NUMERICAL INVERTING OF MATRICES OF HIGH ORDER

JOHN VON NEUMANN AND H. H. GOLDSTINE

ANALYTIC TABLE OF CONTENTS

Presented to the Society, September 5, 1947; received by the editors October 1, 1947. Published with the invited addresses for reasons of space and editorial convenience.

Preface

The purpose of this paper is to derive rigorous error estimates in connection with the inverting of matrices of high order. The reasons for taking up this subject at this time in such considerable detail are essentially these: First, the rather widespread revival of mathematical interest in numerical methods, and the introduction of new procedures and devices which make it both possible and necessary to perform operations of this type on matrices of much higher orders than was even remotely practical in the past. Second, the fact that considerably diverging opinions are now current as to the extremely high or extremely low precisions which are required when inverting matrices of orders $n \geqq 10$. (Cf. in this connection footnotes 10, 11, 12 below.)

It has been our aim to provide a rigorous discussion of this rather involved problem in estimation. Our estimates of errors have furthermore been carried out in strict observance of these two rules, which seem to us to be essential: To produce no numbers (final or intermediate) that lie outside a given finite interval (for which we chose $-1, 1$), and to treat these numbers solely as aggregates of a fixed number of digits, given in advance.

The reader will find a complete enumeration and interpretation of our results in Chapter VII, especially in §§7.7, 7.8. He may find it convenient to consult these first. We conclude there, for example, matrices of the orders 15, 50, 150 can usually be inverted with a (relative) precision of 8, 10, 12 decimal digits less, respectively, than the number of digits carried throughout. By "usually" we mean that if a plausible statistic of matrices is assumed, then these estimates hold with the exception of a low probability minority. These general estimates are based on rigorous individual estimates, valid for all matrices (cf. §§7.4–7.5). If we had been willing to use a probability treatment for individual matrices, too, our estimates could have been improved by several decimal digits (cf. §2.3).

We made no effort to obtain numerically optimal estimates, but we believe that our estimates are optimal at least as far as the practical orders of magnitude are concerned and with respect to the over-all mathematical method used and the principles indicated above.

This work has been made possible by the generous support of the Office of Naval Research, under Contract N7onr-388. Earlier related work will be published elsewhere by V. Bargmann, D. Montgomery and one of us. (Cf. the references in footnote 24 below.)

CHAPTER I. THE SOURCES OF ERRORS IN A COMPUTATION

1.1. **The sources of errors.** When a problem in pure or in applied mathematics is "solved" by numerical computation, errors, that is, deviations of the numerical "solution" obtained from the true, rigorous one, are unavoidable. Such a "solution" is therefore meaningless, unless there is an estimate of the total error in the above sense.

Such estimates have to be obtained by a combination of several different methods, because the errors that are involved are aggregates of several different kinds of contributory, primary errors. These primary errors are so different from each other in their origin and character, that the methods by which they have to be estimated must differ widely from each other. A discussion of the subject may, therefore, advantageously begin with an analysis of the main kinds of primary errors, or rather of the sources from which they spring.

This analysis of the sources of errors should be objective and strict inasmuch as completeness is concerned, but when it comes to the defining, classifying, and separating of the sources, a certain subjectiveness and arbitrariness is unavoidable. With these reservations, the following enumeration and classification of sources of errors seems to be adequate and reasonable.

(A) The mathematical formulation that is chosen to represent the underlying problem may represent it only with certain idealizations, simplifications, neglections. This is even conceivable in pure mathematics, when the numerical calculation is effected in order to obtain a preliminary orientation over the underlying problem. It will, however, be the rule and not the exception in applied mathematics, where these things are hardly avoidable in a mathematical representation. This complex is further closely related to the methodological observation that a mathematical formulation necessarily represents only a (more or less explicit) theory of some phase of reality, and not reality itself.

(B) Even if the mathematical formulation according to (A) is not questioned, that is, if the theoretical description which it represents and the idealizations, simplifications, and neglections which it involves are accepted as final (and not viewed as sources of errors), this further point remains: The description according to (A) may involve parameters, the values of which have to be derived directly or indirectly (that is, through other theories or calculations) from observations. These parameters will be affected with errors, and these underlying errors will cause errors in the result of our calculation.

(C) Now let (A), (B) (mathematical formulation and observational data) go unquestioned. The next stumbling block is this: The mathematical formulation of (A) will in general involve transcendental operations (for example, functions like sin or log, operations like integration or differentiation, and so on) and implicit definitions (for example, solutions of algebraical or transcendental equations, proper value problems of various kinds, and so on). In order to be approached by numerical calculation, these have to be replaced by elementary processes (involving only those elementary arithmetical operations which the computer can handle directly) and explicit definitions, which correspond to a finite, constructive procedure that resolves itself into a linear sequence of steps.[1]

[1] This applies directly to all digital computing schemes: Digital computing by human operators, by "hand" and by semi-automatic "desk" machines, also computing by the large modern fully automatic, "self-sequenced," computing machines. Fundamentally, however, it applies equally to those "analogy" machines which can perform certain operations directly, that are "transcendental" or "implicit" from the digital point of view. Thus for machines of the genus of the "differential analyser" differentiating, integrating and solving certain (essentially implicit) differential equations are elementary, explicit operations. While a digital procedure must replace a total

Similarly every convergent, limiting process, which in its strict mathematical form is infinite, must in a numerical computation be broken off at some finite stage, where the approximation to the limiting value is known to have reached a level that is considered to be satisfactory. It would be easy to give further examples.

All these replacements are, as stated, approximative, and so the *strict* mathematical statement of (A) is now replaced by an *approximate* one. This constitutes our third source of errors.

(D) Finally, let not only (A), (B), but even the approximation process of (C) pass unchallenged. There still remains this limitation: No computing procedure or device can perform the operations which are its "elementary" operations (or, at least, all of them) rigorously and faultlessly. This point is most important, and is best discussed separately for digital and for "analogy" procedures or devices.

The case of the analogy devices is immediate and clear: An analogy device which represents two numbers x and y by two physical quantities $\bar{x}$ and $\bar{y}$ will form the sum $x+y$ or the product xy as two physical quantities $\bar{x}\oplus\bar{y}$ or $\bar{x}\otimes\bar{y}$. Yet $\bar{x}\oplus\bar{y}$ or $\bar{x}\otimes\bar{y}$ will unavoidably be affected with the (more or less) random "noise" of the computing instrument, that is, with the errors and imperfections inherent in any physical, engineering embodiment of a mathematical principle. Hence $\bar{x}\oplus\bar{y}$ and $\bar{x}\otimes\bar{y}$ will correspond not to the true $x+y$ and xy, but to certain $x+y+\epsilon^{(s)}$ and $xy+\epsilon^{(p)}$, where $\epsilon^{(s)}$, $\epsilon^{(p)}$ are (more or less) random, "noise" variables, of which only the probable (and possibly the maximum) size is known in advance. Also, $\epsilon^{(s)}$, $\epsilon^{(p)}$ assume new and (usually in the main) independent values with every new execution of an operation $+$ or $\times$. The same goes for the other operations which the device can perform, any or all of $-$, $/$, $\sqrt{}$, $\int$, d/dx, and possibly others.

differential equation by finite difference equations (to make it elementary) and possibly use iterative, trial-and-error methods (to make it explicit), the "differential analyser" may be able to treat such a problem "directly." But for a partial differential equation (where a digital procedure requires the same circumventory measures as in the above case of a total differential equation), the "differential analyser" can give its "direct" treatment only to one (independent) variable, while on the other (independent) variable or variables it will have to resort to finite-difference and possibly iteration and trial-and-error methods, very much like a digital procedure has to on all variables.

Thus the differences are only in degree (number of processes that rate as "elementary" and "explicit") but not in kind. Such differences, by the way, exist even among digital devices: Thus one may treat square rooting as an "elementary," "explicit" process, and another one not, and so on.

For digital procedures or devices, we must note first that they represent (continuous, real) numbers x by finite digital aggregates $\bar{x}=(\alpha_1, \cdots, \alpha_s)$ where each $\alpha_\sigma=0, 1, \cdots, \beta-1$. Here $\beta(=2, 3, \cdots)$ is supposed to be the *basis* of the digital representation,[2] s the number of *places* used, while we need not for the moment pay attention to the position t of the β-adic point.[3] Now for two s-place numbers $\bar{x}$, $\bar{y}$ the sum and the difference $\bar{x}\pm\bar{y}$ are again s-place numbers,[4] but their product $\bar{x}\bar{y}$ is not. $\bar{x}\bar{y}$ is, of course, a $2s$-place number. One might carry along this $\bar{x}\bar{y}$ in subsequent computations as a $2s$-place number, but later multiplications will increase the number of places further and further, if such a scheme is being followed consistently. With any practical procedure or device a line has to be drawn somewhere, that is, a maximum number of places in a number x set. In our discussion we might as well assume then that s has already reached this maximum. Hence for two s-place numbers $\bar{x}$, $\bar{y}$ the true $2s$-place product $\bar{x}\bar{y}$ must be replaced by some s-place approximation. Let us denote this s-place approximation by $\bar{x}\times\bar{y}$, and call $\bar{x}\bar{y}$ the *true* and $\bar{x}\times\bar{y}$ the *pseudo-product.*

The same observations apply to the quotient $\bar{x}/\bar{y}$, except that the true $\bar{x}/\bar{y}$ will in general have infinitely many places, and not only $2s$. We call $\bar{x}/\bar{y}$ the *true* quotient, and a suitable s-place approximation $\bar{x}\div\bar{y}$ the *pseudo-quotient.*[5] Similar pseudo-operations should be introduced for other operations if they are "elementary" for the device under consideration (for example, the square root), but we need not consider these matters here.

The transition from the true operations to their pseudo-operations is effected by any one of the familiar methods of *round off.*[6] Thus the true $\bar{x}\bar{y}$, $\bar{x}/\bar{y}$ are replaced by the pseudo $\bar{x}\times\bar{y}=\bar{x}\bar{y}+\eta^{(p)}$, $\bar{x}\div\bar{y}=(\bar{x}/\bar{y})+\eta^{(q)}$, with the *round off errors* $\eta^{(p)}$, $\eta^{(q)}$.

There is a good deal of similarity between these $\eta^{(p)}$, $\eta^{(q)}$ and the $\epsilon^{(s)}$, $\epsilon^{(p)}$ that we encountered above (for analogy devices): While the

[2] The most probable choices are $\beta=10$ and $\beta=2$.

[3] This is between t and $t+1$ $(t=0, 1, \cdots, s)$: $\bar{x}=(\alpha_1, \cdots, \alpha_s)=\sum_{\sigma=1}^{s}\beta^{t-\sigma}\alpha_\sigma$.

[4] Unless they exceed the permissible limits of size $\bar{x}+\bar{y}\geqq\beta^t$ or $\bar{x}-\bar{y}<0$. Regarding this, cf. footnote 17. We shall not discuss this complication here, nor the connected one, that the digital aggregates have to be provided with a sign. The latter point is harmless, and both are irrelevant at this point, cf., however, (a) in 2.1, particularly (2.1).

[5] We omit again discussions of size at this point. Cf. footnote 4 and its references.

[6] The simplest method consists of omitting all digits beyond place s. A more elaborate one required adding $\beta/2$ units of place $s+1$ first, and then (having effected the carries which are thus caused) omitting as above. There exist still other procedures.

η are strictly very complicated but uniquely defined number theoretical functions (of $\bar{x}$, $\bar{y}$), yet our ignorance of their true nature is such that we best treat them as random variables. We know their average and maximum sizes: with the usual round off method (the second one in footnote 6), the maximum of $|\eta|$ is $\beta^{-s}/2$ (that is, $\beta\cdot\beta^{-(s+1)}/2$ in the sense of loc. cit.), and if we assume η to be equidistributed in $-\beta^{-s}/2, +\beta^{-s}/2$ then

$$\text{Mean}(\eta) = 0, \tag{1.1}$$

$$\text{Dispersion}(\eta) = (\text{Mean}(\eta^2))^{1/2} = (\beta^{-2s}/12)^{1/2} = .29\beta^{-s}. \tag{1.2}$$

Finally, $\eta^{(p)}$, $\eta^{(q)}$ assume new and (usually in the main) independent values with every new execution of an operation × or ÷.

Thus the "digital" η are in many essential ways "noise" variables, just as the "analogy" ϵ.

These *noise variables* or *round off variables* ϵ and η, which are injected into the computation every time when an "elementary" operation is performed (excepting $\pm$ in the digital case) constitute our fourth and last source of errors.

1.2. **Discussion and interpretation of the errors (A)–(D). Stability.** The errors described in (A) are the *errors due to the theory.* While their role is clearly most important, their analysis and estimation should not be considered part of the mathematical or of the computational phase of the problem, but of the underlying subject, in which the problem originates. There can be little doubt regarding this methodological position. We will therefore not concern ourselves here with (A) any further.

The errors described in (B) are essentially the *errors due to observation.* To this extent they are, strictly construed, again no concern of the mathematician. However, their influence on the result is the thing that really matters. In this way, their analysis requires an analysis of this question: What are the limits of the change of the result, caused by changes of the parameters (data) of the problem within given limits? This is the question of the *continuity of the result as a function of the parameters of the problem*, or, somewhat more loosely worded, of the *mathematical stability of the problem.* This question of continuity or stability is actually not the subject matter of this paper, but it has some influence on it (cf. the discussion in §1.3), and we can therefore not let it slip out of sight completely.

The errors described in (C) are those which are most conspicuous as *errors of approximation* or *truncation.* Most discussions in "approximation mathematics" are devoted to analysis and estimation of these errors: Numerical methods to obtain approximate solutions of algebraical equations by iterative and interpolation processes, numerical methods to evaluate definite integrals, stepwise (finite-difference) methods to integrate differential equations correctly to varying "orders in the differential," and so on. Just because these errors have formed the subject of the major part of the existing literature, we shall not consider them here to any great extent. In fact, we have selected the specific problem of this paper in such a way that this source of errors has no direct part in it. (Cf. §1.5.)

There is, however, one phase of this part of the subject about which a little more should be said, just in view of the distribution of emphases in our present work: This is its relation to the question of *stability*, to which we have already referred in our above remarks on (B). Let us therefore consider this matter, before we go on to the discussion of (D).

1.3. **Analysis of stability. The results of Courant, Friedrichs' and Lewy.** The point is this: (B) dealt with the continuity or stability of (A), that is, of its result when viewed as a function of the parameters. This applies not only to the errors in the values of those parameters due to the causes mentioned in (B) (observational), but also to any other perturbations which may affect the values of any of the parameters which enter into the mathematical formulation of (A). Such perturbations equally affect quantities which are usually not interpreted as parameters at all, because they are not of observational origin. (This aspect of the matter will be relevant in connection with (D), cf. below in §1.4.)

Now (C) replaces the (strict) mathematical problem of (A) by a different one (the approximate problem). The considerations of (C) must establish that the problem of (C) differs quantitatively but little from the problem of (A). This does, however, not guarantee necessarily that the continuity or stability of (A) implies that of (C) as well. (Cf. below.) Yet, the actual computation deals with the problem of (C), and not with the problem of (A); consequently it is the continuity or stability of the former (of which the latter is a limiting case) that is really required.

That the stability of the strict problem need not imply that of an arbitrarily close approximant was made particularly clear by some important results of R. Courant, K. Friedrichs, and H. Lewy.[7] They showed, among other things, that although the partial differential

[7] *Über die partiellen Differenzengleichungen der Mathematischen Physik,* Math. Ann. vol. 100 (1927) pp. 32–74.

equation

$$\frac{\partial^2 y}{\partial t^2} = \frac{\partial}{\partial x}\left(F\left(\frac{\partial y}{\partial x}\right)\right) \tag{1.3}$$

is usually stable,[8] its stepwise, finite-difference approximant

$$\begin{aligned} &\frac{y(t+\Delta t, x) - 2y(t, x) + y(t-\Delta t, x)}{\Delta t^2} \\ &\qquad = \frac{1}{\Delta x}\left\{F\left(\frac{y(t, x+\Delta x) - y(t, x)}{\Delta x}\right)\right. \\ &\qquad\qquad \left. - F\left(\frac{y(t, x) - y(t, x-\Delta x)}{\Delta x}\right)\right\} \end{aligned} \tag{1.4}$$

need not be, no matter how small Δt, Δx are.[9] The necessary and sufficient condition for the stability of (1.4) is

$$\Delta x \geqq c\Delta t \qquad \text{where } c = \left(\frac{dF(v)}{dv}\right)^{1/2} \tag{1.5}$$

in the entire domain of integration.

Thus (C) requires an extension of the stability considerations of (B) from the original, strict problem of (A) to the secondary, approximant problem of (C).

1.4. **Analysis of "noise" and round off errors and their relation to high speed computing.** We now come to the errors described in (D). As we saw, they are due to the inner "noise level" of the numerical computing procedure or device—in the digital case this means: to the round off errors.

These differ from the perturbations of the apparent or the hidden parameters of the problem, to which we referred in §1.3, in this significant respect: Those perturbations will cause a parameter to deviate from its ideal value, but this deviation takes place only once, and is then valid with a constant value throughout the entire problem. The perturbations of (D), on the other hand, take place anew, and essentially independently, every time an "elementary" operation is performed. (Cf. the discussion in §1.1.) They form, therefore, a

[8] This is the Lagrangean form of the equation of motion of a one-dimensional, compressible, isentropic, nonviscous, nonconductive flow. It need not be linear, that is, it may go beyond the "acoustic" approximation.

[9] This approximant is correct up to second order terms in the differentials Δt, Δx, and it is the one that is most frequently used in numerical work.

constantly renewed source of contaminations, and are likely to be more dangerous than the single perturbations of §1.3 (that is, of (B)). Their influence increases with the number of "elementary" operations that have to be performed. They are therefore especially important in long computations, involving many such operations. Such long computations will undoubtedly be normal for very high speed computing devices.[10] It is therefore just for the highest speed devices that the source (D) will prove to be most important. We propose to concentrate on it in this paper.

The errors which the source of (D) is continuously injecting into a computation will be individually small, but appear in large numbers. The decisive factor that controls their effect is therefore a continuity or stability phenomenon of the type discussed in §1.3 above. And it is the stability of the approximant procedure of (C), and not of the strict procedure of (A), which matters—just as we saw in §1.3. For this reason stability discussions in the sense of §1.3 should play an important part in this phase of the problem.

1.5. **The purpose of this paper. Reasons for the selection of its problem.** On the basis of what has been stated so far we can define the purpose of this paper. We wish to analyze the stability of a computational procedure in the sense of (D), that is, with respect to the "inner noise" of the computation—in the digital situation: with respect to the round off errors. We shall attempt to isolate the phase of the problem that we want to analyze from all other, obscuring influences as much as possible. We shall therefore select a problem in which the difficulties due to (D) are a maximum and all others are a minimum. In other words, we shall choose a problem which is strictly "elementary," that is, where no transcendental or limiting processes occur, and where the result is defined by purely algebraical formulae. On the other hand the problem should lead with ease to very large numbers of "elementary" operations. This points towards problems with a high order iterative character. Finally, it should be of inherently low, or rather insecure, stability. Errors committed

[10] Fully automatic electronic computing machines which multiply two real numbers (full size digital aggregates) in 10^{-4} to 10^{-3} seconds, and which are sufficiently well organized to be able to have a duty-cycle of 1/10 to 1/5 with respect to multiplication, will probably come into use in a not too distant future. Single problems consuming 2 to 20 hours on such a machine should be the norm.

Taking average figures: $3 \cdot 10^{-4}$ second multiplier, 1/7 duty cycle and a 6 hour problem, gives 10^7 multiplications for a single problem. This number may serve as an orientation regarding the orders of magnitude that are likely to be involved. For more specific figures in the problem of matrix inversion cf. the remarks at the end of §7.8.

(that is, noise introduced) in an earlier stage of the computation should be exposed to a possibility of considerable amplification by the subsequent operations of the computation.

For this purpose the problem of solving n (simultaneous) linear equations in n variables seems very appropriate, when n assumes large values.[11] Besides this problem will be a very important one when the fast digital machines referred to in footnote 10 become available; those machines will create a prima facie possibility to attack a wide variety of important problems that require matrix manipulations, and in particular inversions, for unusually large values of n.[12]

1.6. **Factors which influence the errors (A)–(D). Selection of the elimination method.** It should be noted that the four error sources (A)–(D) show an increasing dependence on procedural detail: (A) depends only on the strict mathematical statement of the problem, and this is still true of (B), although observational elements begin to appear. (C) introduces the dependence on the mathematical approximations used. (D), finally, depends even on the actual algorithm according to which the equations of (C) are processed: The order in which they are taken up, whether an expression $(a+b)c$ is formed in this order or as $ac+bc$, whether an expression ab/c is formed in this order or as $(a/c)b$ or $a(b/c)$ or $(a/c^{1/2})(b/c^{1/2})$ (regarding a set of such alternatives cf. §6.1), and so on.

Since we wish to study the role of (D) in the problem of matrix inversion, it is necessary to decide which of the several available algorithms is to be used. We select the well known *elimination method*, or rather a new variant of it that we shall develop, because we conclude, from the results that we shall obtain, that this method is superior to the other known methods. (For the details of the procedure cf. the preliminary discussion of §§5.1, 5.2; the more specific procedures at the end of §6.1, especially (6.3), (6.4); the first part of §6.9; and the final discussion together with formally complete references in §7.6. Regarding the value of the method, cf. §§7.7, 7.8.)

1.7. **Comparison between "analogy" and digital computing methods.** We conclude this chapter with a general remark regarding the comparison between digital and "analogy" machines, from the point of view of the "noise variables" of (D) in §1.1.

We have noted the fact that these two categories of devices do not differ very essentially in that respect, where one might prima facie

[11] The difficulties of present day numerical methods in the problem of matrix-inversion begin to assume very serious dimensions when n increases beyond 10.

[12] We anticipate that $n\sim 100$ will become manageable. Cf. the end of §7.8.

look for the main difference: we mean the circumstance that "analogy" machines are undoubtedly "approximate" in their effecting the "elementary" operations, while the digital devices might be viewed as rigorous. This is not so, or at least not so in the sense in which it really matters: "Analogy" devices are, of course, affected in their "elementary" operations by a genuine, physical "noise." In digital devices, on the other hand, the round off errors are unavoidable by the intrinsic nature of things, and they play exactly the same role as the true "noise" in an "analogy" device. It is therefore best to talk of "noise" in both cases. In digital devices this "noise" affects only multiplication and division, but not addition and subtraction—but this circumstance does not cause a very important differentiation from the "analogy" devices.

The circumstance which is important is that the "noise level" in a digital device can be made much lower than in an "analogy" device. For s-place, base β numbers it is $\sim .29\beta^{-s}$. (Cf. (1.2). This is, of course, relative to the maximum numerical size allowed.) A typical situation is $\beta=10$, $s=10$,[13] that is, the dispersion of the "noise variable" η is $\sim 3\cdot 10^{-11}$. Even the best "analogy" devices that are possible with present techniques have dispersions greater than or equal to 10^{-5} for their "noise variable" ϵ (again relative to the maximum size allowed).

In addition, a conventional "analogy" device which is built for extreme precision is naturally working in an area of "decreasing returns" for precision: Cutting the size of the dispersion of ϵ by an additional factor of, say, 2 gets the more difficult, the smaller this dispersion is already. In a digital machine, on the other hand, cutting the size of the dispersion of η by an additional factor 2 (or 10) is equivalent to building the machine with one more binary (or decimal) digit, and this addendum gets percentually less when the number of digits increases, that is, when the attained dispersion η decreases.

Thus the digital procedure may be best viewed as the most effective means yet discovered to reduce the "inner noise level" of computing. This aspect becomes increasingly important as the rate at which this "noise" is injected into the computation increases, that is, as the computations assume larger sizes (consist of greater numbers of "elementary" operations), and the machines which carry them out get faster.

[13] All existing machines (or almost all) are decimal, that is, have $\beta=10$. With rare exceptions $s=7$ to 10, for example, on the familiar "desk" machines $s=8$ or 10. The "Mark I" computer at Harvard University has $s=11$ or 23.

Non-decimal machines of the future are likely to adhere, at least at first, to the same standard: for example, $\beta=2$, $s=30$ to 40.

CHAPTER II. ROUND OFF ERRORS AND ORDINARY ALGEBRAICAL PROCESSES

2.1. **Digital numbers, pseudo-operations. Conventions regarding their nature, size and use: (a), (b).** In Chapter I, and more particularly in §§1.5 and 1.6, we defined our purpose in this paper: We wish to determine the precision and the stability of the familiar *elimination method* for the *inversion of matrices* of order n, when n is large, with the primary emphasis on the effects of the "inner noise" of the digital computing procedure caused by the round off errors, that is, we want to determine how many (base β) places have to be carried in order to obtain significant results (meeting some specified standard of precision) in inverting a matrix of order n by the elimination method. We should thus obtain a lower limit for the number s of (base β) places in terms of the matrix order n. In doing this, we are prepared to accept as standard even such a variant of the elimination method which may not be among the commonly used ones, provided that it permits us to derive more favorable estimates of precision—that is, lower limits of s in terms of n. (This will actually happen, cf. the references at the end of §1.6.)

The main tools of our analysis will therefore be real members $\bar{x}$ which are represented by s-place, base β, digital aggregates in the sense of (D) in §1.1. We shall call them *digital numbers*, to distinguish them from the *ordinary* (real) *numbers*, which will also play a certain role in the discussions. When we deal with digital numbers, we shall observe certain rigid conventions, which facilitate an unequivocal and rigorous treatment, and which seem to us to be simple and reasonable, both in manipulation and in interpretation. It will, furthermore, always be permissible to view (in an appropriate part of the discussion) a number which was introduced as a digital number as an ordinary real number. To the extent to which we do this, the conventions in question will not apply.

We now enumerate these conventions:

(a) A *digital number* $\bar{x}$ is an s-place, base β, digital aggregate with sign:[14]

$$\bar{x} = \epsilon(\alpha_1, \cdots, \alpha_s);$$

$$(2.1)\qquad \epsilon = \left\{\begin{matrix} +, \text{ that is, } +1 \\ -, \text{ that is, } -1 \end{matrix}\right| ; \qquad \alpha_1, \cdots, \alpha_s = 0, 1, \cdots, \beta - 1.$$

The *sum* and the *difference* have their ordinary meaning, and will be denoted by $\bar{x} \pm \bar{y}$. The *product* and the *quotient*, on the other hand,

[14] This is our first step beyond the limitations of footnote 4.

will be rounded off to s places (cf. (D) in §1.1), and the quantities which result in this way will be called *pseudo-product* and *pseudo-quotient*, and will be denoted by $\bar{x}\times\bar{y}$ and $\bar{x}\div\bar{y}$. In addition to these *pseudo-operations*, others could be introduced, for other "elementary" operations (for example, for the square root), too. This, however, does not seem necessary for our present purposes.

Occasionally a digital number $\bar{x}$ has to be multiplied by an integer $l\ (=0, \pm 1, \pm 2, \cdots):l\bar{x}$. This should be thought of in terms of repeated additions or subtractions, and it is therefore not a pseudo-operation and involves no round offs.

(b) The position of the β-adic point in representing $\bar{x}$ was already referred to before (cf. (D) in §1.1, particularly footnote 3). It seems to us simplest to fix it at the extreme left (that is, $t=0$ in the notations of loc. cit. above). Any other position can be made equivalent to this one by the use of appropriate *scale factors*.[15] Besides, other positions of the β-adic point are of advantage only in relation to particular and very specifically characterized problems or situations, while the position at the extreme left permits a considerable uniformity in discussing very general situations. Finally, this positioning has the effect that the maximum size of any digital number is 1, so that absolute and relative error sizes[16] coincide, which simplifies and clarifies all assessments. This positioning requires, of course, a careful and continuing check on all number sizes which develop in the course of the computation, and the introduction of scale factors when they threaten to grow out of range.[17] It should be noted, how-

[15] These scale factors are of considerable importance. They are usually integer factors, most conveniently powers β^p $(p=0, \pm 1, \pm 2, \cdots)$ of the base β. (Cf. in this respect the further analysis of §2.5.) Their main purpose is to keep the numbers resulting from intermediate operations within the operating range of the machine (cf. footnote 17 below), and also to avoid that they get crowded into a small segment of this interval (usually near to 0) with an attendant loss of "significant digits," that is, of ultimate precision.

They are by no means characteristic of digital machines. They are equally necessary in "analogy" machines. Thus, in differential analyzers appropriate gears are essential to insure that no integrator runs off its wheel, and that none should be limited systematically to insignificant movements, and so on.

For a proper appreciation of the importance of these scale factors it should be realized that no computing scheme or estimation of errors and of validity in a computing scheme is complete without a precise accounting for their role. We shall have to do with them again subsequently: $2^{p_{ij}}$ in, §6.4; 2^{r_i}, 2^q, 2^{q_0} in §6.7; 2^{q_1} in §6.10; 2^p, $2^{p'}$ in §7.3.

[16] Relative to the maximum number size.

[17] Owing to this positioning all $|\bar{x}| \leqq 1$, $|\bar{y}| \leqq 1$, cf. below. Hence automatically $|\bar{z}| \leqq 1$ for $\bar{z}=\bar{x}\times\bar{y}$, but not necessarily for $\bar{z}=\bar{x}\pm\bar{y}$ or $\bar{z}=\bar{x}\div\bar{y}$. For $\bar{z}=\bar{x}\pm\bar{y}$ a scale factor β^{-1} will always be adequate, for $\bar{z}=\bar{x}\div\bar{y}$ a scale factor β^{-u} with any $u=1$,

ever, that this would be equally necessary for any other, fixed positioning of the β-adic point.[18]

This choice of the position of the β-adic point permits us to expand (2.1) to

$$(2.1') \quad \begin{gathered} \bar{x} = \epsilon(\alpha_1, \cdots, \alpha_s) = \epsilon \sum_{\sigma=1}^{s} \beta^{-\sigma}\alpha_\sigma; \\ \epsilon = \left\{\begin{matrix} +, \text{ that is, } +1 \\ -, \text{ that is, } -1 \end{matrix}\right| ; \qquad \alpha_1, \cdots, \alpha_s = 0, 1, \cdots, \beta - 1, \end{gathered}$$

and to assert that

(2.2) a digital number $\bar{x}$ lies necessarily in the interval -1, 1.

2.2. Ordinary real numbers, true operations. Precision of data. Conventions regarding these: (c), (d).

To these convention-setting remarks (a), (b) we add in a more discursive sense:

(c) We shall also use *ordinary real* numbers x. We shall even reinterpret, whenever it is convenient and for any appropriate part of the discussion, a number, which was introduced as a digital number, as an ordinary real number.

Ordinary real numbers are subject to no restrictions in size, and to them the *true operations* $x \pm y$, xy, x/y, and so on, apply.

(d) The parameters of our problem (that is, the elements of the matrix to be inverted) will usually be introduced as digital numbers. The question arises, as to what ordinary real numbers they replace.[19] The effects that these replacements, that is, errors, in the parameters have on the result are properly the subject of (B) and not of (D). It is therefore justified to view them separately, and to discuss (D) itself under the assumption that the (digital) parameter values are strictly correct. Regarding (C) cf. also §1.3.

2.3. Estimates concerning the round off errors.

Two further remarks regarding the technique and character of the round off:

2, $\cdots$ may be called for. (Our first reference to these possibilities was made in footnote 4.)

[18] We shall not discuss here the possibilities of a movable and self-adjusting, "floating" β-adic point. From the point of view of the precision of the calculation they do not differ from those of the continuous size-check-and-scale-factor procedure, to which we propose to adhere. Indeed, these two procedures bear to each other simply the relationship of automatic vs. mathematically conscious application of the same arithmetical principles.

[19] Possibly, but not necessarily, by round off. Cf., for example, the discussion of §7.5.

(e) We pointed out in (D) that the round off errors[20] η behave, as far as is known at present, essentially like independent random variables, although they are actually uniquely defined number-theoretical functions. Taking the probabilistic view of η we have (by (1.1), (1.2))

$$(2.3) \qquad \text{Mean } (\eta) = 0, \qquad \text{Dispersion } (\eta) = .29\cdot\beta^{-s};$$

taking the strict view, on the other hand, we can only assert that

$$(2.4) \qquad \text{Max } (\,|\,\eta\,|\,) = .5\cdot\beta^{-s}.$$

This discrepancy becomes even more significant when we deal with a sum of, say, m such quantities $\eta_1, \cdots, \eta_m$: Probabilistically we may infer from (2.3) that

$$(2.5) \qquad \text{Mean}\left(\sum_{l=1}^{m}\eta_l\right) = 0, \ \text{Dispersion}\left(\sum_{l=1}^{m}\eta_l\right) = .29\cdot m^{1/2}\cdot\beta^{-s}$$

while strictly we can infer from (2.4) only that

$$(2.6) \qquad \text{Max}\left(\left|\sum_{l=1}^{m}\eta_l\right|\right) \leqq .5\cdot m\cdot\beta^{-s}.$$

The estimate (2.6) is inferior to the estimate (2.5) by a factor $.5m/.29m^{1/2} = 1.7m^{1/2}$!

This creates a strong temptation to use *probabilistic estimates* instead of *strict estimates*, especially because expressions of the form

$$(2.7.a) \qquad \sum_{l=1}^{m}\bar{x}_l\bar{y}_l,$$

which give rise to round off errors

$$(2.7.b) \quad \sum_{l=1}^{m}\bar{x}_l\bar{y}_l - \sum_{l=1}^{m}\bar{x}_l\times\bar{y}_l = \sum_{l=1}^{m}(\bar{x}_l\bar{y}_l - \bar{x}_l\times\bar{y}_l) = \sum_{l=1}^{m}\eta_l$$

of the type in question, will be particularly frequent in our deductions. We shall, nevertheless, adhere to strict estimates throughout this paper (with some specified exceptions in §3.5).

(f) There is an alternative method which reduces the total round off error in the situations (2.7.a)–(2.7.b), and which deserves consideration. In fact, it effects an even greater reduction of the round off error in question than the probabilistic view of (e), and it does so on the basis of strict estimates. It requires, however, an actual change

[20] We mean $\eta^{(p)} = \bar{x}\bar{y} - \bar{x}\times\bar{y}$ and $\eta^{(q)} = (\bar{x}/\bar{y}) - (\bar{x}\div\bar{y})$. The considerations which follow are primarily significant for $\eta^{(p)}$.

in the computing technique—but this is a change to which most existing and most planned digital computing devices lend themselves readily.

This method may be described as follows:

In multiplying two s-place numbers, most computing machines do actually form the true $2s$-place product, and the rounding off to s-places is a separate operation, which may (and usually is) effected subsequently, but which can also be omitted. The s-place character of the machine finds its expression at a different point: The machine can accept s-place factors only, that is, it cannot form the product of two $2s$-place numbers (neither the $2s$-place pseudo-product, nor the $4s$-place true product). In addition, it can accept s-place addends (or minuends and subtrahends) only. It is easy, however, to use such a machine to add or to subtract $2s$-place numbers, but it would be considerably more involved to use it to obtain products of $2s$-place numbers.

It is therefore usually quite feasible and convenient to do this: Maintain the definition of digital numbers as s-place aggregates, that is, maintain (a)–(b) in (2.1). When the situation (2.7.a)–(2.7.b) arises, that is, when an expression (2.7.a) has to be computed, then do not form in the conventional way

$$(2.7.\text{a}') \qquad \sum_{l=1}^{m} \bar{x}_l \times \bar{y}_l,$$

that is, do not round off each term of (2.7.a) separately to s places. Instead, form the true $2s$-place products $\bar{x}_l\bar{y}_l$ of the s-place factors $\bar{x}_l$, $\bar{y}_l$, form their sum $\sum_{l=1}^{m}$ correctly to $2s$-places, and then (at the end) round off to s places. The result is a digital number in the original sense, that is, s-place, to be denoted by

$$(2.7.\text{a}'') \qquad \sum_{l=1}^{m}{}^{*} \bar{x}_l\bar{y}_l.$$

This (2.7.a″) is a much better approximant of (2.7a) than (2.7.a′). Indeed, for the latter we have only the estimate

$$(2.7.\text{b}') \qquad \left| \sum_{l=1}^{m} \bar{x}_l\bar{y}_l - \sum_{l=1}^{m} \bar{x}_l \times \bar{y}_l \right| \leqq \frac{m\beta^{-s}}{2}$$

while for the former clearly

$$(2.7.\text{b}'') \qquad \left| \sum_{l=1}^{m} \bar{x}_l\bar{y}_l - \sum_{l=1}^{m}{}^{*} \bar{x}_l\bar{y}_l \right| \leqq \frac{\beta^{-s}}{2} \cdot$$

Thus, the estimate (2.7.b′) is inferior to the estimate (2.7.b″) by a factor m. Note that both estimates are strict (not probabilistic).

This is the *double precision procedure*. There are several places in this paper where it could be used to improve our estimates. We shall, however, not use it in this paper (with some specified exceptions in §3.5).

2.4. **The approximative rules of algebra for pseudo-operations.** The pseudo-operations with which we shall have to work affect the ordinary laws of algebra in a manner which deserves some comment.

The laws to which we refer are these: Distributivity, commutativity, and associativity of multiplication, and the inverse relation between multiplication and division. When we replace true multiplication and division by pseudo-multiplication and division, then all of these, with the sole exception of the commutative law of multiplication, cease to be strictly valid. They are replaced by inequalities involving the round off error $\beta^{-s}/2$.

The basic inequalities in this field are

$$|\bar{a} \times \bar{b} - \bar{a}\bar{b}| \leqq \beta^{-s}/2, \tag{2.8}$$

$$|\bar{a} \div \bar{b} - \bar{a}/\bar{b}| \leqq \beta^{-s}/2. \tag{2.9}$$

From these we derive further inequalities as follows:

(2.8) implies $|(\bar{a}+\bar{b}) \times \bar{c} - (\bar{a} \times \bar{c} + \bar{b} \times \bar{c})| \leqq 3\beta^{-s}/2.$

However, the left-hand side is an integer multiple of β^{-s}, hence

$$|(\bar{a}+\bar{b}) \times \bar{c} - (\bar{a} \times \bar{c} + \bar{b} \times \bar{c})| \leqq \beta^{-s}. \tag{2.10}$$

We mentioned already

$$\bar{a} \times \bar{b} = \bar{b} \times \bar{a}. \tag{2.11}$$

Next

$$\bar{a} \times (\bar{b} \times \bar{c}) - \bar{a}\bar{b}\bar{c} = (\bar{a} \times (\bar{b} \times \bar{c}) - \bar{a}(\bar{b} \times \bar{c})) + \bar{a}(\bar{b} \times \bar{c} - \bar{b}\bar{c}),$$

hence

$$|\bar{a} \times (\bar{b} \times \bar{c}) - \bar{a}\bar{b}\bar{c}| \leqq (1 + |\bar{a}|)\beta^{-s}/2 \leqq \beta^{-s}. \tag{2.12}$$

Interchanging $\bar{a}$, $\bar{c}$ and adding gives

$$\begin{aligned} &|\bar{a} \times (\bar{b} \times \bar{c}) - (\bar{a} \times \bar{b}) \times \bar{c}| \\ &\qquad \leqq (2 + |\bar{a}| + |\bar{c}|)\beta^{-s}/2 \leqq 2\beta^{-s}. \end{aligned} \tag{2.13}$$

In addition, if either $|\bar{a}| \neq 1$ or $|\bar{c}| \neq 1$, then this is less than $2\beta^{-s}$, and since the left-hand side is an integer multiple of β^{-s}, it is necessarily less than or equal to β^{-s}. For $|\bar{a}| = |\bar{c}| = 1$, that is, $\bar{a}, \bar{c} = \pm 1$, the left-hand side is clearly 0. Hence always

$$|\bar{a} \times (\bar{b} \times \bar{c}) - (\bar{a} \times \bar{b}) \times \bar{c}| \leqq \beta^{-s}. \tag{2.13$'$}$$

Finally

$$(\bar{a} \div \bar{b}) \times \bar{b} - \bar{a} = ((\bar{a} \div \bar{b}) \times \bar{b} - (\bar{a} \div \bar{b})\bar{b}) + (\bar{a} \div \bar{b} - \bar{a}/\bar{b})\bar{b},$$

hence

$$|(\bar{a} \div \bar{b}) \times \bar{b} - \bar{a}| \leqq (1 + |\bar{b}|)\beta^{-s}/2 \leqq \beta^{-s}. \tag{2.14}$$

Again, if $|\bar{b}| \neq 1$, then this is less than β^{-s}, and since the left-hand side is an integer multiple of β^{-s} it is necessarily equal to 0. For $|\bar{b}| = 1$, that is, $\bar{b} = \pm 1$, the left-hand side is clearly 0. Hence always

$$(\bar{a} \div \bar{b}) \times \bar{b} = \bar{a}. \tag{2.14$'$}$$

On the other hand

$$(\bar{a} \times \bar{b}) \div \bar{b} - \bar{a} = ((\bar{a} \times \bar{b}) \div \bar{b} - (\bar{a} \times \bar{b})/\bar{b}) + (\bar{a} \times \bar{b} - \bar{a}\bar{b})/\bar{b},$$

hence

$$|(\bar{a} \times \bar{b}) \div \bar{b} - \bar{a}| \leqq (1 + |\bar{b}|^{-1})\beta^{-s}/2 \leqq |\bar{b}|^{-1}\beta^{-s}. \tag{2.15}$$

Note how unfavorably (2.15) compares with (2.14′), or even with (2.14), especially when $\bar{b} \ll 1$. Distinctions of this type will play an important role in our work, and they are worth emphasizing, since they are not at all in the spirit of ordinary algebra.

2.5. **Scaling by iterated halving.** The pseudo-operations that we have discussed so far are probably adequate for our work. It is nevertheless convenient to introduce an additional one. It must be said that both the need for this operation and the optimality of the form in which we introduce it are less cogently established than their equivalents for the pseudo-operations considered so far. The second point is particularly relevant: Better ways of defining and manipulating a new pseudo-operation with essentially the same potentialities may be found. At present, however, the procedure that we propose to follow seems reasonable and adequate.

The operation in question is needed in order to facilitate the manipulation of the scale factors mentioned in (b) in 2.1. If an increase in the size of a (digital) number $\bar{a}$ is wanted, we can multiply it with an integer l $(=2, 3, \cdots)$: $l\bar{a}$. This is not a pseudo-operation (cf. the end of (a) in §2.1). In order to be able to effect a decrease in size, it is desirable to be able to perform the inverse operation: Division by an integer l $(=2, 3, \cdots)$. This is necessarily a pseudo-operation: $\bar{a} \div l$. (Since l does not lie in the interval $-1, 1$, it is not a digital number in the sense of (a) in §2.1.) It will be important to arrange this operation so that it can be iterated with as little extra complication as possible, since scale factors in a calculation are likely to be introduced successively and take cumulative effect. This indicates the desirability of having the "associative law"

$$(\bar{a} \div l) \div m = \bar{a} \div lm. \tag{2.16}$$

It also suggests that it might be sufficient to use only those l which are powers of a fixed integer γ ($=2, 3, \cdots$):

$$l = \gamma^p \qquad (p = 0, 1, 2, \cdots). \tag{2.17}$$

We can then obtain (2.16) with ease, by defining $\bar{a} \div \gamma^p$ not as the result of *a single division of $\bar{a}$ by $l = \gamma^p$*, but of *a p times iterated division of $\bar{a}$ by γ*. We shall adhere to this definition throughout what follows:

$$\bar{a} \div \gamma^p = (\cdots ((\bar{a} \div \gamma) \div \gamma) \cdots) \div \gamma \qquad (p \text{ times}). \tag{2.18}$$

In the choice of γ two considerations intervene. First, the smaller γ, the more precise, that is, the less wasteful, the adjustments of scale will be that we base on it. (Cf., for example, the relationship of (6.50.a) and (6.50.b).) Since $\gamma = 2, 3, \cdots$, this suggests the choice $\gamma = 2$. Second, it simplifies things somewhat if we put γ equal to the base of our digital system: $\gamma = \beta$. Indeed, in this case $\bar{a} \div \gamma$ is merely a shift of $\bar{a}$ by one place to the right. (Or, equivalently, a shift of the β-adic point by one place to the left. We prefer the first formulation, in view of the convention regarding the position of the β-adic point, formulated in (b) in §2.1.)

Thus we have two competing choices: $\gamma = 2$ and $\gamma = \beta$. For $\beta = 2$, that is, in the binary system, the two coincide. Indeed, this seems to be one of the major arguments in favor of the use of the binary system in high speed, automatic computing. It seems preferable, however, to make here no assumptions concerning β, but to dispose of γ only. After taking all factors into account, it seems to us that the choice

$$\gamma = 2 \tag{2.19}$$

is preferable for all β, and we shall therefore use (2.19) throughout what follows.

We conclude with two estimates.

Clearly

$$| \bar{a} \div 2 - \bar{a}/2 | \leqq \beta^{-s}/2. \tag{2.20}$$

The formula

$$\bar{a} \div 2^p - \bar{a}/2^p = \sum_{q=1}^{p} ((\bar{a} \div 2^{q-1}) \div 2 - (\bar{a} \div 2^{q-1})/2)/2^{p-q}$$

now gives

$$\begin{aligned}\left| \bar{a} \div 2^p - \bar{a}/2^p \right| &\leqq \left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{p-1}}\right)\frac{\beta^{-s}}{2} \\ &= \left(2 - \frac{1}{2^{p-1}}\right)\frac{\beta^{-s}}{2} = \left(1 - \frac{1}{2^p}\right)\beta^{-s}.\end{aligned}$$

From this we infer first:

$$(2.21) \qquad \left| \bar{a} \div 2^p - \bar{a}/2^p \right| < \beta^{-s}.$$

Second, if $|\bar{a} - b| \leqq k\beta^{-s}$, then in view of the formula

$$\bar{a} \div 2^p - b/2^p = (\bar{a} \div 2^p - \bar{a}/2^p) + (\bar{a} - b)/2^p$$

we infer

$$\begin{aligned}\left| \bar{a} \div 2^p - b/2^p \right| &\leqq \left(1 - \frac{1}{2^p}\right)\beta^{-s} + \frac{k}{2^p}\beta^{-s} = \left(1 + \frac{k-1}{2^p}\right)\beta^{-s} \\ &\leqq (1 + \operatorname{Max}(0, k-1))\, \beta^{-s} \\ &= \operatorname{Max}(1, k)\, \beta^{-s},\end{aligned}$$

that is,

$$(2.22) \qquad \begin{aligned} |\bar{a} - b| &\leqq k\beta^{-s} \quad \text{implies} \\ \left| \bar{a} \div 2^p - b/2^p \right| &\leqq \operatorname{Max}(1, k)\, \beta^{-s}. \end{aligned}$$

Chapter III. Elementary matrix relations

3.1. **The elementary vector and matrix operations.** Since our discussions will center on nth order matrices

$$A = (a_{ij}),\ B = (b_{ij}),\ \cdots \qquad (i, j = 1, \cdots, n),$$

we have to introduce matrix notations. It will also be convenient to be able to refer to nth order vectors: $\xi = (x_i)$, $\eta = (y_i)$, $\cdots$ $(i = 1, \cdots, n)$. At first we shall discuss these in terms of ordinary real numbers (and true operations) only, but in §3.5 we shall introduce digital numbers (and pseudo-operations), too.

We use, of course, the *sum* and the *scalar product* for vectors and for matrices: $\xi + \eta = (x_i + y_i)$, $a\xi = (ax_i)$, $A + B = (a_{ij} + b_{ij})$, $aA = (aa_{ij})$. We fix the conventions for the *application* of a matrix to a vector: $A\xi = \eta$ with $\sum_{j=1}^{n} a_{ij}x_j = y_i$ and for the *matrix product*: $AB = C$ with

$\sum_{k=1}^{n} a_{ik} b_{kj} = c_{ij}$, so as to have the mixed associative law: $A(B\xi) = (AB)\xi$.

We need further:

For vectors: The *inner product* $(\xi, \eta) = \sum_{i=1}^{n} x_i y_i$, and the *norm* $|\xi| \geqq 0$ with $|\xi|^2 = (\xi, \xi) = \sum_{i=1}^{n} x_i^2$.

For matrices: The *transposed* matrix $A^* = (a_{ji})$, the determinant $D(A)$, and the *trace* $t(A) = \sum_{i=1}^{n} a_{ii}$. Clearly $t(AB) = t(BA) = \sum_{i,j=1}^{n} a_{ij} b_{ji}$; the *norm* $N(A) \geqq 0$ with $(N(A))^2 = t(A^*A) = t(AA^*) = \sum_{i,j=1}^{n} a_{ij}^2$; also the (*upper*) *bound* $|A|$ and the *lower bound* $|A|_l$, which will be defined further below.

The properties of these entities are too well known to require much discussion. We shall only touch briefly on those which link the most crtical ones: $|A|$, $|A|_l$ and $N(A)$.

3.2. **Properties of $|A|$, $|A|_l$ and $N(A)$.** We begin with $|A|$, $|A|_l$. We define:

$$|A| = \operatorname*{Max}_{|\xi|=1} |A\xi|, \tag{3.1.a}$$

$$|A|_l = \operatorname*{Min}_{|\xi|=1} |A\xi|. \tag{3.1.b}$$

It follows immediately, that

(3.2.a) $|A|$ is the smallest c for which $|A\xi| \leqq c|\xi|$ holds for all ξ,

(3.2.b) $|A|_l$ is the largest c for which $|A\xi| \geqq c|\xi|$ holds for all ξ.

Clearly

$$|A| \geqq |A|_l \geqq 0. \tag{3.3}$$

Also:

(3.4) $|A| > 0$ is equivalent to $A \neq 0$.

(3.5) $|A|_l > 0$ is equivalent to this:

(3.5.a) $A\xi = \eta$ is a one-to-one mapping of all vectors ξ on all vectors η, that is to this:

(3.5.b) A^{-1} exists.

This is, of course, equivalent to

$$D(A) \neq 0, \tag{3.5.c}$$

and is termed the *nonsingularity* of A.

For a nonsingular A we have further:

$$|A^{-1}| = |A|_l^{-1}, \tag{3.5.d}$$

$$|A^{-1}|_l = |A|^{-1}. \tag{3.5.e}$$

Other obvious relations are:

$$|aA| = |a||A|, \tag{3.6.a}$$

$$|aA|_l = |a||A|_l, \tag{3.6.b}$$

$$|I| = |I|_l = 1, \tag{3.7}$$

where I is the unit matrix:

$$I = (\delta_{ij}), \qquad \delta_{ij} = \left\{\begin{matrix}1 & \text{for} & i = j,\\ 0 & \text{for} & i \neq j.\end{matrix}\right| \tag{3.7.a}$$

Further

$$|A + B| \begin{cases}\leqq |A| + |B|,\\ \geqq ||A| - |B||,\end{cases} \tag{3.8.a}$$

$$|A + B|_l \begin{cases}\leqq |A|_l + |B|,\\ \geqq |A|_l - |B|\end{cases} \tag{3.8.b}$$

(and the same with A, B interchanged),

$$|A|_l|B|_l \leqq |AB|_l \leqq \begin{Bmatrix}|A||B|_l\\ |A|_l|B|\end{Bmatrix} \leqq |AB| \leqq |A||B|. \tag{3.9}$$

Next

$$|A| \text{ is the smallest } c \text{ for which } |(A\xi, \eta)| \leqq c|\xi||\eta|. \tag{3.10}$$

To see this, it suffices to show that for any given c the validity of $|A\xi| \leqq c|\xi|$ for all ξ is equivalent to the validity of $|(A\xi, \eta)| \leqq c|\xi||\eta|$ for all ξ, η. Now the former implies $|(A\xi, \eta)| \leqq |A\xi||\eta| \leqq c|\xi||\eta|$, that is, it implies the latter; and the latter implies (with $\eta = A\xi$) $|A\xi|^2 = |(A\xi, A\xi)| \leqq c|\xi||A\xi|$, hence $|A\xi| \leqq c|\xi|$ (this obtains by division by $|A\xi|$ when $|A\xi| > 0$, otherwise it is obvious, since $|A\xi| = 0$), that is, it implies the former.

Since $(A^*\xi, \eta) = (A\eta, \xi)$, therefore (3.10) implies

$$|A| = |A^*|. \tag{3.11.a}$$

Since $(A^*)^{-1}$ exists if and only if A^{-1} exists and is then $= (A^{-1})^*$, therefore (3.5) on one hand and (3.5.d), (3.5.e), (3.11.a) on the other give

$$|A|_l = |A^*|_l. \tag{3.11.b}$$

Next

$$|A^*A| = |A|^2. \tag{3.12.a}$$

Indeed: $\leqq$ follows from (3.9), (3.11.a). $\geqq$ obtains by using (3.10) for A^*A and (3.2.a) for A: $|A\xi|^2=(A\xi, A\xi)=(A^*A\xi, \xi)\leqq|A^*A||\xi|^2$, $|A\xi|\leqq(|A^*A|)^{1/2}|\xi|$, hence $|A|\leqq(|A^*A|)^{1/2}$, $|A^*A|\geqq|A|^2$.

If A^{-1} exists, then $(A^*A)^{-1}$ exists and equals $A^{-1}(A^{-1})^*$; if $(A^*A)^{-1}$ exists then A^{-1} exists and equals $(A^*A)^{-1}A^*$. Hence $(A^*A)^{-1}$ exists if and only if A^{-1} exists and is then $A^{-1}(A^{-1})^*$. Therefore, (3.5) on one hand and (3.5.d), (3.5.e), (3.11.b), (3.12.a) on the other give

$$|A^*A|_l = |A|_l^2. \tag{3.12.b}$$

We now pass to the consideration of $N(A)$.

Clearly

$$N(A) = N(A^*), \tag{3.13}$$

and

$$N(aA) = |a|\,N(A). \tag{3.14}$$

For $A=(a_{ij})$ fix $j=1, \cdots, n$ and view $i=1, \cdots, n$ as a vector index, then $A^{\{j\}}=(a_{ij})$ $(i=1, \cdots, n)$ defines a vector $A^{\{j\}}$. Clearly $(N(A))^2=\sum_{j=1}^n |A^{\{j\}}|^2$. Now$(A+B)^{\{j\}}=A^{\{j\}}+B^{\{j\}}$, hence

$$\begin{aligned} N(A+B) &= \left(\sum_{j=1}^n |A^{\{j\}}+B^{\{j\}}|^2\right)^{1/2} \\ &\leqq \left(\sum_{j=1}^n (|A^{\{j\}}|+|B^{\{j\}}|)^2\right)^{1/2} \\ &\leqq \left(\sum_{j=1}^n |A^{\{j\}}|^2\right)^{1/2} + \left(\sum_{j=1}^n |B^{\{j\}}|^2\right)^{1/2} \\ &\leqq N(A)+N(B), \end{aligned}$$

that is

$$N(A+B) \leqq N(A)+N(B). \tag{3.15}$$

Furthermore, $(AB)^{\{j\}}=A(B^{\{j\}})$, hence

$$\begin{aligned} N(AB) &= \left(\sum_{j=1}^n |A(B^{\{j\}})|^2\right)^{1/2} \leqq \left(\sum_{j=1}^n (|A||B^{\{j\}}|)^2\right)^{1/2} \\ &= |A|\left(\sum_{j=1}^n |B^{\{j\}}|^2\right)^{1/2} = |A|\,N(B), \end{aligned}$$

that is,

$$N(AB) \leqq |A|\, N(B). \tag{3.16.a}$$

Applying (3.16.a) to B^*, A^* (in place of A, B) and using $(AB)^* = B^*A^*$ as well as (3.11.a), (3.13) gives

$$N(AB) \leqq |B|\, N(A). \tag{3.16.b}$$

Given a vector $\xi = (x_i)$ $(i = 1, \cdots, n)$ define a matrix

$$\xi^{\sim} = (x_{ij}) \qquad x_{ij} = \begin{cases} x_i & \text{for } j = 1, \\ 0 & \text{for } j \neq 1. \end{cases}$$

Then $N(\xi^{\sim}) = |\xi|$, $(A\xi)^{\sim} = A(\xi^{\sim})$. Hence (3.16.b) gives $|A\xi| \leqq N(A)|\xi|$, that is, $|A| \leqq N(A)$. Combining this with (3.16.a) with $B = I$ (note that $N(I) = n^{1/2}$) gives

$$|A| \leqq N(A) \leqq n^{1/2}|A|. \tag{3.17.a}$$

Both estimates of (3.17.a) are optimal: The second $\leqq$ becomes $=$ for $A = I = (\delta_{ij})$, the first $\leqq$ becomes $=$ for $A = (1)$.

Consider the vectors $I^{\{k\}} = (\delta_{ik})$ $(i = 1, \cdots, n)$. $|I^{\{k\}}| = 1$, $(AI^{\{j\}}, I^{\{i\}}) = a_{ij}$ hence (3.10) gives $|a_{ij}| \leqq |A|$. Again $(N(A))^2 = \sum_{i,j=1}^{n} a_{ij}^2 \leqq n^2 \text{Max}_{i,j=1,\cdots,n}\, a_{ij}^2$, hence (by 3.17.a) (first $\leqq$) $|A| \leqq n\text{Max}_{i,j=1,\cdots,n} |a_{ij}|$. Thus

$$\underset{i,j=1,\cdots,n}{\text{Max}} |a_{ij}| \leqq |A| \leqq n \underset{i,j=1,\cdots,n}{\text{Max}} |a_{ij}|. \tag{3.17.b}$$

Both estimates of (3.17.b) are optimal: The first $\leqq$ becomes $=$ for $A = I$; the second $\leqq$ becomes $=$ for $A = (1)$.

3.3. **Symmetry and definiteness.** We recall further the definitions of *symmetry* and of *definiteness*[21] for matrices. A is *symmetric* if

$$A = A^*, \quad \text{that is, if} \quad a_{ij} = a_{ji} \qquad (i, j = 1, \cdots, n). \tag{3.18}$$

A is *definite* if it is symmetric and if

$$(A\xi, \xi) \geqq 0 \qquad \text{for all } \xi. \tag{3.19}$$

We note:

(3.20) A^*A is always definite.

Indeed: $(A^*A)^* = A^*A^{**} = A^*A$, and $(A^*A\xi, \xi) = (A\xi, A\xi) = |A\xi|^2 \geqq 0$.

[21] Our present concept of definiteness corresponds to what is usually known as "non-negative semi-definiteness."

We define the *proper values* $\lambda_1, \cdots, \lambda_n$ of a matrix A (with multiplicities) as usual: They are the roots (with multiplicities) of the nth order polynomial $D(\lambda I - A)$. We shall only use them when A is symmetric; in this case they are all real. For a symmetric A definiteness is equivalent to

$$\lambda_i \geqq 0 \qquad \text{for all } i = 1, \cdots, n. \tag{3.21.a}$$

In this case we make it a convention to arrange the proper values in a monotone nonincreasing sequence:

$$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n \geqq 0. \tag{3.21.b}$$

For a definite A

$$|A| = \lambda = \lambda_1, \tag{3.22.a}$$

$$|A|_l = \mu = \lambda_n \tag{3.22.b}$$

and therefore (using (3.5))

$$A \text{ is non-singular if and only if } \lambda_n > 0. \tag{3.22.c}$$

Further

$$D(A) = \prod_{i=1}^{n} \lambda_i, \tag{3.23}$$

$$t(A) = \sum_{i=1}^{n} \lambda_i, \tag{3.24}$$ [22]

and by applying (3.24) to $A^*A = A^2$, whose proper values are $\lambda_1^2, \cdots, \lambda_n^2$,

$$N(A) = \left(\sum_{i=1}^{n} \lambda_i^2 \right)^{1/2}. \tag{3.25}$$ [23]

3.4. **Diagonality and semi-diagonality.** To conclude, we refer to the classes of *diagonal*, *upper semi-diagonal* and *lower semi-diagonal* matrices. A matrix $A = (a_{ij})$ belongs to these classes if $a_{ij} = 0$ whenever $i \neq j$, or whenever $i > j$, or whenever $i < j$, respectively. Denote these three classes by $\mathcal{C}_0$, $\mathcal{C}_+$, $\mathcal{C}_-$, respectively. For $\mathcal{C} = \mathcal{C}_0$ or $\mathcal{C}_+$ or $\mathcal{C}_-$ define $\mathcal{C}' = \mathcal{C}_0$ or $\mathcal{C}_-$ or $\mathcal{C}_+$, respectively. Now the following facts are well known:

(3.26) Let A, B belong to $\mathcal{C}$. Then aA, $A \pm B$, AB and (if it exists) A^{-1} belong to $\mathcal{C}$, while A^* belongs to $\mathcal{C}'$. A^{-1} exists if and only if all

[22] (3.23), (3.24) hold, of course, for all matrices A.

[23] Here $A^* = A$ is being used; (3.25) holds only for symmetric matrices A.

diagonal elements of A are unequal to 0. In all these procedures the diagonal elements of A behave as if they formed a diagonal matrix by themselves.

(3.27) For $A=(a_{ij})=(a_i\delta_{ij})$ in $\mathcal{C}_0$ the diagonal elements $a_1, \cdots, a_n$ are the proper values of A (not necessarily monotone) and

$$|A| = \operatorname*{Max}_{i=1,\cdots,n} |a_i|,$$

$$|A|_l = \operatorname*{Min}_{i=1,\cdots,n} |a_i|.$$

These two relations are not valid in $\mathcal{C}_+$ and $\mathcal{C}_-$.

For an $A=(a_{ij})=(a_i\delta_{ij})$ in $\mathcal{C}_0$ the formation of A^{-1} is trivial: $A^{-1}=(a_i^{-1}\delta_{ij})$. For an $A=(a_{ij})$ in $\mathcal{C}_+$ or $\mathcal{C}_-$, A^{-1} still obtains by a fairly simple and explicit algorithm. We shall see subsequently (cf. the end of §4.3) that this is one of the two salient points of the elimination method.

3.5. **Pseudo-operations for matrices and vectors. The relevant estimates.** We now pass to the pseudo-operations for matrices and vectors. We shall actually need the *matrix pseudo-product*, and it is quite convenient, essentially for the purpose of illustration, to introduce the (vector) *inner pseudo-product*, too. Besides, we shall discuss each one of these in two forms: ordinary precision (cf. (a) in §2.1) and double precision (cf. (f) in §2.3).

In (a) in §2.1 we introduced digital numbers $\bar{x}$, which could, however, also be viewed as ordinary real numbers, cf. (c) in §2.2. We introduce now, in the same sense, *digital matrices* $\overline{A}=(\bar{a}_{ij})$, $B=(\bar{b}_{ij}), \cdots$ $(i,j=1, \cdots, n)$ and digital vectors $\bar{\xi}=(\bar{x}_i)$, $\bar{\eta}=(\bar{y}_i), \cdots$ $(i=1, \cdots, n)$—the relevant fact being that the $\bar{a}_{ij}$, $\bar{b}_{ij}, \cdots, \bar{x}_i, \bar{y}_i, \cdots$ are digital numbers. As indicated above, we introduce only two pseudo-operations, but each in two forms:

The (*ordinary precision*) *inner pseudo-product*: $(\bar{\xi}\bigcirc\bar{\eta})=\sum_{i=1}^{n}\bar{x}_i\times\bar{y}_i$; the *double precision inner pseudo-product*: $(\bar{\xi}\bigcirc\bigcirc\bar{\eta})=\sum_{i=1}^{*n}\bar{x}_i\bar{y}_i$; the (*ordinary precision*) *matrix pseudo-product*: $\overline{A}\times\overline{B}=\overline{C}$ with $\bar{c}_{ij}=\sum_{k=1}^{n}\bar{a}_{ik}\times\bar{b}_{kj}$; the *double precision matrix pseudo-product*: $\overline{A}\times\times\overline{B}=\overline{C}$ with $\bar{c}_{ij}=\sum^{*}\bar{a}_{ik}\bar{b}_{kj}$.

The only ordinary law of algebra which is not invalidated by the transition from true operations to pseudo-operations is, as in §2.4, the commutative law of multiplication. It holds for the true inner product, but not for the true matrix product, hence we obtain only these pseudo-relations:

$$(3.28.a) \qquad (\bar{\xi} \bigcirc \bar{\eta}) = (\bar{\eta} \bigcirc \bar{\xi}),$$

$$(3.28.b) \qquad (\bar{\xi} \bigcirc\bigcirc \bar{\eta}) = (\bar{\eta} \bigcirc\bigcirc \bar{\xi}).$$

The other laws are, as in §2.4, replaced by inequalities involving the round off error $\beta^{-s}/2$.

In order to obtain a first orientation concerning these, we begin by restating from (e) and (f) in (2.3):

$$(3.29.a) \qquad |(\bar{\xi}, \bar{\eta}) - (\bar{\xi} \bigcirc \bar{\eta})| \leqq n\beta^{-s}/2 \qquad \text{(strict)},$$

$$(3.29.b) \qquad (\bar{\xi}, \bar{\eta}) - (\bar{\xi} \bigcirc \bar{\eta}) \text{ has a Mean} = 0 \text{ and a Dispersion} \leqq .29n^{1/2}\beta^{-s} \qquad \text{(probabilistic)},$$

$$(3.29.c) \qquad |(\bar{\xi}, \bar{\eta}) - (\bar{\xi} \bigcirc\bigcirc \bar{\eta})| \leqq \beta^{-s}/2 \qquad \text{(strict)}.$$

We now pass to $\bar{A}\times\bar{B}$ and $\bar{A}\times\times\bar{B}$. The elements of these matrices and the corresponding ones in $\bar{A}\bar{B}$ are built exactly like the expressions $(\bar{\xi}\bigcirc\bar{\eta})$, $(\bar{\xi}\bigcirc\bigcirc\bar{\eta})$ and $(\bar{\xi}, \bar{\eta})$. We have, therefore, in complete analogy with (3.29.a)–(3.29.c):

For $\bar{A}\bar{B}-\bar{A}\times\bar{B}=(\rho_{ij})$

$$(3.30.a) \qquad |\rho_{ij}| \leqq n\beta^{-s}/2 \qquad \text{(strict)},$$

$$(3.30.b) \qquad \rho_{ij} \text{ has a Mean } = 0 \text{ and a Dispersion } \leqq .29n^{1/2}\beta^{-s} \qquad \text{(probabilistic)},$$

for $\bar{A}\bar{B}-\bar{A}\times\times\bar{B}=(\sigma_{ij})$

$$(3.30.c) \qquad |\sigma_{ij}| \leqq \beta^{-s}/2 \qquad \text{(strict)}.$$

(3.17.b) permits us to infer from (3.30.a) and (3.30.c):

$$(3.31.a) \qquad |\bar{A}\bar{B} - \bar{A} \times \bar{B}| \leqq n^2\beta^{-s}/2$$

$$(3.31.c) \qquad |\bar{A}\bar{B} - \bar{A} \times\times \bar{B}| \leqq n\beta^{-s}/2 \qquad \text{(strict)}.$$

Drawing a probabilistic inference from (3.30.b) is more difficult. Using some results of V. Bargmann[24] it is possible to show this:

(3.31.b) $|\bar{A}\bar{B}-\bar{A}\times\bar{B}| \leqq kn\beta^{-s}$ has a probability nearly 1 for moderately large values of k.

It seems worth noting that the estimates of (3.31.b) and (3.31.c)

[24] These results are contained in a manuscript entitled *Statistical distribution of proper values*. This work was done under the auspices of the U. S. Navy, Bureau of Ordnance, under Contract NORD9596 (1946), and will be published elsewhere.

In this connection we wish to mention further work done on matrix inversion by the iteration method. It was done under the same contract and appeared in a report by V. Bargmann, D. Montgomery, and J. von Neumann, entitled *Solution of linear systems of high order*.

are of the same order of magnitude (that is, they involve the same power of n), which is not true for the estimates of (3.29.b) and (3.29.c), on which they are based. However, we do not propose to pursue the probabilistic estimates of the type (b) any further in this paper, although they are interesting and practically very relevant. We shall consider them at a later occasion. We shall instead continue here with the analysis of the strict estimates of the types (a) and (c).

(3.31.a), (3.33.c) give

$$(3.32.a)\qquad |\bar{A} \times (\bar{B} + \bar{C}) - (\bar{A} \times \bar{B} + \bar{A} \times \bar{C})| \leqq 3n^2\beta^{-s}/2,$$

$$(3.32.c)\qquad |\bar{A} \times\times (\bar{B} + \bar{C}) - (\bar{A} \times\times \bar{B} + \bar{A} \times\times \bar{C})| \leqq 3n\beta^{-s}/2.$$

Further

$$\bar{A} \times (\bar{B} \times \bar{C}) - \overline{ABC} = (\bar{A} \times (\bar{B} \times \bar{C}) - \bar{A}(\bar{B} \times \bar{C})) + \bar{A}(\bar{B} \times \bar{C} - \overline{BC}),$$

and similarly

$$\bar{A} \times\times (\bar{B} \times\times \bar{C}) - \overline{ABC} = (\bar{A} \times\times (\bar{B} \times\times \bar{C}) - \bar{A}(\bar{B} \times\times \bar{C})) + \bar{A}(\bar{B} \times\times \bar{C} - \overline{BC}),$$

hence (3.31.a), (3.31.c) also give

$$(3.33.a)\qquad |\bar{A} \times (\bar{B} \times \bar{C}) - \overline{ABC}| \leqq (1 + |\bar{A}|)n^2\beta^{-s}/2,$$

$$(3.33.c)\qquad |\bar{A} \times\times (\bar{B} \times\times \bar{C}) - \overline{ABC}| \leqq (1 + |\bar{A}|)n\beta^{-s}/2.$$

Interchanging $\bar{A}$, $\bar{C}$ and adding gives

$$(3.34.a)\qquad \begin{aligned} &|\bar{A} \times (\bar{B} \times \bar{C}) - (\bar{A} \times \bar{B}) \times \bar{C}| \\ &\qquad\qquad \leqq (2 + |\bar{A}| + |\bar{C}|)n^2\beta^{-s}/2, \end{aligned}$$

$$(3.34.c)\qquad \begin{aligned} &|\bar{A} \times\times (\bar{B} \times\times \bar{C}) - (\bar{A} \times\times \bar{B}) \times\times \bar{C}| \\ &\qquad\qquad \leqq (2 + |\bar{A}| + |\bar{C}|)n\beta^{-s}/2. \end{aligned}$$

In comparing (3.33.a)–(3.34.c) with (2.12), (2.13), it should be remembered that we had there $|\bar{a}| \leqq 1$, $|\bar{c}| \leqq 1$, whereas now we have $|\bar{a}_{ij}| \leqq 1$, $|\bar{c}_{ij}| \leqq 1$, but from this we can infer (by (3.17.b)) only $|\bar{A}| \leqq n$, $|\bar{C}| \leqq n$.

More detailed evaluations will be derived when we get to our primary problems in Chapter VI.

Chapter IV. The elimination method

4.1. Statement of the conventional elimination method. In order to have a fixed point of reference, and also in order to introduce the notations that will be used in the subsequent sections of this paper, we described first the conventional elimination method—using true operations, and not yet pseudo-operations.

The elimination method is usually viewed as one for equation-solving and not for matrix-inverting, but this actually amounts to the same thing: Given a nonsingular matrix $A=(a_{ij})$ $(i, j=1, \cdots, n)$ and the corresponding equation system

$$\sum_{j=1}^{n} a_{ij}x_j = y_i \qquad (i = 1 \cdots n), \tag{4.1}$$

that is,

$$A\xi = \eta, \tag{4.1'}$$

the solution

$$\sum_{j=1}^{n} t_{ij}y_j = x_i \qquad (i = 1, \cdots, n), \tag{4.2}$$

that is,

$$T\eta = \xi, \tag{4.2'}$$

is clearly furnishing the desired inverse:

$$T = A^{-1}. \tag{4.3}$$

Given the system of n equations (4.1) with the n unknowns $x_1, \cdots, x_n$, the solution by elimination proceeds in the following, familiar way:

Assume that the $k-1$ first unknowns $x_1, \cdots, x_{k-1}$ $(k=1, \cdots, n-1)$ have already been eliminated, and that, for the remaining $n-k+1$ unknowns $x_k, \cdots, x_n$, $n-k+1$ equations have been derived:

$$\sum_{j=k}^{n} a_{ij}^{(k)} x_j = y_i^{(k)} \qquad (i = k, \cdots, n). \tag{4.4}$$

Then the elimination of the next unknown, x_k, is effected by subtracting the $a_{kk}^{(k)}/a_{ik}^{(k)}$-fold of equation number k from equation number i $(i=k+1, \cdots, n)$. This gives a new set of equations

$$\sum_{j=k+1}^{n} a_{ij}^{(k+1)} x_j = y_i^{(k+1)} \qquad (i = k+1, \cdots, n), \tag{4.5}$$

where

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - a_{ik}^{(k)} a_{kj}^{(k)} / a_{kk}^{(k)} \qquad (i, j = k+1, \cdots, n), \tag{4.6}$$

$$y_i^{(k+1)} = y_i^{(k)} - (a_{ik}^{(k)} / a_{kk}^{(k)}) y_k^{(k)} \qquad (i = k+1, \cdots, n). \tag{4.7}$$

The transition from (4.4) to (4.5) is clearly an inductive step from k to $k+1$. This induction begins, of course, with the original equations (4.1), that is, we have

$$a_{ij}^{(1)} = a_{ij} \qquad (i, j = 1, \cdots, n), \tag{4.8}$$

$$y_i^{(1)} = y_i \qquad (i = 1, \cdots n). \tag{4.9}$$

The induction produces (4.4) successively for $k=1, \cdots, n$, that is, it produces

$$a_{ij}^{(k)} \qquad (k = 1, \cdots, n; i, j = k, \cdots, n), \tag{4.10}$$

$$y_i^{(k)} \qquad (k = 1, \cdots, n; i = k, \cdots, n). \tag{4.11}$$

After all n systems (4.4) have been derived, the first equation of each system is selected, and these are combined to a new system of n equations with the n original unknowns $x_1, \cdots, x_n$:

$$\sum_{j=k}^{n} a_{kj}^{(k)} x_j = y_k^{(k)} \qquad (k = 1, \cdots, n). \tag{4.12}$$

These are now solved by a backward induction over $k=n, \cdots, 1$:

$$x_k = \frac{1}{a_{kk}^{(k)}} y_k^{(k)} - \sum_{j=k+1}^{n} \frac{a_{kj}^{(k)}}{a_{kk}^{(k)}} x_j. \tag{4.13}$$

4.2. **Positioning for size in the intermediate matrices.** Before we undertake to analyze the procedure of §4.1, we note this:

The inductive step from k to $k+1$ (on (4.4)) involves a division by $a_{kk}^{(k)}$, and this division reappears in the k-step of (4.13). Hence it is important, from the abstract point of view, that $a_{kk}^{(k)} \neq 0$ and, from the actual computational point of view, that $a_{kk}^{(k)}$ be essentially as large as possible.

It is, however, perfectly conceivable, that an $a_{kk}^{(k)}$ turns out to be small, or even zero, although A is nonsingular: The simplest example is furnished by the possibility of $a_{11}^{(1)} = a_{11} = 0$ (that is, $k=1$), although A is nonsingular. In the actual, numerical uses of the elimination method this point is fully appreciated: It is customary to make arrangements to have $a_{kk}^{(k)}$ possess the largest absolute value among all $a_{ij}^{(k)}$ $(i,j=k,\cdots,n)$.[25] This is done by permuting the $i=k,\cdots,n$ and the $j=k,\cdots,n$ (separately) in such a way that $\text{Max}_{i,j=1,\cdots,n}\,|a_{ij}^{(k)}|$ is assumed for $i=j=k$. This permutation is effected just before the operations that lead from (4.4) to (4.5) are undertaken, that is, just before the inductive step from k to $k+1$. This occurs $n-1$ times: For $k=1,\cdots,n-1$.

We call these permutations of i and j *positioning for size.*

Note that this positioning for size will produce an $a_{kk}^{(k)}\neq 0$ (in its k-step, $k=1,\cdots,n$), unless $\text{Max}_{i,j=k,\cdots,n}\,|a_{ij}^{(k)}|=0$, that is, unless $a_{ij}^{(k)}=0$ for all $i,j=k,\cdots,n$.[26]

Now we prove:

(4.14) If A is nonsingular, then positioning for size will always produce an $a_{kk}^{(k)}\neq 0$ $(k=1,\cdots,n)$, that is, never $a_{ij}^{(k)}=0$ for all $i,j=k,\cdots,n$.

Assume the opposite: Let $k=k_1$ $(=1,\cdots,n)$ be the smallest k so that $a_{ij}^{(k)}=0$ for all $i,j=k,\cdots,n$. The system of equations (4.1) is clearly equivalent to the system of equations (4.4) with $k=k_1$, together with the system of equations (4.13) with $k=k_1-1,\cdots,1$. Now our assumption amounts to stating that the left-hand sides in the system (4.4) vanish identically. Hence the system (4.4), (4.13), that is, the equivalent system (4.1), cannot have a unique solution $x_1,\cdots,x_n$. This however is in contradiction with the nonsingularity of the matrix $A=(a_{ij})$ of (4.1).

(4.14) is the rigorous justification for the operation of positioning for size. Throughout what follows, we shall keep pointing out whether the positioning for size is or is not assumed to have taken place in any particular part of the discussion.

4.3. Statement of the elimination method in terms of factoring A into semi-diagonal factors C, B'. We return now to the procedure of §4.1, without positioning for size, for the balance of this chapter.

Summing (4.7) over $k=1,\cdots,i-1$, and remembering (4.9), gives

$$y_i = y_i^{(i)} + \sum_{k=1}^{i-1} \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\, y_k^{(k)}. \tag{4.15}$$

[25] Or at least one which has the same order of magnitude as the maximum in question. We propose, however, to disregard this possible relaxation of the requirement. We shall postulate that $|a_{kk}^{(k)}|$ be strictly equal to $\text{Max}_{i,j=k,\cdots,n}|a_{ij}^{(k)}|$.

[26] Positioning for size, as described above, occurs only for $k=1,\cdots,n-1$. For $k=n$, however, $a_{kk}^{(k)}$ is the only $a_{ij}^{(k)}$, hence the assertion is trivial.

We have $\xi=(x_i)$, $\eta=(y_i)$, let us introduce in addition $\zeta=(y_i^{(i)})$. Then (4.12) and (4.15) express two very simple matrix relations between ξ and ζ and between η and ζ. If we define

$$(4.16)\qquad \begin{aligned} B' &= (b'_{ij}) \qquad (i, j = 1, \cdots, n)\\ \text{with}\quad b'_{ij} &= \begin{cases} a_{ij}^{(i)} & \text{for } i \leqq j \\ 0 & \text{for } i > j \end{cases} \end{aligned}$$

$$(4.17)\qquad \begin{aligned} C &= (c_{ij}) \qquad (i, j = 1, \cdots, n)\\ \text{with}\quad c_{ij} &= \begin{cases} a_{ij}^{(j)}/a_{jj}^{(j)} & \text{for } i \geqq j \\ \left[\begin{matrix}\text{hence} & \\ 1 & \text{for } i = j\end{matrix}\right] & \\ 0 & \text{for } i < j \end{cases} \end{aligned}$$

then (4.12), (4.15) become

$$(4.18)\qquad B'\xi = \zeta,$$

$$(4.19)\qquad C\zeta = \eta.$$

Since these are identities with respect to the original variables $x_1, \cdots, x_n$, that is, with respect to ξ, therefore comparison of (4.18), (4.19) with (4.1′) gives

$$(4.20)\qquad A = CB'.$$

From (4.20)

$$(4.21)\qquad A^{-1} = B'^{-1}C^{-1},$$

and (4.16), (4.17) show that B', C are semi-diagonal (upper and lower, that is, in $\mathcal{C}_+$ and $\mathcal{C}_-$ respectively). Furthermore, (4.7), (4.13), which represent the conventional way of expressing the elimination method, are clearly the inductive processes that invert (4.15), (4.12), that is, (4.19), (4.18), that is, they invert the matrices C, B'. C, B' are semi-diagonal, and renewed inspection of (4.7), (4.13) shows at once that these are indeed the inductive processes that are required to invert semi-diagonal matrices. (In this connection cf. the remark at the end of §3.4, and the explicit expressions (4.29), (4.30).)

We may therefore interpret the elimination method as one which bases the inverting of an arbitrary matrix A on the combination of two tricks: First, it decomposes A into a product of two semi-diagonal matrices C, B', according to (4.20), and consequently the inverse of

A obtains immediately from those of C, B', according to (4.21).[27] Second, using the semi-diagonality of C, B', it forms their inverses by a simple, explicit, inductive process.

4.4. **Replacement of C, B' by B, C, D.** The discussion of §4.3 is complete, but it suffers from a certain asymmetry: B', C play quite symmetric roles, being upper and lower semi-diagonal, and the right- and left-factors of the decomposition (4.20) of A; however, all diagonal elements of C are identically 1, whereas those of B' are not.

This is easily remedied: Put

$$\begin{aligned} &D = (d_i\delta_{ij}) && (i, j = 1, \cdots, n) \\ \text{with}\quad &d_i = \overset{(i)}{a_{ii}} && \end{aligned} \tag{4.22}$$

$$\begin{aligned} &B = (b_{ij}) && (i, j = 1, \cdots, n) \\ \text{with}\quad &b_{ij} = \begin{cases} \overset{(i)}{a_{ij}}/\overset{(i)}{a_{ii}} & \text{for } i \leqq j, \\ \left[\begin{matrix}\text{hence} \\ 1\end{matrix}\right. & \left.\begin{matrix} \\ \text{for } i = j,\end{matrix}\right] \\ 0 & \text{for } i > j. \end{cases} \end{aligned} \tag{4.23}$$

Then clearly

$$B' = DB, \tag{4.24}$$

hence (4.20) becomes

$$A = CDB, \tag{4.25}$$

and (4.21) becomes

$$A^{-1} = B^{-1}D^{-1}C^{-1}. \tag{4.26}$$

To sum up:

(4.27) B, C, D fulfill (4.25). They belong to $\mathcal{C}_+$, $\mathcal{C}_-$, $\mathcal{C}_0$, respectively (cf. 3.4). All diagonal elements of B, C are identically 1.

Now (4.26) furnishes the desired A^{-1}, based on B^{-1}, C^{-1}, D^{-1}. D^{-1} is immediately given by

$$\overset{-1}{D} = (\overset{-1}{d_i}\delta_{ij}) \qquad (i, j = 1, \cdots, n), \tag{4.28}$$

and B^{-1}, C^{-1} obtain from simple, explicit, inductive algorithms which involve no divisions:

[27] C, B' could not both belong to the same class $\mathcal{C}_\pm$, since each class $\mathcal{C}_\pm$ is reproduced by multiplication (cf. 3.26), and A is, of course, not assumed to belong to either (to be semi-diagonal). Indeed, C is in $\mathcal{C}_-$ and B' in $\mathcal{C}_+$.

$$B^{-1} = R = (r_{ij}) \qquad (i, j = 1, \cdots, n)$$

$$\text{(4.29)} \qquad \text{with} \quad r_{ij} = \begin{cases} -\sum_{k=i+1}^{j} b_{ik}r_{kj} & \text{for } i < j, \\ 1 & \text{for } i = j, \\ 0 & \text{for } i > j, \end{cases}$$

$$C^{-1} = S = (s_{ij}) \qquad (i, j = 1, \cdots, n)$$

$$\text{(4.30)} \qquad \text{with} \quad s_{ij} = \begin{cases} -\sum_{j+1}^{i} s_{ik}c_{kj} & \text{for } i > j, \\ 1 & \text{for } i = j, \\ 0 & \text{for } i < j. \end{cases}$$

Note that (4.29) obtains from $BR=I$, and gives for every fixed j $(=1, \cdots, n)$ an inductive definition over $i=j, \cdots, 1$; while (4.30) obtains from $SC=I$, and gives for every fixed i $(=1, \cdots, n)$ an inductive definition over $j=i, \cdots, 1$.

4.5. **Reconsideration of the decomposition theorem. The uniqueness theorem.** The decisive relation (4.24) or (4.25) can also be derived directly from (4.6).

Indeed, consider two fixed $i, j=1, \cdots, n$. Put $i'=\text{Min}\ (i, j)$. Form (4.6) for $k=1, \cdots, i'-1$, and note that

$$\text{(4.6}'\text{)} \qquad 0 = \overset{(k)}{a_{ij}} - \overset{(k)}{a_{ik}}\overset{(k)}{a_{kj}}/\overset{(k)}{a_{kk}}$$

for $k=i$ and for $k=j$, that is, for $k=i'$. Summing all these equations, and remembering (4.8), gives

$$\text{(4.31)} \qquad a_{ij} = \sum_{k=1}^{i'} \frac{\overset{(k)}{a_{ik}}\overset{(k)}{a_{kj}}}{a^{(k)}_{kk}}.$$

By (4.17), (4.22), (4.23) this may be written

$$\text{(4.32)} \qquad a_{ij} = \sum_{k=1}^{i'} c_{ik}d_kb_{kj} = \sum_{k=1}^{n} c_{ik}d_kb_{kj},$$

and this is precisely the statement of (4.25).

We give this alternative derivation of (4.25), because it is ex post more direct than the original one (in §§4.3, 4.4), and because our final discussion for pseudo-operations will have to follow this pattern (cf. §§5.2 and 6.1, especially (6.3)).

To conclude, we show:

(4.33) Given A, (4.25) and (4.27) determine B, C, D uniquely.

Let $A=CDB$ and $A=C_1D_1B_1$ be two decompositions that fulfill (4.27). A is nonsingular, hence the same is true for B, C, D, B_1, C_1, D_1. $CDB=C_1D_1B_1$, hence $C_1^{-1}C=D_1B_1B^{-1}D^{-1}$. Now C, C_1 belong to $\mathcal{C}_-$, hence $C_1^{-1}C$ belongs to $\mathcal{C}_-$. B, B_1 belong to $\mathcal{C}_+$, D, D_1 belong to $\mathcal{C}_0$, hence also to $\mathcal{C}_+$, so $D_1B_1B^{-1}D^{-1}$ belongs to $\mathcal{C}_+$. (For this, and what follows, cf. §3.4.) Thus $C_1^{-1}C$ belongs to $\mathcal{C}_-$ and to $\mathcal{C}_+$, hence it belongs to $\mathcal{C}_0$, that is, it is diagonal. C, C_1 are in $\mathcal{C}_-$ and have diagonal elements 1, therefore the same is true for $C_1^{-1}C$. Owing to the above this means that $C_1^{-1}C=I$, that is, $C=C_1$. Similarly (or by interchanging rows and columns) $B=B_1$. Now $CDB=C_1D_1B_1$ gives $D=D_1$. Hence B, C, D coincide with B_1, C_1, D_1, as desired.

Note that all these results were formulated and derived without the assumption of positioning for size.

Chapter V. Specialization to definite matrices

5.1. Reasons for limiting the discussion to definite matrices. We have not so far been able to obtain satisfactory error estimates for the pseudo-operational equivalent of the elimination method in its general form, that is, for the equivalent of §§4.3–4.5. The reason for this is that any such estimate would have to depend on the bounds of some or of all of the matrices B, B^{-1}, C, C^{-1}, D, D^{-1}, that is, on $|B|$, $|B|_l$, $|C|$, $|C|_l$, $|D|$, $|D|_l$ (cf. (3.5. d)). It would be necessary to correlate these quantities, or possibly other, allied ones, to $|A|$, $|A|_l$.[28] As stated above, we have not so far been able to derive such correlations to any adequate extent.[29]

We did, however, succeed in securing everything that is needed in the special case of a definite A. Furthermore, the inverting of an unrestricted (but, of course, nonsingular) A is easily derivable from the inverting of a definite one: Indeed, by (3.20), A^*A is always definite and, by the considerations that preceded (3.12.b), A^{-1} exists if and only if $(A^*A)^{-1}$ exists and then $A^{-1}=(A^*A)^{-1}A^*$.

For these reasons, which may not be absolutely and permanently valid ones, we shall restrict the direct application of the elimination method, or rather of its pseudo-operational equivalent, to definite matrices A.

[28] Cf. the corresponding results in §5.4, where the efforts in this direction prove successful

[29] Such correlations would probably also have to depend on the positioning for size, in the sense of §4.2. Cf. also the discussion following (5.7).

In addition various important categories of matrices are per se definite, for example, all correlation matrices.

The considerations of this chapter will still take place in terms of true, and not of pseudo-operations. They do, however, set the pattern for the subsequent pseudo-operatorial discussion of Chapter VI.

5.2. **Properties of our algorithm (that is, of the elimination method) for a symmetric matrix A. Need to consider positioning for size as well.** We shall show that if A is (nonsingular and) definite, then all matrices

$$(5.1)\qquad A^{(k)} = (a_{ij}^{(k)}) \qquad (k = 1, \cdots, n;\ i, j = k, \cdots, n)$$

are also definite. Let us, however, consider first the connection between A and the $A^{(k)}$ with respect to the weaker property of symmetry.

We continue without positioning for size for a short while yet.

It is clear from (4.6) that $a_{ij}^{(k)} = a_{ji}^{(k)}$ (for all $i, j = k, \cdots, n$) implies $a_{ij}^{(k+1)} = a_{ji}^{(k+1)}$ (for all $i, j = k+1, \cdots, n$), that is, that the symmetry of $A^{(k)}$ implies that of $A^{(k+1)}$ $(k = 1, \cdots, n-1)$. If we begin with $A = A^{(1)}$, then we have:

(5.2) If A is symmetric, then all $A^{(k)}$ $(k = 1, \cdots, n)$ are.

From this we can infer:

(5.3) The symmetry of A is equivalent to having $C = B^*$ in (4.25), that is, to (4.25) assuming the form $A = B^*DB$.

Indeed: If A is symmetric, then by (5.2) always $a_{ij}^{(k)} = a_{ji}^{(k)}$, hence, by (4.23), (4.17), $c_{ij} = b_{ji}$, that is, $C = B^*$. Conversely, $C = B^*$, that is, $A = B^*DB$ implies $A^* = B^*D^*B^{**} = B^*DB = A$.

Let us now introduce positioning for size. This can disrupt the validity of (5.2), (5.3) above. Indeed: If for any k $(=1, \cdots, n-1)$ $\mathrm{Max}_{i,j=k,\cdots,n} |a_{ij}^{(k)}|$ is assumed for no pair i, j with $i = j$, then the required permutations of $i = k, \cdots, n$ and of $j = k, \cdots, n$ are unavoidably different (cf. §4.2). Hence these permutations will disrupt the symmetry of $A^{(k)}$ inasmuch as it determines $A^{(k+1)}$ by (4.6), and therefore they will a fortiori disrupt the symmetry of $A^{(k+1)}$. Thus (5.2) fails, and consequently (5.3) fails, too.

The behavior of the $a_{ij}^{(k)}$, that is, of $A^{(k)}$, to which we refer, is perfectly possible. Clearly $A = A^{(1)}$ itself may be like this.

This discussion shows that it is unsafe to postpone the consideration of the problems of positioning for size any further. We shall therefore face them from now on, and we assume accordingly that positioning for size does take place in the cases which follow.

5.3. **Properties of our algorithm for a definite matrix A.** Consider now the property of definiteness, that is, the case of a (nonsingular and) definite A. It is best to derive a number of intermediate propositions in succession.

(5.4) Let $M=(m_{ij})$ be a definite matrix. Then we have:
(a) Always $m_{ii} \geqq 0$.
(b) Always $m_{ii}m_{jj} \geqq m_{ij}^2$.
(c) $m_{ii}=0$ implies $m_{ij}=0$ for all j.
(d) If Max m_{ii} is assumed for $i=h$, then $m_{hh} \geqq |m_{ij}|$ for all i, j.

Indeed: The diagonal minors of M are definite along with M, and hence their determinants are greater than or equal to 0. Applying this to the first and second order minors gives (a), (b), respectively. (c), (d) are immediate consequences of (b) with (a).

(5.5) If an $A^{(k)}=(a_{ij}^{(k)})$ is definite, and if $\text{Max}_{i=k,\cdots,n} a_{ii}^{(k)}$ is assumed for $i=h$, then $\text{Max}_{i,j=k,\cdots,n} |a_{ij}^{(k)}|$ is assumed for $i=j=h$.[30] We can therefore choose the same permutation for $i=k, \cdots, n$ and for $j=k, \cdots, n$ when we comply with the requirements of positioning for size. We propose to do this in all cases where it is possible. Hence if $A^{(k)}$ is definite, the positioning for size will not disrupt its symmetry.

The first assertion follows from (5.4.d), all other assertions are immediate consequences of the first one.

(5.6) If $A=A^{(1)}$ is (nonsingular and) definite, then the same is true for $A^{(2)}$, and

$$0 < |A|_l = |A^{(1)}|_l \leqq |A^{(2)}|_l \leqq |A^{(2)}| \leqq |A^{(1)}| = |A|.$$

Because of (3.5) the nonsingularity of A implies the assertion $0<|A|_l$ and conversely the assertion $0<|A^{(2)}|_l$ implies the equally asserted nonsingularity of $A^{(2)}$. Of the remaining relations (equalities and inequalities) only

(a) $$|A^{(2)}| \leqq |A^{(1)}|$$

and

(b) $$|A^{(2)}|_l \geqq |A^{(1)}|_l$$

[30] Note that we do not claim that $\text{Max}_{i=k,\cdots,n} a_{ii}^{(k)}$ need be assumed for one $i=h$ only, nor that $\text{Max}_{i,j=k,\cdots,n} |a_{ij}^{(k)}|$ may not also be assumed for pairs i, j with $i \neq j$.

require proof.

Use for the moment the definiteness of $A^{(1)}$, $A^{(2)}$. Then by (3.22.a) and (3.22.b) $|A^{(h)}| \leqq \lambda'$ or $|A^{(h)}|_l \geqq \mu'$ ($h=1$, 2) is equivalent to having all proper values of $A^{(h)} \leqq \lambda'$ or $\geqq \mu'$, respectively; that is, all proper values of $\lambda' \cdot I - A^{(h)}$ or $A^{(h)} - \mu' \cdot I$, respectively, $\geqq 0$; that is, by (3.21.a) to the definiteness of $\lambda' \cdot I - A^{(h)}$ or $A^{(h)} - \mu' \cdot I$, respectively. These, in turn, may be written $(A^{(h)}\xi, \xi) \leqq \lambda' |\xi|^2$ and $(A^{(h)}\xi, \xi) \geqq \mu' |\xi|^2$, respectively. So we see:

(a') $\quad |A^{(h)}| \leqq \lambda'$ is equivalent to $(A\xi, \xi) \leqq \lambda' |\xi|^2 \quad$ for all ξ,

(b') $\quad |A^{(h)}|_l \geqq \mu'$ is equivalent to $(A\xi, \xi) \geqq \mu' |\xi|^2 \quad$ for all ξ.

Put $\lambda' = |A^{(1)}|$, $\mu' = |A^{(1)}|_l$. Then (a'), (b') with $h=1$ show that

(c) $\quad \lambda' |\xi|^2 \geqq (A^{(1)}\xi, \xi) \geqq \mu' |\xi|^2 \quad$ for all ξ,

and (a'), (b') with $h=2$ show that (a), (b) are equivalent to

(d) $\quad \lambda' |\xi|^2 \geqq (A^{(2)}\xi, \xi) \geqq \mu' |\xi|^2 \quad$ for all ξ.

Note that the ξ of (c) are n-dimensional vectors: $\xi = (x_1, \cdots, x_n)$, while the ξ of (d) are $n-1$-dimensional vectors: $\xi = (x_2, \cdots, x_n)$.

Since $A^{(1)} = A$ is definite, it follows that we need to prove only two things: The definiteness of $A^{(2)}$ and (d).

$A^{(1)} = A$ is definite, hence symmetric, hence by (5.5) the positioning for size subjects i and j to the same permutation. Consequently the form of (4.6) is unaffected, and $A^{(1)}$ as well as $A^{(2)}$ remain symmetric. Therefore the definiteness of $A^{(2)}$ is secure if $(A^{(2)}\xi, \xi) \geqq 0$. This, however, follows from (d).

Thus we need to prove (d) only.

Put, with $\xi = (x_2, \cdots, x_n)$,

$$x_i' = \begin{cases} 0 \text{ for } i = 1, \\ x_i \text{ for } i = 2, \cdots, n, \end{cases} \qquad \xi' = (x_1', x_2', \cdots, x_n'),$$

$$x_i'' = \begin{cases} -\sum_{j=2}^{n} \dfrac{a_{1j}^{(1)}}{a_{11}^{(1)}} x_j \text{ for } i = 1, \\ x_i \text{ for } i = 2, \cdots, n, \end{cases} \qquad \xi'' = (x_1'', x_2'', \cdots, x_n'').$$

A simple calculation based on (4.6) gives

$$\sum_{i,j=2}^{n} a_{ij}^{(2)} x_i x_j = \begin{cases} \sum_{i,j=1}^{n} a_{ij}^{(1)} x_i' x_j' - a_{11}^{(1)} (x_1'')^2 \leqq \sum_{i,j=1}^{n} a_{ij}^{(1)} x_i' x_j', \\ \sum_{i,j=2}^{n} a_{ij}^{(1)} x_i'' x_j'' \end{cases}$$

that is,

$$(A^{(2)}\xi, \xi)\begin{cases} \leqq (A^{(1)}\xi', \xi'), \\ = (A^{(1)}\xi'', \xi''). \end{cases}$$

Using (c), the first relation gives

$$(A^{(2)}\xi, \xi) \leqq (A^{(1)}\xi', \xi') \leqq \lambda' \,|\, \xi' \,|^2 = \lambda' \,|\, \xi \,|^2,$$

and the second relation gives

$$(A^{(2)}\xi, \xi) = (A^{(1)}\xi'', \xi'') \geqq \mu' \,|\, \xi'' \,|^2 \geqq \mu' \,|\, \xi \,|^2,$$

establishing together (d), as desired.

(5.6′) If A is (nonsingular and) definite then the same is true for all $A^{(k)}(k=1, \cdots, n)$, and

$$0 < |\, A \,|_l = |\, A^{(1)} \,|_l \leqq |\, A^{(2)} \,|_l \leqq \cdots \leqq |\, A^{(n)} \,|_l \leqq |\, A^{(n)} \,| \leqq \cdots$$
$$\leqq |\, A^{(2)} \,| \leqq |\, A^{(1)} \,| = |\, A \,|.$$

This is immediate, by applying (5.6) to $A = A^{(1)}$ and to $A^{(2)}, \cdots, A^{(n-1)}$ in succession, in place of A.

We interrupt at this point our chain of deductions, in order to make a subsidiary observation.

(5.7) If A is (nonsingular and) definite, then all $a_{ii}^{(k)} > 0$.

Assume that some $a_{ii}^{(k)}$ is not greater than 0. $A^{(k)}$ is definite, hence by (5.4.a) this $a_{ii}^{(k)}$ is 0, and by (5.4.c) $a_{ij}^{(k)} = 0$ for this (k and) i and for all $j = k, \cdots, n$. Hence $A^{(k)}$ is singular, in contradiction with (5.6′).

(5.7) shows that for a (nonsingular and) definite A the elimination method could have been carried out without positioning for size in the sense of 4.2, since all $a_{ii}^{(k)} > 0$ automatically, that is, just in that case where positioning for size creates no difficulties (cf. the remarks at the end of 5.2), it seems to be superfluous.

This, however, is not the complete truth. A (nonsingular and) definite A could indeed be put through the algorithm of 4.1 in the rigorous sense, without positioning for size. However, if pseudo-operations are used, no satisfactory estimates seem to be obtainable, unless positioning for size is also effected. This will become apparent in several instances in Chapter VI, primarily inasmuch as the estimate (6.8), which is identical with (6.23.d′), depends directly on the positioning for size, and this (6.8) is the basis for the decisive estimates (6.12), (6.25). This is our true reason for insisting on positioning for size in the situation that we are going to discuss.

We return now to the main line of our deductions.

(5.8) If A is (nonsingular and) definite and all its elements lie in $-1, 1$, then all $A^{(k)}$ $(k=1, \cdots, n)$ are also definite and all their elements also lie in $-1, 1$.

The matter of definiteness was settled in (5.6′). By (5.4) all $a_{ii}^{(k)} \geqq 0$, by (4.6) $a_{ii}^{(k+1)} = a_{ii}^{(k)} - (a_{ik}^{(k)})^2 / a_{kk}^{(k)} \leqq a_{ii}^{(k)}$, hence $a_{ii}^{(k)} \leqq a_{ii}^{(k-1)} \leqq \cdots \leqq a_{ii}^{(1)} = a_{ii} \leqq 1$. Thus $0 \leqq a_{ii}^{(k)} \leqq 1$. Now (5.4.b) gives $|a_{ij}^{(k)}| \leqq 1$, that is, all $a_{ij}^{(k)}$ lie in $-1, 1$ as desired.

(5.9) Under the same assumptions as in (5.8) we have further:

(a) For all elements d_i of D, $0 < d_i \leqq 1$.

(b) All elements of B lie in $-1, 1$.

Proof of (a): By (4.22) $d_i = a_{ii}^{(i)}$, hence by (5.8) d_i lies in $-1, 1$, and by (4.14) it is not equal to 0. Furthermore, d_i is greater than or equal to 0 by (5.4.a). All these give together $0 < d_i \leqq 1$, as desired.

Proof of (b): Since the positioning for size has taken place, we have $|a_{kk}^{(k)}| \geqq |a_{ij}^{(k)}|$ (for all $i, j = k, \cdots, n$). Hence (4.23) guarantees $|b_{ij}| \leqq 1$ (for all $i, j = k, \cdots, n$), that is all b_{ij} lie in $-1, 1$, as desired.

5.4. **Detailed matrix bound estimates, based on the results of the preceding section.** For the balance of this chapter we assume A to be (nonsingular and) definite, and positioning for size to have taken place in the sense of (5.5). This implies (5.6′), (5.7), and hence (5.3), too. We may therefore restate (4.27) (together with (4.25)) as follows:

(5.10) B, D fulfill

$$A = B^* D B.$$

They belong to $\mathcal{C}_+$, $\mathcal{C}_0$, respectively. All diagonal elements of B are identically 1.

We now proceed to derive estimates for $|B|$, $|B|_l$, $|D|$, $|D|_l$, and some other, allied quantities, in terms of $|A|$, $|A|_l$, in the sense of §5.1.

Let $\lambda_1, \cdots, \lambda_n$ be the proper values of A, ordered in a monotone non-increasing sequence, cf. (3.21.b). We recall (3.22.a), (3.22.b) and define λ, μ:

(5.11.a) $$|A| = \lambda = \lambda_1,$$

(5.11.b) $$|A|_l = \mu = \lambda_n.$$

From (4.22)

$$D^{v} = (d_i^{v}\delta_{ij}) \qquad (i, j = 1, \cdots, n) \tag{5.12}$$

for any exponent v; we shall use this for $v = \pm 1$, $\pm 1/2$.

Now $(D^{1/2}B)^* \cdot D^{1/2}B = B^*D^{1/2} \cdot D^{1/2}B = B^*DB = A$. Hence (3.12.a), (3.12.b) permit us to infer from (5.11.a), (5.11.b)

$$|D^{1/2}B| = \lambda^{1/2}, \tag{5.13.a}$$

$$|D^{1/2}B|_l = \mu^{1/2}. \tag{5.13.b}$$

$D^{1/2}B$ is in $\mathcal{C}_+$ and its diagonal elements are those of $D^{1/2}$, the $d_i^{1/2}$. Consequently $(D^{1/2}B)^{-1}$ is also in $\mathcal{C}_+$ and its diagonal elements are the $d_i^{-1/2}$. (Cf. (3.26).) Hence by (3.17.b) (first half) $|d_i^{1/2}| \leqq |D^{1/2}B| = \lambda^{1/2}$, $|d_i^{-1/2}| \leqq |(D^{1/2}B)^{-1}| = |D^{1/2}B|_l^{-1} = \mu^{-1/2}$. By (5.9.a), $0 < d_i \leqq 1$. Hence

$$\mu \leqq d_i \leqq \lambda \text{ and } 1. \tag{5.14}$$

Now (5.12) and (3.27) give

$$|D^v| \left\{ \begin{array}{lll} \leqq \lambda^v \text{ and } 1 & \text{for} & v \geqq 0 \\ \geqq \mu^v & \text{for} & v \leqq 0 \end{array} \right|, \tag{5.15.a}$$

$$|D^v|_l \left\{ \begin{array}{lll} \leqq \mu^v & \text{for} & v \geqq 0 \\ \geqq \lambda^v \text{ and } 1 & \text{for} & v \leqq 0 \end{array} \right|. \tag{5.15.b}$$

Combining (5.13.a), (5.13.b) with (5.15.a), (5.15.b), $v = \pm 1/2$, gives

$$|DB| \leqq \lambda \text{ and } \lambda^{1/2}, \tag{5.16.a}$$

$$|DB|_l \geqq \mu^{1/2}, \tag{5.16.b}$$

$$|B| \leqq (\lambda/\mu)^{1/2}, \tag{5.17.a}$$

$$|B|_l \geqq (\mu/\lambda)^{1/2} \text{ and } \mu^{1/2}. \tag{5.17.b}$$

The estimates (5.13.a)–(5.13.b) justify this conclusion: The primary estimates, on which all others are based, are those concerning $D^{1/2}B$, that is, (5.13.a), (5.13.b). These are consequently the sharpest ones, as can also be inferred from the fact that they alone are equalities, all others being inequalities. Hence $D^{1/2}B$ is the truly fundamental quantity in preference to B, DB, and even to D.

Now the method of inversion discussed in §4.3 is based on $B' = DB$ and on C, which is now equal to B^*. In §4.4 (specifically: (4.25), (4.26)) we used B, C (which is now equal to B^*) and D. It follows from the above that, if we use these matrices, the methods of estimating should nevertheless emphasize $D^{1/2}B$. It will become apparent in several places throughout §6.6 and in parts of §6.8 how we endeavor to follow this principle.

CHAPTER VI. THE PSEUDO-OPERATIONAL PROCEDURE

6.1. Choice of the appropriate pseudo-procedures, by which the true elimination will be imitated. After the preparations in the foregoing chapters we can now attack our main problem: The pseudo-operational matrix inversion by means of the elimination method, the latter being reinterpreted, modified and specialized in the sense of chapters IV, V.

We consider accordingly a digital matrix $\bar{A}=(\bar{a}_{ij})$ $(i,j=1,\cdots,n)$ (cf. §3.5), of which we assume that it is nonsingular and definite.

We have to begin by performing the pseudo-operational equivalent of the manipulations of §4.1 on $\bar{A}$. This means that we define a sequence of digital matrices $\bar{A}^{(k)}=(\bar{a}_{ij}^{(k)})$ $(i,\ j=k,\cdots,n)$ for $k=1,\cdots,n$. The induction begins for $k=1$ with $\bar{A}^{(1)}=\bar{A}$, that is,

$$(6.1)\qquad \bar{a}_{ij}^{(1)}=\bar{a}_{ij} \qquad (i,j=1,\cdots,n),$$

following (4.8). The inductive step from k to $k+1$ $(k=1,\cdots,n-1)$ has to follow (4.6). This creates a new problem: How are the true operations of the expression $a_{ik}^{(k)}a_{kj}^{(k)}/a_{kk}^{(k)}$ in (4.6) to be replaced by pseudo-operations?

There are obviously several ways of doing this. The simplest ones, which are most economical in the number of operations to be performed, are these:

$$(6.2.a)\qquad (\bar{a}_{ik}^{(k)}\times\bar{a}_{kj}^{(k)})\div\bar{a}_{kk}^{(k)},$$

$$(6.2.b)\qquad \bar{a}_{ik}^{(k)}\times(\bar{a}_{kj}^{(k)}\div\bar{a}_{kk}^{(k)}),$$

$$(6.2.c)\qquad (\bar{a}_{ik}^{(k)}\div\bar{a}_{kk}^{(k)})\times\bar{a}_{kj}^{(k)}.$$

The build of (6.2.b) and (6.2.c) is so similar that it suffices to discuss (6.2.a) and (6.2.b). Comparing these with $\bar{a}_{ik}^{(k)}\bar{a}_{kj}^{(k)}/\bar{a}_{kk}^{(k)}$ which they are supposed to approximate, we obtain

$$(6.3.a)\qquad \begin{aligned}&\left|\,\bar{a}_{ik}^{(k)}\bar{a}_{kj}^{(k)}/\bar{a}_{kk}^{(k)}-(\bar{a}_{ik}^{(k)}\times\bar{a}_{jk}^{(k)})\div\bar{a}_{kk}^{(k)}\right|\\ &\qquad\qquad\leqq(1+|\,\bar{a}_{kk}^{(k)}|^{-1})\beta^{-s}/2,\end{aligned}$$

$$(6.3.b)\qquad \begin{aligned}&\left|\,\bar{a}_{ik}^{(k)}\bar{a}_{kj}^{(k)}/\bar{a}_{kk}^{(k)}-\bar{a}_{ik}^{(k)}\times(\bar{a}_{kj}^{(k)}\div\bar{a}_{kk}^{(k)})\right|\\ &\qquad\qquad\leqq(1+|\,\bar{a}_{kk}^{(k)}|)\beta^{-s}/2.\end{aligned}$$

(Cf. these estimates with the very similar ones which were derived and discussed at the end of §2.4: (2.14), (2.15).) Clearly the estimate (6.3.a) is considerably less favorable than (6.3.b) (by a factor

$|\bar{a}_{kk}^{(k)}|^{-1}$) and altogether unsatisfactory per se: It involves the terms $|\bar{a}_{kk}^{(k)}|^{-1}$ which may be arbitrarily and unpredictably large. We reject, therefore, (6.2.a). (6.2.b), on the other hand, has this flaw: If, in anticipation of the symmetry of $\bar{a}_{ij}^{(k)}$, we write it in the form

$$\text{(6.2.b}'\text{)} \qquad \bar{a}_{ki}^{(k)} \times (\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}),$$

then the pseudo-character of its operations prevents it from being symmetric in i, j. We overcome this difficulty by using (6.2.b′) only when $i \leqq j$, and interchanging i, j when $i > j$. That is, we use this expression:

$$\text{(6.2.b}''\text{)} \qquad \begin{aligned} &\bar{a}_{ki'}^{(k)} \times (\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)}), \\ &\qquad \text{where } i' = \text{Min } (i, j), \ j' = \text{Max } (i, j). \end{aligned}$$

We formulate therefore the inductive step by replacing $\bar{a}_{ik}^{(k)}\bar{a}_{kj}^{(k)}/\bar{a}_{kk}^{(k)}$ in (4.6) by (6.2.b″). The inductive step consequently assumes this form:

$$\text{(6.3)} \qquad \begin{aligned} &\bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - \bar{a}_{ki'}^{(k)} \times (\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)}), \\ &\text{where } i' = \text{Min } (i, j), \ j' = \text{Max } (i, j) \quad (i, j = k+1, \cdots, n) \end{aligned}$$

for $k = 1, \cdots, n-1$.

We also assume that positioning for size takes place, but we anticipate the equivalent of (5.5) by permitting only one joint permutation for i and j. We define accordingly:

(6.4) Before (6.3) is effected, it must be made certain by an appropriate joint permutation of $i, j = k, \cdots, n$ that $\text{Max}_{i=k,\cdots,n}\bar{a}_{ii}^{(k)}$ is assumed for $i = k$.

Thus (6.1) and (6.3), (6.4) define the $\bar{a}_{ij}^{(k)}$ $(k = 1, \cdots, n; \ i, j = k, \cdots, n)$. We also define the digital matrices

$$\text{(6.5)} \qquad \bar{A}^{(k)} = (\bar{a}_{ij}^{(k)}) \qquad (k = 1, \cdots, n; \ i, j = k, \cdots, n).$$

It remains to show that all these definitions are indeed possible, that is, that all numbers produced by (6.1), (6.3), (6.4) (including the intermediate expressions $\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)}$ and $\bar{a}_{ki'}^{(k)} \times (\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)})$ in (6.3)), and designated as digital numbers, are properly formed and, in particular, lie in $-1, 1$. We shall prove this in 6.2 below. The inductive proof which establishes this will also secure the equivalents of (5.5)–(5.7).

6.2. **Properties of the pseudo-algorithm.** Assume that for a $k = 1, \cdots, n$ this is true:

(6.6) (a) It is possible to form $\bar{A}^{(1)}, \cdots, \bar{A}^{(k)}$, obtaining properly formed, digital numbers throughout.

(b) $\bar{A}^{(k)}$ is (nonsingular and) definite.

(c) All $\bar{a}_{ii}^{(k)} \leqq 1$.

By (5.4.a) we have $\bar{a}_{ii}^{(k)} \geqq 0$. From $\bar{a}_{ii}^{(k)} = 0$ we could infer, as in the proof of (5.7), that $\bar{A}^{(k)}$ is singular. Hence

$$0 < \bar{a}_{ii}^{(k)} \leqq 1. \tag{6.7}$$

Next by (5.4.d):

$$| \bar{a}_{ij}^{(k)} | \leqq \bar{a}_{kk}^{(k)}. \tag{6.8}$$

(6.7), (6.8) give together:

$$\text{All } \bar{a}_{ij}^{(k)} \text{ lie in } -1, 1. \tag{6.9}$$

Now assume $k \neq n$, so that the validity of (6.6) for $k+1$ can be considered. (6.7)–(6.9) show that $\bar{A}^{(k+1)}$ can be formed, obtaining properly formed digital numbers throughout, with the one limitation, that it remains to be proved that the elements $\bar{a}_{ij}^{(k+1)}$ lie in -1, 1. Assume, furthermore, that

$$|\bar{A}^{(k)}|_l > (n-k)\beta^{-s}. \tag{6.10}$$

Since $\bar{A}^{(k)}$ is definite, it is also symmetric. (6.4) does not affect the form of (6.3), and $\bar{A}^{(k)}$ as well as $\bar{A}^{(k+1)}$ remain symmetric.

We may rewrite (6.3) as

$$\bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - \bar{a}_{ki'}^{(k)}\bar{a}_{kj'}^{(k)}/\bar{a}_{kk}^{(k)} + \eta_{ij}^{(k)}$$
$$(i, j = k+1, \cdots, n;\ i' = \text{Min }(i, j),\ j' = \text{Max }(i,j)),$$

that is, as

$$\bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - \bar{a}_{ik}^{(k)}\bar{a}_{kj}^{(k)}/\bar{a}_{kk}^{(k)} + \eta_{ij}^{(k)} \qquad (i, j = k+1, \cdots, n), \tag{6.11}$$

where by (6.8), (6.9)

$$| \eta_{ij}^{(k)} | \leqq \beta^{-s}. \tag{6.12}$$

Put

$$\tilde{A}^{(k+1)} = (\bar{a}_{ij}^{(k)} - \bar{a}_{ik}^{(k)}\bar{a}_{kj}^{(k)}/\bar{a}_{kk}^{(k)}) \qquad (i, j = k+1, \cdots, n), \tag{6.11.a}$$

$$H^{(k+1)} = (\eta_{ij}^{(k)}) \qquad (i, j = k+1, \cdots n), \tag{6.11.b}$$

then (6.11) becomes

(6.11′) $$\bar{A}^{(k+1)} = \tilde{A}^{(k+1)} + H^{(k+1)}.$$

Now we may replace $A = A^{(1)}$, $A^{(2)}$ in (5.6) by $\bar{A}^{(k)}$, $\tilde{A}^{(k+1)}$; this gives

$$| \tilde{A}^{(k+1)} |_l \geqq | \bar{A}^{(k)} |_l,$$

and hence, by (d) in the proof of (5.6),

(6.13) $$(\tilde{A}^{(k+1)}\xi, \xi) \geqq \mu^* | \xi |^2 \qquad \text{where } \mu^* = | \bar{A}^{(k)} |_l.$$

Next by (3.17.b) and (6.12) (note that we have here $n-k$ in place of n)

$$| H^{(k+1)} | \leqq (n - k)\beta^{-s},$$

and so

(6.14) $$| (H^{(k+1)}\xi, \xi) | \leqq (n - k)\beta^{-s} | \xi |^2.$$

(6.11′) together with (6.13) and (6.14) gives

(6.15) $$(\bar{A}^{(k+1)}\xi, \xi) \geqq (| \bar{A}^{(k)} |_l - (n - k)\beta^{-s}) | \xi |^2.$$

By (6.10) this implies that

(6.16) $$\bar{A}^{(k+1)} \text{ is definite.}$$

We can now argue, exactly as in the derivation of (b′) in the proof of (5.6), that (6.15) is equivalent to

(6.17) $$| \bar{A}^{(k+1)} |_l \geqq | \bar{A}^{(k)} |_l - (n - k)\beta^{-s}.$$

Now (6.10) secures $| \bar{A}^{(k+1)} |_l > 0$, that is, by (3.5)

(6.18) $$\bar{A}^{(k+1)} \text{ is nonsingular.}$$

For $i = j$ we have $i' = j' = i = j$, hence the two factors of the second term of the right-hand side in (6.3) have the same sign. Hence that subtrahend is greater than or equal to 0. (If two digital numbers have the same sign, then no round off rule will impair the nonnegativity of their pseudo-product and pseudo-quotient. Cf. footnote 6.) Consequently

(6.19) $$\bar{a}_{ii}^{(k+1)} \leqq \bar{a}_{ii}^{(k)}.$$

Now we have established all parts of (6.6) for $k+1$: (a) follows from the remark immediately preceding (6.10) with the one limitation that it remains to be proved that the elements $\bar{a}_{ij}^{(k+1)}$ lie in $-1, 1$; (b) is contained in (6.18), (6.16); (c) is contained in (6.7), (6.19). On the basis of these we can now infer (6.7)–(6.9) for $k+1$, too, and hence we have the last part of (a): All $\bar{a}_{ij}^{(k+1)}$ lie in $-1, 1$. So we see:

(6.20) (6.6) for k and (6.10) imply (6.6) for $k+1$ and (6.17).

Consider now the condition

$$(6.21) \qquad |\bar{A}|_l > (n(n-1)/2)\beta^{-s}.$$

Then applying (6.20), (6.17) for all $k=1, \cdots, n-1$ in succession gives:

(6.22) If (6.6) holds for $k=1$, that is, for $A=A^{(1)}$, and if (6.21) holds, then (6.6) holds for all $k=1, \cdots, n$.

We restate this more explicitly, together with the inferences (6.7)–(6.9) from (6.6):

(6.23) Assume the following:
(a) $\bar{A}$ is (nonsingular and) definite.
(b) All $\bar{a}_{ii} \leqq 1$.
(c) $|\bar{A}|_l > (n(n-1)/2)\beta^{-s}$.
Then we have:
(a′) It is possible to form all $\bar{A}^{(k)}$, obtaining only properly formed, digital numbers throughout.
(b′) All $\bar{A}^{(k)}$ are (nonsingular and) definite.
(c′) Always $0 < \bar{a}_{ii}^{(k)} \leqq 1$.
(d′) Always $|\bar{a}_{ij}^{(k)}| \leqq \bar{a}_{kk}^{(k)}$.
(e′) All $\bar{a}_{ij}^{(k)}$ lie in $-1, 1$.

Note that the assumptions (a), (b) above are reasonable in view of the nature of our problem. (c) will be eliminated (or rather absorbed by a stronger condition) after (6.67).

With (6.23) the question at the end of §6.1 is completely answered: The processes (6.1), (6.3), (6.4) (including those of the intermediate expressions $\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)}$ and $\bar{a}_{ki'}^{(k)} \times (\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)})$ in (6.3)) produce indeed digital numbers, which are properly formed, and, in particular, lie in $-1, 1$. Furthermore, we have the certainty, as we had it in the corresponding situation in (5.5)–(5.6), that all the $\bar{A}^{(k)} (k=1, \cdots, n)$ that we produce are (nonsingular and) definite.

6.3. **The approximate decomposition of $\bar{A}$, based on the pseudo-algorithm.** We now proceed to derive the approximate equivalent of (5.10) (that is, of (4.25)).

Rewrite (6.3) as

$$\bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - (\bar{a}_{ki'}^{(k)} \div \bar{a}_{kk}^{(k)})\bar{a}_{kk}^{(k)}(\bar{a}_{kj'}^{(k)} \div \bar{a}_{kk}^{(k)}) + \theta_{ij}^{(k)}$$
$$(i, j = k+1, \cdots, n;\ i' = \mathrm{Min}\,(i, j),\ j' = \mathrm{Max}\,(i, j)),$$

that is, as

$$\begin{aligned}\bar{a}_{ij}^{(k+1)} &= \bar{a}_{ij}^{(k)} - (\bar{a}_{ki}^{(k)} \div \bar{a}_{kk}^{(k)})\bar{a}_{kk}^{(k)}(\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) + \theta_{ij}^{(k)} \\ &\qquad (k = 1, \cdots, i' - 1;\ i' = \mathrm{Min}\ (i, j)),\end{aligned} \tag{6.24}$$

where by (6.23.d′), (6.23.e′)

$$|\theta_{ij}^{(k)}| \leqq \beta^{-s}. \tag{6.25}$$

(It should be noted that (6.24), (6.25) are analogous to (6.11), (6.12), but not the same.)

We observe next that

$$0 = \bar{a}_{ij}^{(k)} - (\bar{a}_{ki}^{(k)} \div \bar{a}_{kk}^{(k)})\bar{a}_{kk}^{(k)}(\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) + \theta_{ij}^{(k)} \tag{6.26}$$

for $k=i$ and for $k=j$, that is, for $k=i'$, with

$$|\theta_{ij}^{(k)}| \leqq \beta^{-s}/2. \tag{6.27}$$

Summing all these equations (that is, (6.24) for $k=1, \cdots, i'-1$ and (6.26) for $k=i'$), and remembering (6.1), gives

$$\bar{a}_{ij} = \sum_{k=1}^{i'} (\bar{a}_{ki}^{(k)} \div \bar{a}_{kk}^{(k)})\bar{a}_{kk}^{(k)}(\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) + \zeta_{ij} \tag{6.28}$$

where $\zeta_{ij} = \sum_{k=1}^{i'} \theta_{ij}^{(k)}$, hence by (6.25), (6.27)

$$|\zeta_{ij}| \leqq (i' - 1/2)\beta^{-s}. \tag{6.29}$$

The relation (6.28) is the analog of (4.31), therefore we now wish to perform the analog of the transition from (4.31) to (4.32). For this purpose we define first, in analogy with (4.22), (4.23)

$$\begin{aligned} \bar{D} &= (\bar{d}_i\delta_{ij}) \qquad (i, j = 1, \cdots, n) \\ \text{with}\quad \bar{d}_i &= \bar{a}_{ii}^{(i)} \end{aligned} \tag{6.30}$$

$$\begin{aligned} \bar{B} &= (\bar{b}_{ij}) \qquad (i, j = 1, \cdots, n) \\ \text{with}\quad \bar{b}_{ij} &= \begin{cases} \bar{a}_{ij}^{(i)} \div \bar{a}_{ii}^{(i)} & \text{for } i \leqq j, \\ \left[\begin{matrix}\text{hence} & \\ 1 & \text{for } i = j,\end{matrix}\right] & \\ 0 & \text{for } i > j. \end{cases} \end{aligned} \tag{6.31}$$

Now

$$(6.32)\qquad \begin{aligned}&\sum_{k=1}^{i'} (\bar{a}_{ki}^{(k)} \div \bar{a}_{kk}^{(k)})\bar{a}_{kk}^{(k)}(\bar{a}_{kj}^{(k)} \div \bar{a}_{kk}^{(k)}) \\ &\qquad\qquad = \sum_{k=1}^{i'} \bar{b}_{ki}\bar{d}_k\bar{b}_{kj} = \sum_{k=1}^{n} \bar{b}_{ki}\bar{d}_k\bar{b}_{kj}.\end{aligned}$$

We may, therefore, write for (6.28)

$$(6.33)\qquad \bar{A} = \bar{B}^*\bar{D}\bar{B} + Z \qquad\qquad \text{where } Z = (\zeta_{ij}).$$

Next by (6.29)

$$\begin{aligned}(N(Z))^2 &= \sum_{i,j=1}^{n} (\zeta_{ij})^2 \leqq \sum_{i,j=1}^{n} \left(i' - \frac{1}{2}\right)^2 \beta^{-2s} \\ &= \sum_{i'=1}^{n} (2n - 2i' + 1)\left(i' - \frac{1}{2}\right)^2 \beta^{-2s} = \frac{n^2(2n^2+1)}{12}\beta^{-2s},\end{aligned}$$

hence by (3.17.a)

$$|Z| \leqq N(Z) \leqq (n^2(2n^2+1)/12)^{1/2}\beta^{-s}.$$

For $n \to \infty$, $(n^2(2n^2+1)/12)^{1/2} \sim n^2/6^{1/2} = .4084n^2$, and already, for $n \geqq 3$, $(n^2(2n^2+1)/12)^{1/2} \leqq .42n^2$. Hence (we assume from now on that $n \geqq 3$)

$$(6.34)\qquad |\bar{A} - \bar{B}^*\bar{D}\bar{B}| \leqq .42n^2\beta^{-s}.$$

6.4. **The inverting of $\bar{B}$ and the necessary scale factors.** (6.34) is the approximate analog of (5.10). It is, therefore, the point of departure for inverting $\bar{A}$ in the sense of (5.10), that is, of (4.27), that is, of the formulae (4.25), (4.26). We proceed, therefore, in this direction.

In other words: $\bar{A}$ is approximated by $\bar{B}^*\bar{D}\bar{B}$, therefore $\bar{A}^{-1}$ will be approximated by $\bar{B}^{-1}\bar{D}^{-1}(\bar{B}^{-1})^*$, or, rather, by some approximant of this expression. In any case we need $\bar{B}^{-1}$ and $\bar{D}^{-1}$, or approximants of these. Hence, we must analogise the formulae (4.29) and (4.28), which gave B^{-1} and D^{-1} respectively.

We begin by considering $\bar{B}^{-1}$, that is, by analogising B^{-1} and (4.29). The obvious way of analogising (4.29) in terms of pseudo-operations would seem to be to define

$$\bar{X} = (\bar{x}_{ij}) \qquad\qquad (i, j = 1, \cdots, n).$$

$$(6.35')\qquad \text{with}\quad \bar{x}_{ij} = \begin{cases} -\sum_{k=i+1}^{j} \bar{b}_{ik} \times \bar{x}_{kj} & \text{for } i < j, \\ 1 & \text{for } i = j, \\ 0 & \text{for } i > j. \end{cases}$$

This, however, is unfeasible; the use of the notation $\bar{x}_{ij}$ for the quantities obtained from (6.35′), which implies that they are digital numbers, is improper: These quantities will not lie in general in $-1, 1$, and in order to obtain an algorithm in terms of digital numbers it is now necessary to introduce scale factors, in the sense of (b) in §2.1 and of §2.5. In order to effect this efficiently we note the following facts:

First, the $\bar{x}_{ij}$ with $i>j$ present no difficulty at all. They are all zero, and they do not enter into the definitions of the others. Second, as far as the other $\bar{x}_{ij}$, that is, those with $i \leqq j$, are concerned, the definition interrelates only $\bar{x}_{ij}$'s with the same j. Third, for these interrelated families of $\bar{x}_{ij}$'s, that is, the $\bar{x}_{ij}$, $i=1, \cdots, j$, with a fixed $j\ (=1, \cdots, n)$, the definition is inductive in i, but it proceeds in the direction of decreasing i: The $\bar{x}_{ij}$'s are obtained in this order: $i=j, j-1, \cdots, 1$.

The scale factors must, therefore, be introduced separately for every j, and must there be built up successively as i goes through the values $j, j-1, \cdots, 1$ (in this order).

Since the sequence $\bar{x}_{ij}$ (j fixed) begins with $\bar{x}_{jj}=1$ the scale factors must only be used to reduce the size of $\bar{x}_{ij}$. Hence the considerations of §2.5 apply: The scaling of $\bar{x}_{ij}$ will be effected by (pseudo-) division by an appropriate 2^p ($p=0, 1, 2, \cdots$), say $2^{p_{ij}}$. Denote the quantity which corresponds to $\bar{x}_{ij}$ scaled down by the factor $2^{p_{ij}}$ by $\bar{y}_{ij}$. In the formula (6.35′), case $i<j$, the replacement of $\bar{x}_{ij}$ by $\bar{y}_{ij}$ and of the $\bar{x}_{kj}$ by the $\bar{y}_{kj}$ requires, of course, that all of them be affected with the same scale factor. Hence, the scale factor which applies to the left-hand side, $2^{p_{ij}}$, must be a multiple of those which apply to the terms of the right-hand side, the $2^{p_{kj}}$ ($k=i+1, \cdots, j$). That is, $p_{ij} \geqq p_{kj}$ for $k=i+1, \cdots, j$. Then the kj-term of the right-hand side must be "adjusted to scale" by (pseudo-) division by $2^{p_{ij}-p_{kj}}$. Furthermore, p_{ij} must be chosen large enough to make the resulting $\bar{y}_{ij}$ lie in $-1, 1$. After all $\bar{y}_{ij}$, $i=j, j-1, \cdots, 1$, have been obtained, they must all be "adjusted to scale" with each other, by (pseudo-) division (of $\bar{y}_{ij}$) by $2^{p_{1j}-p_{ij}}$. We call this "readjusted" form of $\bar{y}_{ij}$, $\bar{z}_{ij}$. Thus $\bar{z}_{ij}$ corresponds to $\bar{x}_{ij}$ scaled down by $2^{p_{1j}}$.

The dependence of this scale factor on the column-index is worth noting, it appears to be essential for the obtaining of efficient estimates. (This is connected with the use of the vectors $U^{(j)}$ of (6.44).)

We can now reformulate (6.35′), and state it in its corrected and valid form:

$$
(6.35)\quad
\begin{aligned}
&\text{(a)}\quad \begin{aligned} &\bar y_{jj} = 1,\\ &p_{jj} = 0. \end{aligned}\\
&\text{(b)}\quad \begin{cases} \bar y_{ij} = -\displaystyle\sum_{k=i+1}^{j} (\bar b_{ik} \times \bar y_{kj}) \div 2^{p_{ij}-p_{kj}},\\ p_{ij} \text{ is the smallest integer } \geqq p_{i+1,j}\ (\geqq p_{i+2,j} \geqq \cdots \geqq p_{jj})\\ \text{which makes } \bar y_{ij} \text{ lie in } -1,\ 1\ (i = j-1, \cdots, 1). \end{cases}\\
&\text{(c)}\quad \bar z_{ij} = \begin{cases} \bar y_{ij} \div 2^{p_{1j}-p_{ij}} & \text{for } i \leqq j,\\ 0 & \text{for } i > j. \end{cases}
\end{aligned}
$$

As a result of (6.35) all $\bar y_{ij}$ and all $\bar z_{ij}$ lie in -1, 1. By (6.23) all $\bar b_{ij}$ lie in -1, 1, too. Hence the processes (6.35) including those of the intermediate expressions $\bar b_{ik} \times \bar y_{kj}$ and $(\bar b_{ik} \times \bar y_{kj}) \div 2^{p_{ij}-p_{kj}}$ produce properly formed digital numbers, lying in -1, 1.

6.5. **Estimates connected with the inverse of $\bar B$.** Having defined the digital numbers $\bar z_{ij}$ by (6.35), we now form the (upper semi-diagonal) digital matrix

$$(6.36)\qquad \bar Z = (\bar z_{ij}) \qquad (i, j = 1, \cdots n)$$

and proceed to investigate the properties of $\bar Z$ and of the $\bar z_{ij}$.

We begin by proving:

$$(6.37)\qquad 1/2 \leqq \operatorname*{Max}_{i=1,\cdots,j} |\bar z_{ij}| \leqq 1 \qquad (j = 1, \cdots, n).$$

Choose the largest $i = 1, \cdots, j$ with $p_{ij} = p_{1j}$. If $i = j$, then $p_{1j} = p_{jj} = 0$, hence $\bar y_{jj} = \bar x_{jj} = 1$. If $i < j$, then $p_{i+1,j}$ exists, and is necessarily unequal to p_{1j}, that is, $p_{i+1,j} < p_{1j}$. Hence $p_{ij} > p_{i+1,j}$. The reason for choosing $p_{ij} > p_{i+1,j}$ could only have been that this was necessary to make $|\bar y_{ij}| \leqq 1$. Since p_{ij} must have been the smallest integer not less than $p_{i+1,j}$ which has this effect, therefore $p_{ij} - 1$ (which is also not less than $p_{i+1,j}$) cannot have had this property. This excludes $|\bar y_{ij}| < 1/2$. Hence $|\bar y_{ij}| \geqq 1/2$. Now $\bar z_{ij} = \bar y_{ij} \div 2^{p_{1j}-p_{ij}} = \bar y_{ij}$, hence $|\bar z_{ij}| \geqq 1/2$.

Thus we have in any event a $|\bar z_{ij}| \geqq 1/2$. We know that all $|\bar z_{ij}| \leqq 1$. These two facts together establish (6.37).

Let us now evaluate the elements of the matrix $\bar B \bar Z$ (true multiplication!).

The ij-element of this matrix is $\sum_{k=1}^{n} \bar b_{ik} \bar z_{kj}$. Since $\bar B$ and $\bar Z$ are both upper semi-diagonal, we can conclude: For $i > j$ all terms in this sum are zero, so that the sum itself is zero. For $i = j$ the sum has precisely one nonzero term: $\bar{b}_{ii}\bar{z}_{ii}$. Since $\bar{b}_{ii}=1$, $\bar{z}_{ii}=\bar{y}_{ii}\div 2^{p_{1i}-p_{ii}}=1\div 2^{p_{1i}}$, the sum is equal to $1\div 2^{p_{1i}}$. This deviates from $2^{-p_{1i}}$ by not more than $\beta^{-s}/2$. For $i<j$ the only nonzero terms in the sum are those with $i\leqq k\leqq j$. The $k=i$ term is $\bar{b}_{ii}\bar{z}_{ij}=\bar{z}_{ij}$. Hence in this case the sum is

$$\bar{z}_{ij}+\sum_{k=i+1}^{j}\bar{b}_{ik}\bar{z}_{kj} = -\left(\sum_{k=i+1}^{j}(\bar{b}_{ik}\times\bar{y}_{kj})\div 2^{p_{ij}-p_{kj}}\right)\div 2^{p_{1j}-p_{ij}}$$
$$+\sum_{k=i+1}^{j}\bar{b}_{ik}(\bar{y}_{kj}\div 2^{p_{1j}-p_{kj}}).$$

If we replace all pseudo-operations by true ones, then both terms on the right-hand side go over into the same expression: $\sum_{k=i+1}^{j}\bar{b}_{ik}\bar{y}_{kj}/2^{p_{1j}-p_{kj}}$. By (2.21), (2.22) the error caused by these transitions is not more than $(j-i)\beta^{-s}$ in either term. Hence the right-hand side, which is the difference of these two terms, deviates from zero by not more than $2(j-i)\beta^{-s}$.

To sum up: Put

$$\overline{B}\overline{Z}=\Delta+U,\qquad \Delta=(2^{-p_{ij}\delta_{ij}}),\qquad U=(u_{ij}), \tag{6.38}$$

then

$$|u_{ij}|\begin{cases}=0 & \text{for } i>j,\\ \leqq\beta^{-s}/2 & \text{for } i=j,\\ \leqq 2(j-i)\beta^{-s} & \text{for } i<j.\end{cases} \tag{6.39}$$

We are interested in the vectors $U^{\{j\}}=(u_{ij})$ and in the matrix U. We have

$$|U^{\{j\}}|^2=\sum_{i=1}^{n}u_{ij}^2\leqq\left(\sum_{i=1}^{j-1}4(j-i)^2+\frac{1}{4}\right)\beta^{-s}$$
$$=\left(\sum_{h=1}^{j-1}4h^2+\frac{1}{4}\right)\beta^{-2s}$$
$$=\left(4\frac{j(j-1)(2j-1)}{6}+\frac{1}{4}\right)\beta^{-2s},$$

that is,

$$|U^{\{j\}}|^2\leqq(2j(j-1)(2j-1)/3+1/4)\beta^{-2s}. \tag{6.40}$$

The right-hand side is not greater than

$$(2n(n-1)(2n-1)/3+1/4)\beta^{-2s}$$
$$=(2(n-1)(2n-1)/3n^3+1/4n^4)\,n^4\beta^{-2s}.$$

For $n \geqq 10$ this is not greater than $.115\, n^4\beta^{-2s}$. Hence (we assume from now on that $n \geqq 10$)

$$(6.40') \qquad | U^{\{j\}} | \leqq .34 n^2 \beta^{-s}.$$

Next by (6.40)

$$(N(U))^2 = \sum_{j=1}^{n} | U^{\{j\}} |^2 \leqq \left(\sum_{j=1}^{n} \frac{2}{3} j(j-1)(2j-1) + \frac{n}{4} \right) \beta^{-2s}$$
$$= \left(\frac{n^2(n^2-1)}{3} + \frac{n}{4} \right) \beta^{-2s} \leqq \frac{1}{3} n^4 \beta^{-2s},$$

hence

$$| U | \leqq N(U) \leqq (1/3^{1/2}) n^2 \beta^{-s},$$

that is,

$$(6.40'') \qquad | U | \leqq .58 n^2 \beta^{-s}.$$

To conclude this section, we define

$$(6.41) \qquad \alpha = n^2 \beta^{-s} / \mu.$$

Then (6.34), (6.40''), (6.40') become:

$$(6.42) \qquad | \bar{A} - \bar{B}^* \bar{D} \bar{B} | \leqq .42 \mu\alpha,$$

$$(6.43) \qquad | U | \leqq .58 \mu\alpha,$$

$$(6.44) \qquad | U^{\{j\}} | \leqq .34 \mu\alpha.$$

6.6. **Continuation.** (6.42)–(6.44) (together with (6.38)) are the decisive estimates on which we base all others. We proceed as follows. $|A| = \lambda$, $|A|_l = \mu$, hence (6.42) gives $|\bar{B}^*\bar{D}\bar{B}| \leqq \lambda + .42\mu\alpha \leqq \lambda(1+.42\alpha)$, $|\bar{B}^*\bar{D}\bar{B}|_l \geqq \mu - .42\mu\alpha = \mu(1-.42\alpha)$. Next, $(\bar{D}^{1/2}\bar{B})^*(\bar{D}^{1/2}\bar{B}) = \bar{B}^*\bar{D}\bar{B}$, hence $|\bar{D}^{1/2}\bar{B}|^2 = |\bar{B}^*\bar{D}\bar{B}|$, $|\bar{D}^{1/2}\bar{B}|_l^2 = |\bar{B}^*\bar{D}\bar{B}|_l$. Consequently

$$(6.45.\text{a}) \qquad | \bar{D}^{1/2} \bar{B} | \leqq (\lambda(1 + .42\alpha))^{1/2},$$

$$(6.45.\text{b}) \qquad | \bar{D}^{1/2} \bar{B} |_l \geqq (\mu(1 - .42\alpha))^{1/2}.$$

We note two additional, minor facts:

(6.30), (6.23.c') imply

$$(6.46) \qquad | D^v | \leqq 1 \qquad \text{for } v \geqq 0.$$

Since all $|\bar{a}_{ij}| \leqq 1$, therefore $t(A) = \sum_{i=1}^{n} \bar{a}_{ii} \leqq n$. On the other hand $t(A) = \sum_{i=1}^{n} \lambda_i \geqq n\mu$. Hence

$$(6.47) \qquad \mu \leqq 1.$$

We now pass to considering the vectors $\overline{Z}^{\{j\}}=(\bar{z}_{ij})$ and $\Delta^{\{j\}}=(2^{-p_{1j}}\delta_{ij})=2^{-p_{1j}}I^{\{j\}}$. By (6.38)

$$\begin{aligned}\overline{D}^{1/2}\overline{B}(\overline{Z}^{\{j\}}) = \overline{D}^{1/2}(\overline{BZ})^{\{j\}} &= \overline{D}^{1/2}(\Delta^{\{j\}} + U^{\{j\}}) \\ &= \bar{d}_j^{1/2}2^{-p_{1j}}I^{\{j\}} + \overline{D}^{1/2}(U^{\{j\}}).\end{aligned}$$

By (6.45.b), (6.37)

$$\left| \overline{D}^{1/2}\overline{B}(\overline{Z}^{\{j\}}) \right| \geqq \left| \overline{D}^{1/2}\overline{B} \right|_l \left| \overline{Z}^{\{j\}} \right| \geqq (1/2)(\mu(1 - .42\alpha))^{1/2};$$

obviously

$$\left| \bar{d}_j^{1/2}2^{-p_{1j}}I^{\{j\}} \right| = 2^{-p_{1j}}\bar{d}_j^{1/2}$$

and by (6.46), (6.44)

$$\left| \overline{D}^{1/2}(U^{\{j\}}) \right| \leqq \left| \overline{D}^{1/2} \right| \left| U^{\{j\}} \right| \leqq .34\mu\alpha.$$

From all these relations we can infer that

$$\begin{aligned}(\mu(1 - .42\alpha))^{1/2}/2 &\leqq 2^{-p_{1j}}\bar{d}_j^{1/2} + .34\mu\alpha, \\ 2^{-p_{1j}}\bar{d}_j^{1/2} &\geqq (\mu(1 - .42\alpha))^{1/2}/2 - .34\mu\alpha \\ &= ((\mu(1 - .42\alpha))^{1/2} - .68\mu\alpha)/2\end{aligned}$$

and, remembering (6.47),

$$\begin{aligned}2^{-p_{1j}}\bar{d}_j^{1/2} &\geqq (\mu(1 - .42\alpha) - 2\cdot(.68)\mu\alpha)^{1/2}/2 \\ &= (\mu(1 - 1.78\alpha))^{1/2}/2,\end{aligned}$$

that is,

$$2^{-p_{1j}}\bar{d}_j^{1/2} \geqq (\mu(1 - 1.78\alpha))^{1/2}/2. \tag{6.48}$$

Since $\Delta\overline{D}^{1/2}=(2^{-p_{1i}}\bar{d}_i^{1/2}\delta_{ij})$, (6.48) may also be written

$$\left| \Delta\overline{D}^{1/2} \right|_l \geqq (\mu(1 - 1.78\alpha))^{1/2}/2. \tag{6.48'}$$

Next by (6.38), (6.45.b), (6.46), (6.43), (6.48')

$$\begin{aligned}\overline{D}^{1/2}\overline{BZ}\Delta^{-1}\overline{D}^{-1/2} &= \overline{D}^{1/2}(\Delta + U)\Delta^{-1}\overline{D}^{-1/2} = I + \overline{D}^{1/2}U\Delta^{-1}\overline{D}^{-1/2}, \\ \left| \overline{D}^{1/2}\overline{BZ}\Delta^{-1}\overline{D}^{-1/2} \right| &\geqq \left| \overline{D}^{1/2}\overline{B} \right|_l \left| \overline{Z}\Delta^{-1}\overline{D}^{-1/2} \right| \\ &\geqq (\mu(1 - .42\alpha))^{1/2} \left| \overline{Z}\Delta^{-1}\overline{D}^{-1/2} \right|, \\ \left| \overline{D}^{1/2}U\Delta^{-1}\overline{D}^{-1/2} \right| &\leqq \left| \overline{D}^{1/2} \right| \left| U \right| \left| \Delta\overline{D}^{1/2} \right|_l^{-1} \\ &\leqq .58\mu\alpha\cdot 2/(\mu(1 - 1.78\alpha))^{1/2} \\ &= 1.16\mu^{1/2}\alpha/(1 - 1.78\alpha)^{1/2}.\end{aligned}$$

Hence

$$(\mu(1-.42\alpha))^{1/2}\,|\bar{Z}\Delta^{-1}\bar{D}^{-1/2}| \leqq 1 + 1.16\mu^{1/2}\alpha/(1-1.78\alpha)^{1/2}$$

and by (6.47)

$$\begin{aligned}(\mu(1-.42\alpha))^{1/2}\,&|\bar{Z}\Delta^{-1}\bar{D}^{-1/2}|\\ &\leqq 1 + 1.16\frac{\alpha}{(1-1.78\alpha)^{1/2}} = \frac{(1-1.78\alpha)^{1/2}+1.16\alpha}{(1-1.78\alpha)^{1/2}}\\ &\leqq \frac{(1-1.78\alpha/2)+1.16\alpha}{(1-1.78\alpha)^{1/2}} = \frac{1+.27\alpha}{(1-1.78\alpha)^{1/2}}\,.\end{aligned}$$

Consequently

$$(6.49)\qquad |\bar{Z}\Delta^{-1}\bar{D}^{-1/2}| \leqq \frac{1+.27\alpha}{(1-.42\alpha)^{1/2}(1-1.78\alpha)^{1/2}}\,\frac{1}{\mu^{1/2}}\,.$$

Since $\Delta^{-1}\bar{D}^{-1/2}\bar{Z}^* = (\bar{Z}\Delta^{-1}\bar{D}^{-1/2})^*$ and $\bar{Z}\Delta^{-2}\bar{D}^{-1}\bar{Z}^* = (\Delta^{-1}\bar{D}^{1/2}\bar{Z}^*)^*\cdot(\Delta^{-1}\bar{D}^{1/2}\bar{Z}^*)$ (remember that Δ and $\bar{D}$ commute, because they are diagonal matrices), therefore

$$(6.49')\qquad |\Delta^{-1}\bar{D}^{-1/2}\bar{Z}^*| \leqq \frac{1+.27\alpha}{(1-.42\alpha)^{1/2}(1-1.78\alpha)^{1/2}}\,\frac{1}{\mu^{1/2}}\,,$$

$$(6.49'')\qquad |\bar{Z}\Delta^{-2}\bar{D}^{-1}\bar{Z}^*| \leqq \frac{(1+.27\alpha)^2}{(1-.42\alpha)(1-1.78\alpha)}\,\frac{1}{\mu}\,.$$

6.7. **Continuation.** Choose q $(=0, 1, 2, \cdots)$ minimal with

$$(6.50.\text{a})\qquad 2^q > \frac{4}{1-1.78\alpha}\,\frac{1}{\mu}\,,$$

that is, having in addition the property

$$(6.50.\text{b})\qquad 2^q \leqq \frac{8}{1-1.78\alpha}\,\frac{1}{\mu}\,.$$

By (6.23.c′), (6.30), $0<\bar{d}_j\leqq 1$. Hence we can form the maximal $r=0, 1, 2, \cdots$ with $2^r d_j \leqq 1$, say r_j. We have

$$(6.51)\qquad 1/2 < 2^{r_j}\bar{d}_j \leqq 1.$$

By (6.48), (6.50.a), $2^q\bar{d}_j > {}^{2p_{1j}}$, $2^{q-2p_{1j}}\bar{d}_j > 1$. Hence $q-2p_{1j}>r_j$, that is,

$$(6.52)\qquad q - 2p_{1j} \geqq r_j + 1.$$

From (6.51), (6.52) we can infer:

(6.53) $\bar{e}_j^{(q)} = (1/2 \div 2^{r_j}\bar{d}_j) \div 2^{q-2p_{1j}-r_j-1}$ is, together with all its intermediate expressions, a digital number and well formed.[31]

Next by (6.50.a), (6.49′′)

$$(6.54) \qquad \left| 2^{-q}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^* \right| \leqq \frac{1}{4}\,\frac{(1+.27\alpha)^2}{1-.42\alpha}\,.$$

Furthermore form

$$(6.55) \qquad \overline{W}^{(q)} = (\bar{w}_{ij}^{(q)}), \qquad \bar{w}_{ij}^{(q)} = \sum_{k=1}^{n} (\bar{z}_{i'k} \times e_k^{(q)}) \times z_{j'k}$$

$$(i' = \operatorname{Min}(i, j),\ j' = \operatorname{Max}(i, j)).$$

(We must still establish the digital character of $\overline{W}^{(q)}$ and of the $\bar{w}_{ij}^{(q)}$, which in this case amounts to showing that all $\bar{w}_{ij}^{(q)}$ lie in $-1, 1$. For this cf. (6.58), (6.59).) Now

$$\overline{W}^{(q)} - 2^{-q}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^*$$

$$= \left(\sum_{k=1}^{n} (\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} - \sum_{k=1}^{n} \bar{z}_{i'k} 2^{-q+2p_{ik}} \bar{d}_k^{-1} \bar{z}_{j'k} \right).$$

(Note that we replaced in the second term on the right-hand side i, j by i', j'. This is permissible, since it is symmetric in i, j.) The ij-element of the right-hand side can be written

$$\sum_{k=1}^{n} ((\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} - \bar{z}_{i'k} 2^{-q+2p_{ik}} \bar{d}_k^{-1} \bar{z}_{j'k}),$$

or, since $\overline{Z} = (\bar{z}_{ij})$ is upper semi-diagonal,

$$\sum_{k=j'}^{n} ((\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} - \bar{z}_{i'k} 2^{-q+2p_{ik}} \bar{d}_k^{-1} \bar{z}_{j'k}), \qquad j' = \operatorname{Max}(i, j).$$

For this we can further write

$$\sum_{k=j'}^{n} \left\{ (\bar{z}_{i'k} \times \bar{e}_k^{(q)}) \times \bar{z}_{j'k} - \bar{z}_{i'k} \bar{e}_k^{(q)} \bar{z}_{j'k}) + \bar{z}_{i'k} (\bar{e}_k^{(q)} - 2^{-q+2p_{ik}} \bar{d}_k^{-1}) \bar{z}_{j'k} \right\}.$$

By (6.51),

$$\left| \frac{1}{2} \div 2^{r_j} \bar{d}_j - 2^{-r_j-1} \bar{d}_j^{-1} \right| \leqq \beta^{-s}/2,$$

[31] We are assuming here that 1/2 is a digital number. This is only true when the base β is even. This limitation could be removed with little trouble, but it does not seem worthwhile, since $\beta = 2$ and $\beta = 10$ are both even.

hence, by (2.22) and (6.53), $|\bar{e}_j^{(q)} - 2^{-q+2p_{ij}}\bar{d}_j^{-1}| \leqq \beta^{-s}$. Therefore the second term in the $\{ \cdots \}$ above has an absolute value not greater than β^{-s}. The first term in this $\{ \cdots \}$ has clearly an absolute value not greater than $\beta^{-s}/2 + \beta^{-s}/2 = \beta^{-s}$. Hence this $\{ \cdots \}$ has an absolute value not greater than $2\beta^{-s}$, and so the entire expression in question is not greater than $2(n-j'+1)\beta^{-s}$. Consequently

$$(N(\overline{W}^{(q)} - 2^{-q}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^*))^2$$

$$\leqq \sum_{i,j=1}^{n} 4(n - j' + 1)^2\beta^{-2s} = \sum_{j'=1}^{n} 4(2j' - 1)(n - j' + 1)^2\beta^{-2s}$$

$$= \sum_{h=1}^{n} 4(2n - 2h + 1)h^2\beta^{-2s} = \frac{2n(n + 1)(n^2 + n + 1)}{3}\beta^{-2s}$$

and ($n \geqq 10$, cf. the remark preceding (6.40′)) is not greater than

$$.82n^4\beta^{-2s},$$

hence

$$|\overline{W}^{(q)} - 2^{-q}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^*| \leqq N(\overline{W}^{(q)} - 2^{-q}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^*)$$
$$\leqq .91n^2\beta^{-s} = .91\mu\alpha,$$

that is,

$$(6.56) \qquad |\overline{W}^{(q)} - 2^{-q}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^*| \leqq .91\mu\alpha.$$

From (6.54), (6.56), remembering (6.47), we get

$$(6.57) \qquad |\overline{W}^{(q)}| \leqq \frac{1}{4}\,\frac{(1 + .27\alpha)^2}{1 - .42\alpha} + .91\alpha.$$

We shall see later (cf. (6.67)) that it is reasonable to assume that $\alpha \leqq .1$. This implies that the right-hand side of (6.57) is not greater than .37. Consequently

$$(6.57') \qquad |\overline{W}^{(q)}| \leqq .37.$$

From this, (3.17.b) (first half) permits us to infer

$$(6.57'.\text{a}) \qquad |\bar{w}_{ij}^{(q)}| \leqq 1 \qquad (i, j = 1, \cdots, n).$$

Next, since $\overline{W}^{(q)}(I^{\{j\}}) = (\overline{W}^{(q)})^{\{j\}}$ and $|I^{\{j\}}| = 1$, (6.57′) implies $|(\overline{W}^{(q)})^{\{j\}}| \leqq .37$, that is,

$$\sum_{i=1}^{n} (\bar{w}_{ij}^{(q)})^2 \leqq (.37)^2 \leqq .14.$$

If we replace in this sum $(\bar{w}_{ij}^{(q)})^2$ by $\bar{w}_{ij}^{(q)} \times \bar{w}_{ij}^{(q)}$, then the total change is not greater than $n\cdot\beta^{-s}/2 = n^2\beta^{-s}/2n = \mu\alpha/2n$. Since $\alpha \leqq .1$, $n \geqq 10$ (cf. above) and $\mu \leqq 1$ (by (6.47)), therefore this is not greater than .005. Consequently

$$\sum_{i=1}^{n} \bar{w}_{ij}^{(q)} \times \bar{w}_{ij}^{(q)} \leqq .99 \qquad (j = 1, \cdots, n). \tag{6.57$'$.b}$$

We sum up:

(6.58) For the minimal q ($=0, 1, 2, \cdots$) of (6.50.a) the conditions (6.53), (6.57′) are fulfilled. The latter implies (6.57′.a) and (6.57′.b).

In this section we use (6.57′.a); the need for (6.57′.b) will arise later. (Cf. (6.92).)

We define:

(6.59) Let q_0 be the minimal q ($=0, 1, 2, \cdots$) for which the conditions (6.53), (6.57′.a) are fulfilled.

Note that these are simple, explicit conditions, which permit forming the q_0 in question directly.[32]

(6.58) shows that the minimal q of (6.50.a) is not less than q_0. Hence q_0, too, fulfills (6.50.b), that is,

$$2^{q_0} \leqq \frac{8}{1 - 1.78\alpha} \frac{1}{\mu} \cdot \tag{6.60}$$

We now put

$$\overline{W}_0 = \overline{W}^{(q_0)} = (\bar{w}_{ij}^{(q_0)}). \tag{6.61}$$

The derivation of (6.56) was based on (6.53) alone, hence (6.56) holds for $\overline{W}_0 = \overline{W}^{(q_0)}$, too:

$$\left| \overline{W}_0 - 2^{-q_0}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^* \right| \leqq .91\mu\alpha. \tag{6.62}$$

6.8. **Continuation. The estimates connected with the inverse of $\overline{A}$.** We are now able to effect our final estimates. The relevant auxiliary estimates are (6.42), (6.43), (6.45.a), (6.45.b), (6.46), (6.48′), (6.49′), (6.49″), (6.60), (6.62). The procedure is as follows:

Put

$$A' = \overline{B}^*\overline{D}\overline{B}, \tag{6.63.a}$$

$$W' = 2^{-q_0}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^*. \tag{6.63.b}$$

[32] This would remain true if we replaced (6.57′.a) by (6.57′.b) (cf. (6.92)), but not if we reverted to (6.57′).

Owing to

$$2^{q_0}\overline{A}\,\overline{W}_0 - 2^{q_0}A'W' = (\overline{A} - A')2^{q_0}W' + 2^{q_0}\overline{A}(\overline{W}_0 - W'),$$

we have

$$|\,2^{q_0}\overline{A}\,\overline{W}_0 - 2^{q_0}A'W'\,| \leqq |\,\overline{A} - A'\,|\,|\,2^{q_0}W'\,| + 2^{q_0}|\,\overline{A}\,|\,|\,\overline{W}_0 - W'\,|.$$

By (6.42), (6.49″), (6.60), (6.62) this is less than or equal to

$$.42\mu\alpha\cdot\frac{(1 + .27\alpha)^2}{(1 - .42\alpha)(1 - 1.78\alpha)}\,\frac{1}{\mu} + \frac{8}{1 - 1.78\alpha}\,\frac{1}{\mu}\,\lambda\cdot .91\mu\alpha$$
$$= \left[.42\,\frac{(1 + .27\alpha)^2}{(1 - .42\alpha)(1 - 1.78\alpha)} + 7.28\,\frac{1}{1 - 1.78\alpha}\,\lambda\right]\alpha.$$

Summing up:

$$(6.63)\qquad \begin{aligned} &|\,2^{q_0}\overline{A}\,\overline{W}_0 - 2^{q_0}A'W'\,| \\ &\qquad \leqq \left[.42\,\frac{(1 + .27\alpha)^2}{(1 - .42\alpha)(1 - 1.78\alpha)} + 7.28\,\frac{1}{1 - 1.78\alpha}\,\lambda\right]\alpha. \end{aligned}$$

Next

$$\begin{aligned} 2^{q_0}A'W' - I &= \overline{B}^*\overline{D}\,\overline{BZ}\Delta^{-2}\overline{D}^{-1}\overline{Z}^* - I \\ &= \overline{B}^*\overline{D}(\Delta + U)\Delta^{-2}\overline{D}^{-1}\overline{Z}^* - I \\ &= (\overline{B}^*\Delta^{-1}\overline{Z}^* - I) + \overline{B}^*\overline{D}U\Delta^{-2}\overline{D}^{-1}\overline{Z}^* \end{aligned}$$

(remember that Δ and $\overline{D}$ commute, because they are diagonal matrices). Now

$$\begin{aligned} \overline{B}^*\Delta^{-1}\overline{Z}^* - I &= (\overline{Z}\Delta^{-1}\overline{B} - I)^* = (\overline{B}^{-1}(\overline{BZ}\Delta^{-1} - I)\overline{B})^* \\ &= (\overline{B}^{-1}((\Delta + U)\Delta^{-1} - I)\overline{B})^* = (\overline{B}^{-1}U\Delta^{-1}\overline{B})^* \\ &= ((\overline{D}^{1/2}\overline{B})^{-1}\overline{D}^{1/2}U(\Delta^{-1}\overline{D}^{-1/2})(\overline{D}^{1/2}\overline{B}))^*, \end{aligned}$$

hence, by (6.45.b), (6.46), (6.43), (6.48′), (6.45.a),

$$\begin{aligned} &|\,\overline{B}^*\Delta^{-1}\overline{Z}^* - I\,| \\ &\qquad \leqq |\,\overline{D}^{1/2}\overline{B}\,|_l^{-1}\,|\,\overline{D}^{1/2}\,|\,|\,U\,|\,|\,\Delta\overline{D}^{1/2}\,|_l^{-1}\,|\,\overline{D}^{1/2}\overline{B}\,| \\ &\qquad \leqq \frac{1}{(\mu(1 - .42\alpha))^{1/2}}\,.58\mu\alpha\,\frac{2}{(\mu(1 - 1.78\alpha))^{1/2}}\,(\lambda(1 + .42\alpha))^{1/2} \\ &\qquad = 1.16\,\frac{(1 + .42\alpha)^{1/2}}{(1 - .42\alpha)^{1/2}(1 - 1.78\alpha)^{1/2}}\,\lambda^{1/2}\alpha. \end{aligned}$$

Furthermore

$$\overline{B}^*\overline{D}U\Delta^{-2}\overline{D}^{-1}\overline{Z}^* = (\overline{D}^{1/2}\overline{B})^*\overline{D}^{1/2}U(\Delta^{-1}\overline{D}^{-1/2})(\Delta^{-1}\overline{D}^{-1/2}\overline{Z}^*),$$

hence by (6.45.a), (6.46), (6.43), (6.48′), (6.49′)

$$\begin{aligned}
\left|\overline{B}^*\overline{D}U\Delta^{-2}\overline{D}^{-1}\overline{Z}^*\right| &\leqq \left|\overline{D}^{1/2}\overline{B}\right|\left|\overline{D}^{1/2}\right|\left|U\right|\left|\Delta\overline{D}^{1/2}\right|_l^{-1}\left|\Delta^{-1}\overline{D}^{-1/2}\overline{Z}^*\right| \\
&\leqq (\lambda(1+.42\alpha))^{1/2}.58\mu\alpha\frac{2}{(\mu(1-1.78\alpha))^{1/2}} \\
&\qquad\cdot\frac{1+.27\alpha}{(1-.42\alpha)^{1/2}(1-1.78\alpha)^{1/2}}\frac{1}{\mu^{1/2}} \\
&= 1.16\frac{(1+.42\alpha)^{1/2}(1+.27\alpha)}{(1-.42\alpha)^{1/2}(1-1.78\alpha)}\lambda^{1/2}\alpha.
\end{aligned}$$

Summing up:

$$\begin{aligned}
&\left|2^{q_0}A'W'-I\right| \\
(6.64)\qquad &\leqq 1.16\frac{(1+.42\alpha)^{1/2}}{(1-.42\alpha)^{1/2}}\left(\frac{1}{(1-1.78\alpha)^{1/2}}+\frac{1+.27\alpha}{1-1.78\alpha}\right)\lambda^{1/2}\alpha.
\end{aligned}$$

Combining (6.63), (6.64) we obtain

$$\begin{aligned}
&\left|2^{q_0}\overline{A}\overline{W}_0-I\right| \\
&\leqq\Bigg[.42\frac{(1+.27\alpha)^2}{(1-.42\alpha)(1-1.78\alpha)} \\
(6.65)\qquad &+1.16\frac{(1+.42\alpha)^{1/2}}{(1-.42\alpha)^{1/2}}\left(\frac{1}{(1-1.78\alpha)^{1/2}}+\frac{1+.27\alpha}{1-1.78\alpha}\right)\lambda^{1/2} \\
&+7.28\frac{1}{1-1.78\alpha}\lambda\Bigg]\alpha.
\end{aligned}$$

If we now assume $\alpha\leqq.1$, then the right-hand side of (6.65) is less than or equal to the expression

$$(6.66)\qquad (.56+2.83\lambda^{1/2}+8.86\lambda)\alpha.$$

It is indicated to scale $\overline{A}=(\bar{a}_{ij})$ so that $\mathrm{Max}_{i,j=1,\ldots,n}(\bar{a}_{ij})$ is near 1, hence (by (3.17.b), first half) λ is near or more than 1. Hence the above coefficient of α is presumably greater than or equal to 10. This implies that for $\alpha=.1$ the right-hand side of (6.65) is presumably greater than or equal to 1. This, however, means that $\overline{W}_0$ was not worth constructing since even 0 in place of $\overline{W}_0$ would have given a right-hand side equal to 1. In other words: If $\alpha\leqq.1$ does not hold, then the result (6.65) is without interest. If $\alpha\leqq.1$ holds, then the right-hand side of (6.65) is less than or equal to the expression (6.66). It seems therefore logical to assume

$$\alpha \leqq .1. \tag{6.67}$$

Note that this renders our earlier assumption (6.23.c) superfluous: $\alpha \leqq .1$ means that $n^2\beta^{-s}/\mu \leqq 1/10$, $10n^2\beta^{-s} \leqq \mu$, which implies the relation in question.

We can now incorporate (6.66) into (6.65). In this way we obtain:

$$| 2^{q_0}\bar{A}\bar{W}_0 - I | \leqq (.56 + 2.83\lambda^{1/2} + 8.86\lambda)\alpha. \tag{6.65'}$$

Since $\lambda^{1/2} \leqq (1+\lambda)/2$ we can simplify (6.65′) to

$$| 2^{q_0}\bar{A}\bar{W}_0 - I | \leqq (1.98 + 10.28\lambda)\alpha. \tag{6.65''}$$

6.9. **The general $\bar{A}_I$. Various estimates.** (6.65″) (together with (6.41) and (6.67)) is our final result for the $\bar{A}$ fulfilling the conditions of (6.23) (that is, (a), (b) there, (c) was eliminated, cf. above after (6.67)). That is, this completes our work dealing with the definite $\bar{A}$. This result still requires some discussion and interpretation, but we postpone these until Chapter VII. At this point we turn our attention to the general (that is, nonsingular and not necessarily definite) $\bar{A}$. In order to emphasize the difference, we shall denote the general matrix in question by $\bar{A}_I$ instead of $\bar{A}$.

Let, then

$$\bar{A}_I = (\bar{a}_{I,ij}) \qquad (i, j = 1, \cdots, n) \tag{6.68}$$

be a general (nonsingular), digital matrix.

Since we have solved the problem of inverting a matrix in the case when it is definite (cf. above), we wish to replace the problem of inverting A_I by that one of inverting an appropriate definite matrix. This should be done in analogy with the procedure suggested in §5.1. More specifically: The inverting of $\bar{A}_I$ should be based on that of $\bar{A}_I^*\bar{A}_I$. Since we are dealing with digital matrices, however, we have to consider $\bar{A}_I^* \times \bar{A}_I$ instead of $\bar{A}_I^*\bar{A}_I$. Furthermore, it will prove technically more convenient to use $\bar{A}_I \times \bar{A}_I^*$ rather than $\bar{A}_I^* \times \bar{A}_I$ (cf. the algebraical manipulations leading up to (6.100). We introduce accordingly

$$\bar{A} = (\bar{a}_{ij}) = \bar{A}_I \times \bar{A}_I^*. \tag{6.69}$$

(6.69) indicates that $\bar{A}$ is a digital matrix. This, however, is not immediate: If we assume of $\bar{A}_I$ merely that all its elements $\bar{a}_{I,ij}$ lie in -1, 1, then the elements

$$\bar{a}_{ij} = \sum_{k=1}^{n} \bar{a}_{I,ik} \times \bar{a}_{I,jk} \tag{6.70}$$

of $\bar{A}$ need not lie in $-1, 1$. We shall rectify this shortcoming before long, but we prefer to disregard it for a moment, and discuss a few other matters first.

Put

$$|\bar{A}_I| = \lambda_I, \tag{6.71.a}$$

$$|\bar{A}_I|_l = \mu_I, \tag{6.71.b}$$

$$|\bar{A}| = \lambda, \tag{6.72.a}$$

$$|\bar{A}|_l = \mu. \tag{6.72.b}$$

(Cf. the analogous (5.11.a), (5.11.b).) Furthermore put

$$\alpha_I = n^2\beta^{-s}/\mu_I^2, \tag{6.73}$$

$$\alpha = n^2\beta^{-s}/\mu. \tag{6.74}$$

(Cf. the analogous (6.41). Note, however, that the denominator of the right-hand side of (6.73) is μ_I^2 and not μ_I.)

By (3.31.a)

$$|\bar{A} - \bar{A}_I\bar{A}_I^*| \leqq n^2\beta^{-s}/2 = \mu_I^2\alpha_I/2. \tag{6.75}$$

Hence

$$\begin{aligned} \lambda &= |\bar{A}| \leqq |\bar{A}_I\bar{A}_I^*| + \mu_I^2\alpha_I/2 = |\bar{A}_I|^2 + \mu_I^2\alpha_I/2 \\ &= \lambda_I^2 + \mu_I^2\alpha_I/2 \leqq \lambda_I^2(1 + \alpha_I/2), \\ \mu &= |\bar{A}|_l \geqq |\bar{A}_I\bar{A}_I^*|_l - \mu_I^2\alpha_I/2 = |\bar{A}_I|_l^2 - \mu_I^2\alpha_I/2 \\ &= \mu_I^2 - \mu_I^2\alpha_I/2 = \mu_I^2(1 - \alpha_I/2), \end{aligned}$$

that is,

$$\lambda \leqq \lambda_I^2(1 + \alpha_I/2), \tag{6.76.a}$$

$$\mu \geqq \mu_I^2(1 - \alpha_I/2). \tag{6.76.b}$$

From (6.76.b) and (6.73), (6.74)

$$\alpha \leqq \frac{\alpha_I}{1 - \alpha_I/2}\,. \tag{6.77}$$

Hence the condition (6.67), which we restate

$$\alpha \leqq .1, \tag{6.78}$$

is fulfilled, if

$$(6.78') \qquad \alpha_I \leqq .095.$$

Accordingly, we postulate (6.78′).

(6.78) implies $\mu \neq 0$, that is, the nonsingularity of $\bar{A}$. (6.75) implies

$$\left| (\bar{A}\xi, \xi) - (\bar{A}_I\bar{A}_I^*\xi, \xi) \right| \leqq \mu_I^2\alpha_I \left| \xi \right|^2/2.$$

Next

$$(\bar{A}_I\bar{A}_I^*\xi, \xi) = (\bar{A}_I^*\xi, \bar{A}_I^*\xi) = \left| \bar{A}_I^*\xi \right|^2 \geqq \left| \bar{A}_I^* \right|_l^2 \left| \xi \right|^2.$$

Since $|\bar{A}_I^*|_l = |\bar{A}_I|_l = \mu_I$, this is greater than or equal to

$$\mu_I^2 \left| \xi \right|^2.$$

Hence

$$(\bar{A}\xi, \xi) \geqq \mu_I^2 \left| \xi \right|^2 - \mu_I^2\alpha_I \left| \xi \right|^2/2,$$

and so, by (6.78′), $(\bar{A}\xi, \xi) \geqq 0$. Therefore $\bar{A}$ is definite. We sum up:

(6.79) $\bar{A}$ is (nonsingular and) definite.

We conclude this section by securing the digital character of $\bar{A}$, that is, of all $\bar{a}_{ij}$. What is required is that all $\bar{a}_{ij}$ lie in $-1, 1$. Now by (6.70)

$$\left| \bar{a}_{ij} - \sum_{k=1}^{n} \bar{a}_{I,ik}\bar{a}_{I,jk} \right| \leqq n\beta^{-s}/2 = n^2\beta^{-s}/2n = \mu_I^2\alpha_I/2n,$$

and, since $n \geqq 10$, $\alpha_I < .1$, is less than or equal to

$$\frac{1}{200}\mu_I^2.$$

Hence

$$(6.80) \qquad \left| \bar{a}_{ij} \right| \leqq \sum_{k=1}^{n} \left| \bar{a}_{I,ik} \right| \left| \bar{a}_{I,jk} \right| + \mu_I^2/200.$$

Now

$$(6.81) \qquad \begin{aligned} \sum_{k=1}^{n} \left| \bar{a}_{I,ik} \right| \left| \bar{a}_{I,jk} \right| &\leqq \left(\sum_{k=1}^{n} (\bar{a}_{I,ik})^2 \right)^{1/2} \left(\sum_{k=1}^{n} (\bar{a}_{I,jk})^2 \right)^{1/2} \\ &\leqq \operatorname*{Max}_{h=1,\cdots,n} \sum_{k=1}^{n} (\bar{a}_{I,hk})^2, \end{aligned}$$

and, since $A_I(I^{\{h\}})=A_I^{\{h\}}$ and $|I^{\{h\}}|=1$,

$$\mu_I^2 \leqq |A_I^{\{h\}}|^2 = \sum_{k=1}^{n} (\bar{a}_{I,hk})^2,$$

that is,

$$\mu_I^2 \leqq \operatorname*{Min}_{h=1,\cdots,n} \sum_{k=1}^{n} (\bar{a}_{I,hk})^2. \tag{6.82}$$

By (6.81), (6.82) we obtain from (6.80)

$$|\bar{a}_{ij}| \leqq 1.005 \operatorname*{Max}_{h=1,\cdots,n} \sum_{k=1}^{n} (\bar{a}_{I,hk})^2. \tag{6.83}$$

Again

$$\left| \sum_{k=1}^{n} \bar{a}_{I,hk} \times \bar{a}_{I,hk} - \sum_{k=1}^{n} (\bar{a}_{I,hk})^2 \right| \leqq \frac{n\beta^{-s}}{2}$$

and, as above, is less than or equal to

$$\frac{\mu_I^2}{200}.$$

Hence

$$\sum_{k=1}^{n} \bar{a}_{I,hk} \times \bar{a}_{I,hk} \geqq \sum_{k=1}^{n} (\bar{a}_{I,hk})^2 - \frac{\mu_I^2}{200}$$

and by (6.82) is greater than or equal to

$$.995 \sum_{k=1}^{n} (\bar{a}_{I,hk})^2.$$

Consequently

$$\sum_{k=1}^{n} (\bar{a}_{I,hk})^2 \leqq \frac{1}{.995} \sum_{k=1}^{n} \bar{a}_{I,hk} \times \bar{a}_{I,hk}. \tag{6.84}$$

Thus

$$\sum_{j=1}^{n} \bar{a}_{I,ij} \times \bar{a}_{I,ij} \leqq .99 \qquad (i = 1, \cdots, n) \tag{6.85}$$

implies by (6.84) and (6.82), (6.83)

$$\mu_I \leqq 1, \tag{6.82'}$$

$$|\bar{a}_{ij}| \leqq 1. \tag{6.83'}$$

We assume the validity of (6.85). Then the digital character of $\overline{A}$ and of the $\bar{a}_{ij}$ is secured.

6.10. **Continuation.** By (6.78), (6.79), and the remark after (6.83′), $\overline{A}$ fulfills the conditions (6.23a), (6.23b), (6.67). Hence our results on inverting a definite matrix apply to it, and we can form the q_0 and $\overline{W}_0$ of §§6.7, 6.8 for this $\overline{A}$.

(6.65″) shows that $2^{q_0}\overline{W}_0$ is an approximate inverse of $\overline{A}$, and (6.75) shows that $\overline{A}$ is approximately $\overline{A}_I\overline{A}_I^*$. It is therefore reasonable to expect that $2^{q_0}\overline{A}_I^*\overline{W}_0$ will be usable as an approximate inverse of $\overline{A}_I$. Since we want a digital matrix, we should consider $2^{q_0}\overline{A}_I^*\times\overline{W}_0$ instead of $2^{q_0}\overline{A}_I^*\overline{W}_0$.

The digital character is, of course, desired for $\overline{A}_I^*\times\overline{W}_0$, and not for $2^{q_0}\overline{A}_I^*\times\overline{W}_0$. (Cf. with respect to this the last remark in §7.6.) However, the digital character of $\overline{A}_I^*\times\overline{W}_0$ is open to the same doubts, which we discussed immediately after (6.69) in connection with $\overline{A}_I\times\overline{A}_I^*$: We know that the elements $\bar{a}_{I,ji}$ of $\overline{A}_I^*$ and the elements $\bar{w}_{ij}^{(q_0)}$ of $\overline{W}_0$ lie in $-1, 1$, but this does not guarantee that the elements $\bar{s}_{ij}=\sum_{k=1}^{n}\bar{a}_{I,ki}\times\bar{w}_{kj}^{(q_0)}$ of $\overline{A}_I^*\times\overline{W}_0$ lie also in $-1, 1$. It is therefore necessary that we make certain that the $\bar{s}_{ij}$ do lie in $-1, 1$.

Write q for q_0, and put

$$\bar{s}_{ij}^{(q)} = \sum_{k=1}^{n}\bar{a}_{I,ki}\times\bar{w}_{kj}^{(q)}. \tag{6.86}$$

We argue as in the corresponding part of 6.9: By (6.86)

$$\left|\bar{s}_{ij}^{(q)} - \sum_{k=1}^{n}\bar{a}_{I,ki}\bar{w}_{kj}^{(q)}\right| \leqq n\beta^{-s}/2 = n^2\beta^{-s}/2n = \mu_I^2\alpha_I/2n,$$

and, since $n\geqq 10$, $\alpha_I\leqq .1$, and by (6.82′), is less than or equal to

$$\frac{1}{200}.$$

Hence

$$|\bar{s}_{ij}^{(q)}| \leqq \sum_{k=1}^{n}|\bar{a}_{I,ki}||\bar{w}_{kj}^{(q)}| + \frac{1}{200}. \tag{6.87}$$

Now

$$\sum_{k=1}^{n}|\bar{a}_{I,ki}||\bar{w}_{kj}^{(q)}| \leqq \left(\sum_{k=1}^{n}(\bar{a}_{I,ki})^2\right)^{1/2}\left(\sum_{k=1}^{n}(\bar{w}_{kj}^{(q)})^2\right)^{1/2}. \tag{6.88}$$

Furthermore

$$(6.89.\mathrm{a})\qquad \left|\sum_{k=1}^{n}\bar{a}_{I,ki}\times\bar{a}_{I,ki}-\sum_{k=1}^{n}(\bar{a}_{I,ki})^2\right|\leqq\frac{n\beta^{-s}}{2}\leqq\frac{1}{200},$$

$$(6.89.\mathrm{b})\qquad \left|\sum_{k=1}^{n}\bar{w}_{kj}^{(q)}\times\bar{w}_{kj}^{(q)}-\sum_{k=1}^{n}(\bar{w}_{kj}^{(q)})^2\right|\leqq\frac{n\beta^{-s}}{2}\leqq\frac{1}{200}\cdot$$

Therefore

$$(6.90.\mathrm{a})\qquad \sum_{k=1}^{n}\bar{a}_{I,ki}\times\bar{a}_{I,ki}\leqq .99,$$

$$(6.90.\mathrm{b})\qquad \sum_{k=1}^{n}\bar{w}_{kj}^{(q)}\times\bar{w}_{kj}^{(q)}\leqq .99$$

imply by (6.89.a), (6.89.b) with (6.88) and (6.87), that

$$(6.87')\qquad \left|\bar{s}_{ij}^{(q)}\right|\leqq 1.$$

That is, in this case $\bar{s}_{ij}^{(q)}$ lies in $-1, 1$, as desired.

We propose to treat (6.90.a), that is,

$$(6.91)\qquad \sum_{i=1}^{n}\bar{a}_{I,ij}\times\bar{a}_{I,ij}\leqq .99\qquad (j=1,\cdots,n),$$

like the analogous (6.85), as a postulate concerning $\bar{A}_I$. On the other hand (6.90.b), which coincides with (6.57′.b), will be secured by an appropriate choice of q. In other words:

The q_0, which was the value given to q throughout §6.8, was defined by (6.59) in §6.7. It was the minimal q $(=0, 1, 2, \cdots)$ fulfilling (6.53), (6.57′.a). We have seen that we should now replace (6.57′.a) by (6.57′.b), and thereby give q a new value q_1, instead of q_0.

We define accordingly:

(6.92) Let q_1 be the minimal q $(=0, 1, 2, \cdots)$ for which the conditions (6.53), (6.57′.b) are fulfilled. (Cf. the remark after (6.58) and footnote 32.)

We can now repeat a good deal of the argument following upon (6.58) with practically no change:

(6.58) shows that the minimal q of (6.50.a) is not more than q_1. Hence q_1, too, fulfills (6.50.b), that is,

$$(6.93)\qquad 2^{q_1}\leqq\frac{8}{1-1.78\alpha}\,\frac{1}{\mu}\cdot$$

We put

$$\overline{W}_1 = \overline{W}^{(q_1)} = (\bar{w}_{ij}^{(q_1)}). \tag{6.94}$$

The derivation of (6.56) was based on (6.53) alone, hence (6.56) holds for $\overline{W}_1 = \overline{W}^{(q_1)}$, too:

$$\left| \overline{W}_1 - 2^{-q_1}\overline{Z}\Delta^{-2}\overline{D}^{-1}\overline{Z}^* \right| \leqq .91\mu\alpha. \tag{6.95}$$

(6.93)–(6.95) are the precise equivalents of (6.60)–(6.62). Therefore the entire argument of §6.8 can be repeated unchanged, and we obtain the equivalent of (6.65″):

$$\left| 2^{q_1}\overline{A}\overline{W}_1 - I \right| \leqq (1.98 + 10.28\lambda)\alpha. \tag{6.96}$$

In addition to this we know now that $\overline{A}_I^* \times \overline{W}_1$ is a properly formed digital matrix.

6.11. **Continuation. The estimates connected with the inverse of $\overline{A}_I$.** We are now able to effect the final estimates of the general case. The relevant auxiliary estimates are (6.75), (6.76.a), (6.76.b), (6.77), (6.93), (6.96), as well as the two following ones: By (3.31.a)

$$\left| \overline{A}_I^* \times \overline{W}_1 - \overline{A}_I\overline{W}_1 \right| \leqq n^{2}\beta^{-s}/2 = \mu_I^{2}\alpha_I/2. \tag{6.97}$$

By (6.96)

$$\left| \overline{A} \right|_l \left| 2^{q_1}\overline{W}_1 \right| \leqq 1 + (1.98 + 10.28\lambda)\alpha,$$
$$\left| \overline{A} \right|_l = \mu$$

hence

$$\left| 2^{q_1}\overline{W}_1 \right| \leqq \frac{1 + (1.98 + 10.28\lambda)\alpha}{\mu},$$

and by (6.76.a), (6.76.b), (6.77) this is less than or equal to

$$\frac{1 + (1.98 + 10.28(1 + \alpha_I/2)\lambda_I^{2})\alpha_I/(1 - \alpha_I/2)}{\mu_I^2(1 - \alpha_I/2)}.$$

Since $\alpha_I \leqq .1$, this is less than or equal to

$$\frac{1.20 + 1.20\lambda_I^{2}}{\mu_I^2},$$

that is,

$$\left| 2^{q_1}\overline{W}_1 \right| \leqq \frac{1.20 + 1.20\lambda_I^{2}}{\mu_I^2}. \tag{6.98}$$

These being understood, the procedure is as follows:

Put

$$\overline{S} = \overline{A}_I^* \times \overline{W}_1. \tag{6.99}$$

($\overline{S}$ is digital, cf. the end of §6.10.)

Owing to

$$\begin{aligned} 2^{q_1}\overline{A}_I\overline{S} - I &= 2^{q_1}\overline{A}_I(\overline{A}_I^* \times \overline{W}_1) - I \\ &= 2^{q_1}\overline{A}_I(\overline{A}_I^* \times \overline{W}_1 - \overline{A}_I^*\overline{W}_1) - (\overline{A} - \overline{A}_I\overline{A}_I^*)\cdot 2^{q_1}\overline{W}_1 \\ &\quad + (2^{q_1}\overline{A}\,\overline{W}_1 - I), \end{aligned}$$

we have

$$\begin{aligned} |\, 2^{q_1}\overline{A}_I\overline{S} - I \,| \leqq 2^{q_1} |\, \overline{A}_I \,| \,|\, \overline{A}_I^* \times \overline{W}_1 - \overline{A}_I^*\overline{W}_1 \,| \\ + |\, \overline{A} - \overline{A}_I\overline{A}_I^* \,| \,|\, 2^{q_1}\overline{W}_1 \,| \\ + |\, 2^{q_1}\overline{A}\,\overline{W}_1 - I \,|. \end{aligned} \tag{6.100}$$

By (6.93), (6.97) the first term of the right-hand side is less than or equal to

$$\frac{8}{1 - 1.78\alpha}\,\frac{1}{\mu}\cdot\lambda_I\cdot\frac{1}{2}\,\mu_I^2\alpha_I,$$

and by (6.77), (6.76.b) this is less than or equal to

$$4\,\frac{1}{1 - 1.78\alpha_I/(1 - \alpha_I/2)}\,\frac{1}{1 - \alpha_I/2}\,\lambda_I\alpha_I,$$

and, since $\alpha_I \leqq .1$, is less than or equal to

$$5.20\lambda_I\alpha_I.$$

By (6.75), (6.98) the second term is less than or equal to

$$\frac{1}{2}\,\mu_I^2\alpha_I\cdot\frac{1.20 + 1.20\lambda_I^2}{\mu_I^2} = (.60 + .60\lambda_I^2)\alpha_I.$$

By (6.96) the third term is less than or equal to

$$(1.98 + 10.28\lambda)\alpha,$$

and by (6.76.a), (6.77) this is less than or equal to

$$\left(1.98 + 10.28\left(1 + \frac{\alpha_I}{2}\right)\lambda_I^2\right)\frac{\alpha_I}{1 - \alpha_I/2}\,\cdot$$

Since $\alpha_I \leqq .1$, this is less than or equal to

$$(2.09 + 11.35\lambda_I^2)\alpha_I.$$

Summing up, and using (6.100), we obtain:

$$(6.101) \qquad \left| 2^{q_1}\overline{A}_I\overline{S} - I \right| \leqq (2.69 + 5.20\lambda_I + 11.95\lambda_I^2)\alpha_I.$$

Since $\lambda_I \leqq (1+\lambda_I^2)/2$, we can simplify (6.101) to

$$(6.102) \qquad \left| 2^{q_1}\overline{A}_I\overline{S} - I \right| \leqq (5.29 + 14.55\lambda_I^2)\alpha_I.$$

Chapter VII. Evaluation of the results

7.1. **Need for a concluding analysis and evaluation.** Our final result is stated in (6.65″) for (nonsingular and) definite matrices $\overline{A}$ and in (6.102) for (nonsingular) general matrices $\overline{A}_I$. These statements and the considerations which led up to them form a logically complete whole. Nevertheless a concluding analysis and evaluation of these results, including a connected restatement of their underlying assumptions and of their constituent computations and discriminations, is definitely called for. Indeed, both the assumptions and the computations are dispersed over the length of Chapter VI and are not easy to visualise in their entirety; furthermore the assumptions were repeatedly modified, merged and rearranged. In addition, and this is more important, there entered into the procedure various quantities and properties which cannot be supposed to be known when the problem of inverting a matrix $\overline{A}$ or $\overline{A}_I$ comes up: For some of these the determination involves problems which are at least as difficult as that of inverting $\overline{A}$ or $\overline{A}_I$, and may even be in themselves closely connected or nearly equivalent to that inverting. Examples of such quantities or properties are: $|\overline{A}| = \lambda$, $|\overline{A}|_l = \mu$, $|\overline{A}_I| = \lambda_I$, $|\overline{A}_I|_l = \mu_I$, the nonsingularity of $\overline{A}$ or of $\overline{A}_I$, the definiteness of $\overline{A}$. Indeed, the basic quantities $\alpha = n^2\beta^{-s}/\mu$ and $\alpha_I = n^2\beta^{-s}/\mu_I^2$ belong to this category, and with them the final estimates (6.65″) and (6.102) and their preliminary conditions (6.67) (or (6.78)) and (6.78′).

We shall clarify these matters, and show that our procedure is actually self-consistent and leads directly to those types of results that one can reasonably desire for a problem like ours.

In connection with this we shall also estimate the amount of computation work that our procedure involves, and say something about the standards by which this amount may be judged.

7.2. **Restatement of the conditions affecting $\overline{A}$ and $\overline{A}_I$: $(\mathcal{A})-(\mathcal{D})$.** We assume, as we did throughout Chapter VI, that $n \geqq 10$. Indeed, for smaller values of n the problem of inverting a matrix hardly justifies this thorough analysis.

Let us now consider the hypotheses concerning $\bar{A}$ and $\bar{A}_I$.

First, both are introduced as digital matrices. This secures automatically that all their elements lie in $-1, 1$. This implies, of course, (6.23.b) for $\bar{A}$. In the case of $\bar{A}_I$ we need more: (6.85), (6.91), that is,

$$(7.1_I.\text{a}) \qquad \sum_{j=1}^{n} \bar{a}_{I,ij} \times \bar{a}_{I,ij} \leqq .99 \qquad (i = 1, \cdots, n),$$

$$(7.1_I.\text{b}) \qquad \sum_{i=1}^{n} \bar{a}_{I,ij} \times \bar{a}_{I,ij} \leqq .99 \qquad (j = 1, \cdots, n).$$

Second, nonsingularity is required for both $\bar{A}$ and $\bar{A}_I$, but this means $\mu \neq 0$ and $\mu_I \neq 0$, which is obviously subsumed in (6.67) (or (6.78)) and (6.78′). We restate, however, these latter conditions:

$$(7.2) \qquad \alpha \leqq .1, \qquad \text{that is,} \quad \mu \geqq 10n^2\beta^{-s},$$

$$(7.2_I) \qquad \alpha_I \leqq .095, \quad \text{that is,} \quad \mu_I^2 \geqq 10.5n^2\beta^{-s}.$$

Thus (6.23.c) (cf. the remark after (6.67)) and part of (6.23.a) for $\bar{A}$· are taken care of.

Third, $\bar{A}$ has to be definite, which covers the residual part of (6.23.a).

This is a complete list of our requirements. We restate it.

($\mathcal{A}$) $\bar{A}$ and $\bar{A}_I$ are digital.

($\mathcal{B}$) $\bar{A}_I$ fulfills (7.1_I.a), (7.1_I.b).

($\mathcal{C}$) $\bar{A}$ and $\bar{A}_I$ fulfill (7.2) and (7.2_I), respectively.

($\mathcal{D}$) $\bar{A}$ is definite.

($\mathcal{A}$), ($\mathcal{B}$) are explicit conditions, of which ($\mathcal{A}$) is automatic and ($\mathcal{B}$) immediately verifiable by digital computation. ($\mathcal{C}$), ($\mathcal{D}$), on the other hand, represent the difficult type to which we referred in 7.1.

It is desirable to say a few things in connection with ($\mathcal{A}$), ($\mathcal{B}$) before we begin the analysis of ($\mathcal{C}$), ($\mathcal{D}$).

7.3. **Discussion of ($\mathcal{A}$), ($\mathcal{B}$): Scaling of $\bar{A}$ and $\bar{A}_I$.** We noted already that ($\mathcal{A}$) is automatically fulfilled. ($\mathcal{B}$) can be satisfied by an appropriate "scaling down" of $\bar{A}_I$, for example, by applying the operation $\div 2^p$ with a suitable p ($=0, 1, 2, \cdots$) to all its elements.

On the other hand, we may if necessary "scale up" $\bar{A}$ or $\bar{A}_I$, for example, by multiplying it by $2^{p'}$ with a suitable p'($=0, 1, 2, \cdots$). In the case of $\bar{A}$, by choosing p' maximal without violating ($\mathcal{A}$), we can make $\text{Max}_{i,j=1,\cdots,n} |\bar{a}_{ij}|$ greater than or equal to one-half its permissible maximum, that is,

$$(7.3) \qquad \frac{1}{2} \leqq \operatorname*{Max}_{i,j=1,\cdots,n} |\bar{a}_{ij}| \leqq 1.$$

In the case of $\overline{A}_I$: if no p was needed (that is, if $p=0$), we can choose p' maximal without violating $(\mathcal{A})$, $(\mathcal{B})$; or, if p was needed (that is, if $p=0$ conflicts with $(\mathcal{B})$), we can choose p minimal without violating $(\mathcal{B})$ and omit p' (that is, put $p'=0$). This makes $\operatorname{Max}_{j=1,\cdots,n \text{ or } i=1,\cdots,n}(\sum_{i=1}^n \bar{a}_{I,ij}\times\bar{a}_{I,ij}, \sum_{j=1}^n \bar{a}_{I,ij}\times\bar{a}_{I,ij})$ greater than or equal to one-quarter its permissible maximum, that is,

$$(7.3_I) \qquad \frac{.99}{4} \leqq \operatorname*{Max}_{j=1,\cdots,n \text{ or } i=1,\cdots,n} \left(\sum_{i=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij}, \sum_{j=1}^n \bar{a}_{I,ij} \times \bar{a}_{I,ij} \right)$$
$$\leqq .99.$$

We assume that these scaling operations have been effected, so that we have (7.3) and (7.3_I) for $\overline{A}$ and $\overline{A}_I$, respectively.

(7.3) implies by (3.17.b) (first half)

$$(7.4) \qquad \lambda \geqq 1/2.$$

From (7.3_I), on the other hand, we can infer this. $\sum_{i=1}^n \bar{a}_{I,ij}\times\bar{a}_{I,ij}$ and $\sum_{i=1}^n \bar{a}_{I,ij}^2$ differ by not more than $n\beta^{-s}/2=\mu_I^2\alpha_I/2n$. Since we shall assume $\alpha_I \leqq .1$, we can argue, as we did at the end of 6.9, that this quantity is not greater than 1/200. Hence

$$\sum_{j=1}^n \bar{a}_{I,ij}^2 \geqq .24.$$

Now since $A_I(I^{\{j\}})=A_I^{\{j\}}$ and $|I^{\{j\}}|=1$, so

$$\lambda_I^2 \geqq |A_I^{\{j\}}|^2 = \sum_{j=1}^n \bar{a}_{I,ij}^2 \geqq .24,$$

that is,

$$(7.4_I) \qquad \lambda_I \geqq .49.$$

We sum up:

$(\mathcal{A})$, $(\mathcal{B})$ can and will be satisfied, indeed strengthened to (7.3) and (7.3_I), by scaling $\overline{A}$ and $\overline{A}_I$ by appropriate powers of 2.

7.4. **Discussion of $(\mathcal{C})$: Approximate inverse, approximate singularity.** Let us now consider $(\mathcal{C})$.

Whether $(\mathcal{C})$ is fulfilled or not cannot be decided in advance by any direct method, or to be more precise, it constitutes a problem that is rather more difficult than the inverting of $\overline{A}$ or of $\overline{A}_I$. Accordingly, we do not propose to decide this in general. What we do propose to do instead is this:

Given $\overline{A}$ or $\overline{A}_I$, we wish to obtain an approximate inverse of $\overline{A}$ (or $\overline{A}_I$) by computational methods. Now it is clear that the solution of this problem cannot consist of furnishing such an (approximate) inverse, since $\overline{A}$ (or $\overline{A}_I$) may not have any inverse, that is, since it may be singular. Whether $\overline{A}$ (or $\overline{A}_I$) is singular or not is in general (that is, disregarding certain special situations) a problem of about the same character and depth as the finding of an (approximate) inverse (if one exists). Consequently the proper formulation of our problem is not this: "Find an (approximate) inverse of $\overline{A}$ (or $\overline{A}_I$)," but rather this: "Either find an (approximate) inverse of $\overline{A}$ (or $\overline{A}_I$), or guarantee that none exists."

Let us consider each one of these two alternatives more closely.

An approximate inverse of a matrix $\overline{P}$ might be defined as one which lies close to the exact inverse $\overline{P}^{-1}$. From the point of view of numerical procedure it seems more appropriate, however, to interpret it as the inverse P'^{-1} of a matrix P' that lies close to $\overline{P}$ that is, to permit an uncertainty of, say, ϵ in every element of $\overline{P}$. Thus we mean by an approximate inverse of $\overline{P}$ a matrix $Q = P'^{-1}$, where all elements of $\overline{P} - P'$ have absolute values not greater than ϵ.[33]

The nonexistence of an approximate inverse of a matrix should now be interpreted in the same spirit. From the point of view of numerical procedure it seems appropriate to interpret it as meaning that, with the uncertainty ϵ which affects every element of $\overline{P}$, $\overline{P}$ is not distinguishable from a singular matrix. That is, that there exists a singular matrix P'' such that all elements of $\overline{P} - P''$ have absolute values not greater than ϵ. (Cf. again footnote 33 above.)

We can now correlate our results concerning $\overline{A}$ and $\overline{A}_I$ with $(\mathcal{C})$: We had, for $\overline{A}$, (6.65'') based on (7.2) (that is, (6.67) or (6.78)), and, for $\overline{A}_I$, (6.102) based on (7.2$_I$) (that is, (6.78')). The conditions (7.2) and (7.2$_I$) correspond, of course, to $(\mathcal{C})$.

We restate (6.65'') and (6.102):

$$(7.5) \qquad \left| 2^{q_0}\overline{A}\,\overline{W}_0 - I \right| \leqq (1.98 + 10.28\lambda)n^2\beta^{-s}/\mu,$$

$$(7.5_I) \qquad \left| 2^{q_1}\overline{A}_I\overline{S} - I \right| \leqq (5.29 + 14.55\lambda_I^2)n^2\beta^{-s}/\mu_I^2,$$

respectively. By (7.4), (7.4$_I$) these imply

[33] This corresponds, of course, to the source of errors (B) in §1.1. As we pointed out before, the effects of (B) are not the subject of this paper. It seems nevertheless reasonable to take (B) into consideration at this stage, when we analyse what concept and what degree of approximation is to be viewed as significant.

$$(7.5') \qquad |\, 2^{q_0}\bar{A}\bar{W}_0 - I \,| \leqq 14.24(\lambda/\mu)n^2\beta^{-s},$$

$$(7.5'_I) \qquad |\, 2^{q_1}\bar{A}_I\bar{S} - I \,| \leqq 36.58(\lambda_I^2/\mu_I^2)n^2\beta^{-s},$$

respectively. (7.2) and (7.2$_I$) are (sufficient) conditions for the validity of these relations. We restate instead their negations, which are alternatives to the relations in question. They are:

$$(7.6) \qquad \mu < 10n^2\beta^{-s},$$

$$(7.6_I) \qquad \mu_I^2 < 10.5n^2\beta^{-s},$$

respectively.

Thus we have either (7.5′) or (7.6) for $\bar{A}$, and either (7.5$'_I$) or (7.6$_I$) for $\bar{A}_I$. Now (7.5′), (7.5$'_I$) on one hand and (7.6), (7.6$_I$) on the other correspond just to the two alternatives mentioned above:

(7.5′) and (7.5$'_I$) express that $2^{q_0}\bar{W}_0$ and $2^{q_1}\bar{S}$ are approximate inverses of $\bar{A}$ and $\bar{A}_I$, respectively. (7.6) and (7.6$_I$) express that $\bar{A}$ and $\bar{A}_1$, respectively, are approximately singular. We leave the working out of the details, which can be prosecuted in several different ways, to the reader.

7.5. **Discussion of ($\mathcal{D}$): Approximate definiteness.** We come finally to ($\mathcal{D}$).

Whether ($\mathcal{D}$) is fulfilled or not, that is, whether $\bar{A}$ is definite or not, is again difficult to ascertain. In this case, however, the situation is somewhat different from what it was in the preceding ones.

($\mathcal{D}$) arises for $\bar{A}$ only. Indeed, it is the extra condition by which the inverting of $\bar{A}$ is distinguished from the inverting of $\bar{A}_I$, that is, which justifies the use of the more favorable estimates (7.5′), (7.6) that apply to the former, instead of the less favorable estimates (7.5$'_I$), (7.6$_I$) that apply to the latter. It states that $\bar{A}$ is definite.

One will therefore let the need for ($\mathcal{D}$) arise, that is, want to use the $\bar{A}$-method, only when it is known a priori that ($\mathcal{D}$) is fulfilled, that is, that $\bar{A}$ is definite; that is, only when $\bar{A}$ originates in procedures which are known to produce definite matrices only.[34]

This might seem to dispose of ($\mathcal{D}$), but there is one minor observation that might be made profitably:

$\bar{A}$ will have been obtained by numerical procedures, which are affected by (round off) errors. In spite of this we can assume that $\bar{A}$ is symmetric, but it need not be definite, only approximately definite. That is, there will be an estimate, by virtue of which this can be asserted: For a suitable definite matrix A' all elements of $\bar{A}-A'$ have

[34] For example, when $\bar{A}$ is a correlation matrix.

absolute values, say not greater than ϵ. (This may be interpreted as a violation of the principle stated at the end of (d) in §2.2.)

This does not, of course, guarantee that $\bar{A}$ is definite. It does, however, imply this:

(7.7) If $\bar{A}$ is not definite, then $\mu \leqq n\epsilon$.

Indeed: Assume that $\bar{A}$ is not definite. It is assumed to be symmetric, hence by (3.21.a) it has a proper value $\lambda_i < 0$. Since λ_i is a proper value, there exists a $\xi \neq 0$ with

$$\bar{A}\xi = \lambda_i \xi. \tag{7.8}$$

Hence $(\bar{A}\xi, \xi) = \lambda_i |\xi|^2 < 0$, that is,

$$(\bar{A}\xi, \xi) < 0. \tag{7.9}$$

Since A' is definite, therefore

$$(A'\xi, \xi) \geqq 0. \tag{7.10}$$

By (3.17.b) (second half) $|\bar{A} - A'| \leqq n\epsilon$ hence by (3.10)

$$|((\bar{A} - A')\xi, \xi)| \leqq n\epsilon |\xi|^2. \tag{7.11}$$

(7.9), (7.10), (7.11) imply together

$$- n\epsilon |\xi|^2 \leqq (\bar{A}\xi, \xi) \leqq 0. \tag{7.12}$$

Since $(\bar{A}\xi, \xi) = \lambda_i |\xi|^2$, (7.12) implies $-n\epsilon |\xi|^2 \leqq \lambda_i |\xi|^2 \leqq 0$, hence

$$|\lambda_i| \leqq n\epsilon. \tag{7.13}$$

Now (7.8), (7.13) gives $|\bar{A}\xi| \leqq n\epsilon |\xi|$, and therefore $\mu = |\bar{A}|_l \leqq n\epsilon$, as desired.

The significance of (7.7) is that it produces for ($\mathcal{D}$) the same type of alternative which we obtained, and found satisfactory, in the last part of §7.4 for ($\mathcal{C}$). Indeed, (7.7) guarantees that $\bar{A}$ is either definite, that is, ($\mathcal{D}$) is fulfilled, or that

$$\mu \leqq n\epsilon \tag{7.7'}$$

and (7.7') is exactly of the same type as the alternative conditions (7.6) and (7.6_I) in the part of §7.4 referred to.

7.6. **Restatement of the computational prescriptions. Digital character of all numbers that have to be formed.** We can sum up our conclusions reached so far as follows:

$\bar{A}$ and $\bar{A}_I$ must be scaled by an appropriate power of 2 as indicated in §7.3, that is, according to (7.3) and (7.3_I), respectively. If $\bar{A}$ is to be used, we assume in addition that it is symmetric and approximately definite (in the sense of §7.5, within a termwise error of, say,

ϵ). Our computational prescriptions then furnished the matrices $2^{q_0}\overline{W}_0$ and $2^{q_1}\overline{W}_1$, such that either (7.5) or (7.6) or (7.7′) holds in the case of $\overline{A}$, and either (7.5$_I$) or (7.6$_I$) holds in the case of $\overline{A}_I$. (7.5) and (7.5$_I$) mean that we found an approximate inverse; (7.6) or (7.7′), and (7.6$_I$) mean that the matrix was not inverted, because we found it to be approximately singular.

To this the following further remarks should be added:

The computational prescriptions to which we referred above are:

In the case of $\overline{A}$: Form the $\bar{a}_{ij}^{(k)}$ $(k=1, \cdots, n;\ i, j=k, \cdots, n)$ according to (6.3), (6.4). From these obtain the $\bar{d}_i$ $(i=1, \cdots, n)$ by (6.30) and the $\bar{b}_{ij}$ $(i, j=1, \cdots, n)$ by (6.31). From these obtain the $p_{ij}, \bar{y}_{ij}, \bar{z}_{ij}$ $(i, j=1, \cdots, n)$ by (6.35). Then form the r_j $(j=1, \cdots, n)$ by (6.51). From all these form the $\bar{e}_j^{(q)}$ $(j=1, \cdots, n)$ by (6.53) and the $\bar{w}_{ij}^{(q)}$ $(i, j=1, \cdots, n)$ by (6.55), obtaining $q=q_0$ from (6.59) (that is, with the help of (6.53), (6.57′.a)). Finally put $\overline{W}_0=(\bar{w}_{ij}^{(q_0)})$.

In the case of $\overline{A}_I$: Form $\overline{A}$ by (6.69). Then proceed exactly as in the case of $\overline{A}$ above, with this exception: Instead of $q=q_0$ obtain $q=q_1$ from (6.92) (that is, with the help of (6.53), (6.57′.b)). Then form $\overline{S}$ by (6.99).

All these constructions were carried out and discussed in Chapter VI. We also showed in the course of those discussions that all the numbers to which we referred above, as well as all the intermediate ones which occur in their constructions, are properly formed digital numbers, and, in particular, lie in $-1, 1$ (except q_0, 2^{q_0} and q_1, 2^{q_1}, cf. below). This depended however on our assumptions concerning $\overline{A}$ and $\overline{A}_I$, which were summarized in $(\mathcal{A})-(\mathcal{D})$ in §7.2. Now $(\mathcal{A})$-$(\mathcal{D})$ are either contained in the assumptions made at the beginning of the present section, or expressed by the alternative possibilities (7.5), (7.6), (7.7′) and (7.5$_I$), (7.6$_I$) enumerated there. We can therefore assert this:

Either all the constructions that we enumerated above produce only digital numbers (including all the intermediate ones which occur in these constructions), which are properly formed, and, in particular, lie in $-1, 1$; or one of the alternative conditions must hold: (7.6) or (7.7′) for $\overline{A}$, (7.6$_I$) for $\overline{A}_I$.

We conclude this section by noting: The approximate inverses of $\overline{A}$ and $\overline{A}_I$ are, by (7.5) and (7.5$_I$), $2^{q_0}\overline{W}_0$ and $2^{q_1}\overline{S}$, respectively. $\overline{W}_0$ and $\overline{S}$ are digital matrices, but $2^{q_0}\overline{W}_0$ and $2^{q_1}\overline{S}$ need not be: Their elements need, of course, not lie in $-1, 1$. This is clearly unavoidable for an approximate inverse. Since we want to use only digital numbers, $2^{q_0}\overline{W}_0$, $2^{q_1}\overline{S}$ should be formed and recorded by keeping 2^{q_0}, 2^{q_1} and $\overline{W}_0$, $\overline{S}$ separately. $\overline{W}_0$, $\overline{S}$ are digital matrices, so they offer no difficulty. 2^{q_0}, 2^{q_1} are not digital, but we may form and record the digital $2^{q_0}\beta^{-s}$, $2^{q_1}\beta^{-s}$ or the equally digital $q_0\beta^{-s}$, $q_1\beta^{-s}$ instead. All the computations to which we referred at the beginning of this section deal, however, with digital numbers only, as we pointed out further above.

7.7. **Number of arithmetical operations involved.** It may be of interest to count the arithmetical operations that are involved in our computational prescriptions, as stated at the beginning of §7.6. Since most computations are dominated by the multiplications and divisions they contain, we shall only count these.

Referring back to the enumeration at the beginning of §7.6, we find:

In the case of $\bar{A}$:

	Multiplications	*Divisions*
(6.3)	$n(n+1)(n+2)/6$[35]	$n(n+1)/2$[35]
(6.4)	—	—
(6.30)	—	—
(6.31)	—	known from (6.3)
(6.35)	$(n-1)n(n+1)/6$	—[35]
(6.51)	—[35]	—
(6.53)	—[35]	n[35]
(6.55)	$n(n+1)(n+2)/6$	—
(6.59)	—	—
Total	$(n^3+2n^2+n)/2$	$(n^2+3n)/2$
Additional in the case of $\bar{A}_I$:		
From $\bar{A}$	$(n^3+2n^2+n)/2$	$(n^2+3n)/2$
(6.69)	$n^2(n+1)/2$	—
(6.99)	n^3	—
	$(4n^3+3n^2+n)/2$	$(n^2+3n)/2$

[35] We do not omit trivial multiplications like $\bar{a}\times 1$ and trivial divisions like $\bar{a}\div\bar{a}$, since their numbers are irrelevant compared to the whole. We do, however, omit scaling operations $2^s\bar{a}$ and $\bar{a}\div 2^s$, since these are likely to be effected in simpler ways than by full-sized multiplications and divisions. Besides their numbers, too, are irrelevant.

Since we are interested in large values of n (at least $n \geqq 10$), we can use the asymptotic forms: For $\bar{A}$: $n^3/2$ multiplications, $n^2/2$ divisions. For $\bar{A}_I$: $2n^3$ multiplications, $n^2/2$ divisions. The divisions are presumably irrelevant in comparison with the multiplications. Hence our final result is: For $\bar{A}$: $n^3/2$ multiplications, for $\bar{A}_I$: $2n^3$ multiplications.

Note that an ordinary matrix multiplication consists of n^3 (number) multiplications. Hence the $\bar{A}$-method of inversion is comparable to half a matrix multiplication, while the $\bar{A}_I$-method of inversion is comparable to two matrix multiplications. It is a priori plausible that an inversion should be a more complicated operation than a multiplication. Thus we have in the above a quantitative measure of the high efficiency of inverting a matrix by elimination.

7.8. **Numerical estimates of precision.** In conclusion, it seems desirable to make some effective numerical evaluations of our estimates: Of (7.5′), (7.6) for $\bar{A}$ (definite case) and of ($7.5_I'$), (7.6_I) for $\bar{A}_I$ (general case).

Since the intervening quantities λ, μ and λ_I, μ_I are not known in general (or even, in most special cases, in advance), this cannot be done without some additional hypotheses.

We shall introduce such a hypothesis in the form of the statistical results of V. Bargmann,referred to in footnote 24. According to these, we can assert for a "random" matrix $\bar{A}_I$ (which we may assume to have been scaled in the sense of §7.3, that is, according to (7.3_I)) that λ_I, μ_I have with a probability ~ 1 the following sizes:

$$\lambda_I \sim n^{1/2}, \tag{7.14_I.a}$$

$$\mu_I \sim 1/n^{1/2}, \tag{7.14_I.b}$$

and hence

$$\lambda_I/\mu_I \sim n. \tag{7.14_I.c}$$

In order to reduce the probabilistic uncertainties to reasonably safe levels, we allow a factor 10 in excess of each estimate (7.14_I.a)–(7.14_I.c):

$$\lambda_I \leqq 10n^{1/2}, \tag{$7.14_I'$.a}$$

$$\mu_I \geqq 1/10n^{1/2}, \tag{$7.14_I'$.b}$$

$$\lambda_I/\mu_I \leqq 10n. \tag{$7.14_I'$.c}$$

For the definite (or approximately definite) matrices it seems very unreasonable to introduce any direct "randomness," since they are usually secondary, originating in other, general matrices. It does not seem unreasonable to estimate their λ, μ about as the squares of the above λ_I, μ_I; but we shall not attempt to analyze this hypothesis here any further. If it is accepted, we obtain from (7.14$_I'$.a)–(7.14$_I'$.c):

(7.14′.a) $$\lambda \leqq 100n,$$

(7.14′.b) $$\mu \geqq 1/100n,$$

(7.14′.c) $$\lambda/\mu \leqq 100n^2.$$

Now both estimates (7.6), (7.6$_I$) imply (approximately):

(7.15) $$n \geqq .1\beta^{s/3}.$$

That is, an approximate inverse will usually be found if n does not fulfill (7.15), that is, if

(7.15′) $$n < .1\beta^{s/3}.$$

Furthermore, the right-hand sides of both estimates (7.5′), (7.5$_I'$) become (approximately):

(7.16) $$\leqq 2{,}000n^4\beta^{-s}.$$

(The factor 2,000 should actually have been ~1,500 in the case of (7.5′), and ~3,500 in the case of (7.5$_I'$). We replaced these by the common (and, for the second alternative, low) value 2,000 in order to simplify matters. This change is irrelevant, because in passing to (7.16′) a fourth root is being extracted. Besides, the estimate by which we passed from (7.5$_I$) to (7.5$_I'$) was very generous, because λ_I is likely to be essentially larger than indicated by (7.4$_I$).)

Hence this is less than 1 if

(7.16′) $$n < .15\beta^{s/4},$$

that is, an approximate inverse will usually be found if n fulfills (7.16′). Its (relative) precision is measured by the fourth power of the factor by which n is below the limit of (7.16′), or by the first power of the factor by which β^s is above it.

(7.16′) is more stringent than (7.15′) if $.15\beta^{s/4} \leqq .1\beta^{s/3}$, which is equivalent to $\beta^{s/12} \geqq 1.5$, $\beta^s \geqq 1.5^{12} \approx 130$. This is the case for all precisions at which a calculation of the type that we consider is likely to be carried out. (It is hardly conceivable that there should not be $\beta^s \geqq 10^6$.) We may say therefore:

(7.16′) is the critical condition, regarding the (relative) precision of the approximate inverse, cf. the remark after (7.16′).

Let us now consider some plausible precisions:

(7.17.a) $$\beta^s \sim 10^8 \sim 2^{27},$$

(7.17.b) $$\beta^s \sim 10^{10} \sim 2^{33},$$

(7.17.c) $$\beta^s \sim 10^{12} \sim 2^{40}.$$

(7.16′) becomes:

(7.18.a) $$n < 15,$$

(7.18.b) $$n < 50,$$

(7.18.c) $$n < 150,$$

respectively. As we saw in §7.7, these n correspond to maximally $\sim n^3 \sim 3{,}500$; 120,000; 3,500,000 multiplications. This might provide a basis for estimating what degrees of precision are called for in this problem by various possible types of procedures and equipment.

INSTITUTE FOR ADVANCED STUDY