# CONSERVING CONFLUENCE CURBS ILL-CONDITION

W. Kahan

California University

Prepared for:

Office of Naval Research

4 August 1972

# COMPUTER SCIENCE

UNIVERSITY OF CALIFORNIA

BERKELEY



## TECHNICAL REPORT

Security Classification

**DOCUMENT CONTROL DATA – R & D**

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

| | |
|---|---|
| 1. ORIGINATING ACTIVITY (Corporate author)<br>COMPUTER SCIENCE DEPARTMENT<br>UNIVERSITY OF CALIFORNIA, BERKELEY<br>BERKELEY, CALIFORNIA 94720 | 2a. REPORT SECURITY CLASSIFICATION<br>UNCLASSIFIED<br>2b. GROUP |

3. REPORT TITLE

CONSERVING CONFLUENCE CURBS ILL-CONDITION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

SCIENTIFIC FINAL

5. AUTHOR(S) (First name, middle initial, last name)

W. KAHAN

| 6. REPORT DATE | 7a. TOTAL NO. OF PAGES | 7b. NO. OF REFS |
|---|---|---|
| AUGUST, 1972 | 54 60 | 22 |

| 8a. CONTRACT OR GRANT NO. | 9a. ORIGINATOR'S REPORT NUMBER(S) |
|---|---|
| N00014-69-A-0200-1017<br>b. PROJECT NO.<br>c.<br>d. | 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) |

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited

| 11. SUPPLEMENTARY NOTES | 12. SPONSORING MILITARY ACTIVITY |
|---|---|
| | Mathematics Branch<br>Office of Naval Research<br>Washington, D.C. 20360 |

13. ABSTRACT

Certain problems are ill-conditioned, in the sense that their solutions are hypersensitive to small changes in data, only because a slight change in data could cause those solutions to exhibit singular behaviour associated with various kinds of confluence. For example, an over- or under-determined linear system solved by least-squares can be ill-conditioned only if there exists some small perturbations to its matrix which increase its nullity (i.e. diminish its rank); zeros of a polynomial can be ill-conditioned only if their multiplicities can be increased by very small perturbations of the polynomial's coefficients; eigenvalues of a non-Hermitian matrix can be ill-conditioned only if their algebraic multiplicities can be increased by very small perturbations of the matrix. When perturbations constrained to a small neighbourhood can be further constrained to maximize confluence, i.e. to maximize nullity (minimize rank) or maximize multiplicity, and when that maximized confluence can be increased again only by perturbations far beyond the small neighbourhood, then the slightly perturbed problems exhibit well-conditioned confluent solutions. Beyond these vague statements lie the shadows of numberical methods which may either eliminate ill-condition or, when ill-condition is persistent, illuminate its cause.

DD FORM 1 NOV 68 1473

Security Classification

| 14. KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| Ill-Posed | | | | | | |
| Ill Condition | | | | | | |
| Pseudo-Inverse | | | | | | |
| Zeros of Polynomials | | | | | | |
| Eigerproblems | | | | | | |
| Degenerate Eigenvalues | | | | | | |

COMPUTER SCIENCE

UNIVERSITY OF CALIFORNIA
BERKELEY

# CONSERVING CONFLUENCE CURBS ILL-CONDITION

by

W. Kahan

Technical Report 6



August 4, 1972

# CONSERVING CONFLUENCE CURBS ILL-CONDITION

W. Kahan*

Abstract. Certain problems are ill-conditioned, in the sense that their solutions are hypersensitive to small changes in data, only because a slight change in data could cause those solutions to exhibit singular behaviour associated with various kinds of confluence. For example, an over- or under-determined linear system solved by least-squares can be ill-conditioned only if there exist some small perturbations to its matrix which increase its nullity (i.e. diminish its rank); zeros of a polynomial can be ill-conditioned only if their multiplicities can be increased by very small perturbations of the polynomial's coefficients; eigenvalues of a non-Hermitian matrix can be ill-conditioned only if their algebraic multiplicities can be increased by very small perturbations of the matrix. When perturbations constrained to a small neighbourhood can be further constrained to maximize confluence, i.e. to maximize nullity (minimize rank) or maximize multiplicity, and when that maximized confluence can be increased again only by perturbations far beyond the small neighbourhood, then the slightly perturbed problems exhibit well-conditioned confluent solutions. Beyond these vague statements lie the shadows of numerical methods which may either eliminate ill-condition or, when ill-condition is persistent, illuminate its cause.

---

* Computer Science Department, University of California, Berkeley.
This work was also supported by a grant from the U.S. Office of Naval Research, contract no. N00014-69-A-0200-1017.

# CONSERVING CONFLUENCE CURBS ILL-CONDITION

W. Kahan

August 3, 1972

## Contents

# CONSERVING CONFLUENCE CURBS ILL-CONDITION

W. Kahan

"Mother may I go to swim?"
"Yes, my darling daughter;
Hang your clothes on yon tree limb,
But don't go *near* the water."

Introduction. Numerical calculations generally appear in the form

$$\text{Compute} \quad y \equiv f(x)$$

where $f$ characterizes a class of problems and $x$ represents the particular data. Commonly $f$ is defined implicitly by a set of equations whose coefficients' values constitute $x$, and $y$ is the solution of those equations. The equations are called *ill-conditioned* whenever there exist tiny perturbations $\delta x$ which cause huge changes $\delta y \equiv f(x+\delta x) - f(x)$. To make this notion more precise we imagine $x$ and $y$ to reside in metric spaces -- normed linear spaces are customary -- and define a condition number

$$\gamma \equiv \sup \|\delta y\| / \|\delta x\|$$

where the supremum is taken over all $\delta x$ in some neighbourhood of $x$. Thus, the condition number $\gamma$ is a *Lipschitz* constant; $\|f(x+\delta x) - f(x)\| \leq \gamma \|\delta x\|$. The larger is $\gamma$, the more ill-conditioned is the problem $f$ near $x$. When $\gamma$ is infinite we sometimes say that $f$ is *ill-posed* near $x$, though this term is reserved by some for discontinuous behaviour.

Non-differentiable functions $f$ are so rarely encountered in practice that we might as well exploit the simplification afforded by constraining perturbations $\delta x$ to infinitesimal neighbourhoods. Now

$$f(x+\delta x) - f(x) = (\partial f/\partial x)\delta x$$

wherever the Frechet derivative $\partial f/\partial x$ exists, in which case

$$\gamma = \|\partial f/\partial x\| \quad ;$$

here we use the induced norm for linear operators between two normed linear spaces.

Since $\partial f/\partial x$ is usually differentiable too, it seems natural to guess that an ill-conditioned problem, with $\|\partial f/\partial x\|$ huge, probably has its data $x$ near a place where $\partial f/\partial x$ becomes infinite or fails to exist. The locus of all such places is usually a manifold in $x$'s space, and that manifold is the subject of this paper. Here are three examples:

When $f$ represents solving a system of linear equations $Ay = b$ with square matrix $A$, so each point $x$ in data-space has coordinates $(A,b)$, and when the infinitesimal neighbourhoods are generated by all infinitesimal $(\delta A,\delta b)$ without constraint, then the manifold where $\partial f/\partial x$ becomes infinite consists of just those points $x \sim (A,b)$ with singular $A$ since elsewhere $y = A^{-1}b$ varies by $\delta y = A^{-1}\delta b - A^{-1}(\delta A)A^{-1}b$, a bounded linear function of the infinitesimal perturbation $\delta x \sim (\delta A,\delta b)$. When $f$ represents solving polynomial equations

$$y^n - x_1 y^{n-1} - x_2 y^{n-2} - \cdots - x_{n-1} y - x_n = 0 \quad ,$$

so each point $x \sim (x_1, x_2, \ldots, x_n)$ in data-space is identified with a polynomial $x(y) = y^n - \sum_1^n x_j y^{n-j}$, the manifold where $\partial f/\partial x$ becomes infinite consists of just those polynomials $x$ with some multiple zeros since elsewhere each simple zero $y$ of $x$ varies by $\delta y = \sum_1^n y^{n-j} \delta x_j / x'(y)$. A similar situation arises when $f$ represents solving eigenproblems for square matrices $X$; the eigenvalues and eigenvectors are well-known to be differentiable functions of $X$'s elements only when $X$'s eigenvalues are distinct, so the manifold of interest consists of those matrices $X$ with some multiple eigenvalues.

One might be tempted to assign some pejorative adjective to that manifold on which $\partial f/\partial x$ fails to be finite. (There are precedents; in 1884 Sylvester assigned the word *derogatory* to certain matrices with multiple eigenvalues, and physicists almost universally apply the epithet *degenerate* to eigenvalues whose only flaw is their indistinguishability.) In so far as $f$ is ill-behaved *near* that manifold, the more so as it is approached, the manifold warrants the name *pejorative**. But in the last two examples above $f$ will be found to behave very well *on* the manifold, except as $x$ approaches certain sub-manifold's. More precisely, for almost all $x$ on the pejorative manifold and for all infinitesimally nearby $x+\delta x$ also on that manifold the difference $f(x+\delta x) - f(x)$ is a bounded linear function of $\delta x$, and the bound varies with $x$ on the manifold in such a way that the bound can approach infinity only as $x$ approaches some doubly pejorative sub-manifold on which the same kind of behaviour recurs. That phenomenon is what this paper is about.

---

*Pejorative: from the Latin *pejorare* to make worse.

The paradox, that $f$ can be well-behaved on a manifold in every open neighbourhood of which $f$ is arbitrarily ill-behaved, would be uninteresting but for another property of such pejorative manifolds; they can be characterized ostensibly independently of $f$'s good or ill behaviour. For want of a better term I use the word *confluence* to describe what happens to $f$ on those manifolds. When $f$ represents zeros of polynomials or eigenvalues of matrices the confluence is obvious; some zeros flow together as a polynomial $x$ approaches a pejorative manifold; some eigenvalues flow together as a matrix $X$ approaches a pejorative manifold. Confluence in a linear system is identified with collapse of the range of its matrix as it approaches a pejorative manifold; this manifold in matrix-space is the locus of discontinuities (drops) in the rank function.

Pejorative manifolds are interesting just because they are associated simultaneously with confluence and with an abrupt change from wild mis-behaviour to tame good-behaviour. Consider, for example, a polynomial $x_o$ so constructed as to ensure, in the absence of error, that among its zeros $y_o = f(x_o)$ must be some that are coincident; but because error $\Delta x$ has crept into the data $x_o$ none of the available zeros $y_o + \Delta y = f(x_o+\Delta x)$ are coincident. They may well be nowhere near coincident. Frantic dispersal of perturbed zeros is frequently quite pronounced when $x_o$ is of high degree, and is not surprising when we realize how wildly $f$ must misbehave near a pejorative manifold. Given only $x_o+\Delta x$ and a bound for $\|\Delta x\|$, can we discover a nearby $x_1$ on a pejorative manifold? That $x_1$ will not be unique but,

provided the bound on $\|\Delta x\|$ is small enough to keep $x_1$ well away from a doubly pejorative sub-manifold, we can expect that the multiple zeros among $y_1 = f(x_1)$ will not vary much as $x_1$ runs through those values on the pejorative manifold close to $x_o+\Delta x$. Thus do we substitute a well-conditioned problem $f(x_1)$ for an ostensibly ill-conditioned problem $f(x_o+\Delta x)$. On the other hand, we may discover that $x_o+\Delta x$ is farther from the pejorative manifold than the bound on $\|\Delta x\|$, in which case we infer that something, either the bound or the construction of $x_o$, is wrong (i.e. mistaken).

The properties of pejorative manifolds have many other practical implications but to discuss them here would be premature. First we must verify the foregoing assertions about those properties. Secondly, we should consider how to locate the manifolds computationally; here is where the theory is weak. Only for linear systems do we know how to tell cheaply whether a data-point $x$ is close to or far from a pejorative manifold, and whether there are multiply pejorative sub-manifolds nearby, and where they are. Some of this knowledge is imparted in part I of the paper.

Parts II and III consider polynomials' zeros and matrix eigenproblems respectively. For these problems the simplest pejorative manifolds, corresponding to double zeros and double eigenvalues, are easy enough to locate; but multiply pejorative sub-manifolds are not yet within reach of cheap computation. In particular, we cannot easily tell whether a data point $x$ is far enough from a multiply pejorative sub-manifold that that sub-manifold need not be explored, unless $x$ is very far from every such sub-manifold. Fortunately for our theory, multiply pejorative

sub-manifolds need only rarely be considered; in ordinary language this means that double roots, though rare, are overwhelmingly more common in practice than are roots of higher multiplicity. Consequently, the theory is ripe for exploitation despite its immaturity. The theory's subsequent growth seems likely to depend upon numerical analysts' proficiency with algebraic geometry and metric spaces.

* * * * *

I take pleasure in acknowledging here the assistance and encouragement received, while the foregoing notions were evolving, from several years' discussions with many colleagues and friends. Especially, George Forsythe's continuing interest in those notions considerably stimulated their development. I am indebted too to the organizers of the 5th Gatlinburg Symposium on Numerical Linear Algebra, held at Los Alamos on June 5-10, 1972, for an opportunity to present those notions to a wide audience.

## Part I: The Pseudo-Inverse

The pseudo-inverse $X^+$ of an $m \times n$ matrix $X$ is uniquely defined formally by the familiar equations

$$(+) \quad XX^+X = X \ , \ X^+XX^+ = X^+ \ , \ (X^+X)^* = X^+X \ , \ (XX^+)^* = XX^+ \ ,$$

but a better definition is derived from its principal application, the solution of linear least-squares problems: Given $X$ and an $m$-vector $v$ we seek that $n$-vector $\omega$ which minimizes $\|v - X\omega\|$, and when the minimizing $\omega$ is not unique (as must be the case just when $X$'s columns are linearly dependent) we seek that minimizing $\omega$ with minimal $\|\omega\|$. The vector norm used here is $\|\omega\| \equiv \sqrt{\omega^*\omega}$; we shall also use the induced matrix norm $\|Z\| \equiv \max_{\omega \neq 0} \|Z\omega\|/\|\omega\|$ and the root-sum-squares norm $\|Z\|_2 \equiv \sqrt{\mathrm{tr.}(Z^*Z)}$.

The desired minimizing vector $\omega$ turns out to be $\omega = X^+v$; see R. Penrose (1954,1955). This formula is interesting only when $X$'s columns are linearly dependent or nearly so, since otherwise we could substitute $X^+ = (X^*X)^{-1}X^*$ and ignore the equations $(+)$ above. But just when $X^+$ becomes interesting it becomes numerically exasperating *no matter what method is employed to compute it* because when $X$'s columns are linearly dependent $X^+$ must be a violently discontinuous function of $X$ and hence hypersensitive to small variations, as we shall see.

In what follows we shall discern a nested sequence

$$M_0 \supset M_1 \supset M_2 \supset \cdots$$

of pejorative (for $k \geq 1$) manifolds and sub-manifolds in the space $M_0$ of $m \times n$ matrices $X$; $M_k$ is the manifold of matrices

whose rank does not exceed $\min(m,n) - k$. We shall discover that $X^\dagger$ is a well-behaved function of $X$ provided $X$ is confined to $M_k$ and avoids $M_{k+1}$. More precisely, we shall find that while $X$ and its infinitesimally neighbouring $X + \delta X$ are constrained to $M_k - M_{k+1}$

$$\|X^\dagger\| = 1/\text{'the minimum distance } \|\cdots\| \text{ from } X \text{ to } M_{k+1}\text{'} \quad ,$$

$$\|X^\dagger\|^2 = \sup \|\delta(X^\dagger)\|_2 / \|\delta X\|_2 \quad \text{over} \quad X+\delta X \quad \text{on the same} \quad M_k \quad \text{as} \quad X \ .$$

Some of these discoveries have been seen before, particularly in the works of G.W. Stewart (1969), V. Pereyra (1969), and Golub and Pereyra (1972), whose treatments should be compared with what follows. Finally, we shall consider, given $X$ and a tolerance $\xi > 0$ such that all $X+\Delta X$ with $\|\Delta X\| \leq \xi$ must be regarded as indistinguishable for practical purposes, how to find an approximation $\hat{X}$ indistinguishable from $X$ with the best-behaved $\hat{X}^\dagger$.

Some apparatus is needed. Let us assume $m \geq n$ (otherwise transpose $X$) and denote $X$'s $n$ singular values in order by $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_n \geq 0$. That $\xi_1 = \|X\|$ is well known, as is the fact that $X^\dagger$'s singular values are the re-ordered numbers $\xi_j^\dagger$, where

$$\zeta^\dagger \equiv 1/\zeta \quad \text{except for} \quad 0^\dagger \equiv 0 \quad .$$

Not so well known is the following relation proved by L. Mirsky (1960, theorem 2):

$$\xi_k = \min \|\Delta X\| \quad \text{over} \quad \text{rank}(X+\Delta X) < k \quad .$$

One implication of this relation, to be used later, is that no singular value of $X+\Delta X$ can differ from the correspondingly

numbered singular value of $X$ by more than $\|\Delta X\|$. Another implication obtained via $\|X^\dagger\| = \max_j(\xi_j^+)$ is that

$$\|X^\dagger\| = 1/\min\|\Delta X\| \quad \text{over} \quad \operatorname{rank}(X+\Delta X) < \operatorname{rank}(X) \quad .$$

Consequently, if $X \in M_k$ but $X \notin M_{k+1}$ then

$$\|X^\dagger\| = 1/\min\|\Delta X\| \quad \text{over} \quad X+\Delta X \in M_{k+1} \quad ,$$

which is just what was claimed for $\|X^\dagger\|$ above.

Next we shall exploit a little known formula;

$$X^\dagger - Y^\dagger = -Y^\dagger(X-Y)X^\dagger + (1-Y^\dagger Y)(X-Y)^*X^{\dagger *}X^\dagger + Y^\dagger Y^{\dagger *}(X-Y)^*(1-XX^\dagger) \quad .$$

This formula can be verified by applying the equations (†) above to reduce the right-hand side to its simplest terms. Note that $(1-Y^\dagger Y)$ and $(1-XX^\dagger)$ are orthogonal projectors which annihilate $Y^\dagger$ and $Y^*$, and $X^{\dagger *}$ and $X$ respectively. Consequently we find

$$\begin{aligned}(X^\dagger-Y^\dagger)^*(X^\dagger-Y^\dagger) = {} & X^{\dagger *}(X-Y)^*Y^{\dagger *}Y^\dagger(X-Y)X^\dagger \\ & - X^{\dagger *}(X-Y)^*Y^{\dagger *}Y^\dagger Y^{\dagger *}(X-Y)^*(1-XX^\dagger) \\ & + X^{\dagger *}X^\dagger(X-Y)(1-Y^\dagger Y)(X-Y)^*X^{\dagger *}X^\dagger \\ & - (1-XX^\dagger)(X-Y)Y^\dagger Y^{\dagger *}Y^\dagger(X-Y)X^\dagger \\ & + (1-XX^\dagger)(X-Y)(Y^\dagger Y^{\dagger *})^2(X-Y)^*(1-XX^\dagger) \quad ,\end{aligned}$$

and taking norms yields an important inequality

$$\begin{aligned}\|X^\dagger-Y^\dagger\|^2 &\le \|X-Y\|^2\left(\|X^\dagger\|^4 + \|X^\dagger\|^2\|Y^\dagger\|^2 + 2\|X^\dagger\|\|Y^\dagger\|^3 + \|Y^\dagger\|^4\right) \\ &\le 5\|X-Y\|^2 \max\{\|X^\dagger\|,\|Y^\dagger\|\}^4 \quad .\end{aligned}$$

Now let $Y = X + \Delta X$, and suppose both $X$ and $X+\Delta X$ lie on $M_k$

but not on $M_{k+1}$. As $\Delta X \to 0$ we see that $\|(X+\Delta X)^\dagger\|$ becomes and remains bounded, and then that $\|X^\dagger - (X+\Delta X)^\dagger\| \to 0$. In short, $X^\dagger$ is a continuous function of $X$ on $M_k$ away from $M_{k+1}$. It soon follows that $X^\dagger$ is differentiable too, for we need only set $Y = X + \delta X$, with infinitesimal $\delta X$ constrained to keep $X + \delta X$, like $X$, on $M_k$ away from $M_{k+1}$, to deduce that

$$\delta(X^\dagger) = -X^\dagger(\delta X)X^\dagger + (1-X^\dagger X)(\delta X^*)X^{\dagger *}X^\dagger + X^\dagger X^{\dagger *}(\delta X^*)(1-XX^\dagger) \quad .$$

Next we seek to compute $\sup\|\delta(X^\dagger)\|_2/\|\delta X\|_2$. To this end it is convenient to invoke Autonne's theorem which exhibits $X = P\Lambda Q$ where $P$ is $m \times m$ unitary, $Q$ is $n \times n$ unitary, and $\Lambda$ is $m \times n$ diagonal with the singular values $\xi_j$ on its main diagonal. This singular value decomposition may be computed at modest cost by methods described in Golub and Reinsch (1970), and will be further exploited below. For the present let us partition

$$\Lambda = \begin{pmatrix} \Lambda_o & 0 \\ 0 & 0 \end{pmatrix}$$

in such a way that just $X$'s non-zero singular values $\xi_j$ appear on the diagonal of the square diagonal matrix $\Lambda_o$. Evidently $X^\dagger = Q^* \Lambda^\dagger P^*$ where

$$\Lambda^\dagger = \begin{pmatrix} \Lambda_o^{-1} & 0 \\ 0 & 0 \end{pmatrix}^* \quad .$$

Also $\|X^\dagger\| = \|\Lambda_o^{-1}\|$. Next partition conformally

$$\delta\Lambda \equiv P^*(\delta X)Q^* = \begin{pmatrix} \delta A & \delta B \\ \delta C & \delta D \end{pmatrix} \quad ;$$

by fixing $P$ and $Q$ independently of $\delta X$ we oblige $\delta\Lambda$ to be non-diagonal in general. Since $X+\delta X$ must have the same rank as $X$, $\Lambda+\delta\Lambda$ must have the same rank as $\Lambda$, and this must be the same as the rank of

$$\begin{pmatrix} 1 & 0 \\ -\delta C(\Lambda_o+\delta A)^{-1} & 1 \end{pmatrix} (\Lambda+\delta\Lambda) = \begin{pmatrix} \Lambda_o+\delta A & \delta B \\ 0 & \delta D-\delta C(\Lambda_o+\delta A)^{-1}\delta B \end{pmatrix} .$$

The rank in question is that of $\Lambda_o$, and also of $\Lambda_o+\delta A$ since $\delta A$ is infinitesimal. Therefore we must have

$$\delta D - C(\Lambda_o+\delta A)^{-1}\delta B = 0 \quad ,$$

but this merely says that the infinitesimal $\delta D = 0$. Therefore, infinitesimal perturbations $\delta X$ for which $\Lambda$ and $X+\delta X$ have the same rank must have the form

$$\delta X = P \begin{pmatrix} \delta A & \delta B \\ \delta C & 0 \end{pmatrix} Q \quad .$$

Substitution into the formula above for $\delta(X^\dagger)$ soon leads to the conclusion that

$$\|\delta(X^\dagger)\|_2^2 \leq \|X^\dagger\|^4 \|\delta X\|_2^2$$

with equality possible when $\delta A$, $\delta B$ and $\delta C$ are chosen to have non-zero entries only in rows and columns corresponding to the largest entries in $\Lambda_o^{-1}$. Thus we conclude that pseudo-inversion

can be ill-conditioned with respect to rank-preserving perturbations only if the data-matrix $X$ is very near another of lower rank.

Finally let us discuss how to compute a pseudo-inverse appropriate for a given matrix $X$ when given also a tolerance $\xi > 0$ such that all $X+\Delta X$ with $\|\Delta X\| \leq \xi$ must be regarded as indistinguishable from $X$. Should some of these matrices $X+\Delta X$ have different rank than $X$ there must exist others whose pseudo-inverses differ arbitrarily much among each other. None of those wildly divergent pseudo-inverses can be useful. Instead let us find a matrix $\hat{X} = X+\hat{\Delta X}$ of minimal rank with $\|\hat{\Delta X}\| \leq \xi$. Such a matrix is easily obtained from $\Lambda$ above by annihilating all $\xi_j \leq \xi$; let $\hat{\Lambda}$ denote what results and let $\hat{X} = P\hat{\Lambda}Q$. If $\xi < \xi_n$ then $\Lambda = \hat{\Lambda}$ and $\hat{X} = X$; in this case for all $\Delta X$ with $\|\Delta X\| \leq \xi$ we find that

$$\|(X+\Delta X)^\dagger\| \leq 1/(\xi_n-\xi) \quad \text{and}$$

$$\|(X+\Delta X)^\dagger - X^\dagger\|/\|X^\dagger\| \leq \left(1+\xi_n^2/(\xi_n-\xi)^2\right)^{1/2}\xi/\xi_n \quad .$$

The latter inequality is obtained by substituting $Y = X+\Delta X$ and $X^\dagger X = Y^\dagger Y = 1$ into the formula above for $X^\dagger - Y^\dagger$, and then taking the norm of $(X^\dagger-Y^\dagger)(X^\dagger-Y^\dagger)^*$ with the aid of $\|X^\dagger\| = 1/\xi_n$ and $\|Y^\dagger\| \leq 1/(\xi_n-\xi)$. The point of the inequality is that if $\xi/\xi_n \ll 1$ we may confidently assert that all indistinguishable matrices $X+\Delta X$ have nearly the same pseudo-inverse.

The interesting case occurs when $\xi_{n-k} > \xi \geq \xi_{n+1-k}$ for some $k > 0$. This means that among the matrices $X+\Delta X$ with $\|\Delta X\| \leq \xi$ are some of rank $n-k, n+1-k, \ldots, n$. Every time $X+\Delta X$ changes rank,

$(X+\Delta X)^\dagger$ jumps infinitely violently. But as $X+\Delta X$ runs through matrices on $M_k$ of rank $n-k$ with $\|\Delta X\| \le \xi$, $(X+\Delta X)^\dagger$ varies continuously and

$$\|(X+\Delta X)^\dagger - \hat{X}^\dagger\| \le \sqrt{5}(\xi+\xi_{n+1-k})/(\xi_{n-k}-\xi)^2 \quad .$$

Whenever $\xi/\xi_{n-k} << 1$, the pseudo-inverses of matrices on $M_k$ indistinguishable from $X$ will differ only slightly among each other, although matrices $X+\Delta X$ not on $M_k$ will have huge and wildly varying pseudo-inverses; in this case $\hat{X}^\dagger$ seems to be a reasonable response to the command

$$\text{"Compute } X^\dagger\text{"} \quad .$$

But if $\xi_{n-k}$ is only moderately larger than $\xi$ that command deserves to be questioned.

Another way to appreciate $\hat{X}^\dagger$ when $\xi/\xi_{n-k} << 1 \le \xi/\xi_{n-k+1}$ is geometrical. Consider the image $P$ under the operation $\dagger$ of the ball $B$ of matrices $X+\Delta X$ with $\|\Delta X\| \le \xi$; i.e. consider the set $P$ of pseudo-inverses of all matrices in that ball $B$. $P$ has two disconnected components $P_o$ and $P_\infty$. $P_o$ consists of the pseudo-inverses of matrices in $B \cap M_k$, and looks like a small bent coin roughly centered on $\hat{X}^\dagger$; all the points $(X+\Delta X)^\dagger$ in $P_o$ are close to $\hat{X}^\dagger$ (see the inequality above) and have modest norms not exceeding $1/(\xi_{n-k}-\xi)$. The other component $P_\infty$ has tentacles which reach to $\infty$ starting from far-out points $(X+\Delta X)^\dagger$ which must satisfy $\|(X+\Delta X)^\dagger\| \ge 1/(\xi+\xi_{n-k+1}) >> 1/(\xi_{n-k}-\xi)$.

## Part II: Zeros of Polynomials

Many numerical analysts suffer from a misconception that multiple roots are infinitely more ill-conditioned than simple roots. Actually, a multiple root behaves much better than the clustered simple root-approximations so often accepted in its place. More precisely, we shall find that each zero of a polynomial is a differentiable function of its coefficients provided that zero's multiplicity is conserved; only when multiplicities change can the derivatives become infinite. Moreover we shall find that the condition number of a multiple zero must be inversely proportional to the product of the distances from that multiple zero to all other zeros of the polynomial. For the problem of finding polynomials' zeros the pejorative manifolds and sub-manifolds in the space of polynomials are evidently the loci occupied by polynomials with various combinations of multiple zeros (one double zero, two double zeros, ..., one triple zero, one triple and one double zero, ...). However, given a polynomial $x$ no convenient way is known yet for determining how near $x$ is to a pejorative manifold short of computing laboriously all the points nearest $x$ on each of the various manifolds and sub-manifolds. We shall describe some of the easier such calculations.

### II.1: Differentiability of Multiple Zeros

If $\zeta$ is a simple zero of the monic polynomial

$$x(\tau) \equiv \tau^n - \sum_1^n x_j \tau^{n-j}$$

then $x$'s first derivative $x'(\tau)$ cannot vanish at $\zeta$ and hence

each $\partial\zeta/\partial x_j = \zeta^{n-j}/x'(\zeta)$ must be finite, whence it follows that $\zeta$ must be an analytic function of each coefficient $x_j$ as long as $\zeta$ remains simple. To what extent can this assertion be valid when $\zeta$ is a multiple zero of $x$?

Whenever $x$ has a multiple zero its coefficients $x_j$ must satisfy certain constraints expressible as polynomial equations in those coefficients with the aid of determinants known as *bigradients* or *resultants*; see Bôcher (1907, ch. XV) or Householder (1970, §§1.2-3) or van der Waerden (1950, ch. XI). It suffices to acknowledge these constraints without describing them, and then exploit them with the following result:

Proposition II.1: The constraints satisfied by the coefficients $x_j$ of the monic polynomial

$$x(\tau) \equiv x^n - \sum_1^n x_j \tau^{n-j}$$

when it possesses an $m$-tuple zero $\zeta$ define $\zeta$ and the last $m-1$ coefficients $x_{n+2-m},\ldots,x_n$ to be analytic functions of each of the first $n+1-m$ coefficients $x_1,x_2,\ldots,x_{n+1-m}$ as long as the multiplicity of $\zeta$ remains precisely $m$, irrespective of the other zeros' multiplicities. And then if $\zeta_{m+1},\zeta_{m+2},\ldots,\zeta_n$ are $x$'s other $n-m$ zeros, different from $\zeta$ but otherwise not necessarily distinct,

$$\partial\zeta/\partial x_i = \frac{\zeta^{n-i+1-m}}{m}\binom{n-i}{m-1} \Big/ \prod_{m+1}^{n}(\zeta-\zeta_j) \quad \text{for} \quad 1 \le i \le n+1-m \quad .$$

Proof: Since $\zeta$ is an $m$-tuple zero of $x$, $x^{(m)}(\zeta) \neq 0$ but $x^{(m-1)}(\zeta) = 0$, $x^{(m-2)}(\zeta) = 0$, ..., $x'(\zeta) = 0$ and $x(\zeta) = 0$. The

first two relations imply that $\zeta$, as a simple zero of $x^{(m-1)}$, must be an analytic function of its coefficients $x_1, x_2, \ldots, x_{n+1-m}$. Substituting that function for $\zeta$ in the last $m-1$ equations exhibits the last $m-1$ coefficients in turn as analytic functions of the first $n+1-m$. Then differentiate the equation $x^{(m-1)}(\zeta) = 0$ with respect to $x_i$ to produce

$$x^{(m)}(\zeta)\partial\zeta/\partial x_i - (n-i)!\zeta^{n-i+1-m}/(n-i+1-m)! = 0 \quad ,$$

and apply Leibniz's rule to $x(\tau) = (\tau-\zeta)^m\Pi_{m+1}^n(\tau-\zeta_j)$ to produce $x^{(m)}(\zeta) = m!\Pi_{m+1}^n(\zeta-\zeta_j)$, whence follows the last part of the proposition.

Here are three examples to illustrate the proposition. First, a quadratic $\tau^2-2\alpha\tau+\beta$ has a double zero $\zeta = \alpha$ just when $\beta = \alpha^2$; here $\zeta$ and $\beta$ are analytic functions of $\alpha$ as claimed in the proposition, but if we regarded $\zeta$ and $\alpha$ as functions of $\beta$ they would have a branch-point singularity at $\beta = 0$. This first example provides some excuse for regarding, as does the proposition, the first $n+1-m$ coefficients $x_j$ instead of some other subset as independent variables.

The second example is a quartic

$$\tau^4 - 4\alpha\tau^3 + 6\beta\tau^2 - 4\gamma\tau + \delta$$

which has a triple zero $\zeta = \alpha+\Delta$ whenever $\gamma = (\alpha+\Delta)^2(\alpha-2\Delta)$ and $\delta = (\alpha+\Delta)^3(\alpha-3\Delta)$ where $\Delta \equiv \pm(\alpha^2-\beta)^{1/2}$; evidently $\zeta$, $\gamma$ and $\delta$ are analytic functions of $\alpha$ and $\beta$ except at the branch point where $\Delta = 0$, at which point $\zeta$ becomes a quadruple zero.

The third example is the quartic

$$q(\tau,\lambda) \equiv \tau^4 - (2+\lambda^2)\tau^2 + 2\lambda|\lambda|\tau + 1 - \lambda^2 \quad \text{for real } \lambda$$
$$= (\tau - \text{sign}(\lambda))^2(\tau + \text{sign}(\lambda) + \lambda)(\tau + \text{sign}(\lambda) - \lambda) \quad .$$

$q$ has a double zero $\zeta$ for all real $\lambda$, but $\zeta = 1$ for $\lambda \geq 0$ and $\zeta = -1$ for $\lambda \leq 0$, with ambiguity and discontinuity at $\lambda = 0$ despite that $\zeta$ and the last coefficient $1-\lambda^2$ may appear to be formally analytic functions of the first three coefficients $(0, 2+\lambda^2, 2\lambda|\lambda|)$. But these first coefficients are not free here to vary independently, nor are they analytic functions of the real parameter $\lambda$ near $\lambda = 0$. A better explanation for the apparent anomaly is obtained from a geometrical approach which identifies quartic polynomials with points in a 4-dimensional space. The polynomials with double zeros constitute a 3-dimensional manifold in that space; the manifold intersects itself at points corresponding to polynomials, like $q(\tau,0)$, with two double zeros. As $\lambda$ runs from -1 to 0 to +1, say, $q(\tau,\lambda)$ runs along one sheet of that manifold to a point of self-intersection and then turns a corner to run along the other sheet. The 3-dimensional manifold is pejorative; the corner where $q$'s double zero is discontinuous lies on a multiply pejorative sub-manifold. Little seems to be known about the complicated geometry of these manifolds.

## 11.2: Condition Numbers for Multiple Zeros

The condition of a zero $\zeta$ of a polynomial $x$ is generally a vague notion (cf. Wilkinson (1963, pp.29-32 and 47-48)) partly because the metric by which we measure distance between polynomials

is so often arbitrary. A natural metric for polynomials regarded as points in a linear space is a vector norm $\|\cdots\|$; e.g. for arbitrary weights $w_j > 0$

$$\|\textstyle\sum_0^n x_j \tau^{n-j}\| \equiv \sqrt{(\sum_0^n w_j |x_j|^2)} \quad .$$

Although we shall use just this last norm in what follows, the statements concerning condition numbers will be stated for (and are valid for) any vector norm. Whatever the norm, one corresponding condition number for a zero $\zeta$ of a polynomial $x$ will be defined to be

$$\gamma(\zeta,x,\|\cdots\|) \equiv \sup_{\delta x} |\delta\zeta|/\|\delta x\|$$

where $\delta\zeta$ is the infinitesimal change in $\zeta$ caused by changing the polynomial $x$ infinitesimally to $x+\delta x$. This condition number $\gamma$ is appropriate when absolute variations in $\zeta$ and $x$ are at issue; $\gamma/|\zeta|$ is a more appropriate condition number when relative variations $\delta\zeta/\zeta$ are at issue.

Of course $\gamma$'s definition makes sense only if $\delta x$ is understood to be so constrained that $\zeta$'s multiplicity is conserved; otherwise $\zeta$ loses its identity, disintegrating into a cluster of zeros whose condition numbers approach infinity as the cluster coalesces upon $\zeta$. This assertion, which we have yet to prove, explains why multiple zeros have a bad reputation for ill-condition undeservedly acquired by association with the cluster of closely spaced and therefore ill-conditioned approximate zeros which are so often accepted instead of multiple zeros; cf. Wilkinson (1963, p.41, §8).

Proposition II.2: If $\zeta$ is an $m$-tuple zero of a monic polynomial $x$ whose other zeros are $\zeta_{m+1},\zeta_{m+2},\ldots,\zeta_n$ then its condition number is

$$\gamma(\zeta,x,\|\cdots\|) = K(m,n,\zeta,\|\cdots\|)/\Pi_{m+1}^{n}|\zeta-\zeta_j| \quad ,$$

where $K$ is independent of $x$ and its zeros other than $\zeta$.

Proof (for any norm $\|\cdots\|$): If $x(\tau) = \tau^n - \sum_1^n x_j\tau^{n-j}$ and $\delta x(\tau) = -\sum_1^n \delta x_j\tau^{n-j}$ then by proposition II.1

$$\delta\zeta = \frac{1}{m}\sum_1^{n+1-m}\binom{n-i}{m-1}\zeta^{n-i+1-m}\delta x_i/\Pi_{m+1}^{n}(\zeta-\zeta_j) \quad .$$

Here $\delta\zeta$ is expressed as a linear function of the first $n+1-m$ infinitesimal coefficients $\delta x_j$. The last $m-1$ coefficients are also linear functions of the first $n+1-m$ obtained by solving a triangular system of linear equations derived from the equations

$$x^{(k)}(\zeta) = 0 \quad \text{for} \quad k = 0,1,2,\ldots,m-1 \quad ;$$

$$\delta x^{(k)}(\zeta) + x^{(k+1)}(\zeta)\delta\zeta = 0 \quad \text{for} \quad k = 0,1,2,\ldots,m-1 \quad .$$

The last set reduces simply to

$$\delta x^{(k)}(\zeta) = 0 \quad \text{for} \quad k = 0,1,2,\ldots,m-2 \quad ,$$

which may be solved for $\delta x_{n+2-m},\delta x_{n+3-m},\ldots,\delta x_n$ in turn. Hence there exists some linear operator $Q$ depending upon $\zeta$, $n$ and $m$ alone such that

$$\delta x = Q\delta x^{(m-1)} \quad .$$

This linear operator $Q$ transforms an arbitrary polynomial $p$

of degree $n-m$ into another $q = Qp$ of degree $n-1$ in such a way that

$$q(\zeta) = q'(\zeta) = \cdots = q^{(m-2)}(\zeta) = 0 \quad \text{and} \quad q^{(m-1)} = p \quad ;$$

the last few equations constitute an initial value problem whose solution is

$$q(\tau) = (Qp)(\tau) = \frac{1}{(m-2)!} \int_{\zeta}^{\tau} (\tau-\theta)^{m-2} p(\theta) d\theta \quad .$$

Hence we deduce that $Qp = 0$ only if $p = 0$, and therefore $|||p||| \equiv \|Qp\|$ is another norm on the linear space of polynomials $p$ of degree $n-m$. Now

$$\gamma = \sup |\delta\zeta| / \|\delta x\| \quad \text{over constrained} \quad \delta x = Q\delta x^{(m-1)}$$

and $\delta\zeta = e^* \delta x^{(m-1)} / \Pi_{m+1}^{n} (\zeta-\zeta_j)$ where $e^*$ is the linear functional defined above in the earlier expression for $\delta\zeta$. Hence

$$\gamma = \sup_{\delta x^{(m-1)}} |e^* \delta x^{(m-1)} / \Pi_{m+1}^{n} (\zeta-\zeta_j)| / |||\delta x^{(m-1)}|||$$

$$= K / \Pi_{m+1}^{n} |\zeta-\zeta_j|$$

as claimed, where

$$K \equiv \sup |e^* p| / |||p||| \quad \text{over} \quad (n-m)\text{-degree polynomials} \quad p$$

depends upon $m$, $n$, $\zeta$ and the norm $\|\cdots\|$ but not upon $x$ nor its zeros other than $\zeta$.

Corollary: Proposition II.2 may be applied to non-monic polynomials $x(\tau) \equiv \sum_{o}^{n} x_j \tau^{n-j}$ with $x_o \neq 0$ provided $K$ is replaced by

a different function

$$\tilde{K}(m,n,\zeta,\|\cdots\|)/|x_o| \quad .$$

The foregoing results fail to reflect one important aspect of floating point computation -- independence of scaling. Specifically, we would expect the relative precision of approximations to a zero $\tilde{\zeta} \equiv \sigma\zeta$ of $\tilde{x}(\tau) \equiv \sigma^n x(\tau/\sigma)$ to be independent of $\sigma$ at least as long as the scale factor $\sigma$ is a modest power of the computer's radix. The proposition above appears to give results which are altered by scaling, but it can also be applied in a way independent of scaling. The proposition remains valid when the norm $\|\cdots\|$ varies with $\zeta$, as for example does

$$\left\|\sum_o^n x_j\tau^{n-j}\right\| \equiv \sqrt{\left(\sum_o^n w_j|x_j\zeta^{n-j}|^2\right)} \quad .$$

More generally, whenever the norm $\|\cdots\|$ varies with $\zeta$ in such a way that $\|x(\tau)\|_\zeta = |||\zeta^{-n}x(\zeta\tau)|||$ for some norm $|||\cdots|||$ independent of $\zeta$ we find that

$$\begin{aligned}\gamma(\sigma\zeta,\sigma^n x(\tau/\sigma),\|\cdots\|_{\sigma\zeta})/|\sigma\zeta| &= \gamma(\zeta,x,\|\cdots\|_\zeta)/|\zeta| \\ &= K(m,n,1,\|\cdots\|_1)/\prod_{m+1}^n|1-\zeta_j/\zeta| \quad .\end{aligned}$$

Then the condition number $\gamma/|\zeta|$ is independent of scaling and depends only upon the multiplicity $m$ of $\zeta$ and its *relative* separation from $x$'s other zeros. Consequently, only clusters of relatively closely spaced zeros can be ill-conditioned when such a $\zeta$-dependent norm is used.

The word *cluster* used above has been used very loosely. One might hardly consider the zeros of $x(\tau) \equiv \prod_1^{20}(\tau-j)$ to constitute

a cluster in the usual sense, yet the zeros near 15 have been observed by Wilkinson (1963, pp.41-43) to be ferociously ill-conditioned. This observation does not contradict proposition II.2 and its corollaries; when the constant $K$ is evaluated (for $m = 1$ here) we do get condition numbers of the order of $10^{10}$. This means that one's intuition about clusters is unlikely to be reliable.

Calculations by a student, Mr. David Hough, have shown that one need only perturb each coefficient of $x(\tau) \equiv \Pi_1^{20}(\tau-j)$ by less than one part in $10^{11}$ to construct a nearby polynomial $x+\Delta x$ whose zeros, while still all real, include a double zero. Consequently the polynomial $x$ is very close to a pejorative manifold; in fact, it is almost equally close to several multiply pejorative sub-manifolds. These observations explain Wilkinson's polynomial's ill-condition more convincingly than can any allegation of clustering among its zeros.

## II.3: Where are the Pejorative Manifolds?

When $m$ of a polynomial's zeros are clustered closely in a region well-separated from the rest of the zeros, it is natural to expect that a small perturbation in the polynomial's coefficients should suffice to collapse the cluster into an $m$-tuple zero. That $m$-tuple zero must be a simple zero of the perturbed polynomial's $(m-1)^{\underline{st}}$ derivative, and therefore close to a zero of the original polynomial's $(m-1)^{\underline{st}}$ derivative. Consequently when we wish to substitute what we hope is a well-behaved $m$-tuple zero for a cluster of $m$ ill-behaved zeros, we can approximate

the $m$-tuple zero by a simple zero of the polynomial's $(m-1)^{\underline{st}}$ derivative provided such a simple zero can be found near the cluster. The next result guarantees that such a zero can be found.

Lemma II.3: Suppose the $n^{th}$ degree polynomial $x(\tau)$ has at least $m$ zeros $\zeta_1,\zeta_2,\ldots,\zeta_m$ $(1 \leq m \leq n)$ in some convex region $C$. Then $x^{(m-1)}(\tau)$ must vanish at least once in the star-shaped region $S$ consisting of all points from which $C$ subtends an angle no less than $\pi/(n+1-m)$.

Proof: Let $\Delta^{m-1}x(\zeta_1,\zeta_2,\ldots,\zeta_m)$ be the $(m-1)^{\underline{st}}$ divided difference of $x(\tau)$ over the zeros $\zeta_1,\zeta_2,\ldots,\zeta_m$. Since each $x(\zeta_j) = 0$ that divided difference must vanish. Therefore we obtain

$$\int\int\cdots\int_{\substack{\text{All } \sigma_j \geq 0 \\ \text{and } \Sigma \sigma_j = 1}} x^{(m-1)}\left(\sum_1^m \sigma_j\zeta_j\right)d\sigma_1 d\sigma_2\cdots d\sigma_m = \Delta^{m-1}x(\zeta_1,\zeta_2,\ldots,\zeta_m) = 0$$

from a formula attributed to Hermite and to Genocchi by Milne-Thomson (1933, p.10 and p.18 ex. 6). Let us denote the $n+1-m$ zeros of $x^{(m-1)}$ by $\eta_m,\eta_{m+1},\ldots,\eta_n$ and so infer

$$\int\int\cdots\int_{\substack{\text{All } \sigma_j \geq 0 \\ \text{and } \Sigma \sigma_j = 1}} \prod_{k=m}^{n}\left(\eta_k - \sum_{j=1}^{m}\sigma_j\zeta_j\right)d\sigma_1 d\sigma_2\cdots d\sigma_m = 0 \quad .$$

From this point we pursue an argument similar to Marden's (1966, §24).

Were every $\eta_k$ outside $S$ we could find a $\theta_k$ for each $k = m,m+1,\ldots,n$ such that

$$0 \le \arg(\eta_k - \tau) - \theta_k < \pi/(n+1-m) \quad \text{for all } \tau \text{ in } C \quad .$$

In particular $\sum_1^m \sigma_j \zeta_j$ lies amidst the $\zeta_j$'s, and hence in $C$, for all relevant sets of values $\sigma_1, \sigma_2, \ldots, \sigma_m$; therefore we should deduce that

$$0 \le \arg\Big( \prod_{k=m}^{n} (\eta_k - \sum_{j=1}^{m} \sigma_j \zeta_j) \Big) - \sum_{m}^{n} \theta_k < \pi \quad ,$$

whence it would follow that the last integral, with its integrand confined to a half-plane that excludes zero, could not vanish. This contradiction proves the lemma.

In particular, when $C$ is a circle of radius $\rho$ then $S$ turns out to be a concentric circle of radius $\rho \csc \frac{\pi/2}{n+1-m}$; in general $S$ cannot be enormously larger than $C$, so the desired simple zero of $x^{(m-1)}$ can always be found somewhere near a cluster of $m$ zeros of $x$.

In general one cannot expect ill-conditioned zeros to cluster in an obvious way, and we must search instead for nearby polynomials on pejorative manifolds. Thus one comes to consider problems like this one:

Problem II.3: Given $x(\tau) \equiv \tau^n - \sum_1^n x_j \tau^{n-j}$ find the *nearest* polynomial $x-y$, where $y(\tau) \equiv \sum_1^n y_j \tau^{n-j}$, with an $m$-tuple zero. We interpret "nearest" to mean that

$$\|y\|^2 \equiv \sum_1^n w_j |y_j|^2 \quad ,$$

with given positive weights $w_j$, should be minimized.

This problem can be approached in a conventional way via

Lagrange multipliers. The result is a set of $m$ equations

$$y^{(k)}(\zeta) \equiv \sum_{i=0}^{m-2} \lambda_i \sum_{j=1}^{n-k} \frac{((n-j)!)^2}{(n-j-k)!\,(n-j-i)!} \frac{(\zeta^{n-j-i})^* \zeta^{n-j-k}}{w_j} = x^{(k)}(\zeta)$$

$$\text{for } k = 0,1,2,\ldots,m-1$$

from which we eliminate the Lagrange multipliers $\lambda_i$ by setting a determinant of the coefficients of $(1,\lambda_0,\lambda_1,\ldots,\lambda_{m-2})$ to zero. The result is an equation to be solved for the $m$-tuple zero $\zeta$. The equation is not a polynomial equation because both $\zeta$ and its complex conjugate $\zeta^*$ appear. When $m = 2$ the equation is

$$\zeta x'(\zeta) = x(\zeta)\left(\textstyle\sum_1^{n-1}(n-j)\,|\zeta^{n-j}|^2/w_j\right) / \left(\textstyle\sum_1^{n}|\zeta^{n-j}|^2/w_j\right)$$

and is not hard to solve for $\zeta$, though most of the solutions must be discarded as irrelevant.

The problem becomes more interesting when $x(\tau)$ has real coefficients and, naturally, we require that $y(\tau)$ have real coefficients too.

However ugly these calculations may be, they are worth pursuing whenever $x$ has a badly ill-conditioned $m$-tuple zero $\zeta$. For if $\zeta$'s condition number $\gamma$ is huge then, since proposition II.2 tells us that

$$\gamma = K/\prod_{m+1}^{n} |\zeta-\zeta_j| = m!K/|x^{(m)}(\zeta)|$$

for some modest K, we see that $x$ differs from a polynomial

$$x(\tau) - y(\tau) \equiv x(\tau) - \frac{x^{(m)}(\zeta)}{m!}(\tau-\zeta)^m$$

with an $(m+1)$-tuple zero $\zeta$ by just a little;

$$\|y\| = K\|(\tau-\zeta)^m\|/\gamma \quad .$$

$x$ can be no farther than that from the multiply pejorative submanifold of polynomials with $(m+1)$-tuple zeros.

## Part III: Eigenproblems

"What I tell you three times is true."
Lewis Carroll, *Hunting of the Snark*, Fit 1.

Let $\zeta$ be an $m$-cuple eigenvalue of the $n \times n$ matrix $Z$ and let $\delta Z$ run through infinitesimal perturbations so constrained that $Z+\delta Z$ continues to possess an $m$-tuple eigenvalue $\zeta+\delta\zeta$ near $\zeta$. We define

$$\gamma(\zeta,Z,\|\cdots\|) \equiv \sup |\delta\zeta|/\|\delta Z\| \quad \text{over such constrained } \delta Z$$

to be the condition number of $\zeta$ as an $m$-tuple eigenvalue of $Z$ with respect to some given matrix norm $\|\cdots\|$. The constraints on $\delta Z$ are complicated but indispensable when $m > 1$; without them the condition number $\gamma$ would be either infinite or meaningless.

We shall obtain estimates for $\gamma$ which relate it to the norm of the spectral projector $P$ onto $\zeta$'s $m$-dimensional invariant subspace. $P$ is characterized by the equations

$$P^2 = P \ , \quad PZ = ZP \ , \quad \text{rank}(P) = m \ , \quad P(Z-\zeta)^m = 0 \quad ;$$

$P$ can be computed straightforwardly from the similarity transformation that exhibits $Z$'s Jordan normal form. (For example, when $m = 1$ $\zeta$'s non-zero row and column eigenvectors $x^*$ and $y$, which satisfy $x^*Z = \zeta x^*$ and $Zy = \zeta y$, yield $P = yx^*/x^*y$.) We shall find that, roughly speaking, $\gamma$ is big if and only if $\|P\|$ is big. Since $\gamma$ is appreciably more expensive to compute than $\|P\|$ when $m > 1$, we shall use $\|P\|$ as a measure of $\zeta$'s

ill-condition instead of $\gamma$.

Hypersensitivity to small perturbations, and the consequent risk of numerical instability, always accompany a spectral projector of large norm irrespective of whether it belongs to a multiple eigenvalue or to a cluster of simple eigenvalues of $Z$. The spectral projector $P$ onto an $m$-dimensional invariant subspace belonging to a cluster of $m$ eigenvalues $\zeta_j$ (counting multiplicities) is just the sum of the spectral projectors $P_j$ belonging to the distinct values $\zeta_j$. When $\|P\|/m$ is huge at least one of the $\|P_j\|$'s must be huge too so at least one $\zeta_j$ must be ill-conditioned. We shall see other bad things happen; for example every similarity transformation $Q$, which reduces $Z$ to a diagonal sum

$$Q^{-1}ZQ = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$$

in which the $m \times m$ matrix $B$ has as its spectrum the cluster of $m$ eigenvalues $\zeta_j$, is necessarily ill-conditioned in the sense that $\|Q\| \cdot \|Q^{-1}\|$ must exceed $\|P\|$, roughly. Indeed, when $\|P\|/m$ is huge the cluster's very identity as a cluster of $m$ eigenvalues may be jeopardized by small uncertainties or perturbations in $Z$. Why? Because then to *every* closed contour $\Gamma$ which encloses the cluster and excludes the rest of $Z$'s eigenvalues corresponds at least one small perturbation $\Delta Z$, with $\|\Delta Z\| \leq \kappa \|Z\| / \|P\|^{1/m}$ for a modest constant $\kappa$, such that $Z + \Delta Z$ has either fewer than $m$ or more than $m$ eigenvalues inside $\Gamma$. In the special case when $Z$'s cluster contains just one $m$-tuple

eigenvalue $\zeta$, the small perturbation $\Delta Z$ can be so chosen that that same $\zeta$ is an $(m+1)$-tuple eigenvalue of $Z+\Delta Z$; our proof of this assertion will sharpen and generalize portentous results for $m = 1$ published earlier by Ruhe (1970) and Wilkinson (1972).

So, spectral projectors of huge norm are critical symptoms of hypersensitivity to small perturbations, and no matrix can possess huge projectors unless tiny perturbations to its elements suffice to increase the multiplicities of some of its eigenvalues. Evidently the eigenproblem's pejorative manifolds and sub-manifolds consist of those matrices with various combinations of multiple eigenvalues (one double, one triple, two double, one quadruple, one double and one triple, ...).

Although, given a matrix $Z$, no convenient way is known yet to determine just how near $Z$ is to arbitrary pejorative sub-manifolds, ways are known to find points, close enough to $Z$ for many practical purposes, on some simpler pejorative sub-manifolds. These ways invoke unitary similarity transformations which reduce $Z$ to a block-upper-triangular form with diagonal blocks of small dimensionality. Each block is intended to correspond to a cluster of $Z$'s eigenvalues to which belongs a spectral projector of moderate norm even though the spectral projectors belonging to every sub-cluster of the cluster have huge norms. When such clusters exist, and often they do exist, they may not look like clusters to the naked eye; this is so because the individual eigenvalues in the cluster are very ill-conditioned and disperse frantically in response to most small perturbations of $Z$. The eigenvalues in a cluster can be identified only by the observation that

each eigenvalue's projector, though huge, cancels parts of the others' projectors in such a way that the sum of all the individual projectors has a moderate norm.

Having found suitable clusters and corresponding small blocks, we try to replace each block by its nearest like-dimensioned matrix with just one eigenvalue; this turns out to be tantamount to the construction for each block of the nilpotent matrix nearest to it. Enough is known about that construction to make it cheap for small blocks -- $2 \times 2$ and $3 \times 3$ -- but for larger blocks no cheap construction is known yet.

The theory is extensive but incomplete. Lacking sharp indications of the distance from $Z$ to various pejorative sub-manifolds, we could too often become enmeshed in expensive calculations of nearest nilpotent matrices whenever $Z$ is neither so far from all pejorative sub-manifolds that they are obviously ignorable nor so near to some as to indicate obviously which ones are the only ones worth considering. Yet the theory is attractive. If it can be refined to cover the majority of cases that arise often in practice, it will be complete enough.

## III.1 Some apparatus

Only the following matrix norms will be used;

$$\|X\|_2 \equiv \sqrt{\text{tr.}(X^*X)} = \sqrt{\Sigma\ (\text{singular values of } X)^2}\quad ,$$

$$\|X\| \equiv \max_{Y \neq 0} \|XY\|_2 / \|Y\|_2 = \text{maximum singular value of } X\quad .$$

These norms have been chosen because they are not changed when $X$ is multiplied by a unitary matrix and consequently have many useful

properties which we will invoke with little comment; for details see Mirsky (1960) or Gohberg and Kreĭn (1969).

Given an $n \times n$ matrix $Z$ we shall sometimes identify a cluster of $m$ of its eigenvalues $\zeta_j$ (counting multiplicities) by specifying one of the closed contours $\Gamma$ in the complex plane which enclose all of the cluster's $m$ eigenvalues strictly in their interiors leaving the rest of $Z$'s spectrum strictly outside. Some of the contours may have disconnected components but none of them can pass through an eigenvalue of $Z$. We soon discover, after Kato (1966, p.67), that

$$P \equiv \frac{1}{2\pi i} \oint_\Gamma (\tau - Z)^{-1} d\tau$$

is the spectral projector onto $Z$'s invariant subspace belonging to the cluster of eigenvalues inside $\Gamma$. These eigenvalues are the $m$ non-trivial eigenvalues of $PZ = ZP$, of which the remaining $n-m$ eigenvalues are just 0.

There are other ways to represent $P$. We may aptly select a new (generally not orthogonal) coordinate system, or equivalently perform an apt similarity transformation, which will exhibit $Z$ in the reduced form $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ in which the $m \times m$ matrix $B$ has as its eigenvalues just those inside the cluster and $A$'s eigenvalues are outside. In that coordinate system $P$ appears as $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$.

Alternatively we may invoke Schur's theorem to obtain a new orthogonal coordinate system, or equivalently perform a unitary similarity, which will exhibit $Z$ in the (block-) upper triangular form

$$\begin{pmatrix} A & AR-RB \\ 0 & B \end{pmatrix}$$

in which $A$ and $B$ have the same spectra as before. The block $AR-RB$ is written that way for more convenient correlation with $P$ which, in the same coordinate system, has the form

$$\begin{pmatrix} 0 & -R \\ 0 & 1 \end{pmatrix} \quad .$$

If we do not insist that $A$ and $B$ be upper triangular we can instead arrange with the aid of Autonne's theorem that $R$ be an $(n-m)\times m$ diagonal matrix exhibiting its singular values. Either way, because the similarity transformation is unitary we have $\|P\| = \|\binom{-R}{1}\|$ and, incidentally, $\|1-P\| = \|(1 \;\; R)\| = \|P\|$ (cf. Kato (1960)). Finally, a non-unitary similarity which relates the triangular form to a block diagonal form is

$$\begin{pmatrix} A & AR-RB \\ 0 & B \end{pmatrix} = \begin{pmatrix} 1 & -R \\ 0 & 1 \end{pmatrix}\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}\begin{pmatrix} 1 & R \\ 0 & 1 \end{pmatrix} \quad .$$

When the cluster inside $\Gamma$ contains only one $m$-tuple eigenvalue $\zeta$ the $m\times m$ block $B$ must have only $\zeta$ as an eigenvalue; consequently $(B-\zeta)^m = 0$. When $(B-\zeta)^{m-1} = 0$ too $B$ is called *derogatory* for reasons that will be clear soon. To simplify matters let us temporarily set $\zeta = 0$ as we digress to study nilpotent $m\times m$ matrices; these are characterized by the equation $B^m = 0$.

Lemma III.1.1: $B^m = 0$ if and only if $\mathrm{tr.}(B^k) = 0$ for $k = 1,2,\ldots,m$.

Proof: Apply Newton's identities (cf. Householder (1970) p.37) to sums of powers of $B$'s eigenvalues to deduce that they must all vanish.

What conditions upon an infinitesimal perturbation $\delta B$ ensure that both $B$ and $B+\delta B$ are nilpotent? Another way to think of this question is to imagine that $B = B(\tau)$ is an analytic function of $\tau$ that stays nilpotent for all $\tau$; what characterizes $\dot{B} \equiv dB/d\tau$ for all such functions? The question is not trivial because, although we may differentiate the equations $B^m = 0$ and $\mathrm{tr.}(B^k) = 0$ to get respectively

$$\sum_1^m B^{j-1}\cdot\delta B\cdot B^{m-j} = 0 \quad \text{or} \quad \sum_1^m B^{j-1}\dot{B}B^{m-j} = 0 \quad \text{and}$$

$$\mathrm{tr.}(B^{k-1}\delta B) = 0 \quad \text{or} \quad \mathrm{tr.}(B^{k-1}\dot{B}) = 0 \quad \text{for} \quad k = 1,2,\ldots,m \ ,$$

those are merely necessary conditions upon $\delta B$ and $\dot{B}$; when $B$ is a derogatory nilpotent matrix those conditions fail to be sufficient. For example, when $B = 0$ those conditions impose almost no constraint upon $\delta B$ and $\dot{B}$ whereas they ought to satisfy $(\delta B)^m = 0$ and $\dot{B}^m = 0$.

Lemma III.1.2: When $B_o$ is a non-derogatory nilpotent $m \times m$ matrix the following three conditions are equivalent and characterize the derivative $\dot{B}_o = \dot{B}(0)$ of every nilpotent analytic function $B(\tau)$ which satisfies $B(0) = B_o$ :

1) $\dot{B}_o = \dot{S}_o B_o - B_o \dot{S}_o$ is solvable for $\dot{S}_o$ .

2) $\sum_1^m B_o^{k-1}\dot{B}_o B_o^{m-k} = 0$

3) $\mathrm{tr.}(B_o^k \dot{B}_o) = 0$ for $k = 0,1,2,\ldots,m-1$ .

Proof: Without interpreting the dot as a derivative, we observe trivially that 1) implies 2) and 3). To deduce 3) implies 1), define the linear operator $\mathcal{B}$ thus; $\mathcal{B}X \equiv XB_o - B_oX$. Any linear functional $L$ on the range of $\mathcal{B}$ must have the form $L\mathcal{B}X = \mathrm{tr.}(LBX)$ for some matrix $L$. But $\mathrm{tr.}(LBX) = \mathrm{tr.}(LXB_o - LB_oX)$ $= \mathrm{tr.}(B_oLX - LB_oX) = -\mathrm{tr.}((\mathcal{B}L)X)$. From Fredholm's theorem of the alternative (cf. Dunford-Schwarz (1958) p.609) we know that the equation $\mathcal{B}\dot{S}_o = \dot{B}_o$ is solvable (perhaps not uniquely) for $\dot{S}_o$ only if $L\dot{B}_o = 0$ for every $L$ which satisfies $L\mathcal{B} = 0$, and we have just seen that $L\mathcal{B} = 0$ means $\mathcal{B}L = 0$, which implies $B_oL = LB_o$, which implies that $L$ is a polynomial in $B_o$ since $B_o$ is non-derogatory (cf. Gantmacher (1959) p.222). So every $L$ which satisfies $L\mathcal{B} = 0$ has the form $L\dot{B}_o = \mathrm{tr.}(L\dot{B}_o) =$ $= \mathrm{tr.}((\text{polynomial in } B_o)\dot{B}_o)$, and this must vanish because of 3) and the fact $B_o^k = 0$ for $k \geq m$. Therefore 3) implies 1). Next let us deduce 2) implies 3). Since $B_o$ is non-derogatory and nilpotent it must be similar to $J = \begin{pmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & \ddots & \\ & & & 0 \end{pmatrix}_{m\times m}$. If the similarity that takes $B_o$ to $J$ takes $\dot{B}_o$ to $X = (x_{ij})$ then, by 2), $X$ must satisfy

$$\sum_1^m J^{k-1}XJ^{m-k} = 0 \quad ;$$

i.e. $\sum_{m+1-j}^{m+1-i} x_{i+k-1,j-m+k} = 0$ for $1 \leq i \leq j \leq m$ .

This is soon recognized as equivalent to

$$\sum_{1}^{m-k} x_{j+k,j} = 0 \quad \text{for} \quad 0 \leq k \leq m-1 \quad ;$$

i.e. $$\mathrm{tr}.(J^k X) = 0 \quad \text{for} \quad 0 \leq k \leq m-1 \quad .$$

Reversing the similarity yields 3).

Finally we demonstrate the existence of an analytic nilpotent $B(\tau)$ that interpolates $B(0) = B_o$ and $\dot{B}(0) = \dot{B}_o$. Solve 1) for $\dot{S}_o$ and set $S(\tau) \equiv 1 + \tau \dot{S}_o$ and $B(\tau) \equiv S(\tau) B_o S(\tau)^{-1}$. Now $B(\tau)$ is nilpotent (and non-derogatory), since it is similar to $B_o$, at least for $\tau$ small enough. And $\dot{B}(0) = \dot{S}(0)B_o - B_o\dot{S}(0) = \dot{B}_o$.

Lacking anything comparable to lemma III.1.2 for derogatory nilpotent matrices, we should like to avoid them. That is not difficult to do. In the manifold of nilpotent matrices the non-derogatory ones constitute a dense open set; that this is true can be inferred from the Jordan normal form in a way that will be left to the reader.

## III.2: The condition number of a multiple eigenvalue

Let $\zeta$ be an $m$-tuple eigenvalue of an $n \times n$ matrix $Z$ and let $P$ be $\zeta$'s spectral projector. We shall estimate the condition number

$$\gamma = \gamma(\zeta, Z, \|\cdots\|_2) = \sup |\delta\zeta| / \|\delta Z\|_2$$

where the supremum is taken over all infinitesimal $\delta Z$ such that $Z+\delta Z$ continues to possess an $m$-tuple eigenvalue $\zeta+\delta\zeta$ near $\zeta$. We shall show that

$$\gamma \leq \|P\|_2/m \quad .$$

Furthermore, provided the restriction of $Z$ to $P$'s range is not derogatory, i.e. provided $Z$ has only one eigenvector belonging to $\zeta$ or, equivalently, provided $P(Z-\zeta)^m = 0 \neq P(Z-\zeta)^{m-1}$. we shall show that $\gamma$ can be computed straightforwardly though expensively by solving a linear least-squares problem;

$$\gamma = \min_{\lambda_k} \|P(1 - \sum_1^{m-1} \lambda_k (Z-\zeta)^k)\|/m \quad .$$

In this case, we shall conclude,

$$\gamma \geq m^{-1/2}(\|P\|_2^2 + 1 - m)^{1/(2m)} \quad .$$

Although the upper and lower bounds for $\gamma$ are far apart when $m > 1$ and $\|P\|_2$ is big, each bound can be achieved by an appropriate and non-trivial example.

Here is how those claims are proved. Recall that, provided no eigenvalue of $Z$ lies on the closed contour $\Gamma$,

$$P \equiv \frac{1}{2\pi i} \oint_\Gamma (\tau - Z)^{-1} d\tau$$

is the spectral projector upon $Z$'s invariant subspace belonging to the eigenvalues inside $\Gamma$. Suppose there are $m$ such eigenvalues. Then their average value is

$$\mu \equiv \mathrm{tr}.(PZ)/m \quad .$$

Since no eigenvalue of $Z$ lies on $\Gamma$, we find that both $P$ and $\mu$ are continuously differentiable functions of $Z$; in fact an

infinitesimal perturbation $\delta Z$ causes $P$ and $\mu$ to change by (cf. Kato (1966) pp. 76 and 79)

$$\delta P = \frac{1}{2\pi i}\oint_\Gamma (\tau - Z)^{-1}\delta Z(\tau - Z)^{-1}d\tau \quad \text{and}$$

$$\delta\mu = \text{tr.}(P\delta Z)/m \quad \text{since} \quad \text{tr.}(Z\delta P) = 0 \quad .$$

We are interested in the special case when all $m$ eigenvalues inside $\Gamma$ are coincident at $\zeta$, and when the perturbation $\delta Z$ is so constrained that all $m$ perturbed eigenvalues inside $\Gamma$ stay coincident at $\zeta+\delta\zeta$. In this case $\mu = \zeta$ and $\delta\mu = \delta\zeta$, so $\zeta$'s condition number $\gamma(\zeta, Z, \|\cdots\|_2)$ satisfies

$$\begin{aligned}
\gamma &= \sup |\delta\mu| / \|\delta Z\|_2 \quad \text{over constrained } \delta Z \\
&= \frac{1}{m} \sup |\text{tr.}(P\delta Z)| / \|\delta Z\|_2 \quad \text{over constrained } \delta Z \\
&\leq \frac{1}{m} \sup |\text{tr.}(P\delta Z)| / \|\delta Z\|_2 \quad \text{over all } \delta Z \\
&= \|P\|_2 / m
\end{aligned}$$

Thus we conclude that an ill-conditioned eigenvalue must have a spectral projector of large norm. After we show to what extent the converse is true we shall show how, given $m$ and a value $\|P\|_2$, to construct a matrix $Z$ with $\gamma = \|P\|_2/m$.

To obtain a sharper estimate for $\gamma$ we must take the constraints upon $\delta Z$ into account, and we shall now do that just in the non-derogatory case when $P(Z-\zeta)^m = 0 \neq P(Z-\zeta)^{m-1}$. By lemma III.1.1 the equation $P(Z-\zeta)^m = 0$ is equivalent to

$$\text{tr.}(P(Z-\zeta)^k) = 0 \quad \text{for} \quad k = 1,2,\ldots,m$$

which, when differentiated, yields

$$k \operatorname{tr.}(P(Z-\zeta)^{k-1}(\delta Z-\delta\zeta)) + \operatorname{tr.}((Z-\zeta)^{k}\delta P) = 0 \quad .$$

The last term vanishes because

$$\operatorname{tr.}((Z-\zeta)^{k}\delta P) = \frac{1}{2\pi i}\oint_{\Gamma} \operatorname{tr.}((Z-\zeta)^{k}(\tau-Z)^{-1}\delta Z(\tau-Z)^{-1})d\tau$$

$$= \frac{1}{2\pi i}\operatorname{tr.}(\oint_{\Gamma}(Z-\zeta)^{k}(\tau-Z)^{-2}d\tau\ \delta Z) = 0 \quad .$$

Furthermore the coefficient of $\delta\zeta$, $-k \operatorname{tr.}(P(Z-\zeta)^{k-1})$, already vanishes when $k > 1$. Therefore $\delta Z$ necessarily satisfies

$$\operatorname{tr.}((Z-\zeta)^{k-1}P\delta Z) = 0 \quad \text{for} \quad k = 2,3,\ldots,m$$

and what remains to be shown is that these conditions upon $\delta Z$ are also sufficient to ensure that $\zeta+\delta\zeta$, with $\delta\zeta = \operatorname{tr.}(P\delta Z)/m$, is an $m$-tuple eigenvalue of $Z+\delta Z$.

Let us choose a coordinate system in which $Z-\zeta = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ with non-singular $A$ and an $m\times m$ matrix $B$ which must satisfy $B^{m} = 0 \neq B^{m-1}$. Now $P = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. Let $\delta Z = \begin{pmatrix} \delta Z_{11} & \delta Z_{12} \\ \delta Z_{21} & \delta Z_{22} \end{pmatrix}$ satisfy the conditions in question;

$$\operatorname{tr.}((Z-\zeta)^{k}P\delta Z) = \operatorname{tr.}(B^{k}\delta Z_{22}) = 0 \quad \text{for} \quad k = 1,2,\ldots,m-1 \quad .$$

We wish to infer that $Z+\delta Z$ has an $m$-tuple eigenvalue $\zeta+\delta\zeta$, and shall do so by constructing a non-singular matrix

$$1-\delta S = \begin{pmatrix} 1-\delta S_{11} & -\delta S_{12} \\ -\delta S_{21} & 1-\delta S_{22} \end{pmatrix}$$

(differing infinitesimally from 1) for which

$$(1-\delta S)^{-1}(Z+\delta Z-\zeta-\delta\zeta)(1-\delta S) = \begin{pmatrix} A+\delta A & 0 \\ 0 & B \end{pmatrix} .$$

When this similarity relation is pre-multiplied by $(1-\delta S)$ we find that $\delta S$ and $\delta A$ must satisfy

$$\delta A = \delta S_{11}\cdot A - A\delta S_{11} + \delta Z_{11} - \delta\zeta \qquad A\delta S_{12} - \delta S_{12}\cdot B = \delta Z_{12}$$

$$B\delta S_{21} - \delta S_{21}\cdot A = \delta Z_{21} \qquad B\delta S_{22} - \delta S_{22}\cdot B = \delta Z_{22} - \delta\zeta .$$

These equations are obviously solvable for $\delta A$, $\delta S_{11}$ (arbitrary), $\delta S_{21} = -\sum_0^{m-1} B^j \delta Z_{21}\cdot A^{-j-1}$ and $\delta S_{12} = \sum_0^{m-1} A^{-j-1}\delta Z_{12}\cdot B^j$ ; but the solution $\delta S_{22}$ of the last equation is not so obvious. However, lemma III.1.2 provides assurance that a solution $\delta S_{22}$ does exist provided $\delta\zeta = \mathrm{tr.}(\delta Z_{22})/m = \mathrm{tr.}(P\delta Z)/m$, in which case the conditions 3) of lemma III.1.2 are satisfied with $B_o = B$ and $\dot{B}_o = \delta Z_{22} - \delta\zeta$.

Of course, the foregoing manipulations with infinitesimals $\delta S_{ij}$ can be re-interpreted in terms of derivatives along the lines of lemma III.1.2 and the matrix $S(\tau)$ constructed there.

Now that we know the necessary and sufficient constraints upon $\delta Z$ etc., namely

$$\mathrm{tr.}(P(Z-\zeta)^k) = 0 \quad \text{and} \quad \mathrm{tr.}(P(Z-\zeta)^{k-1}(\delta Z-\delta\zeta)) = 0 \quad \text{for} \quad k = 1,2,\ldots,m$$

provided $P(Z-\zeta)^{m-1} \neq 0$, we return to the computation of

$$\gamma = \sup |\delta\zeta|/\|\delta Z\|_2 \quad \text{over constrained} \quad \delta Z .$$

The computation will be carried out in a new orthogonal coordinate system in which

$$Z - \zeta = \begin{pmatrix} A & AR-RB \\ 0 & B \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & -R \\ 0 & 1 \end{pmatrix}$$

and $A$ is non-singular and $B^m = 0 \neq B^{m-1}$. Once again we set $\delta Z = \begin{pmatrix} \delta Z_{11} & \delta Z_{12} \\ \delta Z_{21} & \delta Z_{22} \end{pmatrix}$ but now the constraints take the form

$$\delta\zeta = \text{tr.}(\delta Z_{22} - \delta Z_{21}R)/m \quad \text{and}$$

$$\text{tr.}((\delta Z_{22} - \delta Z_{21}R)B^k) = 0 \quad \text{for} \quad k = 1,2,\ldots,m-1 \ .$$

Therefore

$$\gamma^2 = \sup\ |\delta\zeta|^2/\|\delta Z\|_2^2 \quad \text{over constrained} \quad \delta Z$$

$$= m^{-2}\sup\ |\text{tr.}(\delta Z_{22}-\delta Z_{21}R)|^2/\Sigma\Sigma\|\delta Z_{ij}\|_2^2) \quad \text{over} \ \ldots$$

$$= m^{-2}\sup\ |\text{tr.}(\delta Z_{22}-\delta Z_{21}R)|^2/(\|\delta Z_{21}\|_2^2 + \|\delta Z_{22}\|_2^2) \quad \text{over} \ \ldots$$

where we have set $\delta Z_{11} = 0$ and $\delta Z_{12} = 0$ because any other values diminish the quotient we are trying to maximize. The desired supremum may now be located by standard variational techniques which we shall merely summarize and verify, though first we shall drop the $\delta$ in front of $\delta Z_{22}$ and $\delta Z_{21}$ since the quotient and the constraints are homogeneous functions.

Let $C \equiv (1 - \sum_1^{m-1}\lambda_j B^j)^*(-R^* \quad 1)$ with the coefficients $\lambda_j$ so chosen that $\|C\|_2^2 = \text{tr.}(CC^*)$ is minimized. The $\lambda_j$'s are the solutions of the normal equations

$$\text{tr.}(C\begin{pmatrix} -R \\ 1 \end{pmatrix}B^k) = 0 \quad \text{for} \quad k = 1,2,\ldots,m-1$$

which are linear in $\{\lambda_j\}$ and non-singular too because, since $B^m = 0 \neq B^{m-1}$, the polynomial $\sum_1^{m-1}\lambda_j B^j$ cannot vanish unless

all $\lambda_j$'s vanish. The normal equations for $C$ coincide with the constraints that $(Z_{21} \quad Z_{22})$ must satisfy (recall that the $\delta$'s have been dropped), so $C$ is a permissible choice for $(Z_{21} \quad Z_{22})$ and differs from any other choice by a matrix $Y \equiv (Z_{21} \quad Z_{22}) - C$ which must satisfy the same constraints, namely

$$\mathrm{tr.}\left(Y\binom{-R}{1}B^k\right) = 0 \quad \text{for} \quad k = 1,2,\ldots,m-1 \quad .$$

We are about to discover that only when $(Z_{21} \quad Z_{22})$ is a non-zero scalar multiple of $C$ can the following quotient achieve its supremum:

$$\begin{aligned} m^2|\delta\zeta|^2/\|\delta Z\|_2^2 &= |\mathrm{tr.}((Z_{21} \quad Z_{22})\binom{-R}{1})|^2/\|(Z_{21} \quad Z_{22})\|_2^2 \\ &= |\mathrm{tr.}((C+Y)(C^*+\binom{-R}{1}\sum_1^{m-1}\lambda_j B^j))|^2/\|C+Y\|_2^2 \\ &= |\|C\|_2^2 + \mathrm{tr.}(YC^*) + 0|^2/\|C+Y\|_2^2 \\ &= \|C\|_2^2 - \left(\|Y\|_2^2\|C\|_2^2 - |\mathrm{tr.}(YC^*)|^2\right)/\|C+Y\|_2^2 \\ &\le \|C\|_2^2 \quad , \end{aligned}$$

with equality only when $Y$ is a scalar multiple of $C$. Therefore we have proved that, in the non-derogatory case,

$$\begin{aligned} \gamma(\zeta,Z,\|\cdots\|_2) &= \|C\|_2/m \\ &= \min_{\lambda_k} \|(1-\textstyle\sum_1^{m-1}\lambda_k B^k)^*(-R^* \quad 1)\|_2/m \\ &= \min_{\lambda_k} \|P\left(1-\textstyle\sum_1^{m-1}\lambda_k(Z-\zeta)^k\right)\|_2/m \quad . \end{aligned}$$

Our next task is to secure a lower bound for $\gamma$. Write $U \equiv 1-\sum_1^{m-1}\lambda_k B^k$ and let $U$'s singular values in order be

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m > 0$. Also write $\rho^2 \equiv \|R\|_2^2 = \|P\|_2^2 - m$. We seek a lower bound in terms of $\rho$ and $m$ for $\|U^*(-R^* \quad 1)\|_2/m$. Evidently

$$\begin{aligned} \|U^*(-R^* \quad 1)\|_2^2 &= \|U^*R^*\|_2^2 + \|U^*\|_2^2 \\ &\geq \sigma_m^2\rho^2 + \mathrm{tr.}(U^*U) \quad ; \end{aligned}$$

the last inequality is achieved just when $UU^*R^* = \sigma_m^2 R^*$. What do we know about $U$'s singular values $\sigma_j$? Since $B^m = 0$, $\det.U = 1$ and hence $\sigma_1\sigma_2\cdots\sigma_m = 1$ while $\sigma_1^2+\sigma_2^2+\cdots+\sigma_m^2 = \mathrm{tr.}(U^*U)$. Therefore

$$\begin{aligned} \|U^*(-R^* \quad 1)\|_2^2 &\geq \sigma_1^2+\sigma_2^2+\cdots+\sigma_{m-1}^2+(1+\rho^2)\sigma_m^2 \\ &\geq m\left(\sigma_1^2\sigma_2^2\cdots\sigma_{m-1}^2\sigma_m^2(1+\rho^2)\right)^{1/m} \\ &= m(1+\rho^2)^{1/m} \end{aligned}$$

with equality in the inequality between arithmetic and geometric means just when $\sigma_1 = \sigma_2 = \cdots = \sigma_{m-1} = \sigma_m\sqrt{1+\rho^2}$.

Assembling the relevant relations above yields

$$\gamma \geq m^{-1/2}\left(\|P\|_2^2 + 1 - m\right)^{1/2m}$$

as claimed. The final tasks are to demonstrate that the bounds are achievable. Briefly, to achieve the upper bound $\gamma \leq \|P\|_2/m$ it suffices that $R^*R$ above be diagonal. To achieve the lower bound it suffices that

$$B = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}_{m\times m}$$

and that $R$ have the form $R = yx^*$ where

$$x^* = (\sigma^{1-m}, \sigma^{2-m}, \ldots, \sigma^{-2}, \sigma^{-1}, 1) \quad \text{and}$$

$$y^*y = \sigma^{2m-2}(\sigma^2-1)$$

for some arbitrary $\sigma > 1$. It will turn out that $U = 1 - (\sigma^2-1)\sum_1^{m-1}\sigma^{-j}B^j$ and $R^*R + 1 = \sigma^2(UU^*)^{-1}$. The details are, once again, left to the reader.

Since $\gamma$ is huge if and only if $\|P\|_2$ is huge, even though they may still be orders of magnitude apart, we shall henceforth dispense with $\gamma$ and use only $\|P\|_2$ or $\|P\|$ as our measure of ill-condition.

## III.3 What happens when $\|P\|$ is huge?

We shall consider now some of the ugly phenomena associated with spectral projectors of huge norm.

Proposition III.3.1: If $\zeta$ is an $m$-tuple eigenvalue of $Z$ and $P$ its spectral projector, then there exists a perturbation $\Delta Z$ such that $\zeta$ is an $(m+1)$-tuple eigenvalue of $Z+\Delta Z$ and $\|\Delta Z\| \leq \|Z-\zeta\|/\|P\|^{1/m}$, so $\|\Delta Z\|$ is small if $\|P\|$ is huge.

Proof: By a unitary similarity exhibit

$$Z - \zeta = \begin{pmatrix} A & AR-RB \\ 0 & B \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & -R \\ 0 & 1 \end{pmatrix}$$

where $A$ is non-singular and $B^m = 0$. Evidently $(Z-\zeta)^m = \begin{pmatrix} A^m & A^mR \\ 0 & 0 \end{pmatrix}$ so we find that

$$\|P\| = \|(1 \quad R)\| \leq \|A^{-m}\| \|(A^m \quad A^m R)\|$$

$$= \|A^{-m}\| \|(Z-\zeta)^m\| \leq \|A^{-1}\|^m \|Z-\zeta\|^m \quad .$$

How close is $A$ to its nearest singular neighbour $A+\Delta A$? We know (see part I) that such a $\Delta A$ can be found with $\|\Delta A\| \leq 1/\|A^{-1}\|$, and the previous inequality shows that that $\Delta A$ satisfies $\|\Delta A\| \leq \|Z-\zeta\| / \|P\|^{1/m}$. Therefore we can use $\Delta Z = \begin{pmatrix} \Delta A & 0 \\ 0 & 0 \end{pmatrix}$ to achieve what has been claimed.

This proposition slightly sharpens one of Wilkinson's (1972) when $m = 1$, in which case the proposition is best-possible without more information about $Z$ than $\|Z-\zeta\|$ and $\|P\|$. Even so, it can be somewhat misleading. Consider an example used by G.E. Forsythe:

$$Z = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ \zeta^n & & & & 0 \end{pmatrix}_{n \times n} \quad .$$

Here we may assume $n > 10$, and $\zeta$ small and positive, say $\zeta < 1/10$. $Z$ has $n$ distinct eigenvalues equally spaced around a circle of radius $\zeta$ all with the same condition number

$$\gamma = \|Projector\| = n^{-1/2} \zeta^{1-n} (1-\zeta^{2n}) / (1-\zeta^2) \quad .$$

Consequently, the proposition says that $Z$ is no farther from a matrix with a double eigenvalue then roughly $n^{1/2} \zeta^{n-1}$. In fact, $Z$ is within $\zeta^n$ of a matrix with an $n$-tuple eigenvalue.

When $m > 1$ proposition III.3.1 probably far over-estimates

the distance to the nearest matrix $Z+\Delta Z$ with an $(m+1)$-tuple eigenvalue.

We now turn to the spectral projectors belonging to clusters of eigenvalues of unspecified multiplicities, and demonstrate why projectors of large norm are to be avoided.

Proposition III.3.2: Let $\Gamma$ be a closed contour in the complex plane which separates $Z$'s spectrum into two parts; $m$ eigenvalues (counting multiplicities) strictly inside $\Gamma$ and the rest strictly outside. And let $P$ be the spectral projector onto $Z$'s invariant subspace belonging to the $m$ eigenvalues inside $\Gamma$. Whenever $\|P\|$ is huge, in particular whenever $\|P\| > \sqrt{m+1}$, there exists a small perturbation $\Delta Z$ satisfying

$$\|\Delta Z\|_2 / \|Z\|_2 \leq 1.22/(\|P\|^2 - 1)^{1/(2m)}$$

such that $Z-\Delta Z$ has at least one eigenvalue on the boundary $\Gamma$.

Proof: Once again use a unitary similarity to exhibit

$$Z = \begin{pmatrix} A & AR-RB \\ 0 & B \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & -R \\ 0 & 1 \end{pmatrix}$$

where $B$ is an $m \times m$ matrix whose spectrum lies inside $\Gamma$ and $A$'s spectrum lies outside. Furthermore, we may exploit Autonne's theorem to exhibit $R$ as an $(n-m) \times m$ diagonal matrix with its singular values $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_m \geq 0$ on its main diagonal. (It is convenient here to assume $n-m \geq m$; otherwise swap the rôles of $A$ and $B$.) Note that $\|R\| = \rho_1$ and $\|P\| = \sqrt{1+\rho_1^2}$.

For any $k$ in $1 \le k \le m$ we may partition

$$R = \begin{pmatrix} \Lambda & 0 \\ 0 & M \end{pmatrix} \quad \text{with square } \Lambda = \operatorname{diag}(\rho_1, \rho_2, \ldots, \rho_k)$$
$$\text{and } M = \operatorname{diag}(\rho_{k+1}, \ldots) \text{ or null,}$$

and conformally partition

$$\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} = X \equiv AR - RB = \begin{pmatrix} A_{11}\Lambda - \Lambda B_{11} & A_{12}M - \Lambda B_{12} \\ A_{21}\Lambda - MB_{21} & A_{22}M - MB_{22} \end{pmatrix} .$$

We shall examine a distinguished $\hat{\Delta}Z \equiv \begin{pmatrix} \hat{\Delta}A & 0 \\ 0 & \hat{\Delta}B \end{pmatrix}$ where

$\hat{\Delta}A \equiv \begin{pmatrix} X_{11}\Lambda^{-1}/2 & 0 \\ A_{21} & 0 \end{pmatrix}$ and $\hat{\Delta}B \equiv \begin{pmatrix} -\Lambda^{-1}X_{11}/2 & B_{12} \\ 0 & 0 \end{pmatrix}$ are so chosen that

$$A - \hat{\Delta}A = \begin{pmatrix} (A_{11} + \Lambda B_{11}\Lambda^{-1})/2 & A_{12} \\ 0 & A_{22} \end{pmatrix} \text{ and } B - \hat{\Delta}B = \begin{pmatrix} (\Lambda^{-1}A_{11}\Lambda + B_{11})/2 & 0 \\ B_{21} & B_{22} \end{pmatrix} .$$

have in common the $k$ common eigenvalues of

$$(A_{11} + \Lambda B_{11}\Lambda^{-1})/2 = \Lambda(\Lambda^{-1}A_{11}\Lambda + B_{11})\Lambda^{-1}/2 .$$

Consequently, using $\Delta Z = \tau\hat{\Delta}Z$ with $0 < \tau < 1$ we shall find that the eigenvalues of $Z - \Delta Z$ move continuously, as $\tau$ increases from 0 to 1, until $k$ eigenvalues that started inside $\Gamma$ coalesce with $k$ that started outside $\Gamma$. For some $\tau$ between 0 and 1 one of those eigenvalues must cross $\Gamma$, and then $\|\Delta Z\|_2 = \tau\|\hat{\Delta}Z\|_2 \le$ $\le \|\hat{\Delta}Z\|_2$.

Thus, all that remains to be shown is that $\|\hat{\Delta}Z\|$ satisfies the inequality claimed in the proposition for at least one $k \ge 1$.

$$\begin{aligned}
\|\hat{\Delta}Z\|_2^2 &= \|\hat{\Delta}A\|_2^2 + \|\hat{\Delta}B\|_2^2 \\
&= \tfrac{1}{4}\|X_{11}\Lambda^{-1}\|_2^2 + \|(X_{21}+MB_{21})\Lambda^{-1}\|_2^2 + \|\Lambda^{-1}(A_{12}^{H}-X_{12})\|_2^2 + \tfrac{1}{4}\|\Lambda^{-1}X_{11}\|_2^2 \\
&\le \rho_k^{-2}\left(\|X_{11}\|_2^2/2 + (\|X_{21}\|_2 + \rho_{k+1}\|B_{21}\|_2)^2 + (\|X_{12}\|_2 + \rho_{k+1}\|A_{12}\|_2)^2\right) \\
&\le \rho_k^{-2}(1+\rho_{k+1}^2)\left(\|X_{11}\|_2^2 + \|X_{21}\|_2^2 + \|B_{21}\|_2^2 + \|X_{12}\|_2^2 + \|A_{12}\|_2^2\right) \\
&\le \rho_k^{-2}(1+\rho_{k+1}^2)\|Z\|_2^2 \quad .
\end{aligned}$$

Let us now choose $k$ to minimize the factor $\rho_k^{-2}(1+\rho_{k+1}^2)$. Suppose $\theta$ is that minimum value; i.e.

$$\rho_k^{-2}(1+\rho_{k+1}^2) \le \theta \quad \text{for} \quad k = 1,2,\ldots,m \quad (\rho_{m+1} \equiv 0) \quad .$$

Then $$\rho_m^2 \le \theta^{-1}$$

$$\rho_{m-1}^2 \le \theta^{-1}(1+\rho_m^2) \le \theta^{-1} + \theta^{-2} \quad ,$$

$$\rho_1^2 \le \theta^{-1}(1+\rho_2^2) \le \theta^{-1} + \theta^{-2} + \cdots + \theta^{-m} \quad ;$$

evidently $\theta$ is no bigger than the positive root $\Theta$ of

$$\|P\|_2^2 = 1 + \rho_1^2 = 1 + \Theta^{-1} + \Theta^{-2} + \cdots + \Theta^{-m} \quad .$$

When $\rho_1^2 > m$ we must have $\Theta^{-1} > 1$ and hence $\rho_1^2 < m\Theta^{-m}$, whence $\Theta < m^{1/m}\rho_1^{-2/m} < e^{1/e}(\|P\|_2^2-1)^{-1/m}$, where $e^{1/e} = 1.445\cdots$. The claimed result follows.

This proposition seems to overestimate $\|\Delta Z\|$ grossly. Indeed, if $P$ has $k$ large singular values and the rest small, say $\sqrt{1+\rho_k^2}/\sqrt{1+\rho_{k+1}^2} >> 1$, then the proof above yields $\|\Delta Z\| \le \rho_k^{-1}\sqrt{1+\rho_{k+1}^2}\|Z\|$, which is far smaller than claimed in the proposition. Another example of overestimation arises when a

similarity (perhaps not unitary) of modest condition number (see below) succeeds in diagonalizing $A$ and $B$ without erasing the block $AR - RB$. It is possible to show then that $\|\Delta Z\|$ need not much exceed $\|Z\|_2/\|P\|_2$ when $\|P\|$ is large; this claim will not be proved here.

Next we shall consider the condition number $\kappa(Q) = \|Q\| \cdot \|Q^{-1}\|$ of similarity transformations that reduce $Z$ to the block diagonal form

$$Q^{-1}ZQ = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} .$$

There are many similarities which reduce $Z$ to this form, and we shall be particularly interested in the ones whose condition numbers are roughly minimal. Experience teaches us that if the minimal condition number is huge then the reduction of $Z$ will be hypersensitive to rounding errors and other perturbations and uncertainties; see Wilkinson (1965) p.87.

Proposition III.3.3: Let $\Gamma$, $Z$, $m$ and $P$ be as in the previous proposition III.3.2. When $\|P\|$ is huge every similarity $Q^{-1}ZQ$, which reduces $Z$ to block diagonal form with one block for the $m$ eigenvalues inside $\Gamma$ and the other block for those outside, must be ill-conditioned; $\kappa(Q) \geq \|P\|$. Conversely, if every similarity is ill-conditioned then $\|P\|$ must be big because for some such similarities $\|P\| > \kappa(Q)/4$.

Proof: Once again use a unitary similarity (which does not aggravate the condition numbers) to exhibit $Z$ in the block triangular

form used in proposition III.3.2. Any eligible similarity $Q$ must exhibit two blocks, one similar to $A$ and the other to $B$. Consequently, every such $Q$ must have the form

$$Q^{-1}ZQ = \begin{pmatrix} S^{-1}AS & 0 \\ 0 & T^{-1}BT \end{pmatrix} ,$$

whence $Q = \begin{pmatrix} S & -RT \\ 0 & T \end{pmatrix}$ and $Q^{-1} = \begin{pmatrix} S^{-1} & S^{-1}R \\ 0 & T^{-1} \end{pmatrix}$. Now $\|Q\| \geq \|S\|$ and $\|Q\| \geq \|\begin{pmatrix} -R \\ 1 \end{pmatrix} T\| \geq \|P\|/\|T^{-1}\|$, and $\|Q^{-1}\| \geq \|T^{-1}\|$ and $\|Q^{-1}\| \geq \|S^{-1}(1 \quad R)\| \geq \|S^{-1}\|/\|S\|$. Therefore

$$4\kappa(Q) = 4\|Q\|\cdot\|Q^{-1}\| \geq \left(\|S\| + \|P\|/\|T^{-1}\|\right)\left(\|T^{-1}\| + \|P\|/\|S\|\right)$$
$$\geq 4\|P\| , \text{ as claimed.}$$

On the other hand, if we choose for $S$ and $T$ any matrices which satisfy $S^*S = \sigma^2$ and $T^*T = \tau^2$ for constants $\sigma$ and $\tau$ that satisfy $\sigma/\tau = \|P\|$, we find that

$$\kappa(Q) = \|Q\|\cdot\|Q^{-1}\| \leq \left(\|S\| + \|P\|\cdot\|T\|\right)\left(\|T^{-1}\| + \|S^{-1}\|\cdot\|P\|\right)$$
$$= (\sigma + \tau\|P\|)(\tau^{-1} + \sigma^{-1}\|P\|)$$
$$= 4\|P\| , \text{ as claimed.}$$

## III.4: The nearest nilpotent matrix

Suppose we have identified every cluster of $Z$'s eigenvalues to which belongs a spectral projector of moderate norm, and no such cluster may be broken up without introducing huge spectral projectors. We could perform a unitary similarity which exhibits $Z$ in block-upper-triangular form with one diagonal block for each cluster. What should be done next?

In a sense, each block resists further reduction as if it were an approximation to a truly irreducible block, namely a block with only one multiple eigenvalue. The purpose of what follows is to discuss how to locate that irreducible block in the hope that we may replace each ill-behaved cluster of eigenvalues by a well-behaved multiple eigenvalue without appreciably changing the given matrix.

Problem III.4.1: Given an $m \times m$ block $B$, find the *nearest* matrix $\beta + C$ with only one eigenvalue $\beta$; $C$ must be nilpotent. By "nearest" we mean to minimize $\|B - \beta - C\|_2$.

It is not hard to find the best value for $\beta$; write $\beta = \text{tr.}(B)/m + \xi$ and observe $\|B - \beta - C\|_2^2 = \|B - \text{tr.}(B)/m - C\|_2^2 + |\xi|^2$ since $\text{tr.}(C) = 0$. Therefore the best value for $\beta$ is

$$\beta = \text{tr.}(B)/m \qquad \text{(cf. } \mu \text{ in III.2)} ;$$

and from the same observation we deduce that the nilpotent matrix $C$ nearest to $B - \beta$ is independent of $\beta$. That at least one such nearest nilpotent $C$ exists follows from the fact that we need only search for the matrix in the compact set of nilpotents $C$ which also satisfy

$$\|B - \beta - C\|_2 \leq \|B - \beta - 0\|_2 \quad ,$$

since there is no need to look at anything farther away than the nilpotent 0.

Let us imagine that the best $C$ has been found, and choose a new set of orthogonal coordinates to exhibit $C$ in upper

triangular form. Since $C$ is nilpotent it is strictly upper triangular. Since $C$ is closest to $B-\beta$, $B-\beta-C$ must be lower triangular in that coordinate system, and that lower triangle must have the minimum norm of all lower triangles of matrices unitarily similar to $B-\beta$. Since the norm of all of $B-\beta$ is unchanged by unitary similarity, we have the following result:

Proposition III.4.2: Given an $m \times m$ matrix $B$, the nearest matrix $\beta+C$ with only one eigenvalue $\beta$ can be constructed as follows. Of all matrices $U^*BU$ unitarily similar to $B$, choose one whose super-diagonal elements have the largest sum of squared magnitudes; call it $E = U^*BU$. Annihilate all the sub-diagonal elements of $E$ to get $F$. Its diagonal elements will all be the same, namely $\beta$ (this is not obvious -- see below). Then $\beta+C = UFU^*$.

To prove that all the diagonal elements of $E$ are the same we need only consider its $2 \times 2$ principal submatrices with adjacent rows and columns. Each such submatrix must be such that no $2 \times 2$ unitary similarity can increase its super-diagonal element. A modest calculation shows that this implies its two diagonal elements are equal. I am indebted to Alan J. Hoffman for suggesting this simple approach to what used to be a much more complicated proof. That proof, which used variational methods, also showed that $(B-\beta-C)^*$ must be a polynomial in $C$, and that if $C^k = 0$ then $k \geq (m+1)/2$, but these facts seem not to help the search for $C$.

Proposition III.4.2 suggests that $C$ might be constructed via a sequence of $2 \times 2$ Jacobi rotations each designed to enhance the magnitudes of super-diagonal elements. Such a scheme works immediately when $m = 2$, may work well when $m = 3$, and seems to be intolerably slow for $m \geq 4$. There is ample scope for further research.

## References


M. Böcher (1907) "Introduction to Higher Algebra" MacMillan, New York.

N. Dunford and J. T. Schwarz (1958) "Linear Operators, Part I" Interscience/ Wiley, New York.

F. R. Gantmacher (1959) "The Theory of Matrices, vol. 1" translated by K. A. Hirsch, Chelsea Publishing Co., New York.

I. C. Gohberg and M. G. Krein (1969) "Introduction to the Theory of Linear Nonselfadjoint Operators" translated by A. Feinstein. "Translations of Math. Monographs vol. 18," Amer. Math. Soc., Providence, R. I.

G. H. Golub and V. Pereyra (1972) "The Differentiation of Pseudoinverses and Nonlinear Least Squares Problems whose Variables Separate" Stanford University Computer Science Department report STAN-CS-72-261, SU326 P30-15.

G. H. Golub and C. Reinsch (1970) "Singular Value Decomposition and Least Squares Solutions" Numer. Math. 14 403-420.

A. S. Householder (1970) "The Numerical Treatment of a Single Nonlinear Equation" McGraw-Hill, New York.

T. Kato (1960) "Estimation of Iterated Matrices, with Application to the von Neumann Condition" Numer. Math. 2 22-29.

T. Kato (1966) "Perturbation Theory for Linear Operators" Springer-Verlag, New York.

M. Marden (1966) "Geometry of Polynomials" Amer. Math. Soc. "Math. Surveys no. 3," Providence, R. I.

L. M. Milne-Thomson (1933) "The Calculus of Finite Differences" MacMillan, London.

L. Mirsky (1960) "Symmetric Gauge Functions and Unitarily Invariant Norms" Quart. J. Math., Oxford (2) 11 50-59.

R. Penrose (1954) "A Generalized Inverse for Matrices" Proc. Camb. Phil. Soc. 51 406-413.

(1955) "On Best Approximate Solutions of Linear Matrix Equations" ibid. 52 17-19.

V. Pereyra (1969) "Stability of General Systems of Linear Equations" Aequat. Math. 2 194-206.

A. Ruhe (1970) "Properties of a Matrix with a Very Ill-Conditioned Eigenproblem" Numer. Math. 15 57-60.

G. W. Stewart (1969) "On the Continuity of the Generalized Inverse" SIAM J. Appl. Math. 17 33-45.

B. L. van der Waerden (1950) "Modern Algebra, vol. II" translated by T. J. Benac, Ungar, New York.

R. J. Walker (1950) "Algebraic Curves" Princeton U. P., reprinted in 1962 by Dover, New York.

J. H. Wilkinson (1963) "Rounding Errors in Algebraic Processes" National Physical Lab. "Notes on Applied Science No. 32" Her Majesty's Stationery Office, London.

J. H. Wilkinson (1965) "The Algebraic Eigenvalue Problem" Oxford Univ. Press.

J. H. Wilkinson (1972) "Note on Matrices with a Very Ill-Conditioned Eigenproblem" Numer. Math. 19 176-8.