generalized iterative method is,

$$C_{m+1} = C_m\{1 + (1 - AC_m) + (1 - AC_m)^2 + \cdots + (1 - AC_m)^{k-1}\}. \tag{1.3}$$

For every k, the condition for convergence is that the roots of the matrix (1.2) all be less than unity in absolute value.

A method of comparing the efficiency of these different iterative methods arises from the following considerations. Since

$$C_0 = A^{-1}(AC_0), \tag{1.4}$$

which is equivalent to

$$C_0 = A^{-1}(1 - D), \tag{1.5}$$

it follows that

$$A^{-1} = C_0(1 - D)^{-1}. \tag{1.6}$$

When the roots of D are all less than unity in absolute value, (1.6) has the infinite expansion,

$$A^{-1} = C_0(1 + D + D^2 + D^3 + \cdots). \tag{1.7}$$

The general iterative process (1.3) generates the infinite series in the following manner,

$$(1 + D + D^2 + \cdots + D^{k-1})(1 + D^k + \cdots + D^{k(k-1)}) \cdot (1 + D^{k^2} + \cdots + D^{k^2(k-1)}) \cdots . \tag{1.8}$$

Each parentheses corresponds to one iteration. Hence k^m terms are generated by m iterations. In order to achieve the accuracy of n terms in (1.7), $m = \log_e n/\log_e k$ iterations are required. Each iteration involves k matrix multiplications, so that $km = k \log_e n/\log_e k$ is the total number of matrix multiplications necessary to achieve this degree of accuracy.

The integer for which this is a minimum is three. Therefore the "most efficient" method of iteration is,

$$C_{m+1} = C_m\{1 + (1 - AC_m) + (1 - AC_m)^2\}. \tag{1.10}$$

If the desired degree of accuracy can be achieved by one application of (1.1), or by two applications of (1.1) but not by one application of (1.10), then (1.1) is preferable.

2. The condition for convergence. The sequence,

$$C_1, \quad C_2, \quad C_3, \quad \cdots \tag{2.1}$$

obtained from (1.3) will converge to the inverse of A if the roots of

$$D = 1 - AC_0, \tag{2.2}$$

are all less than unity in absolute value. The following assumptions determine the nature of D.

We assume that the expected value of each element of the first approximation C_0 is equal to the corresponding element of the exact inverse of A. The actual values of the elements of C_0 will deviate from their expected values. We will consider two important cases. If the deviations are entirely due to the fact that the elements of C_0 are only accurate to a limited number of decimal places, say k, then the deviations may be regarded as distributed with constant density over a range of length 10^{-k}. It will be assumed that the deviations of the elements of C_0 from their expected values are independent. While this case arises in practice, we will first treat a closely related case, which lends itself to exact treatment more readily. We assume that the deviations of the elements of C_0

are normally distributed about their expected values, with variance $\mu = 10^{-2k}/12$. The variance μ is the same as that which arises if the probability density is uniform with range 10^{-k}.

The elements of E, the matrix of deviations,

$$E = A^{-1} - C_0 , \tag{2.3}$$

are independently and normally distributed. Combining (2.2) and (2.3) we obtain

$$D = 1 - AC_0 = A(A^{-1} - C_0) = AE. \tag{2.4}$$

Let p be the order of the matrix A. Each element of D will be a linear combination of p independently and normally distributed variables, and therefore will itself be normally distributed. A sufficient condition for all the roots of D to be less than unity in absolute value, and hence for the process of iteration to converge, is for the sum of the squares of the elements of D to be less than unity in absolute value. We will use the following notation

(2.5) d_{ij}: the element of D in the ith row and jth column,

$$N_D^2 = \sum_i \sum_j d_{ij}^2 .$$

A procedure suggested by this relationship is to determine the probability distribution of N_D^2, so that probability statements concerning the absolute value of the roots of D can be made. Because the elements of D are not all independent, no multiple of N_D^2 can be expected to have the $\chi^2_{(p^2)}$ distribution.[1]

The distribution of N_D^2 is shown to be closely related to the chi-square distribution in the next section, and on the basis of this relationship, lower bounds to the probability of convergence of the iterative process are developed in section 4. In section 5 the exact distribution of the norm is obtained for a general class of cases. The final section is concerned with the validity of applying the results of this study to a practical situation, where the deviations of the elements of C_0 from their expected values are uniformly, rather than normally, distributed.

3. An equivalance. Let e_{ij} be the element of E in the ith row and the jth column, and a_{ij} be the element of A in the ith row and the jth column. From (2.4) and (2.5), we find that

$$d_{ij} = \sum_k a_{ik} e_{kj} .$$

Since the elements of E are independently and normally distributed with variance $\mu = 10^{-2k}/12$ it follows readily that

$$E[e_{ij}e_{kh}] = \delta_{ik}\delta_{jh}\mu. \tag{3.2}$$

[1] The number in the parentheses will indicate the number of degrees of freedom of the chi-square distribution.

Making use of (3.1) and (3.2), we find that for two d_{ij} in the same column,

$$E[d_{ij}\,d_{kj}] = \mu \sum_t a_{it}\,a_{kt}\,, \tag{3.3}$$

while for any two d_{ij} in different columns,

$$E[d_{ij}d_{kh}] = 0. \tag{3.4}$$

From (3.3) and (3.4) it follows that it is permissible to regard the elements of the p columns of D as the coordinates of p independently selected points from a multivariate normal universe with covariance matrix $\sigma = \mu AA'$. We will let $\lambda = \sigma^{-1}$.

The moment generating function of the sum of squares of the coordinates of any point is

$$\frac{|\lambda|^{\frac{1}{2}}}{|\lambda - 2t|^{\frac{1}{2}}}. \tag{3.5}$$

This can also be written as

$$\frac{1}{(1 - 2\sigma_1 t)^{\frac{1}{2}}(1 - 2\sigma_2 t)^{\frac{1}{2}} \cdots (1 - 2\sigma_p t)^{\frac{1}{2}}}, \tag{3.6}$$

where $\sigma_1, \cdots, \sigma_p$ are the characteristic roots of σ.

Since N_D^2 is the sum of p independent expressions of this type, its moment generating function is the pth power of (3.6),

$$\frac{1}{(1 - 2\sigma_1 t)^{\frac{1}{2}p} \cdots (1 - 2\sigma_p t)^{\frac{1}{2}p}}. \tag{3.7}$$

This expression is the moment generating function of

$$\sigma_1 \chi^2_{(p)1} + \sigma_2 \chi^2_{(p)2} + \cdots + \sigma_p \chi^2_{(p)p}\,, \tag{3.8}$$

where the $\chi^2_{(p)i}$ are all independent.

Writing the roots as

$$\sigma_0\,,\ \sigma_0 - k_1\,,\ \cdots,\ \sigma_0 - k_{p-1}\,, \tag{3.9}$$

where σ_0 is the largest root of σ, and all $k_i > 0$, it follows that N_D^2 has the same distribution as

$$\sigma_0 \sum_i \chi^2_{(p)i} - \sum_{j=1}^{p-1} k_j \chi^2_{(p)j}\,. \tag{3.10}$$

Therefore, making use of the reproductive power of χ^2, we obtain

$$\begin{aligned} P\{N_0 < 1\} &= P\left\{\sigma_0 \sum_i \chi^2_{(p)i} < 1 + \sum_{j=1}^{p-1} k_j \chi^2_{(p)j}\right. \\ &= P\left\{\sigma_0 \chi^2_{(p^2)} < 1 + \sum_{j=1}^{p-1} k_j \chi^2_{(p)j}\right\}. \end{aligned} \tag{3.11}$$

By making special assumptions about the k_i, close approximations to the probability that N_D will be less than one, and hence that the process of iteration will converge, can be obtained. Instead of following this procedure, it is more desirable to have definite lower bounds for the probability that N_D will be less than one. This will lead to an overstatement in the number of decimal places of accuracy necessary in the first approximation C_0 to assure convergence, but it will practically eliminate the possibility of having to recalculate the first approximation, and hence will lead to greater efficiency in the long run.

4. The derivation of the formula for determining the required degree of accuracy. The inequality used in this section is derived in two steps from (3.10). Since $k_i > 0$ $(i = 1, \cdots, p-1)$ it follows immediately that

$$P\{N_D < 1\} > P\{\sigma_0 \chi^2_{(p^2)} < 1\}. \tag{4.1}$$

In order to use this inequality, the upper bound for σ_0

$$\sigma_0 \leq (\operatorname{tr} \sigma^t)^{1/t} \bullet \tag{4.2}$$

can be used. For $t = 1$,

$$\sigma_0 \leq \operatorname{tr} \sigma = \operatorname{tr} \mu A A' = \mu \operatorname{tr} AA' = \mu N_A^2 . \tag{4.3}$$

Dr. Wald pointed out that using (4.2) for $t = 1$ reduces the amount of information retained in (4.1) to that which is contained in the inequality,

$$N(D) \leq N(A)N(E). \tag{4.4}$$

A closer upper bound is feasible in any particular case, and can be introduced at this point by letting $t = 2$ or $t = 3$. The following formula will be developed for the general case, making use of (4.3).
Substituting (4.3) in (4.1), we obtain

$$P\{N_D < 1\} > P\left\{\chi^2_{(p^2)} < \frac{1}{\mu N_A^2}\right\}. \tag{4.5}$$

It is desirable to separate the effects of the order of the matrix A on convergence, and the order of magnitude of the elements. Hence we introduce as a measure of the average size of the a_{ij} their root mean square m, so that

$$m^2 = \sum_i \sum_j a_{ij}^2/p^2. \tag{4.6}$$

Hence

$$N_A = pm. \tag{4.7}$$

The final form of the inequality is

$$P\{N_D < 1\} < P\left\{\chi^2_{(p^2)} < \frac{12 \cdot 10^{2k}}{p^2 m^2}\right\}. \tag{4.8}$$

First we will obtain an expression for the number of decimal places required in the first approximation to make the probability of convergence at least .999. Then the expression will be checked directly by means of (4.8) and tables of the chi-square function.

For large values of p, $\sqrt{2\chi^2_{(p^2)}}$ is approximately normally distributed with mean value $\sqrt{2p^2 - 1}$ and unit variance [2], [5]. Applying this transformation to the right hand side of (4.8), and noting that 3.1 standard deviations is slightly greater than the deviation corresponding to .999, we obtain as the condition for

$$P\left\{\chi^2_{(p^2)} < \frac{12 \cdot 10^{2k}}{p^2 m^2}\right\} \geq .999 \tag{4.9}$$

or

$$P\left\{2\chi^2_{(p^2)} < 2 \cdot \frac{12 \cdot 10^{2k}}{p^2 m^2}\right\} \geq .999 \tag{4.10}$$

that it is sufficient that

$$\sqrt{\frac{24 \cdot 10^{2k}}{p^2 m^2}} - \sqrt{2p^2 - 1} \geq 3.1. \tag{4.11}$$

This is equivalent to

$$k \geq \log_{10} p + \log_{10} m + \log_{10}\left(\sqrt{p^2 - \tfrac{1}{2}} + \frac{3.1}{\sqrt{2}}\right) - \tfrac{1}{2}\log_{10} 24 + \log_{10}\sqrt{2}. \tag{4.12}$$

Since the characteristic of a logarithm is insensitive to the argument, rounding off will introduce a negligible error, and we finally obtain an upper limit to the lower bound of k,

$$k > \log_{10} m + \log_{10} p + \log_{10} (p + 3) - .55. \tag{4.13}$$

In order to verify the accuracy of (4.13) for small values of p, certain values of p, k and m are chosen and the probabilities associated with (4.8) determined [2]. The entries in brackets are the corresponding values of k determined from (4.13).

A typical example will illustrate the use of table on facing page. Let the matrix A to be inverted be a fourth order correlation matrix. The mean magnitude m is about $\frac{1}{2}$ and $p = 4$. If the first approximation C_0 is obtained to one place accuracy, then the probability that the sequence $C_1, C_2, \cdots$ will converge to A^{-1} will be greater than .999. Using formula (4.13), we obtain $k = .53$. Since one is the first integer greater than .53, the table verifies the use of the formula.

Although the formula was developed on the assumption that p is large, every value calculated is consistent with the table. This lends support to its use for small values of p.

*The Probability of Convergence of the Iterative Process**

	k \ p	2	3	4	5
$m = \frac{1}{2}$	−1	0+	0+	0+	0+
	0	[.05].982	[.33].199	[.53]0+	[.70]0+
	1	1−	1−	1−	1−
$m = 2$	−1	0+	0+	0+	0+
	0	[.85].051	[.93]0+	0+	0+
	1	1−	1−	[1.13].715	[1.30]0−
	2	1−	1−	1−	1−
$m = 10$	0	0+	0+	0+	0+
	1	[1.35].439	[1.63]0+	1[.83]0+	0+
	2	1−	1−	1−	[2.00]1−

* "1−" means greater than .999.

It has already been pointed out that k is not sensitive to rounding off of the argument of the logarithm. Thus for $p = 20$ and $m = 2$, we can let $\log_{10} m = .3$, $\log_{10} p = 1.3$, $\log_{10}(p + 3) = 1.36$ and obtain

$$k = .3 + 1.3 + 1.36 - .55 = 2.41,$$

from which it follows that three decimal place accuracy in C_0 will practically insure convergence of the iterative process.

5. The mean, variance, and exact distribution. To obtain the moments of N_D^2, the most convenient form to use is (3.8). Since the $\chi^2_{(p)i}$ are independent

$$E[N_0^2] = E[\sum_i \sigma_i \chi^2_{(p)i}] = p \sum_i \sigma_i . \tag{5.1}$$

$$\begin{aligned} \sigma_{N_D^2} &= E[N_D^2] - (E[N_D^2])^2 \\ &= E\left[\sum_{i=1}^{p} \sigma_i^2 (\chi^2_{(p)i})^2 + 2 \sum_{i<j} \chi^2_{(p)i} \chi^2_{(p)j} \sigma_i \sigma_j\right] - (p \sum_i \sigma_i)^2 \\ &= (2p + p^2) \sum_i \sigma_i^2 + 2p^2 \sum_{i<j} \sigma_i \sigma_j - p^2 \sum_i \sigma_i^2 - 2p^2 \sum_{i<j} \sigma_i \sigma_j \\ &= 2p \sum_i \sigma_i^2 . \end{aligned} \tag{5.2}$$

These can be expressed in terms of the elements of A and the variance of the elements of E, since

$$\begin{aligned} \sum_i \sigma_i &= \text{tr}\,(\sigma) = \mu \,\text{tr}\,(AA') = \mu N_A^2 , \\ \sum_i \sigma_i^2 &= \text{tr}\,\sigma^2 = \mu^2 \,\text{tr}\,(AA'AA'). \end{aligned} \tag{5.3}$$

The exact distribution of N_D^2 can be obtained readily when p is even. In this case the infinite integral,

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{e^{-i\,t\,N_D^2}\,dt}{(1 - i2\sigma_1 t)^{p/2} \cdots (1 - i2\sigma_p t)^{p/2}}, \tag{5.4}$$

can be evaluated by contour integration. The integral satisfies the conditions given in Whittaker and Watson [3, sec. 622], if the semicircle of the contour is taken on the lower half of the complex plane.

For the case $p = 2$, for example, there are simple poles, at $t = \frac{-i}{2\sigma_1}, \frac{-i}{2\sigma_2}$. The sum of the residues at these poles, multiplied by i yields the exact distribution:

$$\frac{\sigma_1 e^{-N_D^2/2\sigma_1}}{2(\sigma_1 - \sigma_2)} + \frac{\sigma_2 e^{-N_D^2/2\sigma_2}}{2(\sigma_2 - \sigma_1)}. \tag{5.5}$$

For even values of p greater than 2, the values of the residues can be obtained by repeated differentiation.

6. Summary. We are now in a position to discuss the applicability of the results of this paper to the problem which arises most frequently in practice. The elements of the first approximation to the inverse will deviate from their expected values only because the first approximation is carried to a limited number of places, say k. In this case the deviations will be distributed with constant density over a range of length 10^{-k}. The elements of E, the matrix of deviations,

$$E = A^{-1} - C_0, \tag{6.1}$$

are now each independently distributed, but with uniform density, range 10^{-k} and mean equal to zero. From (2.4)

$$D = AE, \tag{6.2}$$

we observe that each element of D will be a linear combination of p independently and rectangularly distributed variables, each with mean zero and range 10^{-k}. The analysis of sections 3, 4, and 5 will be valid if d_{ij} can be considered to be normally distributed.

There is much experimental evidence and theoretical justification for assuming that the elements of D are normally distributed. A sufficient condition that the d_{ij} approach normality as p increases is that the sum of the a_{ij}^2 in any row of A be divergent as the order of the matrix approach infinity, while at the same time every element be less than some constant value independent of the order of the matrix [4].

The experimental and theoretical evidence supporting the approach of the d_{ij} to normality, the fact that the logarithms are insensitive to errors of approxi-

mation in their arguments and the fact that the lower bounds to the probability of convergence of the iterative process are used, all lend support to the formula

$$k > \log_{10} m + \log_{10} p + \log_{10} (p + 3) - .55.$$

for determining the number of places (k) necessary in the first approximation (C_0) to the inverse of A, a matrix of order p whose elements have mean size m, to make the probability at least .999 that the process of interation will yield a sequence of matrices which will converge to the true inverse. The ultimate justification of the use of this formula can only be by the results of its application in practice.

REFERENCES

[1] HAROLD HOTELLING, "Some new methods in matrix calculation," *Annals of Math. Stat.*, Vol. 14 (1943) pp. 1–34.

[2] FISHER and YATES, *Statistical Table for Biological, Agricultural and Medical Research*, p. 31.

[3] WHITTAKER and WATSON, *Modern Analysis*, 4th edition, pp. 113–114.

[4] J. V. USPENSKY, *Introduction to Mathematical Probability*, 1935, Chapter 14.

[5] E. B. WILSON and M. M. HILFERTY, "The distribution of chi-square," *Proc. Nat. Acad. Science*, Vol. 17 (1931), pp. 684–688.