TECHNICAL MEMORANDUM

98

Estimates of round off error in the inversion of positive definite matrices

A. Hollingsworth and G.J. Cats

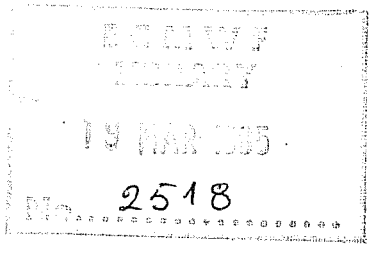
Research Department

March 1985

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

ABSTRACT

A re-examination of the test for ill-conditioning of the analysis equations suggests that the currently implemented test is unnecessarily severe. A revised test which is effective even in the most unlikely circumstances and which is less restrictive than the present one is proposed. It is noted that even this test could be substantially relaxed if account is taken of the statistical distribution of round-off errors.



C O N T E N T S

1. INTRODUCTION

The purpose of this note is to review an earlier error analysis by Cats and Robertson (1980) on the growth of round-off error in the solution of the linear equations which occur in the Optimum Interpolation (O/I) analysis algorithm; the matrices in the problem are symmetric and positive definite. Cats and Robertson (1980) suggested that the O/I matrix is ill-conditioned if

$$K^{-1} < \frac{n^3 p}{\gamma} \qquad (1)$$

where K^{-1} is an appropriate condition number, n is the order of the matrix, p is the machine precision, 2^{-48}, and γ is the relative accuracy which is demanded of the solution.

We suggest that the test be revised to read

$$\max_{\substack{i=1,n \\ j=1,n}} \left| \frac{1}{b_{ij}} \right| \leq \frac{2n^4 p}{\gamma} \qquad (2)$$

where b_{ij} are the elements of the inverse of the O/I index. This is a safe criterion which is less restrictive than (1).

Two widely used condition numbers in error analyses of these kinds are those based on the L_2 and L_∞ norms. The L_2 norm is more convenient for analytical work than the L_∞ norm; on the other hand the L_∞ is easier to use in practical work.

The condition of the O/I equations is discussed theoretically using the L_2 norm in Sect. 2. The results indicate that the equations are well-conditioned if some suggestions by Cats (1981) are followed. Sect. 3 gives estimates of relative error for the solutions of the O/I equations using both the L_2 and L_∞ norms, and justifies the suggestion (2) for improving the current operational test on the condition of the equations. Sect. 4 summarises the results.

2. CONDITION NUMBERS FOR THE O/I MATRIX

The discussions of the influence of round-off error in the O/I context given by Cats and Robertson (1980), and Cats (1981) follows closely that given in Rabinowitz and Ralston (1978, hereafter called **RR**) for the Crout and Doolittle methods for the solution of linear equations. The method used in the analysis algorithm is in fact the Cholesky method, which is substantially more economical, and therefore less sensitive to round-off error. (The Cholesky method requires the decomposition of a symmetric positive definite matrix A to a form $A=L^{T}L$, where L is lower triangular).

The estimation of the probable error of such algorithms is sometimes rather difficult. **RR**(page 11) make the point that when estimating the maximum round-off error in a calculation, the result will be r.p where r is the operation count and p is the machine precision. On the other hand, the probable error will be of order $r^{\frac{1}{2}}.p$. They remark that a common way to estimate round-off error in a long calculation is to find a maximum bound on the error and then replace r (where r is an operation count) by $r^{\frac{1}{2}}$.

2.1 Definition of the condition number

Much of the discussion about the numerical stability of the algorithms hinges on the condition number K(A) of the matrix A in the problem

$$\underline{\underline{A}}\underline{x} = \underline{b}$$

where $\underline{x}$ is the unknown. K is defined to be

$$K = \|A\| \ \|A^{-1}\| \qquad (3)$$

in some suitable norm. We will work first with the L_2 norm and assume that A is positive definite - hence

$$\|A\| = \lambda_{max} \text{ and } \|A^{-1}\| = 1/\lambda_{min}$$

where λ_{max} and λ_{min} are the maximum and minimum eigenvalues of A and

$$K_2 = \frac{\lambda_{max}}{\lambda_{min}}$$

2.2 Calculation of the L_2 condition number

For the O/I matrix the condition number is rather easy to estimate in simple cases. We begin with the definition

$$\underline{\underline{A}} = \underline{\underline{P}} + \underline{\underline{\sigma}}\ \underline{\underline{\theta}}\ \underline{\underline{\sigma}}$$

where $\underline{\underline{P}}$ and $\underline{\underline{\theta}}$ are the correlation matrices for prediction error and observation error, while the diagonal matrix $\underline{\underline{\sigma}}$ gives the non-dimensional rms observation error. In the simple case of equal uncorrelated errors we have

$$\underline{\underline{\sigma}}\ \underline{\underline{\theta}}\ \underline{\underline{\sigma}} = \sigma^2\ \underline{\underline{I}}$$

a) The L_2 norm of A^{-1}

The following lemma was used by Cats - the proof below is due to Wilkinson, 1964 (p.99).

Lemma If $\underline{\underline{B}}$ and $\underline{\underline{C}}$ are positive semi-definite then

$$\tau\ (\underline{\underline{B}} + \underline{\underline{C}}) \geq \tau\ (\underline{\underline{B}}) + \tau\ (\underline{\underline{C}})$$

where τ denotes the absolute value of the minimum eigenvalue.

Proof

$$\tau\ (\underline{\underline{B}} + \underline{\underline{C}}) = \min_{|X|=1} \{\underline{X}^T, (\underline{\underline{B}} + \underline{\underline{C}})\ \underline{X}\}$$

$$= \min_{|X|=1} \{\underline{X}^T,\ \underline{\underline{B}}\ \underline{X} + \underline{\underline{C}}\ \underline{X}\}$$

$$= \min_{|X|=1} [\{\underline{X}^T,\ \underline{\underline{B}}\ \underline{X}\} + \{\underline{X}^T,\ \underline{\underline{C}}\ \underline{X}\}]$$

$$\geq \min_{|X|=1} \{\underline{X}^T,\ \underline{\underline{B}}\ \underline{X}\} + \min_{|X|=1} \{\underline{X}^T,\ \underline{\underline{C}}\ \underline{X}\} \text{ by positiveness}$$

$$= \tau\ (\underline{\underline{B}}) + \tau\ (\underline{\underline{C}}) \quad \text{Q.E.D.}$$

This lemma is useful because $\tau(\underline{\underline{\sigma}}\ \underline{\underline{\theta}}\ \underline{\underline{\sigma}})$ will be bounded away from zero, even though $\tau\ (\underline{\underline{P}})$ may be very close to zero. The latter case can occur when there are nearly coincident observations, or when the height streamfunction correlation is very close to 1, so that height and wind data are regarded as redundant.

We can bound the norm of A^{-1} by $1/\tau(\underline{\underline{\sigma}}\ \underline{\underline{\theta}}\ \underline{\underline{\sigma}})$. There are three cases to consider:

(a) In the case of equal uncorrelated errors the bound is just $1/\sigma^2$.

(b) If the observational errors are not equal, but are still uncorrelated, the bound is $1/\sigma^2_{min}$ where σ_{min} is the rms error of the most accurate observation.

(c) If the observational errors are correlated (e.g. radio-sonde heights or winds in the vertical, or satellite thicknesses in the horizontal), the situation is more complicated. In the case of SATEMS, the occurrence of many SATEMS separated by the correlation length could lead to very small eigenvalues. Similar problems can arise in the vertical. A practical solution (Cats, 1981) is to require that the observational error correlation matrix can be split into two parts

$$\underline{\underline{\sigma}}\ \underline{\underline{\theta}}\ \underline{\underline{\sigma}} = \sigma_1^2\ \underline{\underline{I}} + \underline{\underline{\sigma}}_2\ \underline{\underline{\theta}}\ \underline{\underline{\sigma}}_2$$

the first of which is positive definite and the second is positive semi-definite. Physically this can be done by requiring that two reports cannot have perfectly correlated observational error at zero separation. As implemented by Cats, the current algorithm limits this correlation to a maximum value of $\kappa = 0.8$. Thus we have

$$\|A^{-1}\| \leq \frac{1}{\sigma^2_{min}\ (1-\kappa)} \qquad (4)$$

b) The L_2 norm of A

An upper bound for $\|A\|$ is easily found:

$$\|\underline{\underline{A}}\| = \rho(\underline{\underline{A}}) = \lambda_{max} \leq \Sigma\lambda_i = tr(A) = N + \sum_i \sigma_i^2,$$

since $\underline{\underline{P}}$ and $\underline{\underline{\theta}}$ are correlation matrices. We then have

$$\|\underline{\underline{A}}\| \leq N\ (1 + \bar{\sigma}^2) \text{ where } \bar{\sigma}^2 = \frac{1}{N}\sum_i \sigma_i^2$$

This bound will be attained if all the observations are coincident and the observation errors are perfectly correlated. In more general cases there will be $\sqrt{N}$ eigenvalues which account for most of the variance, with typical magnitude $\sqrt{N}(1+\bar{\sigma}^2)$. Therefore we expect to have

$$\|\underline{\underline{A}}\| \sim \sqrt{N}\ (1 + \bar{\sigma}^2), \tag{5}$$

while in all but the most pathological cases we will certainly have

$$\|A\| < N\ (1 + \bar{\sigma}^2) \tag{6}$$

c) Estimates of the L_2 condition number

Using (5) and (6) in (3) gives

$$K \leq \frac{N(1 + \bar{\sigma}^2)}{\sigma_{min}^2(1 - \kappa)} \tag{7}$$

whilst substituting (4) and (5) leads to

$$K \sim \frac{\sqrt{N}\ (1 + \bar{\sigma}^2)}{\sigma_{min}^2\ (1 - \kappa)} \tag{8}$$

As noted by **RR**, (7) is an extreme upper bound while (8) is a realistic bound.

3. ESTIMATES OF RELATIVE ERROR ARISING FROM ROUND OFF

RR consider the estimates of relative error in the problem

$$\underline{\underline{A}}\ \underline{x} = \underline{b}$$

in two parts, viz response to round-off error in the data b, and response to the round-off errors in $\underline{\underline{A}}$ and $\underline{x}$. Cats considers both simultaneously, but his

results are controlled only by the second problem, i.e. the response to round-off error in $\underline{\underline{A}}$ and x.

RR show that for the algorithms we use, the computed solution $\underline{x}$ satisfies exactly the equation $(\underline{\underline{A}} + \underline{\underline{E}})\ \underline{x}_c = \underline{b}$ from which we deduce the result

$$\frac{\|x_t - x_c\|}{\|x_t\|} \leq \frac{K(A)\|E\|/\|A\|}{1 - K(A)\|E\|/\|A\|} \tag{9a}$$

$$= \frac{\|A^{-1}\|\ \|E\|}{1 - \|A^{-1}\|\ \|E\|} \quad , \tag{9b}$$

where x_t is the true solution and $\|E\|$ is a matrix whose norm can be bounded.

RR deduce upper bounds for the norm of $\underline{\underline{E}}$ under the assumption that $\underline{\underline{A}}$ is "row-equilibrated", which is defined to mean that the L_∞ norm of each row of $\underline{\underline{A}}$ is unity:

i.e. for each i: $\max_j \|A_{ij}\| = 1$

An artibrary matrix A can be equilibrated by dividing each row by the maximum absolute element in the row. (For positive definite matrices this maximum element is not necessarily on the diagonal - we only know that $a_{ij} \leq \max(|a_{ii}|, |a_{jj}|)$. Symmetric matrices lose their symmetry in this process.

In our case we know that the range of the infinity norm of the rows of $\underline{\underline{A}}$ is between $(1 + \sigma^2_{min})$ and $(1 + \sigma^2_{max})$, where σ_{max} is the biggest observation error, and σ_{min} the smallest. If this range is not too big, $\underline{\underline{A}}$ may be row-equilibrated for the purpose of the estimation of $\underline{\underline{E}}$, by dividing all its elements by its maximum element $(1 + \sigma^2_{max})$ - denoted by ε.

RR show that for all standard norms $\|E\|$

$\|E\| < g(2n+1)\, p\varepsilon$ with $g \leq 2$ (the factor ε comes from equilibration) if all the dot products are done in double precision; if the entire computation is carried out in single precision there is the much weaker bound

$$\|E\|_\infty \leq 1.06\,(3n^2 + n^3)g\, p\varepsilon.$$

They further comment that these are extreme upper bounds which take no account of the statistical distribution of round-off errors.

3.1 Estimated relative error using the L_2 norm

Since the second of the above bounds does not apply for the L_2 norm we will use the first bound and assume that the final result is truncated to half precision; in effect we have done the calculation in double precision for half words. Then a realistic estimate of the relative error of the solution, taking note of earlier comments on operation counts (Section 2), is given by the following expression with $g = 2$ and $p = 2^{-24}$.

$$RE = \frac{\|\delta x\|}{\|x\|}_t \leq (\|A^{-1}\|\,\|E\|)/(1-\|A^{-1}\|\,\|E\|) \sim \|A^{-1}\|\,\|E\|\,\varepsilon$$

$$\sim \frac{2\,(2\sqrt{n}+1)2^{-24}}{\sigma^2_{min}\,(1-\kappa)}\,(1+\sigma^2_{max})$$

$$\sim \frac{4\sqrt{n}\,2^{-24}}{\sigma^2_{min}\,(1-\kappa)}\,(1+\sigma^2_{max})$$

Typically $\sigma^2_{min} \sim 0.1$ (giving $\sigma_{min} \geq 0.3$), but in exceptional circumstances $\sigma^2_{min} \sim 0.01$. Therefore

$$RE < \frac{4}{.2 \times 10^{-2}}\,2^{-24}\sqrt{n}\,(1+\sigma^2_{max})$$

$$= \frac{2^{-24}}{5}\sqrt{n}$$

$$= 1.2 \times 10^{-4}\sqrt{n}\,(1+\sigma^2_{max})$$

Thus the relative error will be less than 4×10^{-3} for matrices of order 10^3 in normal cases, and will be of order 10^{-1} for matrices of order 10^3 only in the most pathological cases.

The conclusion is that there should be no reason for concern about round-off errors, if the matrix size is less than 10^3, so long as we are satisfied with relative errors between 5×10^{-3} and 10^{-1}. This result is reassuring about the stability of the algorithm, but not the accuracy. As shown in the next section, the accuracy estimates can be much improved using the L_∞ norm.

3.2 Estimated relative error using the L_∞ norm

The L_2 norm is not readily computable in a real situation and so we estimate the relative errors using the L_∞ norm.

As shown in standard texts

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|$$

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

and $\|A\|_1 = \|A^T\|_\infty$.

The condition number based on the infinity norm is

$$K_\infty = \left(\max_{i=1,n} \sum_{j=1}^{n} |a_{ij}| \right) . \left(\max_{i=1,n} \sum_{j=1}^{n} |b_{ij}| \right)$$

where b_{ij} are the elements of A^{-1}. Cats and Robertson define the quantity K_m as

$$K_m = \left(\max_{\substack{i=1,n \\ j=1,n}} |a_{ij}| \right) . \left(\max_{\substack{i=1,n \\ j=1,n}} |b_{ij}| \right) n^2$$

and they note that $K_\infty \leq K_m$.

Since all the matrices are positive definite the largest element is on the diagonal, and so there is an easy test for an upper bound for K_∞.

From (9a) the relative error RE is given by

$$RE = \frac{\|\delta x\|}{\|x_t\|} \leq \left\{ \frac{K_\infty(A)\|E\|}{\|A\|} \right\} \Big/ \left\{ 1 - \frac{K_\infty(A)\|E\|}{\|A\|} \right\}$$

$$\sim \frac{K_\infty(A)\|E\|}{\|A\|}$$

provided we require the relative error to be small.

If γ_o is an upper bound for the acceptable relative error then we must have

$$\frac{K_\infty(A)\|E\|}{\|A\|} < \gamma_o .$$

Thus the matrix is ill-conditioned if

$$\frac{\|A\|}{K_\infty(A)} < \frac{\|E\|}{\gamma_o}$$

$$\leqslant \frac{1.06\ (n^3+3n^2)\ g.p}{\gamma_o \max\limits_{i,j} |a_{ij}|}$$

or $$\frac{1}{\max\limits_{i,j} |a_{ij}| \|A^{-1}\|} \leqslant \frac{1.06\ (n^3+3n^2)\ g.p}{\gamma_o}$$

As currently implemented in the code the test on the condition is

$$\frac{1}{K_m} < \frac{n^3 p}{\gamma_o}$$

which is too severe.

A reasonable modification is to estimate $\|A^{-1}\|$ by

$$n \cdot \max_{\substack{i=1,n \\ j=1,n}} |b_{ij}|$$

and use that estimate in the condition test.

Thus the condition test would be revised to

$$\frac{1}{\max\limits_i |a_{ii}| \max\limits_i |b_{ii}|} < \frac{n^4 p}{\gamma_o} \qquad (10)$$

rather than the present version:

$$\frac{1}{\max_i |a_{ii}| \max_i |b_{ii}|} < \frac{n^5 p}{\gamma_o}$$

As pointed out by **RR**,(10) is an extreme upper bound which could be replaced by

$$\frac{1}{\max_i |a_{ii}| \max_i |b_{ii}|} < \frac{n^2 p}{\gamma_o} \qquad (11)$$

in most circumstances. Taking even the pessimistic result (10) for matrices of order 200 and relative errors of 10% the maximum element in the inverse matrix would have to exceed 1×10^4 before ill-conditioning became a problem.

Note that all this is under the assumption $(1+\sigma^2_{min})/(1+\sigma^2_{max}) = o(1)$; it can be as low as 1-5% (the code starts objecting at 1/1025). We have not examined the **RR** estimate to see if this is sufficiently close to 1. Cats (1981) used a property that can lead to much more realistic bounds, namely $(\gamma x)^2 \leq xcx$ for all x, where c is the correlation matrix and γ the rms. This property is particular to O/I because it is derived from the fact that γ is a correlation vector; **RR** do not have this available.

4. DISCUSSION

A re-examination of the text for ill-conditioning of the analysis equations suggests that the currently implemented test is unnecessarily severe. A revised test which is effective even in the most unlikely circumstances and which is less restrictive than the present one is proposed. It is noted that even this test could be substantially relaxed if account is taken of the statistical distribution of round-off errors.

REFERENCES

Cats, G.J. and D.A. Robertson, 1980: Condition of the covariance matrix and stability of its inversion algorithm. ECMWF Working Paper No1/6/E GJC/071/1980.

Cats, G.J., 1981: A method of solving a system of linear equations efficiently in order to optimize the analysis code. ECMWF Tech.Memo.No.26.

Ralston, A. and P. Rabinowitz, 1978: A first course in numerical analyses, McGraw-Hill, New York.

Wilkinson, J.H., 1964: The algebraic eigenvalue problem. Oxford University Press.