I. B. Kuperman

APPROXIMATE LINEAR ALGEBRAIC EQUATIONS

APPROXIMATE LINEAR ALGEBRAIC EQUATIONS

I. B. KUPERMAN
University of the Witwatersrand
Johannesburg

VAN NOSTRAND REINHOLD COMPANY

LONDON

NEW YORK CINCINNATI TORONTO MELBOURNE

VAN NOSTRAND REINHOLD COMPANY LTD
Windsor House, 46 Victoria Street, London, S.W.1

INTERNATIONAL OFFICES
New York Cincinnati Toronto Melbourne

Library of Congress Catalog Card No. 70–160200

ISBN 0 442 04546 8

First Published 1971

Paper supplied by
P. F. Bingham Limited, Croydon

Made and printed in Great Britain by
Butler and Tanner Ltd, Frome and London

Preface

Whenever the coefficients or right-hand constants of a system of linear algebraic equations are not known exactly, we have a system of *approximate linear algebraic equations.* This is the case, for example, when the coefficients and right-hand constants are obtained from measurement.

A number of methods are given in this book for obtaining the uncertainties in the unknowns due to the uncertainties in the coefficients and constants.

There is, however, no simple answer to the question: Which is the best, most practical, or recommended method? This depends on the magnitude of the given uncertainties, on the given coefficient matrix, on how accurately the uncertainties in the unknowns are required, and on the order of the given system of equations. For in certain cases, the volume of computation required to obtain the true intervals of uncertainty (Method IX) can become very large and indeed prohibitive. Under these circumstances, one must be satisfied with the results of methods leading to intervals containing the true intervals of uncertainty. Of these, Method VI or VIII leads to the best results, i.e., gives intervals of smallest width containing the true intervals of uncertainty.

In particular, it will be seen that the methods that give the best estimates of the true intervals of uncertainty require the most computation.

There is thus no simple answer as to which is the best method. But in many cases the statistical approach to approximate linear algebraic equations is the most appropriate (Chapter 12).

In order to make the book self-contained, certain mathematical topics with which all readers may not be familiar are dealt with briefly, namely, vector and matrix norms and the convergence of matrix series. Also, brief but adequate introductions are given to interval arithmetic, linear programming, and statistics, so as to make the important topic of approximate linear algebraic equations more easily accessible to a large readership.

Finally, let us say that the style was chosen so as to be best suited for

the average reader on his first reading, bearing in mind that the book contains much source material.

Department of Applied Mathematics and Computer Science,
University of the Witwatersrand,
May, 1970 Israel B. Kuperman

Contents

CHAPTER 1

Introduction

1.1 Introduction

We shall say that a system of linear algebraic equations is approximate if any or all of the coefficients and right-hand constants are not known exactly. If the values of the coefficients and right-hand constants depend upon physical measurements then clearly the values are not known exactly. If the measured values are expressed in decimal notation then it may safely be assumed that there is an uncertainty of at least one half-unit in the least significant position given, and usually by convention the uncertainty is one unit in the least significant position given. The uncertainties in the measured values may of course be considerably larger, but then the values of the uncertainties should be clearly stated. We call the study of the effects of the uncertainties in the coefficients and constants on the solution *approximate equation analysis.*

During the process of solution, the approximate equations are assumed to be exact, a solution having to be found which satisfies the accuracy requirements set, i.e., correct to the number of significant figures required. Solving a system of equations by two different methods or by using two different precisions, i.e., wordlengths, and comparing two such solutions may enable one to determine empirically the number of correct significant figures. Alternatively, interval arithmetic or theoretical error bounds may be used to determine the noise introduced by rounding. Thus, in solving the approximate system of equations, it is assumed that a sufficiently long wordlength is used, so that allowing for the effects of rounding noise leaves us with a solution correct at least to the required number of significant figures. Then, deciding to what accuracy the solution can be meaningfully used in view of the approximate nature of the equations is approximate equation analysis.

The uncertainties in the values of the unknowns are estimated from the uncertainties in the coefficients and constants. This facet of error analysis should form an integral part of a program for solving approximate

linear algebraic equations and this is particularly important when small changes in the coefficients cause relatively large changes in the unknowns, i.e., in systems of equations which tend to become *ill-conditioned*, a term which we define later.

Now, it is well known that if we have a singular coefficient matrix, i.e., one whose determinant is zero, then the existence and nature of the solutions depend on whether the system is consistent or not. If the equations are inconsistent then there is no solution. And if the equations are consistent then there is an infinity of solutions, it being possible to choose the value of at least one of the unknowns arbitrarily. Thus, in a consistent system of equations with a singular coefficient matrix, at least one unknown can be chosen as large as we please in magnitude. We therefore assume the coefficient matrix of the approximate system of equations to be nonsingular, for otherwise the problem of finding a solution with finite uncertainties in the unknowns fails at the very beginning.

But suppose it is possible to find a singular coefficient matrix within the limits of the uncertainties in the coefficients. Then we say that such a system of equations is *critically ill-conditioned*; and in this case the true coefficient matrix may be singular within the limits of our knowledge.

Therefore, we ignore any solution of a critically ill-conditioned system of equations if a solution is sought with finite uncertainties and we say that for the given uncertainties in the coefficients no worthwhile solution can be found. We should then make quite sure that the physical situation giving rise to the equations can be expected to give n linearly independent equations in n unknowns. If this is the case, then a usable solution can only be obtained if the coefficients can be found more accurately, i.e., with smaller uncertainties.

1.2 A Critically Ill-conditioned System of Equations

As a numerical example of a system of equations

$$\mathbf{Ax} = \mathbf{c}, \qquad \mathbf{A} = (a_{ij}), \qquad \mathbf{x} = (x_j), \qquad \mathbf{c} = (c_i),$$

in which there is an uncertainty of one unit in the least significant position given, consider the system of equations

$$\begin{aligned} 0{\cdot}974x_1 + 0{\cdot}790x_2 + 0{\cdot}311x_3 &= 2{\cdot}075 \\ -0{\cdot}631x_1 + 0{\cdot}470x_2 + 0{\cdot}251x_3 &= 0{\cdot}090 \\ 0{\cdot}455x_1 + 0{\cdot}975x_2 + 0{\cdot}425x_3 &= 1{\cdot}855, \end{aligned} \tag{1.1}$$

the uncertainty in each coefficient and right-hand constant being 0·001. The solution by Gaussian elimination (pivotal condensation) is given in Table 1.1 and is self-explanatory.

TABLE 1.1 *Solution of (1.1) by Gaussian Elimination*†

x_1	x_2	x_3	c	Row and explanation
0·974	0·790	0·311	2·075	R_1
−0·631	0·470	0·251	0·090	R_2
0·455	0·975	0·425	1·855	R_3
1	0·81109	0·31930	2·13039	$R_4 = R_1 \div 0{\cdot}974$
	0·98180	0·45248	1·43428	$R_5 = R_2 + 0{\cdot}631R_4$
	0·60595	0·27972	0·88567	$R_6 = R_3 - 0{\cdot}455R_4$
	1	0·46087	1·46087	$R_7 = R_5 \div 0{\cdot}98180$
		0·00045	0·00045	$R_8 = R_6 - 0{\cdot}60595R_7$
		1	1·00000	$R_9 = R_8 \div 0{\cdot}00045$
	1		1·00000	$R_{10} = R_7 - 0{\cdot}46087R_9$
1			1·00000	$R_{11} = R_4 - 0{\cdot}31930R_9 - 0{\cdot}81109R_{10}$

† The solution is rounded to 5 decimal places, the solution having been obtained using a wordlength of 10 significant figures.

Referring to the table, we see that division of rows occurred on three occasions (rows R_4, R_7, and R_9). The divisors are called the *pivots* of the Gaussian elimination process, the three pivots in our case being 0·974, 0·98180, and 0·00045. Our procedure is to choose the coefficient of largest magnitude in the first column as the first pivot (underlined in row R_1) and to eliminate the unknown x_1 from the other equations, thereby obtaining the *reduced system of equations* in rows R_5 and R_6. Then we choose the coefficient of largest magnitude in the first column of the reduced system of equations (underlined in row R_5) and eliminate x_2 from the other equation in the reduced system of equations.

The third pivot is now 0·00045 in row R_8, and dividing by this pivot we obtain the value of x_3 in row R_9. This is the end of the so-called *forward procedure* of the Gaussian elimination process. And we may mention that the above method of choosing the pivot at each stage as the coefficient of largest magnitude in the first available column is called *partial pivoting*.

In rows R_{10} and R_{11} we in effect substitute the values of the unknowns obtained at each stage into previous equations until all the unknowns are found, the solution in our case being $x_1 = x_2 = x_3 = 1{\cdot}00000$. This part of the solution is known as the *back-substitution procedure* of the Gaussian elimination process.

Now, it is instructive in our example to examine the effect of the uncertainty in a_{33} on the pivots. A little consideration of Table 1.1

shows that a change δa_{33} in the coefficient $a_{33} = 0{\cdot}425$ changes the value of the last pivot 0·00045 in row R_8 by δa_{33} and leaves the other pivots unchanged.

Denoting the three pivots in rows R_1, R_5, and R_8 by p_1, p_2, and p_3, respectively, let us then consider the effect of a change δa_{33} given by

$$\delta a_{33} \in [-0{\cdot}001, 0{\cdot}001],$$

i.e., let us consider the effect of an uncertainty of 0·001 in a_{33}. Because a change δa_{33} in a_{33} produces a change δa_{33} in p_3 and leaves p_1 and p_2 unchanged, it follows that for

$$a_{33} \in [0{\cdot}425 - 0{\cdot}001, 0{\cdot}425 + 0{\cdot}001]$$

and all the other coefficients exact we have

$$p_3 \in [0{\cdot}00045 - 0{\cdot}001, 0{\cdot}00045 + 0{\cdot}001],$$

i.e.,

$$p_3 \in [-0{\cdot}00055, 0{\cdot}00145], \tag{1.2}$$

while $p_1 = 0{\cdot}974$ and $p_2 = 0{\cdot}98180$.

Thus, it is possible for the last pivot to be zero because the interval in (1.2) includes zero.

But in any Gaussian elimination process, the product of the pivots is, apart from sign, equal to the determinant of the coefficient matrix.

For one of the methods of evaluating a determinant is to reduce it by elementary row operations to the unit matrix, the determinant of the unit matrix being 1. And it may be recalled that in this procedure of evaluating determinants:

1. a P (permutation) elementary operation involves interchanging rows and changes the sign of the determinant,
2. an M (multiplication) elementary operation involves multiplying a row by a scalar and this multiplies the value of the determinant by the scalar, and
3. an A (addition) elementary operation involves adding to any one row multiples of any of the other rows; this does not change the value of the determinant.

But these elementary operations are involved in the Gaussian elimination process. The P elementary operation occurs if the coefficient chosen as pivot is not the first one in its column, and this then changes the sign of the determinant. The M elementary operation occurs when a pivotal row (i.e., a row containing a pivot) is divided by the pivot, the value of the determinant being divided by the pivot. And the A elementary operations occur during the elimination of the unknowns. These do

not change the value of the determinant. Comparing Tables 1.1 and 1.2 may clarify the above. In Table 1.2, the whole system of equations is rewritten after each elementary operation, the original coefficient matrix being finally reduced by elementary operations to the unit matrix.

TABLE 1.2 *The Gaussian Elimination Process in Full for (1.1)*†

Coefficients			Constants	Corresponding operation in Table 1.1
0·974	0·790	0·311	2·075	R_1
−0·631	0·470	0·251	0·090	R_2
0·455	0·975	0·425	1·855	R_3
1	0·81109	0·31930	2·13039	$R_4 = R_1 \div 0{\cdot}974$
−0·631	0·470	0·251	0·090	
0·455	0·975	0·425	1·855	
1	0·81109	0·31930	2·13039	
0	0·98180	0·45248	1·43428	$R_5 = R_2 + 0{\cdot}631R_4$
0·455	0·975	0·425	1·855	
1	0·81109	0·31930	2·13039	
0	0·98180	0·45248	1·43428	
0	0·60595	0·27972	0·88567	$R_6 = R_3 - 0{\cdot}455R_4$
1	0·81109	0·31930	2·13039	
0	1	0·46087	1·46087	$R_7 = R_5 \div 0{\cdot}98180$
0	0·60595	0·27972	0·88567	
1	0·81109	0·31930	2·13039	
0	1	0·46087	1·46087	
0	0	0·00045	0·00045	$R_8 = R_6 - 0{\cdot}60595R_7$
1	0·81109	0·31930	2·13039	
0	1	0·46087	1·46087	
0	0	1	1·00000	$R_9 = R_8 \div 0{\cdot}00045$
1	0·81109	0·31930	2·13039	
0	1	0	1·00000	$R_{10} = R_7 - 0{\cdot}46087R_9$
0	0	1	1·00000	
1	0	0	1·00000	$R_{11} = R_4 - 0{\cdot}31930R_9 - 0{\cdot}81109R_{10}$
0	1	0	1·00000	
0	0	1	1·00000	

† The system of equations is repeated after each elementary operation, showing clearly that, apart possibly from sign, the determinant of the coefficient matrix is equal to the product of the pivots. (In our example there is, however, no change of sign.)

Thus, in going from the given coefficient matrix to the unit matrix in Table 1.2, the value of the determinant is altered in magnitude whenever we divide by a pivot. Since the value of the determinant of the final unit matrix is unity, it follows that apart from sign (which depends on our choice of pivots) the value of the determinant of the coefficient matrix is equal to the product of the pivots.

It therefore follows that:

> If any of the pivots in a Gaussian elimination process can become zero within the limits of the uncertainties in the coefficients then the system of equations is critically ill-conditioned. (1.3)

In our example, the last pivot in Table 1.1 can certainly become zero for changes in the coefficients within the limits of their uncertainties; in fact, the last pivot can become zero within the limits of the uncertainty in a_{33} alone. Hence, the given system of equations in (1.1) is critically ill-conditioned for an uncertainty of 0·001 in each coefficient.

But it is not always possible to test for critical ill-conditioning by changing only one coefficient; simultaneous changes may have to be introduced in all the coefficients. We investigate this problem in Section 1.4, while we now introduce notation and state our problem more precisely in the next section.

1.3 Statement of Problem

Suppose that we are given a system of n linear algebraic equations in n unknowns

$$\mathbf{Ax} = \mathbf{c}, \qquad \mathbf{A} = (a_{ij}), \quad \mathbf{x} = (x_i), \quad \mathbf{c} = (c_i) \tag{1.4}$$

in which the coefficients a_{ij} and the right-hand constants c_i are the approximate values, the true values not being known exactly. Then, restricting ourselves to the case where $\mathbf{A}$ is nonsingular, the solution of (1.4) is

$$\mathbf{x} = \mathbf{Bc} \tag{1.5}$$

where

$$\mathbf{B} = \mathbf{A}^{-1}, \qquad \mathbf{B} = (b_{ij}).$$

If, in fact, the true system of equations corresponding to (1.4) is

$$\mathbf{A}^*\mathbf{x}^* = \mathbf{c}^* \tag{1.6}$$

let us suppose that the coefficients and right-hand constants of the true system of equations are known no more precisely than that given by

$$a_{ij}^* \in [a_{ij} - \varepsilon_{ij}, a_{ij} + \varepsilon_{ij}], \qquad c_i^* \in [c_i - \varepsilon_i, c_i + \varepsilon_i], \qquad i, j = 1, 2, \ldots, n, \tag{1.7}$$

where the ε_{ij} and the ε_i are clearly nonnegative quantities. We call the ε_{ij} the *uncertainties in the coefficients* and the ε_i the *uncertainties in the right-hand constants*. And we call the intervals in (1.7) the *intervals of uncertainty in the coefficients* and the *intervals of uncertainty in the right-hand constants*, respectively.

We have thus chosen the approximate system of equations in (1.4) to correspond to the midpoints of the intervals in (1.7).

It may be pointed out that had the intervals of uncertainty in the coefficients and constants been given in the form

$$a_{ij}^* \in [f_{ij}, g_{ij}], \qquad c_i^* \in [u_i, v_i], \qquad i, j = 1, 2, \ldots, n, \tag{1.8}$$

then, to correspond to the form in (1.7), we must take

$$a_{ij} = \tfrac{1}{2}(f_{ij} + g_{ij}), \qquad \varepsilon_{ij} = \tfrac{1}{2}(g_{ij} - f_{ij}), \qquad i, j = 1, 2, \ldots, n, \tag{1.9}$$

and

$$c_i = \tfrac{1}{2}(u_i + v_i), \qquad \varepsilon_i = \tfrac{1}{2}(v_i - u_i), \qquad i = 1, 2, \ldots, n. \tag{1.10}$$

Now, our first task in approximate equation analysis is clearly to satisfy ourselves that the true coefficient matrix $\mathbf{A}^*$ cannot become singular within the limits of the uncertainties ε_{ij}.

If, in fact, the approximate system of equations is not critically ill-conditioned, suppose that the x_i^* are given by

$$x_i^* \in [x_i - e_i, x_i + d_i], \qquad i = 1, 2, \ldots, n. \tag{1.11}$$

And let us note that the d_i and e_i in the intervals are clearly nonnegative because one possible set of values of the x_i^* is $x_i^* = x_i$ $(i = 1, 2, \ldots, n)$ (see (9.2)).

Then, we call the intervals in (1.11) the *intervals of uncertainty in the unknowns* and we denote them by U_i $(i = 1, 2, \ldots, n)$, i.e.,

$$U_i = [x_i - e_i, x_i + d_i], \qquad i = 1, 2, \ldots, n. \tag{1.12}$$

Now, the width or length of an interval $[a, b]$ is $(b - a)$ (see (9.4)). Hence the *widths of the intervals of uncertainty* denoted by

$$w(U_i), \qquad i = 1, 2, \ldots, n$$

are given by

$$w(U_i) = d_i + e_i, \qquad i = 1, 2, \ldots, n. \tag{1.13}$$

And for each unknown x_i we call the larger of e_i and d_i in (1.11) the uncertainty Δx_i in the unknown, i.e., the *uncertainties in the unknowns* are

$$\Delta x_i = \max(e_i, d_i), \qquad i = 1, 2, \ldots, n. \tag{1.14}$$

Clearly, we have by (1.12) and (1.14) that

$$U_i \subseteq [x_i - \Delta x_i, x_i + \Delta x_i], \qquad i = 1, 2, \ldots, n \tag{1.15}$$