Honeywell

Advances in Robust Control

Final Technical Report
For the period November 1991 through February 1995
Contract No. F49620-92C-0007

January 1995

Prepared for:
Air Force Office of Scientific Research
Bolling Air Force Base, DC 20332

Honeywell Technology Center
3660 Technology Drive
Minneapolis, Minnesota 55418

C950028

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1/10/95	Final Report, 11/91 to 2/95

4. TITLE AND SUBTITLE
Advances in Robust Control

5. FUNDING NUMBERS

6. AUTHOR(S)
Blaise Morton

AFOSR-TR- 95 0089

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Honeywell Technology Center
Honeywell Inc.
3660 Technology Drive
Minneapolis, Minnesota 55418

8. PERFORMING ORGANIZATION REPORT NUMBER
C950028

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
Bolling Air Force Base, DC 20332

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unlimited

12b. DISTRIBUTION CODE
Unlimited

13. ABSTRACT (Maximum 200 words)
This report describes recent developments in the algebraic theory of structured singular values.

14. SUBJECT TERMS
Robust control, structured singular values

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Honeywell

Advances in Robust Control

Final Technical Report
For the period November 1991 through February 1995
Contract No. F49620-92C-0007

Prepared for:

Air Force Office of Scientific Research
Bolling Air Force Base, DC 20332

Prepared by:

Honeywell Technology Center
3660 Technology Drive
Minneapolis, Minnesota 55418

January 1995

PREFACE

This report on the algebraic theory of structured singular values wraps up an era of mathematical research on μ sponsored by the Air Force Office of Scientific Research (AFOSR). During the last thirteen years, the structured singular value has developed from a fledgling concept to a household word in the aerospace-control community. Today we have commercially supported software enabling control designers to use μ-tools in practical, full-scale applications. Though there remain unanswered theoretical questions, from an aerospace-control practitioner's point of view it is time to declare the problem solved.

Still, as a mathematician, I find it difficult to walk away from a problem as interesting and challenging as the structured singular value, especially when there remain basic, unanswered questions. Most practical problems we face can be solved by using the upper-bound estimate (see [1]) for μ, but this bound gives no information about the worst-case parameter sets.

The algebraic theory presented in this report is the product of my efforts, over the last five years, to compute the value of the μ-function exactly (for a special structure) and to construct worst-case parameter sets. Happily, I seem to have made some real progress in this direction, though there remains a troublesome gap in the theory for general values of N. At least, it seems that the case $N = 4$ can been solved by the dialytical method presented here.

While the theory was being developed for practical applications, I have always felt that structured singular values should be of interest to the theoretical mathematical community, though I have seen little evidence of such interest. From a fundamental point of view, the structured singular value is a natural generalization of the operator-theoretic concepts of norm and spectral radius, measuring the size of a linear operator. In addition, as is shown in this report, the computation of structured singular values leads directly to computational problems in intersection theory and invariant theory. Surely there are more general uses for such a natural and interesting concept, going beyond the control-theoretic applications that the aerospace community has found for it.

This report is written primarily for mathematicians -- to explain a little about the practical control applications and to describe the status of the algebraic theory. Industry could benefit from further progress in this area, especially if significant simplification in the computational approach could be found. I hope that someone with the right blend of interest, energy and talent will choose this theory as an object of study and improve on the results presented here.

I would like to thank John Doyle, Allen Tannenbaum, Dave Morrison, Joel Roberts and my colleagues at Honeywell (especially Mike Elgersma) for numerous helpful discussions during the development of the theory. Thanks also to Marc Jacobs at AFOSR who made it possible for me to spend some of my time working on this research topic.

Blaise Morton

9 January 1995

TABLE OF CONTENTS

1 Introduction

The subject of this report is a special family of functions defined on the set of square matrices with complex entries. Each of these functions measures the size of the matrix, according to some criterion. The operator norm $\| M \|$ of the matrix M defined by

$$\| M \| = max\{\| Mv \| \mid \| v \| = 1\} \tag{1}$$

is a special example in the family of functions we will be working with.

The functions of interest are now defined. Let $\mathcal{M}(N)$ denote the set of $N \times N$ complex matrices, and let $M \in \mathcal{M}(N)$. Let $J = \{j_1, \ldots, j_m\}$ be a partition of N, that is

$$\sum_{k=1}^{m} j_k = N \tag{2}$$

Let $\mathcal{D}_J$ denote the set of block diagonal matrices:

$$\mathcal{D}_J = \left\{ \begin{bmatrix} \Delta(i_1) & 0 & \cdots & 0 \\ 0 & \Delta(i_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Delta(i_m) \end{bmatrix} \mid \Delta(i_j) \in \mathcal{M}(i_j) \right\} \tag{3}$$

Let $\mathcal{D}_J(\delta)$ denote the set of all $\Delta \in \mathcal{D}_J$ of operator norm less than or equal to δ. Suppose for some value of δ there is a $\Delta \in \mathcal{D}_J(\delta)$ such that $Det(I+M\Delta) = 0$, where $I \in \mathcal{M}(N)$ is the identity matrix. In that case, denote by δ_0 the minimum such δ; note that $\delta_0 > 0$. John Doyle [1] defined the function $\mu_J(M)$:

$$\mu_J(M) = 1/\delta_0 \tag{4}$$

If $Det(I + M\Delta) \neq 0$ for all $\Delta \in \mathcal{D}_J$ then $\mu_J(M) = 0$.

The construction above defines a function μ_J for each positive integer N and each partition J of N. When N and J are fixed one writes $\mu(M)$ to denote $\mu_J(M)$. The function $\mu(M)$ is called the structured singular value of M.

When discussing the structured singular value for a particular partition J we often refer to the problem of computing $\mu_J(M)$ by the number of blocks on the diagonal of Δ, i.e. the cardinality of J. For example, if $N = 8$ and

$J = \{1,3,4\}$ we would refer to the computation of $\mu(M)$ as a three-block problem.

A more general definition of μ (allowing repeated blocks in Δ) and some of its fundamental properties were first presented in [1]. The definition above is satisfactory for our purposes here.

It is easy to check that for the single-block partition, $J = \{N\}$, the function μ is simply the operator norm. As is well known, the operator norm of the matrix M is computable by algebraic methods. Form the polynomial $p(r)$:

$$p(r) = Det(r^2 I - MM^*) \tag{5}$$

and the operator norm of M, $\mu(M)$, is the largest real root of the polynomial $p(r)$. The operator norm of M is the same thing as the maximum singular value $\overline{\sigma}(M)$. Numerically robust software for computing $\overline{\sigma}(M)$ (and all the other singular values) has been commercially available for nearly twenty years, predating the definition of the structured singular value. The structured singular value function $\mu(M)$ is a generalization of the maximum singular value function, as the nomenclature suggests.

The problem we consider in this report is the computation of the value of $\mu(M)$ for general values of N and the N-block partition $J = \{1,\ldots,1\}$. Our objective is to derive a polynomial $p(r)$ whose largest real root is the value $\mu(M)$. Such a polynomial is a generalization of the one in equation (5).

1.1 Computation of μ

The primary motivation for our research is to compute the value of μ. While much effort has gone into computing bounds for the various μ-functions, there has been relatively little progress toward computing μ exactly. A related problem of practical interest, which we also address here, is to determine the worst-case parameter sets, i.e. specific matrices $\Delta \in \mathcal{D}_J(\delta_0)$ for which $Det(I + M\Delta) = 0$. In the following we embark on an algebraic theory of structured singular values, the goal of which is to solve these two problems by algebraic methods. To minimize the complexity of the analysis we concentrate on the particular case in which the matrix Δ is an ordinary diagonal matrix. This line of research, started four years ago [2], has evolved to a point where existing computer tools are adequate to perform the computations for small N. By these methods, for the first time we have been able to compute

μ and find worst-case parameter sets for a four-parameter problem (i.e. four blocks, with each block a scalar parameter). See Section 8 for an example.

Unfortunately, the approach leads to impractical computational algorithms for large N. We do not know whether a more efficient algorithm is possible. The current theory leads to a conjecture on the growth in complexity as the number of parameters increases, but we do not have definite results. See the formula at the end of Section 4.

The μ-function, developed to solve practical engineering problems, motivates some interesting mathematical theory. The theory is now beyond the stage where practicing engineers are equipped to contribute, so it is hoped that mathematicians will take over and (perhaps) develop some new theory to answer the outstanding questions. Of special interest is the question whether a polynomial-time solution algorithm (polynomial in N) can be found. To help motivate the problem for non-engineers, a brief history of μ and its engineering significance is presented in the next introductory section.

We hope that mathematicians and engineers alike will find something of value in this presentation.

1.2 The Significance and History of μ

Modern control engineers approximate real systems with finite-dimensional linear time-invariant (FDLTI) models. These models are in the form of an inhomogeneous O.D.E:

$$\frac{dx}{dt} = Ax + Bu \tag{6}$$

$$y = Cx + Du \tag{7}$$

where x is the state-vector, u is the control input vector, y is the output vector, and A, B, C, D are constant matrices.

The technique of representing the system by the matrices A, B, C, D is convenient from a mathematical viewpoint, but its limitations must be recognized.

First, during the design phase, the parameters in these matrices cannot be predicted exactly. One often supposes a nominal system model, derived from physical principles, but the behavior of a real-world system will not coincide exactly with its nominal model. To account for this type of uncertainty, the designer may think of the model as an unknown point in a

specified multidimensional neighborhood of the nominal point in the space of A, B, C, D matrices.

Second, even after the system is built, one usually cannot measure all the matrix coefficients in a real-world situation. The frequency response (transfer function) of a physical system is a more practical thing to measure. For this second reason (and various other reasons), many practical engineers prefer frequency-domain models of their systems. To account for uncertainty in frequency-domain models, engineers augment their nominal models in two ways:

1. with exogenous noise inputs assumed to lie in a frequency-weighted unit ball in the Hardy space H_2

2. with perturbative transfer functions assumed to lie in a frequency-weighted unit ball in the Hardy space H_∞.

The reader should be aware that uncertainties in both time-domain and frequency-domain models often play important roles in the same control system design. The construction of a perturbation structure, to account for both types of uncertainties in a real-world system, is a key part of the control-engineering art.

The μ-approach to control theory uses both time-domain and frequency-domain concepts. First, a multidimensional box is constructed (mathematically) in the A, B, C, D space. The assumption is made that the system model could lie anywhere within this box. Next, by algebraic manipulations, a parametric representation of the entire box-worth of systems is derived. The associated parametric set of frequency-domain models is then augmented with exogenous inputs and perturbative operators to produce the perturbation structure. Finally, a controller is found (if possible) that guarantees good stability and performance properties for every system model contained in the perturbation structure. This process is called robust control design. There is a substantial body of theory underlying this construction (see [3] and the references contained there), here we shall only describe the basic concept behind the frequency-domain stability criterion.

The researcher in practical control theory should have a firm grasp of the frequency-domain theory and its practical significance. It is no exaggeration to say that the standard time-domain theory, by itself, is inadequate for practical applications. For the benefit of those who want more background,

the remainder of this sub-subsection is a quick introduction to the frequency-domain approach.

Begin by assuming that you are operating a real physical device, with knobs (control input u) and dials (measurement vector y). Suppose the physical system has been given to you at a steady-state condition – the input vector, output vector and internal system states are all constant. This assumption of steady-state condition cannot be verified by direct physical observation because the notion of the internal state is a theoretical construct. Even so, in many practical system, if the knobs are all fixed and the dials all indicate constant outputs, and if the system seems to be behaving properly in all other respects, the steady-state assumption is made.

Now wiggle one of the control inputs by a (small) sinusoidal signal and measure the (small) variations of each output signal. Let the perturbing input signal be $A_k sin(\omega t)$ in the k^{th} input channel and measure the additive perturbation on the j^{th} output signal. The j^{th} output will be perturbed by a signal that looks close to $B_{j,k} sin(\omega t + \epsilon_{j,k})$. The ratio $B_{j,k}/A_k$ is called the gain and the angle $\epsilon_{j,k}$ is called the phase shift of the system from the k^{th} input to the j^{th} output at the frequency ω. The data obtainable in this fashion can be collected into a family of complex matrices $F(\omega)$ parametrized by the frequency ω:

$$F(\omega) = [F_{j,k}(\omega)] = \left[\frac{B_{j,k} e^{\sqrt{-1}\epsilon_{j,k}}}{A_k}\right] \tag{8}$$

Call this matrix $F(\omega)$ the frequency response of the linear system. Assuming a linear system response to (small) perturbations, an analytic expression for $F(\omega)$ can be derived from the associated A, B, C, D matrices by using the Laplace transform. In deriving such an expression, it is customary in the engineering literature to let the Laplace transform variable s denote $\sqrt{-1}\omega$ and use the argument s instead of ω for the transfer function $T(s)$. The transfer function $T(s)$ is defined for all complex values of the parameter s as follows:

$$T(s) = D + C(sI - A)^{-1}B \tag{9}$$

For values of s on the imaginary axis, $s = \sqrt{-1}\omega$, we have $T(s) = F(\omega)$. The transfer function $T(s)$ is the basic object of attention in frequency-domain methods. The system is stable if and only if all the poles of the transfer function lie in the open left-hand plane of the complex s-domain.

1.2.1 Frequency-Domain Uncertainty – The Small Gain Theorem

From an analytical point of view, it is essential to recognize that a frequency-domain model (transfer function) $T(s)$, whether computed mathematically or measured experimentally, is an approximation. The principal weaknesses of such a model vary with application, although all practical models suffer from three basic limitations:

1. They vary as a function of steady-state condition
2. They are accurate only for small (perturbation) signal inputs
3. They are accurate only for a bounded range of frequencies

These modeling limitations complicate the analysis of real-world systems having significant nonlinearities, range of operating point and frequency-dependent model uncertainty. The μ-theory was developed primarily to address systems in which this last class of problems is the primary concern. We will be addressing problems associated with frequency-domain uncertainty in FDLTI systems.

One basic tenet of frequency-domain uncertainty is: model uncertainty tends to increase at high frequencies. The range of frequencies for which the model is accurate depends on the physical properties of the system elements. There are many factors that contribute to high-frequency uncertainty – two important examples are sensor limitations and actuator/power-supply limitations. With increasing frequency it becomes increasingly difficult and expensive to produce sensors and actuators that work close to any predictable analytical model. Because cost is a vital factor in system design, mathematical models are often not valid at frequencies above the range required for practical system operation.

Before proceeding, it is worth observing that model uncertainty is the primary motivation for feedback control. Considering the issue abstractly, if our models (including knowledge of the initial state) were perfect, there would be no need to consider adjustment of a control input based on sensor measurements. The control designer could include a simulation of the perfect model in his control laws and use a simulated value in place of any physical measurement. Thus, feedback control strategy depends in a fundamental way on the uncertainty characteristics of the model.

Let us consider briefly some practical issues (there are many) associated with uncertainty. First, if the open-loop system is inherently unstable (as many aerospace vehicles are) there is the issue of robust stabilization. There have been real system designs (poor ones) where the nominal closed-loop system is stable but a small change in model parameters produces an unstable closed-loop system. We want to insure that our systems remain stable for all parameter variations within a specified range of values.

Second, even if robust stability is not a problem, there is the issue of robust performance. Parameter variations, large or small, can influence the performance of a closed-loop system. Ideally, the closed-loop system should be relatively insensitive to variation of parameters within a set of anticipated ranges.

Third, whether the open-loop system is stable or not, we are concerned with the potential destabilizing effects of perturbations to the system state. State perturbations arise when the outside world interacts with our system, causing changes in state not predicted by our nominal model. A wind gust acting on an airplane is a typical example.

Finally, we are concerned with uncertainties in system dynamics, whether because of internal dynamics neglected in our design models or because of subsystem failures. Uncertainties of all four types are considered in a typical control design. We have a limited ability to represent them and to design control systems accommodating them, but these are the real issues that drive robust control design.

We now address analytic representation of frequency-domain uncertainty. Consider a nominal transfer function $T(s)$ subject to uncertainty. Much work has been devoted to modeling various types of uncertainty (an early reference is [4]), let us take additive uncertainty as one simple example. The resulting structure will be applicable to many other types of uncertainty.

Suppose the nominal model $T(s)$ has n inputs and m outputs. A feedback controller $K(s)$ of general type will have n outputs and $m+k$ inputs. The n outputs of $K(s)$ are identified with the inputs of $T(s)$, the last m inputs of $K(s)$ are identified with the outputs of $T(s)$, and the first k inputs of $K(s)$ are identified with command inputs from an external source (e.g. the pilot of an airplane). The closed-loop system now has only k inputs (the externally generated commands) but it still has the same m outputs as the original open-loop system $T(s)$.

We assume the transfer function $T(s)$ is perturbed additively by some unknown, stable function $W(s)\Delta(s)$ where $W(s)$ is a specified $m \times m$ transfer function (weighting matrix) and $\Delta(s)$ is an unknown transfer function in the unit ball (relative to the operator norm) of the space of $m \times n$ matrices with H_∞ entries. The open-loop system becomes

$$T_{pert}(s) = T(s) + W(s)\Delta(s) \tag{10}$$

Consider what happens when the fixed controller $K(s)$ is used to close the loop for $T_{pert}(s)$. The robust stability question is: is it true that for all $\Delta(s)$ in the unit ball of H_∞, the closed-loop system obtained by replacing $T(s)$ with $T_{pert}(s)$ is stable?

To answer the robust stability question we first perform an algebraic transformation to the problem. We assume that the nominal closed-loop system ($\Delta(s) = 0$) is stable. Then it is an elementary exercise to construct a stable transfer function $M(s)$, independent of $\Delta(s)$, with $m + k$ inputs and $n + m$ outputs with the following property: when the first n outputs of $M(s)$ are closed through $\Delta(s)$ to the first m inputs of $M(s)$, the closed loop system is the same as that obtained by closing the bottom loops of $T_{pert}(s)$ through the last m inputs of $K(s)$. Partitioning M into blocks, we find that the closed loop transfer function has the form:

$$\mathcal{F}_u(M, \Delta) = M_{22} + M_{21}\Delta(I - M_{11}\Delta)^{-1}M_{12} \tag{11}$$

From this expression we see that the closed-loop system will be stable for all Δ of norm less than 1 if and only if the factor $(I - M_{11}\Delta)^{-1}$ is stable for all such Δ. Clearly, if the H_∞ operator-norm of $M_{11}(s)$ is less than 1 we can conclude robust stability. In the case where Δ has no additional structure, this sufficient condition turns out to be necessary – that is the small gain theorem.

In those cases where the uncertainty is known to have block-diagonal structure, however, the sufficient condition $\| M_{11}(s) \| < 1$ is no longer necessary. Block-diagonal conditions on Δ arise naturally in many situations; for example, when a collection of physically-isolated, uncertain systems $T_j(s)$ are cascaded. Associated with each $T_j(s)$ will be a separate $\Delta_j(s)$, and the overall $\Delta(s)$ for the cascaded system will have block-diagonal form.

The small-gain theorem can be applied in the case of block-diagonal Δ, but the test is too conservative for many practical applications. Often, the

designer is forced to sacrifice too much performance in order to pass the small-gain test, so a better criterion is needed. From this need was born the structured singular value test, which is evaluated by computing the μ-function.

In applications, the transfer function $M_{11}(s)$ is represented in the computer via a state-space realization (A,B,C,D matrices). For each $s_j = \sqrt{-1}\omega_j$ in a grid on the imaginary axis the transfer function $M_{11}(s_j)$ is computed. The value of $\mu(M_{11}(s_j))$ (or some bounding function) is then computed. For example, the maximum singular value $\overline{\sigma}(M_{11}(s_j))$ is the upper bound corresponding to the small gain theorem. The values of the function are plotted on a log-magnitude vs. log-frequency graph and displayed to the design engineer. If it is found that $\mu(M_{11}(s_j))$ is less than one for all values on the grid, robust stability is concluded. Implicit in this approach is the assumption that the grid-size is fine enough to make it apparent whether the μ function exceeds 1 at any point along the imaginary axis.

If the μ-function is smaller than 1 at all points on the imaginary axis, robust stability follows from the properties of continuity of roots of a polynomial equation (with respect to its coefficients) and the maximum modulus theorem.

A different approach is required if one wants to consider only real values for some of the uncertain parameters, but we shall be concerned with complex parameters (complex entries in the blocks Δ_j) in this report.

1.2.2 The Origin of μ

The concept of the Structured Singular Value function, $\mu(M)$, is now more than a decade old. The early developmental stage of the concept can be traced back to 1977 when singular values were applied by a group of control-design engineers at Honeywell's Systems and Research Center to the analysis of multivariable linear time-invariant systems [5]. Their goal was to find a multiloop generalization of the famous small-gain theorem, so useful in the robust-stability analysis of single-input single-output (SISO) systems. At that time, the accepted practical technique for evaluating robust stability of multi-loop systems was to open a single loop at a time and apply the established SISO criteria (gain and phase margins). The Honeywell group recognized the inadequacy of this one-loop-at-a-time approach and aimed at a more reliable robust stability test. By analogy with the small-gain theorem,

the solution they sought was an analytic tool for frequency domain analysis.

Shortly after its theoretical development, the singular value approach was applied to a helicopter flight-control design [6]. During that study it was found that the methodology displayed some significant weaknesses. The difficulty was that representation of uncertainty in an unstructured way can lead to an overly conservative robustness criterion. In general, the robust stability test in the singular-value methodology was sufficient but far from necessary. A control system with adequate robust stability could fail the test. This shortfall was the reverse of the problem inherent with the single-loop-at-a-time approach, where each loop might look good individually but the multiloop system as a whole might lack robustness.

The overconservative nature of the singular value approach was easy to understand but not so easy to fix. A better methodology was sought: to be acceptable it had to be computable for problems of realistic size and its criteria for robust stability had to be as close to "necessary and sufficient" as possible. After several years, an acceptable solution was found in the form of the Structured Singular Value (SSV).

In 1981 the mathematical theory of the SSV was introduced by Doyle in the landmark paper [1]. At about the same time, an engineering application paper [4] appeared, showing how a wide variety of practical robust stability problems could be reduced to computing (or bounding) the SSV of a matrix transfer function, called the perturbation structure. From a theoretical point of view the SSV was a complete success: a system is robustly stable if and only if the unperturbed system is stable and the SSV function $\mu(M(s))$ of its associated perturbation structure $M(s)$ is less than 1 for all values of s on the imaginary axis. The general MIMO robust stability problem was reduced to a single class of numerical problems: given a complex $N \times N$ matrix, find a sharp, computable upper bound for $\mu(M)$. A computable upper bound, $\bar{\mu}(M)$, which turned out to be good for many applications (early examples provided in [7] and [8]), was provided by Doyle in [1], and a powerful methodology for robust control design and analysis was born.

It is worth emphasizing that the upper-bound function $\bar{\mu}$, not $\mu(M)$, is the function used in today's μ-methodology. This upper bound is the solution of a convex optimization problem and so is easily evaluated on a computer. It represents a significant improvement on the singular value test (maximum singular value $\bar{\sigma}(M)$), which is itself an upper bound on $\mu(M)$. The robust control synthesis methodology, μ-synthesis, is based on a weighted

H_∞ optimization theory associated with the upper bound $\bar{\mu}$.

Since the introduction of μ [1], this concept has been implemented in a variety of computer tools to quantify the robust stability of feedback control systems. Engineers have applied these tools successfully in the design phases of control systems for many advanced aerospace vehicles: the B-2 Bomber, Space Station, and the F-15 STOL Technology Demonstrator to name just a few. The μ-analysis and synthesis tools are now a standard part of modern aerospace system design.

1.3 Structure of the Report

After this introduction we move directly to the algebraic theory. In the second section we derive the basic set of algebraic equations, and in the third section we show how the elimination can be performed for the cases of 2, 3 or 4 parameters. Some abstract theory associated with the elimination for general numbers of parameters is postponed until Section 9.

In Section 4 we redirect attention to a related algebraic problem about which much is known. The link between the basic equations of Section 2 and the hyperdeterminant of a three-dimensional matrix is shown, so that we can apply the known results to our problem.

In Section 5 we introduce a third algebraic problem, also related to the basic equations. This approach was the starting point of the algebraic theory of [2], it has definite computational advantages in the cases of two and three parameters. The general results for this approach in the low-dimensional cases are described in Section 6. Some extensions of these results to higher dimension are presented in Section 7.

In Section 8 we illustrate the techniques described in the first seven sections by computing μ and worst-case parameter sets for numerical examples. Section 9 is an abstract theoretical presentation of the general approach, intended for more advanced researchers. Section 10 is a discussion and summary of results and outstanding issues.

2 Derivation of the Basic Equations

Let M be a complex, $N \times N$ matrix. As was shown in [1], the structured singular value $\mu(M)$ is given by:

$$\mu(M) = \sup_{\Theta \in \mathcal{D}} \rho(e^{i\Theta}M) \tag{12}$$

where ρ is the spectral radius function and $\mathcal{D}$ is the set of real-diagonal $N \times N$ matrices.

The sup function in equation 12 is (effectively) taken over a compact set, so there is some nonzero N-vector z such that, for some Θ,

$$e^{i\Theta}Mz = \mu(M)z \tag{13}$$

Our immediate goal is to determine a polynomial expression whose largest real root is the value $\mu(M)$. For that purpose we introduce the variable parameter r, and define the system of Hermitian forms $H_k(r)$:

$$H_k(r) = M_k^* M_k - r^2 e_k^* e_k \tag{14}$$

where M_k is the k^{th} row of the matrix M and e_k is the row vector whose k^{th} entry is 1, all others 0.

Lemma 1 *$\mu(M)$ is the largest real value of r for which there is a nonzero N-vector z satisfying*

$$z^* H_k(r) z = 0 \tag{15}$$

for $k = 1, \ldots, N$

Proof of Lemma 1: First we show that the conditions of the lemma are satisfied if $\mu(M)$ is substituted for r. Select $z \neq 0$ satisfying equation 13. Compute the squares of the magnitudes of the k^{th} entries on each side of equation 13:

$$(M_k z)^*(M_k z) = \mu(M)^2 z_k^* z_k \tag{16}$$

But this set of equations for $k = 1, \ldots, N$ is equivalent to equation 15.

Conversely, suppose r_0 is the largest real number such that some nonzero z satisfies equation 15. Then r_0 is the largest real number for which there are $\Theta \in \mathcal{D}$ and $z \neq 0$ such that

$$e^{i\Theta}Mz = r_0 z \tag{17}$$

That is, $\mu(M) = r_0$. □

If we think of $z, \bar{z}$ as a vector in a $2N$-dimensional real vector space, the system of equations 15 gives only N polynomial equations in the $2N+1$ real variables $z, \bar{z}, r$. Our goal is to eliminate z and $\bar{z}$ in order to obtain a single polynomial equation in r. Additional polynomial equations are needed for the elimination: these are derived from the condition that the value of r we seek is extremal.

Lemma 2 *Let $J(r, z, \bar{z})$ denote the $N \times 2N$ matrix*

$$J(r, z, \bar{z}) = \begin{bmatrix} z^* H_1(r) & z^T \overline{H_1(r)} \\ \cdots & \cdots \\ z^* H_N(r) & z^T \overline{H_N(r)} \end{bmatrix} \tag{18}$$

If $(r_0, z, \bar{z})$ is a solution to the system of equations 15 and r_0 is extremal among such solutions, then the rank of $J(r_0, z, \bar{z})$ is less than N.

Proof of Lemma 2: Consider the function F: $\mathbf{R}^{2N+1} \rightarrow \mathbf{R}^N$ defined by:

$$F(r, z, \bar{z}) = \begin{bmatrix} z^* H_1(r) z \\ \cdots \\ z^* H_N(r) z \end{bmatrix} \tag{19}$$

Observe that the $N \times 2N$ matrix $J(r, z, \bar{z})$ is the matrix of partial derivatives of F with respect to $z, \bar{z}$, that is:

$$J(r, z, \bar{z}) = \left[\frac{\partial F}{\partial z} \frac{\partial F}{\partial \bar{z}} \right] \tag{20}$$

At a point $(r, z, \bar{z})$ where F vanishes and the matrix J of partial derivatives has full rank N, the implicit function theorem [9] implies that, locally, the zero set of F can be parametrized smoothly by r and an N-dimensional subset of $(z, \bar{z})$. But then r cannot be extremal. This contradiction proves the lemma. □

The previous two lemmas lead directly to a system of polynomial equations in $(r, z, \bar{z})$ that must be satisfied at a solution of the equations 15 for which r is extremal. We use the symbol C to denote this set.

The set C contains:

1. the system of N equations 15

2. the $\left\langle \begin{matrix} 2N \\ N \end{matrix} \right\rangle$ size-N determinantal minors from the matrix $J(r, z, \bar{z})$.

It appears that, for a general class of matrices M, the system $\mathcal{C}$ generates a system of polynomials from which the variables $(z, \bar{z})$ can be eliminated. The eliminant of this system is a polynomial $p(r)$ whose coefficients are polynomials in the coefficients of M and $\overline{M}$. We call this polynomial $p(r)$ the r-polynomial. The roots of the r-polynomial are called r-values, and those r-values that appear at local maxima are called μ-values. The largest r-value is a μ-value which is equal to the value of the function $\mu(M)$.

The procedure required to perform the elimination depends on N. In the following we will show how this elimination can be performed for values of $N \leq 4$.

3 Performing the Elimination

We begin this section with a description of the general elimination technique used to derive the r-polynomial. The details of the computations in the cases $N \leq 4$ are then presented in subsections.

Recall the set of equations $\mathcal{C}$ defined in the previous section. We will use the polynomials in $\mathcal{C}$ to generate a system of polynomials bihomogeneous in $(z, \bar{z})$.

Definition: A polynomial $q(z, \bar{z})$ is called *bihomogeneous* of bidegree (i,j) if it is homogeneous of degree i when considered as a function of the vector z and homogeneous of degree j when considered as a function of the vector $\bar{z}$.

For example, the expressions $z_1\bar{z}_1$, $z_3\bar{z}_2$ and $z^*H_k(r)z$ are all bihomogeneous of bidegree (1,1).

For each N, let $P_N(i,j)$ denote the set of bihomogeneous polynomials of bidegree (i,j). The set $P_N(i,j)$ forms a finite-dimensional vector space over the real number field $\mathcal{R}$. Also, if $p_1(z,\bar{z}) \in P_N(i_1,j_1)$ and $p_2(z,\bar{z}) \in P_N(i_2,j_2)$ then $p_1(z,\bar{z})p_2(z,\bar{z}) \in P_N(i_1+i_2, j_1+j_2)$.

Our elimination approach makes use of an elementary combinatorial lemma, stated here without proof.

Lemma 1 *The dimension of $P_N(i,j)$ is:*

$$dim(P_N(i,j)) = \begin{pmatrix} N+i-1 \\ N-1 \end{pmatrix} \begin{pmatrix} N+j-1 \\ N-1 \end{pmatrix} \tag{21}$$

The notion of bihomogeneous polynomials extends in the obvious way when the coefficients of the polynomials lie in a general ring. Considered over the ring of real-polynomials in the variable r, all of the polynomials in $\mathcal{C}$ are bihomogeneous. Those in equation (15) are of bidegree (1,1), while each of the size-N minors of the the matrix $J(r,z,\bar{z})$ has bidegree (i,j) for a pair of non-negative integers i,j such that $i+j=N$.

Our strategy for obtaining the r-polynomial is a follows. Pick a pair of positive integers i_T, j_T with both i_T and j_T sufficiently large (depending on N). Over the ring of real-polynomials in r, consider the space of bihomogeneous polynomials $P_N(i_T, j_T)$. Now each polynomial q in $\mathcal{C}$ is bihomogeneous, let $bidegree(q) = (i_q, j_q)$. If $i_T \geq i_q$ and $j_T \geq j_q$, q can be multiplied

by any polynomial $h \in P_N(i_T - i_q, j_T - j_q)$ (considered over $\mathcal{R}$) to obtain $qh \in P_N(i_T, j_T)$. In this way the original set of polynomials in $\mathcal{C}$ can be used to generate a larger system of polynomials in the module $P_N(i_T, j_T)$ over the ring of polynomials in r. Each polynomial generated in this way must vanish on the set of points we seek. We can represent the entire set of equations in $P_N(i_T, j_T)$ as a single matrix equation:

$$A(r)\Phi(z, \bar{z}) = 0 \tag{22}$$

where $A(r)$ is a matrix of real polynomials in r and $\Phi(z, \bar{z})$ is a vector of basis monomials of the vector space $P_N(i_T, j_T)$ over $\mathcal{R}$.

The condition that remains to be verified is that, for a generic value of r, the rank of the matrix $A(r)$ is equal to the dimension of $P_N(i_T, j_T)$. Under this condition the r-polynomial $p(r)$ is nontrivial and theoretically well defined – it can be obtained (in principle) by computing the greatest common divisor of the maximal minors of the matrix $A(r)$.

The procedure is illustrated in the examples below. As will be seen, once the roots of the r-polynomial have been found we can recover the z-vector as well.

3.1 The Case $N = 2$

First we determine the polynomials in $\mathcal{C}$. There are two types:

1. The pair of hermitian forms $z^* H_1(r) z$, $z^* H_2(r) z$
2. The size-2 minors of the 2×4 matrix J

All the equations are bihomogeneous. The two equations of the first type are independent, of bidegree (1,1). As for the equations of type two, observe that the 2×4 matrix J has the form:

$$J(r, z, \bar{z}) = \begin{bmatrix} z^* H_1(r) & z^T \overline{H_1(r)} \\ z^* H_2(r) & z^T \overline{H_2(r)} \end{bmatrix} \tag{23}$$

Considering all the size-two minors of J, we find that there are the six equations of the second type: one of bidegree (2,0), four of bidegree (1,1), and one of bidegree (0,2). Consequently, the set $\mathcal{C}$ consists of eight equations.

It turns out that we do not need all the equations in $\mathcal{C}$ to perform the elimination. Let $\mathcal{B}$ consist of those six equations of bidegree (1,1). In fact, there is no loss of generality working with $\mathcal{B}$ instead of $\mathcal{C}$. To see this, consider the 2×2 submatrix formed by the first two columns of J. Note that

$$\begin{bmatrix} z^* H_1(r) z \\ z^* H_2(r) z \end{bmatrix} = \begin{bmatrix} z^* H_1(r) \\ z^* H_2(r) \end{bmatrix} z \tag{24}$$

The two forms in the vector on the left-hand side of equation 24 are in $\mathcal{B}$. Therefore, if z is in the zero set of $\mathcal{B}$, the determinant of the matrix on the right-hand side of equation 24 must be zero. The determinant of that matrix is one of the two equations in $\mathcal{C} \backslash \mathcal{B}$. Similarly, by taking the conjugate of equation 24, we see that the determinant of the last two columns of J (the other equation in $\mathcal{C} \backslash \mathcal{B}$) must vanish as well. We conclude that the zero set of $\mathcal{C}$ is the same as the zero set of $\mathcal{B}$.

Define $\Phi(z, \bar{z})$ by

$$\Phi(z, \bar{z}) = \begin{bmatrix} z_1 \bar{z}_1 \\ z_1 \bar{z}_2 \\ z_2 \bar{z}_1 \\ z_2 \bar{z}_2 \end{bmatrix} \tag{25}$$

There is a 6×4 matrix $A(r)$ consisting of (in general) quadratic functions of r such that the six equations in $\mathcal{B}$ may be written in the form of equation 22. For general matrices M the derived matrix $A(r)$ will be rank four except for those values of the parameter r in a finite set, denoted Σ. From the matrix $A(r)$ one can derive a polynomial $p(r)$ of minimal degree whose roots are the r-values in Σ. In Appendix A the polynomial $p(r)$ is derived.

Suppose the set of real r-values in Σ is known. Let r_0 be a value in Σ. We will show how the vector z can be recovered.

Let V be a nonzero vector such that

$$A(r_0) V = 0 \tag{26}$$

Using the four entries of V, form the 2×2 matrix Q:

$$Q = \begin{bmatrix} V_1 & V_2 \\ V_3 & V_4 \end{bmatrix} \tag{27}$$

If the value r_0 corresponds to a solution of $\mathcal{B}$, the matrix Q is rank-one. If this condition is satisfied, perform the dyadic decomposition:

$$Q = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} [\bar{z}_1 \bar{z}_2] \tag{28}$$

and the vector (z_1, z_2) is a solution of the equations $\mathcal{B}$ for $r = r_0$.

3.2 The Case $N = 3$

For $N = 3$ the approach is similar to the case $N = 2$. Again, we define a set $\mathcal{C}$ consisting of two types of equations:

1. The three hermitian forms $z^* H_1(r) z,\ z^* H_2(r) z,\ z^* H_3(r) z$
2. The size-3 minors of the 3×6 matrix J

Again, the equations of the first type are independent, of bidegree (1,1). This time, however, the equations of type two have bidegree (i,j), where $i + j = 3$.

We derive from $\mathcal{C}$ a set $\mathcal{B}$ of bihomogeneous equations of bidegree (2,1). First, we have the nine equations obtained by multiplying each of the type-one equations by each of z_1, z_2, z_3. To this set we add the nine minor determinants from J of bidegree (2,1). Constructed in this way, the set $\mathcal{B}$ consists of 18 equations of bidegree (2,1).

By equation 21, the dimension of $P_3(2,1)$ is 18. The equations in $\mathcal{B}$ form a system of 18 equations in the 18 variables that span $P_3(2,1)$. Picking a basis, let $\Phi(z, \bar{z})$ denote

$$\Phi(z, \bar{z}) = [z_1^2 \bar{z}_1, z_1^2 \bar{z}_2, z_1^2 \bar{z}_3, z_1 z_2 \bar{z}_1, \ldots, z_3^2 \bar{z}_3]^T \tag{29}$$

With respect to this vector Φ we can write the matrix $A(r)$ of equation 22. It is 18×18, nine of its rows are affine in r while the other nine are cubic. The determinant of the matrix $A(r)$ is a polynomial $p(r)$ of degree 36.

We have shown by numerical computations that, for some matrices M, the polynomial p(r) obtained in this way is not identically zero. For such M, let Σ denote the set of real roots of $p(r)$. For a given $r_0 \in \Sigma$ we can recover the vector z by a process similar to the case $N = 2$.

Let V be a nonzero vector such that

$$A(r_0) V = 0 \tag{30}$$

Using the 18 entries of V, form the 6×3 matrix Q:

$$Q = \begin{bmatrix} V_1 & V_2 & V_3 \\ V_4 & V_5 & V_6 \\ V_7 & V_8 & V_9 \\ V_{10} & V_{11} & V_{12} \\ V_{13} & V_{14} & V_{15} \\ V_{16} & V_{17} & V_{18} \end{bmatrix} \tag{31}$$

If the value r_0 corresponds to a solution of $\mathcal{B}$, the matrix Q is rank-one, hermitian. If this condition is satisfied, perform the dyadic decomposition:

$$Q = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} [\bar{z}_1 \bar{z}_2 \bar{z}_3] \tag{32}$$

The solution vector (z_1, z_2, z_3) can be obtained by taking the conjugate of the right dyadic component of Q, or by computing the dyadic factors of the matrix P given by:

$$P = \begin{bmatrix} y_1 & y_2 & y_3 \\ y_2 & y_4 & y_5 \\ y_3 & y_5 & y_6 \end{bmatrix} \tag{33}$$

The matrix P should be rank-one, symmetric, if r_0 corresponds to a solution of $\mathcal{B}$. In that case the decomposition

$$P = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} [z_1, z_2, z_3] \tag{34}$$

should yield a vector z that is proportional to the conjugate of the right dyadic component of Q.

3.3 The Case $N = 4$

The procedure for $N = 4$ is similar to that required for $N = 2$ and $N = 3$. This time, the set $\mathcal{B}$ consists of bihomogeneous polynomials of bidegree

(3,3). One subset of equations is obtained by multiplying the four eqution of bidegree (1,1) by the 100 basis monomials in $P_4(2,2)$, yielding 400 equations in $\mathcal{B}$. The second subset is obtained by multiplying the 36 4×4 minors of J of bidegree (2,2) by the 16 basis monomials of $P_4(1,1)$, yielding another 576 equations of the same type.

The space $P_4(3,3)$ has dimension 400, so the matrix $A(r)$ has size 976×400. By numerical computation we have verified that there are 4×4 matrices M for which the matrix $A(r)$ has rank 400 for generic values of r. There is a polynomial $p(r)$ whose roots form the set Σ of r-values for which the rank of $A(r)$ is less than 400. At the moment, we do not know the degree of this polynomial, though we suspect (for reasons that will be presented later) the degree is 272.

The value of the z-vector can be computed by methods similar to those described in the cases $N = 2$ and $N = 3$. A numerical example is presented in Section 8.

3.4 The Case $N > 4$

We have not attempted to compute any examples for the case $N = 5$ or larger. We suspect these cases can be handled by a similar computational procedure, but we have no proof.

The degree of the polynomial $p(r)$ in the case $N = 5$ is believed to be 2150. A formula for the (suspected) degree of $p(r)$ for $N > 4$ is presented at the end of the next section on hyperdeterminants.

4 Hyperdeterminants

The r-polynomial has a close tie with the theory of hyperdeterminants. Hyperdeterminants were invented by Cayley [10], a modern treatment is presented in [11]. In this report we will discuss only those facts about hyperdeterminants needed to help understand the r-polynomial. The reader is referred to [11] for proofs of the results stated here.

First, some basic definitions. For an integer $N > 0$, suppose we are given a collection of N^3 numbers indexed by integers i, j and k each running from 1 to N. Let $W = [W_{ijk}]$ denote such a collection. We call W a three-dimensional matrix of size N. The numbers that constitute W can be pictured in a three-dimensional array – a generalization of the planar array of a standard (two-dimensional) matrix. Three-dimensional matrices of size N often arise when a three-index tensor is written relative to a basis of the underlying N-dimensional vector space.

We will be using basic facts about discriminant polynomials. We consider a general form f over $\mathbf{C^N}$ and its zero-set, the hypersurface $Z(f)$ of the projective space P^{N-1}. If f is a form of degree d then the gradient of f, ∇f, is a vector of forms homogeneous of degree $d-1$. For a *generic* form f there are no points $z \in Z(f)$ such that $\nabla f(z) = 0$. In that case, the hypersurface $Z(f)$ is a smooth manifold of dimension $N-2$ in P^{N-1}.

For some forms f the hypersurface $Z(f)$ is not smooth (i.e. $Z(f)$ is singular). There is an irreducible polynomial in the coefficients of f, called the discriminant of f, which vanishes if and only if there is a point in $Z(f)$ where the gradient of f vanishes.

The general theory of discriminants has been well studied; much is known about them. It is a classic result that the discriminant of a homogeneous form f of degree d over $\mathbf{C^N}$ is a polynomial of degree $N(d-1)^{N-1}$ in the $\binom{N+d-1}{N}$ coefficients of f (see [12], p. 99).

We are now prepared to discuss hyperdeterminants of three-dimensional matrices. Though we do not present an explicit construction, the hyperdeterminant of W is a polynomial function $q(W)$ of the values $\{W_{ijk}\}$. The significance of the hyperdeterminant polynomial is as follows:

Lemma 1 *There is a polynomial $q(W)$ in the variables $\{W_{ijk}\}$, called the hyperdeterminant polynomial, which vanishes if and only if there are three*

nonzero N-vectors x,y,z such that:

$$\forall k \sum_{i,j=1}^{N} W_{ijk} x_i y_j = 0 \tag{35}$$

$$\forall j \sum_{i,k=1}^{N} W_{ijk} x_i z_k = 0 \tag{36}$$

$$\forall i \sum_{j,k=1}^{N} W_{ijk} y_j z_k = 0 \tag{37}$$

For more details concerning this lemma and the other results stated in this section, the reader is referred to [11].

If the entries of W are polynomial functions of a variable r, the hyperdeterminant of W is a polynomial in r.

The motivation for the nomenclature "hyperdeterminant" can be understood by a comparison with the ordinary determinant function for (two-dimensional) matrices. Given an $N \times N$ matrix A there is an associated bilinear form α:

$$\alpha(x, y) = \sum_{i,j=1}^{N} A_{ij} x_i y_j \tag{38}$$

Suppose there are nonzero vectors x, y such that $\alpha(x, y) = 0$ and

$$\partial\alpha(x, y)/\partial x = 0 \qquad \partial\alpha(x, y)/\partial y = 0 \tag{39}$$

or, equivalently:

$$\forall j \sum_{i=1}^{N} A_{ij} x_i = 0 \qquad \forall i \sum_{j=1}^{N} A_{ij} y_j = 0 \tag{40}$$

Either of the two equations in 40 imply that the determinant of the matrix A must vanish. Conversely, given that the determinant of A is zero, we can find a pair of nonzero vectors x, y such that equation 40 is satisfied.

By the argument just given, we see that the determinant of the matrix A is exactly the discriminant of the associated bilinear form α. The analogy with

hyperdeterminants can now be made. Associated with the three-dimensional matrix W of size N is a tri-linear form $\omega(x, y, z)$:

$$\omega(x, y, z) = \sum_{i,j,k=1}^{N} W_{ijk} x_i y_j z_k \tag{41}$$

The discriminant of the form ω is the hyperdeterminant of W.

Having vaguely defined the hyperdeterminant of a three-dimensional matrix, we now show how the basic equations of Section 2 lead to a hyperdeterminantal condition.

Recall the basic equations:

$$\forall k \; z^* H_k(r) z = 0 \tag{42}$$

and the rank condition, that $J(r, z, \bar{z})$ defined by:

$$J(r, z, \bar{z}) = \begin{bmatrix} z^* H_1(r) & z^T \overline{H_1(r)} \\ \cdots & \cdots \\ z^* H_N(r) & z^T \overline{H_N(r)} \end{bmatrix} \tag{43}$$

should be rank less than N.

From the rank condition we can derive the following lemma

Lemma 2 *Let $(r, z, \bar{z})$, $r \in \mathbf{R}$, $z \neq 0$, be a point where the matrix $J(r, z, \bar{z})$ is rank $< N$. Then there is a nonzero, real vector $t \in \mathbf{R^N}$ such that*

$$\sum_{k=1}^{N} t_k H_k(r) z = 0 \qquad \sum_{k=1}^{N} t_k z^* H_k(r) = 0 \tag{44}$$

Proof of Lemma 2: Because J is rank deficient, there exists a nonzero $t \in \mathbf{C^N}$ such that

$$\sum_{k=1}^{N} t_k z^* H_k(r) = 0 \qquad \sum_{k=1}^{N} t_k z^T \overline{H_k(r)} = 0 \tag{45}$$

Equivalently, taking the conjugate-transpose of the first equation and the transpose of the second (and using the fact that $H_k(r)$ is hermitian):

$$\sum_{k=1}^{N} \bar{t}_k H_k(r) z = 0 \qquad \sum_{k=1}^{N} t_k H_k(r) z = 0 \tag{46}$$

Considering these two equations together, we see that the vector t can be replaced equally well by both its real part and its imaginary part in all the equations above. Because t is nonzero, one of these two real vectors is nonzero. We conclude that for some nonzero real vector t both equations 44 are satisfied. □

Define the three-dimensional matrix W by:

$$[W_{ijk}] = [H_k(r)]_{i,j} \tag{47}$$

Now replace $\bar{z}$ by an independent variable y. The original N equations then have the form:

$$\forall k \sum_{i,j=1}^{N} W_{ijk} y_i z_j = 0 \tag{48}$$

With this same substitution, the two equations in the previous lemma become:

$$\forall i \sum_{j,k=1}^{N} W_{ijk} z_j t_k = 0 \qquad \forall j \sum_{i,k=1}^{N} W_{ijk} y_i t_k = 0 \tag{49}$$

We see that the hyperdeterminant of the three-dimensional matrix W must vanish. The hyperdeterminant of W is a polynomial $q(r)$, one of whose roots is $\mu(M)$. We conclude that the r-polynomial $p(r)$ divides the hyperdeterminant $q(r)$.

Remark 1 In our argument to show that the hyperdeterminant of W vanishes, no use was made of the fact that the vector t may be chosen real. The reality condition on t is significant, however, when we consider whether the hyperdeterminant $q(r)$ is the same as the r-polynomial $p(r)$. Consider a real value r_0 at which the hyperdeterminant of W vanishes, and let Σ_{r_0} be the associated set of (y, z, t) satisfying the system of equations in Lemma []. Then r_0 is a root of $p(r)$ if and only if some point in Σ_{r_0} satisfies the reality conditions:

$$0 \neq y = \bar{z} \qquad 0 \neq t \in \mathbf{R^N} \tag{50}$$

In general, we do not know whether the degree of $q(r)$ will equal the degree of the r-polynomial $p(r)$.

We are now in a position to take advantage of theoretical results concerning hyperdeterminants. One such result is:

Lemma 3 *Let $q(W)$ denote the hyperdeterminant of the three-dimensional matrix W of size N. The degree of $q(W)$ is:*

$$deg(q) = \sum_{0 \le j \le (N-1)/2} \frac{(j+N)!}{j!^3 (N-1-2j)!} 2^{N-1-2j} \tag{51}$$

This formula is Corollary 2.9 on page 456 of [11]. For $N = 2$, 3, 4 and 5 the values are 4, 36, 272 and 2150 respectively.

It remains an interesting open question whether the r-polynomial is equivalent to the hyperdeterminant polynomial for all N.

5 Families of Hypersurfaces

The hyperdeterminant approach discussed in the previous section has certain theoretical advantages, but so far we have not been able to make use of it in an efficient computational method. An approach that lends itself more readily to computations is the subject of this section. The method described below was first applied in [2].

We adopt the notation used in the statement of Lemma 2 in Section 4. Suppose r is a real variable, z is a complex vector variable and t is a real vector variable. Define $H(t,r)$ to be the $N \times N$ matrix of forms:

$$H(t,r) = \sum_{k=1}^{N} t_k H_k(r) \tag{52}$$

An immediate corollary of Lemma 4.2 can be stated:

Corollary 1 *Let $f_r(t)$ denote the polynomial function:*

$$f_r(t) = Det(H(t,r)) \tag{53}$$

Consider r to be fixed, so that f_r is a form of degree N in the vector t. If there exists a nonzero vector z such that $(r,z,\bar{z})$ satisfies the conditions of Lemma 2, then the discriminant of f_r vanishes. That is, there is a nonzero vector t_0 such that:

$$\forall i \quad \frac{\partial f_r(t_0)}{\partial t_i} = 0 \tag{54}$$

Proof of Corollary: Fix z and r as above, and select an invertible $N \times N$ matrix U in which z is the first column. Consider the matrix of forms $L(t)$ (we suppress the dependence on r, which is fixed):

$$L(t) = U^* H(t,r) U \tag{55}$$

and note:

$$Det(L(t)) = |Det(U)|^2 Det(H(t,r)) = |Det(U)|^2 f_r(t) \tag{56}$$

Observe that $L(t)$ is of the form:

$$L(t) = \begin{pmatrix} 0 & \alpha^*(t) \\ \alpha(t) & A(t) \end{pmatrix} \tag{57}$$

where α and α^* are linear forms with values in $\mathbf{C^{N-1}}$ such that $\alpha(t_0) = 0$ and $\alpha^*(t_0) = 0$. An elementary computation reveals, for all t:

$$Det(L(t)) = \alpha^*(t)A(t)^\dagger\alpha(t) \tag{58}$$

where $A(t)^\dagger$ is the matrix of cofactors (adjoint matrix) of $A(t)$. Examining the first-order behavior of this last expression at t_0 we find that

$$\forall i \quad \frac{\partial Det(L(t_0))}{\partial t_i} = 0 \tag{59}$$

It follows that the discriminant of f_r vanishes. □

This corollary provides an advantage from a computational viewpoint. The advantage is that the discriminant condition on the form f_r leads to an elimination of the N variables t_i from a system of N homogeneous equations. The goal of the elimination is a polynomial $p_d(r)$ whose roots we compute numerically. By comparison, the basic equations derived in Section 2 require elimination of $2N$ variables, while the hyperdeterminant approach of Section 4 requires elimination of $3N$ variables. In general, the computational difficulty in performing an elimination increases rapidly with the number of variables to be eliminated, so the advantage of computing $p_d(r)$ instead of the other eliminants is clear. The disadvantage is that this approach requires modification in the case where $N > 3$.

This corollary provides a nice geometric view of the problem. Thinking of f_r as a family of forms, parametrized by the variable r, there is the associated 1-parameter family of algebraic zero sets V_r of the projective space $\mathbf{P^{N-1}}$. The nonvanishing of the discriminant $p_d(r)$ is associated with the geometric condition that V_r is a smooth submanifold (nonsingular variety). As the parameter r varies, we can imagine the family V_r deforming continuously through smooth varieties for an open set of values of r, with occasional singular sets occurring at the roots of $p_d(r)$. This simple geometric picture is not exactly correct, as we shall see, but the concept can be made precise.

Because the converse of the corollary is not true, the polynomial $p_d(r)$ is not the same as the μ-polynomial $p(r)$. All we can conclude is that the polynomial we really care about, $p(r)$, is a factor of the more easliy computed polynomial $p_d(r)$.

This lemma is applied in Section 6 to the two tractable cases $N = 2$ and $N = 3$. The difficulty in the case $N > 3$ is discussed along with some related general theory in Section 7.

6 Families of Hypersurfaces – Low Dimensional Computations

In this section we apply the family-of-hypersurfaces approach to look at the cases $N = 2$ and $N = 3$. At the end we explain a complication that arises in applying this technique to the cases $N > 3$. As in the previous section, r is a real variable and t is a real vector variable. Given the $N \times N$ matrix of forms:

$$H(t,r) = \sum_{k=1}^{N} t_k H_k(r) \tag{60}$$

we are interested in the polynomial $f_r(t)$ defined by

$$f_r(t) = Det(H(t,r)) \tag{61}$$

and its gradient with respect to t.

6.1 The Case $N = 2$

In the case $N = 2$ the polynomial $f_r(t)$ has the form:

$$f_r(t) = c_{20}(r)t_1^2 + c_{11}(r)t_1t_2 + c_{02}(r)t_2^2 \tag{62}$$

where each of the functions $c_{ij}(r)$ is quadratic in r^2.

For each value of r the zero-set of f_r defines a pair of points in the projective line $\mathbf{P^1}$. The vanishing of the discriminant is equivalent to the geometric condition that the two points are coincident.

The two algebraic equations are:

$$0 = \frac{\partial f_r(t)}{\partial t_1} = 2c_{20}(r)t_1 + c_{11}(r)t_2 \tag{63}$$

$$0 = \frac{\partial f_r(t)}{\partial t_i} = c_{11}(r)t_1 + 2c_{02}(r)t_2 \tag{64}$$

This pair of equations can have a nontrivial solution only if the 2×2 coefficient matrix $C(r)$ defined by:

$$C(r) = \begin{pmatrix} 2c_{20}(r) & c_{11}(r) \\ c_{11}(r) & 2c_{02}(r) \end{pmatrix} \tag{65}$$

is singular. The discriminant polynomial $p_d(r)$ is:

$$p_d(r) = Det(C(r)) = 4c_{20}(r)c_{02}(r) - c_{11}(r)^2 \tag{66}$$

which is quartic (degree 4) in r^2. There are at most four nonnegative real roots r_j of p_d, one of which is the value $\mu(M)$. It can be shown that, in this case, $\mu(M)$ is the largest real root of p_d (see Appendix A).

6.2 The Case $N = 3$

In the case $N = 3$ the polynomial $f_r(t)$ has the form:

$$\begin{aligned} f_r(t) &= c_{210}(r)t_1^2t_2 + c_{201}(r)t_1^2t_3 + c_{120}(r)t_1t_2^2 + c_{111}(r)t_1t_2t_3 + \\ &\quad c_{102}(r)t_1t_3^2 + c_{021}(r)t_2^2t_3 + c_{012}(r)t_2t_3^2 \end{aligned} \tag{67}$$

where each of the functions $c_{ijk}(r)$ is cubic in r^2. There are no terms of the form t_j^3 because each matrix coefficient of t_j in the form $H(t,r)$ has rank at most 2 (as can be verified by examining the construction of $H(t,r)$).

For each value of r the zero-set of f_r defines a cubic curve in the projective plane $\mathbf{P^2}$. The vanishing of the discriminant is equivalent to the geometric condition that the curve is singular.

The three algebraic equations are:

$$0 = \frac{\partial f_r(t)}{\partial t_1} = 2c_{210}(r)t_1t_2 + 2c_{201}(r)t_1t_3 + c_{120}(r)t_2^2 + c_{111}(r)t_2t_3 + c_{102}(r)t_3^2 \tag{68}$$

$$0 = \frac{\partial f_r(t)}{\partial t_2} = c_{210}(r)t_1^2 + 2c_{120}(r)t_1t_2 + c_{111}(r)t_1t_3 + 2c_{021}(r)t_2t_3 + c_{012}(r)t_3^2 \tag{69}$$

$$0 = \frac{\partial f_r(t)}{\partial t_3} = c_{201}(r)t_1^2 + c_{111}(r)t_1t_2 + 2c_{102}(r)t_1t_3 + c_{021}(r)t_2^2 + 2c_{012}(r)t_2t_3 \tag{70}$$

Each of the three equations above can be used to derive four linear equations in the set W of monomials:

$$W = \{t_1^3t_2, t_1^3t_3, t_1^2t_2^2, t_1^2t_2t_3, t_1^2t_3^2, t_1t_2^3, t_1t_2^2t_3, t_1t_2t_3^2, t_1t_3^3, t_2^3t_3, t_2^2t_3^2, t_2t_3^3\} \tag{71}$$

For example, equation 68 can be multiplied by each of the four monomials $\{t_1^2, t_1t_2, t_1t_3, t_2t_3\}$ with the result linear in the monomials of W. By this

dialytic technique we obtain a set of twelve linear equations in the twelve monomials of W.

These twelve equations can have a nontrivial solution only if the 12×12 coefficient matrix $C(r)$ defined by:

$$\begin{pmatrix}
2c_{210} & 2c_{201} & c_{120} & c_{111} & c_{102} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2c_{210} & 2c_{201} & 0 & c_{120} & c_{111} & c_{102} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2c_{210} & 2c_{201} & 0 & c_{120} & c_{111} & c_{102} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 2c_{210} & 2c_{201} & 0 & c_{120} & c_{111} & c_{102} \\
c_{210} & 0 & 2c_{120} & c_{111} & 0 & 0 & 2c_{021} & c_{012} & 0 & 0 & 0 & 0 \\
0 & c_{210} & 0 & 2c_{120} & c_{111} & 0 & 0 & 2c_{021} & c_{012} & 0 & 0 & 0 \\
0 & 0 & c_{210} & 0 & 0 & 2c_{120} & c_{111} & 0 & 0 & 2c_{021} & c_{012} & 0 \\
0 & 0 & 0 & c_{210} & 0 & 0 & 2c_{120} & c_{111} & 0 & 0 & 2c_{021} & c_{012} \\
c_{201} & 0 & c_{111} & 2c_{102} & 0 & c_{021} & 2c_{012} & 0 & 0 & 0 & 0 & 0 \\
0 & c_{201} & 0 & c_{111} & 2c_{102} & 0 & c_{021} & 2c_{012} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & c_{201} & 0 & 0 & c_{111} & 2c_{102} & 0 & c_{021} & 2c_{012} & 0 \\
0 & 0 & 0 & 0 & c_{201} & 0 & 0 & c_{111} & 2c_{102} & 0 & c_{021} & 2c_{012}
\end{pmatrix}$$

is singular. The discriminant polynomial $p_d(r)$ is:

$$p_d(r) = Det(C(r)) \tag{72}$$

which is degree 36 in r^2. There are at most 36 nonnegative real roots r_j of p_d, one of which is the value $\mu(M)$.

Unfortunately, in this case the rank deficiency of $C(r)$ at $r = r_0$ is necessary but not sufficient for r_0 to be a root of the μ-polynomial. There is a Zariski-closed set of 3×3 matrices M for which the matrix $C(r)$ is rank-deficient for all r. It is true, however, that a converse can be derived for 3×3 matrices M in the open set for which $C(r)$ is full-rank for at least one value of r. This issue and the complexity of the general situation is discussed further in the next section.

6.3 The Cases $N > 3$

In the case $N = 4$ the polynomial $f_r(t)$ has the form:

$$f_r(t) = c_{2200}(r)t_1^2t_2^2 + c_{2110}(r)t_1^2t_2t_3 + \cdots + c_{0022}(r)t_3^2t_4^2 \tag{73}$$

where each of the functions $c_{ijkl}(r)$ is quartic in r^2. As in the case $N = 3$ there are no monomials including terms of the form t_j^m for $m > 2$ because each matrix coefficient of t_j in the form $H(t, r)$ has rank at most 2.

For each value of r the zero-set of f_r defines a quartic surface in the projective space $\mathbf{P^3}$. The vanishing of the discriminant is equivalent to the geometric condition that the surface is singular.

If we try to proceed as in the case $N = 3$, we eventually find that the dialytic method applied to the discriminant of f_r does not work. The problem is that, for all r, the zero set of f_r is a singular hypersurface in $\mathbf{P^3}$. The coordinate vertices (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) are always singular points.

The coordinate vertices are singular points for the $Z(f_r)$-hypersurfaces for all $N > 3$. For $N = 4$ they are double points, for $N = 5$ they are triple points, etc.

There are techniques for dealing with this situation; we shall discuss some of them in the next section.

7 Families of Hypersurfaces – General Results

In this section we examine more closely the family-of-hypersurfaces approach to general cases. As in the previous section, r is a real variable and t is a real vector variable. Given the $N \times N$ matrix of forms:

$$H(t,r) = \sum_{k=1}^{N} t_k H_k(r) \tag{74}$$

we are interested in the polynomial $f_r(t)$ defined by

$$f_r(t) = Det(H(t,r)) \tag{75}$$

and its gradient with respect to t.

7.1 The Case $N = 4$

As mentioned in the previous section, in the case $N = 4$ the polynomial $f_r(t)$ has the form:

$$f_r(t) = c_{2200}(r)t_1^2t_2^2 + c_{2110}(r)t_1^2t_2t^3 + \cdots + c_{0022}(r)t_3^2t_4^2 \tag{76}$$

where each of the functions $c_{ijkl}(r)$ is quartic in r^2.

For each value of r the zero-set of f_r defines a singular quartic surface $Z(f_r)$ in the projective space $\mathbf{P^3}$. Fix a value of r and consider the geometry of this surface $V = Z(f_r)$. The values of r associated with the r-polynomial are associated with singular points on V of a special type. In the following we will determine more precise conditions that this singular point must satisfy to be associated with the r-polynomial.

As we observed earlier, for all values of r the coordinate vertices (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) are singular points of V so the discriminant of f_r vanishes identically. Consequently, Corollary 1 of Section 5 provides no information about the value of μ. Corollary 1 can be strengthened by adding a condition that eliminates these coordinate-vertex singular points and others like them. In this subsection we show how the result can be strengthened in the case $N = 4$.

First we consider some general fact about matrices. Denote by $M(k,l)$ the set of complex $k \times k$ matrices of rank no more than l. Inside of $M(4,4)$ lies the codimension-4 algebraic subset $M(4,2)$ of those matrices of rank 2 or less. (In general, the codimension of the set $M(k,l)$ inside of $M(k,k)$ for $k \geq l$ is $(k-l)^2$.) Let us work in the projective space $\mathbf{P}^{15}$ associated with the nonzero 4×4 matrices.

Inside of $\mathbf{P}^{15}$ the form $H(t,r)$ provides an imbedded space Q isomorphic to $\mathbf{P}^3 \times \mathbf{P}^1$. This space Q, being four-dimensional, should be expected to intersect $M(4,2)$ in a nonempty set of points. We observe:

Lemma 1 *At each point (t,r) in $Q \cap M(4,2)$ the gradient of the function $f_r(t)$ vanishes.*

Proof of Lemma: Let $M_i(t,r)$ denote the i^{th} column of $H(t,r)$. The partial derivative of the determinant with respect to t_j can be written:

$$\frac{\partial det}{\partial t_j} = det(\frac{\partial M_1}{\partial t_j} M_2 M_3 M_4) + \cdots + det(M_1 M_2 M_3 \frac{\partial M_4}{\partial t_j}) \qquad (77)$$

Each term in the sum on the right-hand side vanishes because any three columns $M_i(t,r)$ are linearly dependent. □

It is the presence of points (t,r) for which $H(t,r)$ has excess rank deficiency that complicates the converse of Corollary 1, Section 5. The next two lemmas, true for general values of N, clarify the problem.

Lemma 2 *Consider the case of general N. Suppose that for some real value r_0 the gradient of f_{r_0} vanishes at a real vector t_0 for which $H(t_0, r_0)$ has rank $N-1$. Then there is a nonzero complex vector z such that $(r_0, z, \overline{z})$ satisfies the conditions of Lemma 2 of Section 2; that is, r_0 is an r-value.*

Proof of Lemma: Let (t_0, r_0) satisfy the hypotheses of the lemma, and let $z(t)$ be any column of the adjoint matrix of $H(t,r)$ such that $z(t_0)$ is nonzero (there is such a column because of the rank $N-1$ assumption). Now consider the function $g(t)$ defined by:

$$g(t) = z(t)^* H(t, r_0) z(t) \qquad (78)$$

The polynomial $g(t)$ has $f_{r_0}(t)$ as a factor, so we know that $g(t_0)$ and $\frac{\partial g(t_0)}{\partial t_i} = 0$ for all t_i. On the other hand,

$$\frac{\partial g(t_0)}{\partial t_i} = \frac{\partial z(t)^* H(t, r_0) z(t)}{\partial t_i}|_{t=t_0} \tag{79}$$

and the right hand side is exactly $z^* H_i(r_0) z$. □

The situation where there is a real point (t_0, r_0) at which the matrix $H(t_0, r_0)$ loses rank by more than 1 is more complicated. To understand this situation we need to introduce two additional concepts. We restrict attention to the case where the rank of $H(t_0, r_0)$ is $N-2$.

First, we have the $N \times N$ Hessian matrix $Hess(f_r)$ defined by:

$$Hess(f_r)_{ij} = \frac{\partial^2 f_r}{\partial t_i \partial t_j} \tag{80}$$

Because of the rank $N-2$ assumption the matrix $Hess(f_{r_0})_{ij}(t_0)$ is nonzero. Because (t_0, r_0) is real, the hessian matrix is Hermitian.

Second, for an $N \times N$ matrix H denote by $c_{N-2}(H)$ the sum of the principal-minor determinants of size $(N-2) \times (N-2)$. Again, by the assumptions made we know that $c_{N-2}(H(t_0, r_0))$ is a nonzero real number.

Lemma 3 *Consider the case of general N. Suppose that for some real value r_0 the gradient of f_{r_0} vanishes at a real vector t_0 for which $H(t_0, r_0)$ has rank $N-2$. Then r_0 is an r-value if and only if the normalized Hessian:*

$$Hess_0 = \frac{Hess(f_{r_0}(t_0))}{c_{N-2}(H(t_0, r_0))} \tag{81}$$

is negative-semidefinite, rank no more than 2.

Proof of Lemma: Let (t_0, r_0) satisfy the hypotheses of the lemma. Then we may choose a 4×4 complex matrix S of determinant 1 such that the matrix of forms $\Phi(t)$ defined by:

$$\Phi(t) = S^* H(t, r_0) S = \begin{bmatrix} A(t) & B(t) \\ B^*(t) & D(t) \end{bmatrix} \tag{82}$$

satisfies the conditions $A(t_0) = 0$, $B(t_0) = 0 = B^*(t_0)$ where $A(t)$ is 2×2, $B(t)$ is $2 \times (N-2)$ and $B^*(t)$ is the conjugate transpose of the form $B(t)$.

Observe that r_0 is an r-value if and only if for some real t_0 as above the $(1,1)$ entry of $A(t)$ is the zero form (the first column of the matrix S is the vector z satisfying $z^* H_j(r_0) z = 0$).

It is a simple computation to verify that the normalized Hessian matrix $Hess_0$ is exactly the Hessian of the 2×2 determinant function $Det(A(t))$. We will show that a constant S exists making $A_{11}(t)$ the zero form if and only if the Hessian of $Det(A(t))$ is negative semi-definite, rank no greater than 2.

First, if $A_{11}(t) = 0$, observe that

$$Det(A(t)) = -A_{12}(t)A_{21}(t) = -|A_{12}(t)|^2 \tag{83}$$

is a real, negative semidefinite quadratic form. It vanishes on the set of t such that $A_{12}(t) = 0$. Because A_{12} is a complex valued linear form, the dimension of its kernel is at least $N - 2$ – hence the rank of the Hessian form is no greater than 2.

Conversely, suppose $A(t)$ is a 2×2 Hermitian matrix of linear forms defined on t in R^N, and that $Det(A(t))$ is negative semidefinite, rank no more than 2. We want to show there is a 2×2 matrix L of determinant 1 such that

$$L^* A(t) L = \begin{bmatrix} 0 & \beta(t) \\ \beta^*(t) & \gamma(t) \end{bmatrix} \tag{84}$$

Without loss of generality we can write $A(t)$ in the form:

$$A(t) = \begin{bmatrix} a(t) + d(t) & b(t) + \sqrt{-1}c(t) \\ b(t) - \sqrt{-1}c(t) & a(t) - d(t) \end{bmatrix} \tag{85}$$

where $a(t), b(t), c(t), d(t)$ are real linear forms. With this notation,

$$Det(A(t)) = a(t)^2 - b(t)^2 - c(t)^2 - d(t)^2 \tag{86}$$

So the determinant of $A(t)$ is the square of the Minkowski pseudo-norm of the vector of real linear forms (a, b, c, d). For any 2×2 complex matrix L of determinant 1 the transformation g_L defined by:

$$g_L(A(t)) = L^* A(t) L = \begin{bmatrix} a'(t) + d'(t) & b'(t) + \sqrt{-1}c'(t) \\ b'(t) - \sqrt{-1}c'(t) & a'(t) - d'(t) \end{bmatrix} \tag{87}$$

induces a mapping

$$G_L(a, b, c, d) = (a', b', c', d') \tag{88}$$

which preserves the Minkowski pseudo-norm. In fact, this induced mapping is the covering map of the Lie group $SL(2,C)$ over the identity component of the Lorentz group.

The condition that the form $Det(A(t)$ has rank no more than 2 implies that the space of forms (a,b,c,d) spans a space W of dimension no greater than 2. Suppose the subspace is two dimensional. By the negative semidefine assumption we know that the Minkowski inner product is non-positive on W. The converse is proved if we can we find a Lorentz transformation that maps W into the three-dimensional subspace $a(t) - d(t) = 0$. But such a transformation is easy to find using the properties of the Lorentz group. □

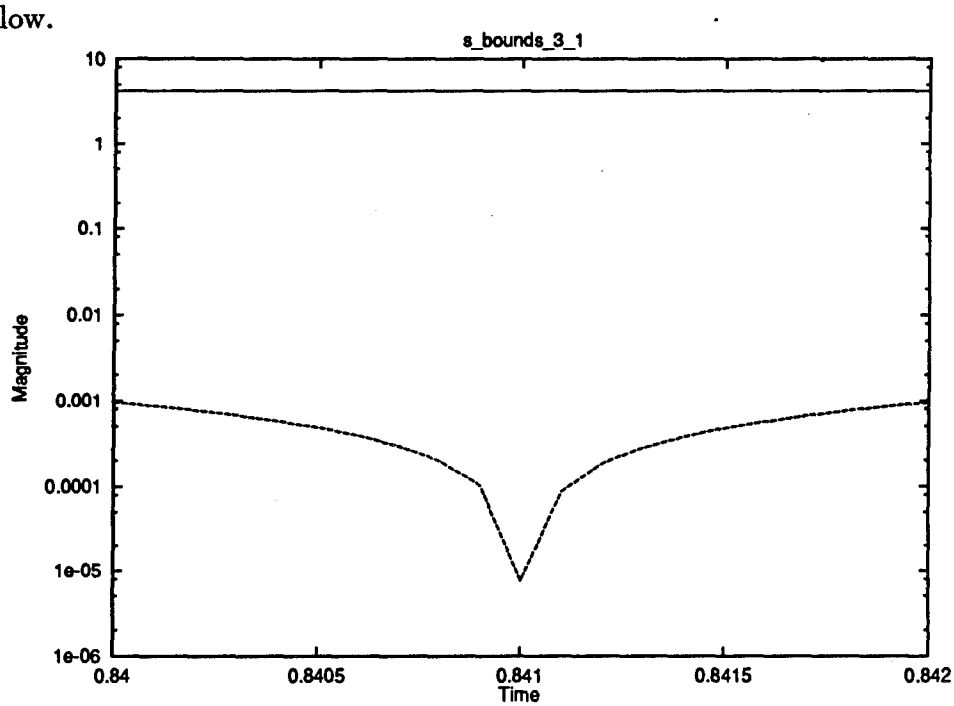
8 A Computational Example

In this section we present a numerical example of the computation of μ for a particular 4×4 matrix M. This matrix M is one member of a family discussed in [13] – its significance is discussed in that reference. Among other things, it is rank 2, with both nonzero singular values equal to 1.

To represent complex numbers we adopt the notation (x, y) to denote the number $x + \sqrt{-1}y$. The matrix M is:

$$M = \begin{bmatrix} (0.0000, 0.0000) & (0.2989, 0.0000) & (0.2989, 0.0000) & (0.3493, 0.3493) \\ (0.2989, 0.0000) & (0.0000, 0.0000) & (0.2113, 0.2113) & (0.4278, 0.2470) \\ (0.0000, 0.2989) & (0.2113, -0.2113) & (0.0000, 0.0000) & (0.2470, 0.4278) \\ (0.3493, 0.3493) & (-0.4278, -0.2470) & (-0.4278, 0.2470) & (0.0000, 0.0000) \end{bmatrix}$$

According to the algorithm described at the end of Section 3 a computer program was used to generate the 976×400 matrix $A(r)$ at values along a grid in the $r-$space. A plot of the two singular values σ_1 and σ_{400} is shown below.



Observe the dip in the value of the 400^{th} singular value at the value $r_0 = 0.841$. The minimum singular vector of $A(r_0)$ was computed and the derived

matrices were found to be (approximately) rank 1. The corresponding vector z at r = 0.841 is:

$$z = \begin{bmatrix} (0.12997134685973169010, 0.00000000000000000000) \\ (0.51950242727553730404, 0.00000052358896364930) \\ (0.51950242727553708200, -0.00000052358896368336) \\ (-0.66583924869638788646, 0.00000000000000019122) \end{bmatrix}$$

The deltas are:

$$\delta = \begin{bmatrix} (-0.84071373809592109261, 0.84071373809592109261) \\ (1.17970685258911189841, -0.14884516617420087692) \\ (1.17970685258911145432, 0.14884516617420112672) \\ (-0.84079248581452115108, 0.84079248581452159517) \end{bmatrix}$$

The matrix $I - M\Delta$ is:

$$\begin{bmatrix} (1.0000, 0.0000) & (-0.3526, 0.0445) & (-0.3526, -0.0445) & (0.5874, 0.0000) \\ (0.2513, -0.2513) & (1.0000, 0.0000) & (-0.2178, -0.2807) & (0.5674, -0.1520) \\ (0.2513, 0.2513) & (-0.2178, 0.2807) & (1.0000, 0.0000) & (0.5674, 0.1520) \\ (0.5873, 0.0000) & (0.5414, 0.2277) & (0.5414, -0.2277) & (1.0000, 0.0000) \end{bmatrix}$$

To verify that $I - M\Delta$ is singular we compute its singular values:

$$\begin{bmatrix} 2.13191170117315254018 \\ 1.13696819871065679664 \\ 0.99498302424291718005 \\ 0.00000000000000004737 \end{bmatrix}$$

Examination of the plots for values of $r > 0.841$ showed that $A(r)$ is nonsingular for larger values (we needed only to check up to $r = 1$, because that is the norm of M). We conclude that $\mu(M) = 0.841$.

The example just presented illustrates an important point about the computation of r-values. The reader might recall the goal stated in Section 1, to derive a polynomial whose largest real root is the value $\mu(M)$. Now that we have come to an example, no polynomial was produced.

The point is that the r-polynomial is associated with some special matrix problems which allow special methods of computations.

Though we have not examined the numerical properties of the solution algorithms for these special types of problem, we have had good experience with them in practice. Note how close to zero the smallest singular value of $I - M\Delta$ is in the example. More is said about this issue in Section 10.1

9 Abstract Interpretation

Using a geometric construction, the problem of determining the degree of the r-polynomial for general N can be formulated in terms of intersection theory [14]. In this subsection we develop that formulation.

We begin by modifying the basic equations

$$z^* H_k(r) z = 0 \quad k = 1, \ldots, N \tag{89}$$

The first step is to replace $\bar{z}$ by an independent variable w, to give:

$$w^T H_k(r) z = 0 \quad k = 1, \ldots, N \tag{90}$$

The second step is to introduce homogeneous coordinates (r_1, r_2) in place of the single affine coordinate r:

$$H_k(r) = r_1 M_k^* M_k - r_2 e_k^* e_k \tag{91}$$

These two changes of variable make the equations 90 tri-linear in the three vectors z, w and r. We now interpret these equations on the product X of three projective spaces:

$$X = \mathbf{P}^{N-1} \times \mathbf{P}^{N-1} \times \mathbf{P}^1 \tag{92}$$

where the last factor corresponds to the r-vector, the second to the w-vector, and the first to the z-vector. By the Segre imbedding i we realize X as a smooth submanifold of the projective space $\mathbf{P}^{2N^2-1}$.

$$i : X \to \mathbf{P}^{2N^2-1} \tag{93}$$

The system of equations 90 cuts out a codimension-N linear subspace L of $\mathbf{P}^{2N^2-1}$. The intersection of X with L is a subvariety of dimension $N-1$, denoted Y.

To proceed with the analysis we need to introduce some vector bundles. First, we have the bundle $T(X)$, the bundle of tangent vectors of X. Contained in $T(X)$ is the codimension-one subbundle E consisting of those vectors that are tangent to the $\mathbf{P}^{N-1} \times \mathbf{P}^{N-1}$ factors of X. Under the projection map $\pi_3 : X \to \mathbf{P}^1$, E is the kernel of the induced morphism $\pi_{3*} : T(X) \to T(\mathbf{P}^1)$. Now the bundle $T(X)$ is itself contained in the larger

bundle $T_X(\mathbf{P}^{2N^2-1})$, the restriction of the tangent bundle of the ambient space to the subvariety X. Finally, over the space L we have the bundle F defined to be $T_L(\mathbf{P}^{2N^2-1})/T(L)$. The bundle F is N-dimensional, its dual is usually called the normal bundle to L in $\mathbf{P}^{2N^2-1}$.

All of the bundles just described can be restricted to Y. Over Y, we have

$$E \to T(X) \to T(\mathbf{P}^{2N^2-1}) \to F \tag{94}$$

The composition morphism $\sigma : E \to F$ of vector bundles on Y is full rank N at most points of Y. There is a special subset of Y, denoted $D_{N-1}(\sigma)$, at which the rank of σ is less than or equal to $N-1$. Let us call this set D_{N-1} the degeneracy locus. For generic data the degeneracy locus D_{N-1} is a codimension $N-1$ subvariety of Y, i.e. a finite set of points. It is that finite set of values (z_i, w_i, r_i) we are after. The r-polynomial $p(r)$ is obtained by eliminating w and z and has as its roots $\{r_i\}$.

The general theory does not tell how to find the points in the degeneracy locus, but (under certain conditions of genericity) the Porteus-Thom formula [14] can be used to compute the number of points in that set. In the remainder of this section we explain the general computational procedure and apply the formula for small values of N.

The notation is as in Chapter 14 of [14]. The degeneracy class of σ is an integral cohomology class in $H^*(Y, \mathbf{Z})$:

$$\Delta_1^{N-1}(c(F-E)) = Det \begin{bmatrix} c_1 & c_0 & \cdots & c_{3-N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N-1} & c_{N-2} & \cdots & c_1 \end{bmatrix} \tag{95}$$

The number of points in the degeneracy locus is:

$$card(D_{N-1}(\sigma)) = \Delta_1^{N-1}(c(F-E)) \bigcap [Y] \tag{96}$$

We will compute this intersection number by pulling back to the space X. Recall that $H^*(X, \mathbf{Z})$ is a truncated polynomial ring on three generators h_1, h_2, h_3 subject to the relations:

$$h_1^N = 0 \quad h_2^N = 0 \quad h_3^2 = 0 \tag{97}$$

Geometrically, each h_j is Poincaire-dual to a hyperplane in its respective projective-space factor. In this ring we have the formula

$$c(T(\mathbf{P}^{N-1} \times \mathbf{P}^{N-1}) = (1+h_1)^N(1+h_2)^N \tag{98}$$

so the pullback of the class of $F - E$ is:

$$i^*(c(F-E)) = (1+(h_1+h_2+h_3))^N/((1+h_1)(1+h_2))^N \tag{99}$$

Finally, by Poincaire duality and the naturality of Chern classes we have:

$$\Delta_1^{N-1}(c(F-E))\bigcap[Y] = \left((h_1+h_2+h_3)^N \bigcup i^*\Delta_1^{N-1}(c(F-E))\right)[X] \tag{100}$$

The right-hand side of this last equation can be computed from the (finite) power series expansion for the rational function in the classes h_1, h_2, h_3, where we have

$$h_1^{N-1}h_2^{N-1}h_3[X] = 1 \tag{101}$$

and all other monomial expressions in h_1, h_2 and h_3 vanish when applied to [X].

9.1 Computation for $N = 2$

In the case $N = 2$ the expression $\Delta_1^1(c(F-E))$ is the determinant of the 1×1 matrix $c_1(F-E)$. From equation **:

$$i^*(c(F-E)) = (1+(h_1+h_2+h_3))^2/((1+h_1)(1+h_2))^2 \tag{102}$$

from which it is easily shown that $c_1(F-E) = 2h_3$. Then

$$\left((h_1+h_2+h_3)^N \bigcup i^*\Delta_1^1(c(F-E))\right)[X] = 4h_1h_2h_3[X] = 4 \tag{103}$$

We conclude there are four r-values for the case $N = 2$

9.2 Computation for $N = 3$

In the case $N = 3$ the expression $\Delta_1^2(c(F-E))$ is the determinant of the 2×2 matrix

$$\Delta_1^2(c(F-E)) = Det\begin{bmatrix} c_1 & c_0 \\ c_2 & c_1 \end{bmatrix} = c_1^2 - c_2c_0 \tag{104}$$

Now

$$i^*(c(F-E)) = (1+(h_1+h_2+h_3))^3/((1+h_1)(1+h_2))^3 \qquad (105)$$

from which it is easily shown that:

$$c_0 = 1 \quad c_1 = 3h_3 \quad c_2 = -3(h_1h_2 + h_1h_3 + h_2h_3) \qquad (106)$$

Thus

$$\Delta_1^2(c(F-E)) = c_1^2 - c_2c_0 = 3(h_1h_2 + h_1h_3 + h_2h_3) \qquad (107)$$

Then

$$\left((h_1+h_2+h_3)^3 \bigcup i^*\Delta_1^3(c(F-E))\right)[X] = 36h_1^2h_2^2h_3[X] = 36 \qquad (108)$$

We conclude there are 36 r-values for the case $N = 3$

9.3 Computation for $N > 3$

Using the Porteus-Thom formula and a symbolic manipulation program (Mathematica) we have performed the computations for the values $N = 4$, $N = 5$ for which the answers are 272 and 2150. These numbers agree with those produced by formula (51) in Section 4. On geometric grounds we suspect that the number produced by the hyperdeterminantal formula will always agree with the number provided by Porteus-Thom – we believe both numbers are the degree of the r-polynomial for generic $N \times N$ complex matrices M. A rigorous proof of this identity, if true, requires a deeper understanding of the theory than we now have.

10 Closing Remarks

Various approaches were taken to generate the r-polynomial and to derive formulas for its degree. The original problem was to find extremal points in a parametrized system of quadratic hermitian forms on the complex vector z for which the system has a nonzero solution. Because hermitian forms are involved, the problem is intrinsically one on the real vector space of z and $\overline{z}$ parameters. Our general approach was to enlarge the space, replacing $\overline{z}$ with a new vector w, to obtain a parametrized system of bilinear forms on z and w. Methods appropriate for general systems of bilinear forms were then employed to tackle the problem.

The order in which we came upon the results in the various sections might be of interest. The earliest computational approach we had was the families-of-hypersurfaces approach described in Sections 5, 6 and 7 [2]. That method was adequate for the 2 and 3-block problems, but it fell short of finishing off the four-block problem (example in Section 8), so we continued to look for an improved approach. In the process we found (through a lead from Dave Morrison) the hyperdeterminantal theory of Section 4. The hyperdeterminantal theory generalizes the families-of-hypersurfaces approach, giving a formula for the degree of the hyperdeterminant associated with the r-polynomial, but it does not provide a (reasonable) computational procedure for finding the roots. After looking at the hyperdeterminants for a while, we finally tried the successful dialytic approach described in Sections 2 and 3. The abstract interpretation of Section 9 was the last piece of the puzzle to fall into place. It is the geometric idea behind the dialytic method of Sections 2 and 3.

Each part of the theory tells something different. The dialytic approach seems to be a general computational procedure but it is not clear why it should work. The abstract interpretation provides the theoretical basis for it. The hyperdeterminantal approach is the most direct tie with the classical theory, while the families-of-hypersurfaces approach (as a special case of the hyperdeterminantal approach) is the easiest computational method for the 2 and 3-block problems.

At this point the picture we have could be complete, but there are some theoretical gaps (next subsection) that could lead to surprises in higher dimensions.

10.1 Theoretical Gaps

The examples we have checked suggest that we have been successful, but in fact we still do not have a rigorous proof. The problem is that the parametrized system of bilinear forms (used in different ways in the dialytic elimination, the hyperdeterminant and the abstract interpretation) is not generic by construction, so there remains some doubt that the theorems we have applied hold true for general N. To see the problem, note that we start with a complex $N \times N$ matrix ($2N^2$ real parameters) from which we generate a system of N hermitian quadratic forms of size N (N^3 real parameters). Strictly speaking, our use of the hyperdeterminantal formula of Section 4 and the Porteus-Thom formula of Section 9 is valid only for generic families of forms. Some additional conditions must be checked for each value of N to see if any matrices M of size N produce a system of forms for which the formulas are correct.

To illustrate the potential problem, let us consider a different approach to the special case $N = 2$. We have a parametrized pair of forms $H_j(r)$ on the complex 2-vector z, let us work in the real four-dimensional space of real and imaginary parts of z. For each value of r the zero set of the pair of forms is represented geometrically by the intersection of two quadric hypersurfaces in $\mathbf{P}^3$ (this is real projective space now). The parameter r_0 is extremal for this family when the hypersurfaces meet tangentially at some point on the intersection curve.

The condition that two quadric surfaces should meet tangentially at some point is well known in the classic literature. The procedure for computing the associated invariant is presented in Chapter 9, article 202 of [15]. For a generic pair of forms, the invariant is a polynomial of total degree 24 in the coefficients of the forms. How do we reconcile this result with the established fact that the degree of the r-polynomial for $N = 2$ is 4?

As a first step, we might claim that the polynomial of degree 24 derived in this manner should have the r-polynomial as a factor. This first claim turns out to be correct. As a second step, we might claim that one can compute μ by finding the 24 roots and then selecting from them the four roots associated with the r-polynomial. This second claim is incorrect. The problem is that for quadrics of the type we have to work with, the degree-24 invariant vanishes identically. Thus, the number 24 that is the right answer for generic families of forms is irrelevant to our problem. If we did not already

know the answer to be 4, we might easily be misled by the computation for generic data.

From this simple illustration we see that there is danger in believing our formulas for the degree of the r-polynomial for general N without more analysis. It would be nice to find a complete, definitive solution to this problem.

10.2 Do We Really Want to Find the r-Polynomial?

Let us reconcile the results obtained with the stated goals of Section 1. We originally said our goal was to find a polynomial p(r), whose largest root was $\mu(M)$, as a generalization of the polynomial in equation (5) for the operator norm of M.

In fact, what we have obtained (by a variety of methods) are matrices $T(r)$ of polynomials in the single variable r that are, in general, full rank but become rank deficient at a finite set of r-values. This finite set of r-values for which $T(r)$ drops rank is the set of values we seek.

In theory, a single polynomial of the type specified can be derived from the matrix $T(r)$ – one finds the greatest common divisor of the determinants of all the maximal square submatrices of $T(r)$. In the classical literature this polynomial is called the highest invariant factor of $T(r)$, see [16]. From a computational point of view it is undesirable to evaluate even one such determinant, let alone find the greatest common divisor of a large collection of them.

In fact, the situation is better than it seems. There are numerical methods for computing the roots of the highest invariant factor of $T(r)$ without computing the determinantal polynomials.

For example, if T(r) is a 3×2 matrix of quadratic polynomials, we can compute the roots as follows. First, write

$$T(r) = T_2 r^2 + T_1 r + T_0 \tag{109}$$

Observe that $T(r)$ drops rank at r_0 only if there is a nonzero vector v such that

$$T(r_0)v = 0 \tag{110}$$

Form the 5×4 affine matrix $S(r)$ defined by:

$$S(r) = \begin{bmatrix} T_2 r + T_1 & T_0 \\ -I & rI \end{bmatrix} \tag{111}$$

where I is the 2×2 identity matrix. Note that the equation

$$S(r_0)w = 0 \tag{112}$$

has a nonzero solution w if and only if r_0 is a root of the highest invariant factor of $T(r)$. So we have reduced the problem to one of finding the invariant factors of an affine matrix function of r.

To find the roots of the highest invariant factor of an affine matrix function we proceed as follows:

1. Pick a maximal square submatrix (4×4 for our example)
2. Solve the generalized eigenvalue problem for that matrix
3. Pick another maximal square submatrix
4. Solve the generalized eigenvalue problem for that matrix
5. Compare the solutions for the two problems
6. Iterate as needed

In theory one might have to solve many generalized eigenvalue problems to reduce the set of common roots to the set of desired solutions. In practice, one should get the desired set of values after only a few trial problems.

In fact, one can get away with solving only one generalized eigenvalue problem if the finite set of roots obtained can be checked by back substitution. The dialytic procedure described in Sections 2 and 3 can be checked in this way, so the issue of not deriving a single polynomial is not a real inconvenience.

Of course, there is no denying that our algorithm produces large matrices in the generalized eigenvalue computations for large values of N. The question of a more efficient computation for the set of r-value remains open.

10.3 Reducibility of the r-Polynomial

As a final note, it is interesting that in the case $N = 2$ the r-polynomial is the difference between two squares, hence it is a reducible polynomial (see Appendix A).

Bibliography

[1] Doyle, J. C., " Analysis of Feedback Systems with Structured Uncertainty," IEE Proc. 129 (pt. D, No. 6): 242-250, 1982.

[2] Morton, B., et al., *Advanced Topics in Robust Control: Volume II: Invariants and Structured Singular Values,* Final Technical Report to ONR for Period April 1982 to July 1990, Contract Number N00014-82-C-0157, Prepared by Honeywell Systems and Research Center, October 1990.

[3] Balas, G., J. Doyle, K. Glover, A. Packard and R. Smith, *μ-Analysis And Synthesis Toolbox, User's Guide,* The MathWorks, 1991.

[4] Doyle, J., J. Wall and G. Stein, "Performance and Robustness Analysis for Structured Uncertainty," IEEE Proceedings of the Conference on Decision and Control, 1982.

[5] Doyle, J., "Robustness of Multiloop Linear Feedback Systems," Appendix E in *Optimal Linear Control* by C. Harvey and J. Doyle, Final technical Report to ONR, Contract Number N00014-75-C-0144 for period April 1977 through July 1978, Prepared by Honeywell Systems and Research Center, August 1978. Also in the IEEE Proceedings of the Conference on Decision and Control, 1978.

[6] Stein, G. and J. Doyle "Singular Values and Feedback: Design Examples, " Allerton Conference, Urbana, IL, 1978.

[7] Morton, B. and R. McAfoos, "A μ-test for Robustness Analysis of a Real-Parameter Variation Problem," Proceedings of the American Control Conference, Boston, MA, 1985.

[8] Morton, B., " New Applications of μ to Real-Parameter Variation Problems," CDC Proceedings, Ft. Lauderdale, FL, 1985.

[9] Spivak, M., *Calculus on Manifolds,* W.A. Benjamin, Co., 1965.

[10] Cayley, A. "On the Theory of Elimination," *Collected Papers, Volume 1,* Cambridge University Press, 1889, pp. 370-374.

[11] Gelfand, I., M. Kapranov and A. Zelevinsky, *Discriminants, Resultants and Multidimensional Determinants,* Birkhauser 1994.

[12] Salmon, G., *Lessons Introductory to the Modern Higher Algebra,* Fifth Edition, Chelsea.

[13] Packard, A. and J. Doyle, *Robust Control of Multivariable and Large Scale Systems,* Final Technical Report to AFOSR for Period October 1985 to March 1988, Contract Number F49620-86-C-0001, Prepared by Honeywell Systems and Research Center, March 1988.

[14] Fulton, W., *Intersection Theory,*, Springer-Verlag, 1984.

[15] Salmon, G., *Analytic Geometry of Three Dimensions, Vol. I,* Seventh Edition, Chelsea, 1927.

[16] Wedderburn, J. *Lectures on Matrices*, Volume XVII of the Colloquium Publications of the American Mathematical Society, 1934.

APPENDIX A

The Mu-Polynomial for Two-by-Two Matrices

The Mu-Polynomial for Two-by-Two Matrices

Consider the 2×2 complex matrix M:

$$M = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \tag{A1}$$

We take as a definition:

$$\mu(M) = \max \rho(e^{i\Theta} M) \tag{A2}$$

where the maximization is over all real diagonal 2×2 matrices Θ. This note is to establish how the function $\mu(M)$ can be computed by finding the largest real root of a real polynomial in one variable with coefficients that are functions of the complex matrix M. The polynomial, written as a function of the variable r, is:

$$p(r) = -\left[r^4 - (|m_{11}|^2 + |m_{22}|^2)\, r^2 + |\det(M)|^2\right]^2 + 4\, r^4\, |m_{12}|^2\, |m_{21}|^2 \tag{A14}$$

The derivation follows.

Define the pair of 2×2 matrices $H^1(r)$, $H^2(r)$ by the formulas:

$$H^1(r) = \begin{bmatrix} \overline{m}_{11} \\ \overline{m}_{12} \end{bmatrix} \begin{bmatrix} m_{11} & m_{12} \end{bmatrix} - \begin{bmatrix} r^2 & 0 \\ 0 & 0 \end{bmatrix} \tag{A3}$$

$$H^2(r) = \begin{bmatrix} \overline{m}_{21} \\ \overline{m}_{22} \end{bmatrix} \begin{bmatrix} m_{21} & m_{22} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & r^2 \end{bmatrix} \tag{A4}$$

Then $\mu(M)$ is the largest real value of r such that there is a nonzero complex vector z for which

$$\begin{bmatrix} \overline{z}_1 & \overline{z}_2 \end{bmatrix} H^j(r) \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = 0, \quad j = 1,2 \tag{A5}$$

The complex vector z is an eigenvector of $e^{i\Theta}M$ associated with the eigenvalue $\mu(M)$.

In previous work it has been shown that there is a nonzero real vector t^0 such that:

$$\left[t_1^0\, H^1(\mu(M)) + t_2^0\, H^2(\mu(M))\right] z = 0 \tag{A6}$$

Therefore, if we consider the matrix function H(t,r) defined by:

$$H(t,r) = t_1 H^1(r) + t_2 H^2(r) \tag{A7}$$

we find that the matrix $H(t^0, \mu(M))$ is rank deficient, with the vector z in its kernel. Now let W be a unitary transformation whose first column is proportional to z. Then:

$$W^* H^1(\mu(M)) W = \begin{bmatrix} 0 & \bar{a}_{12}^1 \\ a_{12}^1 & x^1 \end{bmatrix} \tag{A8}$$

$$W^* H^2(\mu(M)) W = \begin{bmatrix} 0 & \bar{a}_{12}^2 \\ a_{12}^2 & x^2 \end{bmatrix}$$

and

$$W^* H(t^0,\mu(M)) W = \begin{bmatrix} 0 & 0 \\ 0 & x^0 \end{bmatrix} \tag{A9}$$

Therefore:

$$W^* H(t,\mu(M)) W = \begin{bmatrix} 0 & 0 \\ 0 & x^0 \end{bmatrix} + (t_1 - t_1^0) \begin{bmatrix} 0 & \bar{a}_{12}^1 \\ a_{12}^1 & x^1 \end{bmatrix} + (t_2 - t_2^0) \begin{bmatrix} 0 & \bar{a}_{12}^2 \\ a_{12}^2 & x^2 \end{bmatrix} \tag{A10}$$

From this expression, we conclude that the polynomial function f(t,r) defined by:

$$f(t,r) = \det(H(t,r)) = -r^2 |m_{12}|^2 t_1^2 + \tag{A11}$$

$$[(|m_{11}|^2 - r^2)(|m_{22}|^2 - r^2) + |m_{21}|^2|m_{12}|^2 - 2^*\mathrm{Re}(\bar{m}_{12}m_{11}\bar{m}_{21}m_{22})]\, t_1t_2 - r^2 |m_{21}|^2 t_2^2,$$

which reduces to:

$$f(t,r) = -r^2 |m_{12}|^2 t_1^2 + [r^4 - (|m_{11}|^2 + |m_{22}|^2) r^2 + |\det(M)|^2]\, t_1t_2 - r^2 |m_{21}|^2 t_2^2$$

has the following properties:

$$f(t^0,\mu(M)) = 0 \tag{A12a}$$

$$\frac{\partial f(t^0,\mu(M))}{\partial t_j} = 0, \quad j = 1,2 \tag{A12b}$$

From the pair of equations in (A12b) it is possible to eliminate the unknown parameters t_j^0 and arrive at a polynomial in the coefficients of M and $\bar{M}$ that must be satisfied by $\mu(M)$. To this end, consider the pair of polynomial equations:

$$\frac{\partial f(t,r)}{\partial t_j} = 0, \quad j = 1,2 \tag{A13}$$

and eliminate t_1 and t_2. This elimination is performed in the appendix. The final result is: equation (A13) can be solved for some nonzero vector t^0 if and only if r is a root of the polynomial p(r) defined by

$$p(r) = -\left[r^4 - (|m_{11}|^2 + |m_{22}|^2)\, r^2 + |\det(M)|^2\right]^2 + 4\, r^4\, |m_{12}|^2\, |m_{21}|^2 \tag{A14}$$

It is also part of the result shown in the appendix that if r is real then the nonzero vector t^0 is real. By what we have shown so far, $\mu(M)$ is a real root of p(r). What we would like to show is the following:

Claim: The polynomial p(r) has at least one nonnegative real root, and the largest real root is $\mu(M)$.

Proof of Claim: By the construction of p(r) we know that $\mu(M)$, a nonnegative real number, is a root, so the first part of the claim must be true if equation (A14) is correct. It turns out that it is easy to verify the first part of the claim directly, as follows.

The polynomial p(r) can be factored as the difference between two squares. One of the factors is:

$$p_1(r) = r^4 - (|m_{11}|^2 + |m_{22}|^2 + 2\, |m_{12}||m_{21}|)\, r^2 + |\det(M)|^2 \tag{A15}$$

Considering (A15) as a quadratic polynomial in r^2, the discriminant is:

$$\mathrm{disc}(M) = \left[|m_{11}|^2 + |m_{22}|^2 + 2\, |m_{12}||m_{21}|\right]^2 - 4\, |\det(M)|^2 \tag{A16}$$

which is itself a difference between two squares. One of the two factors is obviously nonnegative (being the sum of nonnegative quantities), the other is:

$$|m_{11}|^2 + |m_{22}|^2 + 2\, |m_{12}||m_{21}| - 2\, |\det(M)| \tag{A17}$$

By the triangle inequality, the expression in (A17) is greater than or equal to

$$|m_{11}|^2 + |m_{22}|^2 - 2\, |m_{11}||m_{22}| \tag{A18}$$

which is a square, hence a nonnegative quantity. It follows that (A15) has two real roots (for r^2), both of which are positive because the coefficient of r^2 is negative and the constant term is positive. Then there are at least two nonnegative real roots of p as a function of r, counted with multiplicity, so the first part of the claim has been verified.

We now consider the second part of the claim. Suppose r_0 is the largest real root of (A14). By the computation in the appendix, there is a nonzero real vector t^0 such that

$$\frac{\partial f(t^0, r_0)}{\partial t_j} = 0, \quad j = 1,2 \tag{A19}$$

Because f(t,r) is homogeneous in t, we also have $f(t^0, r_0) = 0$, so $H(t^0, r_0)$ is singular. We can find a

unitary matrix V such that (compare with equation (A10)):

$$V^* H(t, r_0) V = \begin{bmatrix} 0 & 0 \\ 0 & x^0 \end{bmatrix} + (t_1 - t_1^0) \begin{bmatrix} y^1 & \bar{a}_{12}^1 \\ a_{12}^1 & x^1 \end{bmatrix} + (t_2 - t_2^0) \begin{bmatrix} y^2 & \bar{a}_{12}^2 \\ a_{12}^2 & x^2 \end{bmatrix} \tag{A20}$$

First suppose x^0 is nonzero. Then equation (A19) implies $y^1 = 0$ and $y^2 = 0$, so the first column of V is a vector z satisfying (A5), and $\mu(M) = r_0$.

Next, suppose x_0 is 0. Then $H(t^0, r_0) = 0$, and if both t_1^0 and t_2^0 are nonzero then the two matrices $H^1(r_0)$ and $H^2(r_0)$ are proportional. Then any nonzero vector z satisfying

$$\begin{bmatrix} \bar{z}_1 & \bar{z}_2 \end{bmatrix} H^1(r_0) \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \tag{A21}$$

will also satisfy the same equation with H^1 replaced with H^2. Now the matrix $H^1(r)$ is indefinite (signature 0) for all real r, so there is always a nonzero vector z satisfying (A21), so if both t_1^0 and t_2^0 are nonzero we may conclude that $\mu(M) = r_0$.

Finally, if (say) t_1^0 is zero, then t_2^0 is nonzero and so $H^2(r_0)$ must be zero (recall we are assuming $H(t^0, r_0) = 0$) In this case m_{21} is zero. It can be verified directly from the definition (A2) that $\mu(M)$ is the maximum of $\{|m_{11}|, |m_{22}|\}$. Also, in this special case the polynomial p(r) is:

$$p(r) = -[r^4 - (|m_{11}|^2 + |m_{22}|^2)\, r^2 + |m_{11}|^2 |m_{22}|^2]^2 \tag{A22}$$

and the largest real root r_0 is the maximum of $\{|m_{11}|, |m_{22}|\}$ which is $\mu(M)$. A symmetric argument applies if t_2^0 is zero. The claim is verified.

Computation

We derive the result stated between (A13) and (A14) in the above text.

Consider equation (A13):

$$\frac{\partial f(t,r)}{\partial t_j} = 0, \qquad j = 1,2 \tag{A13}$$

The function f(t,r) is homogeneous, quadratic in t_1, t_2 so equation (A13) can be rewritten:

$$Q(r^2)\, t = \begin{bmatrix} q_{11}(r^2) & q_{12}(r^2) \\ q_{21}(r^2) & q_{22}(r^2) \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = 0 \tag{A23}$$

where (use the expressions in (A11)):

$$q_{11}(r^2) = -2\ |m_{12}|^2\ r^2 \tag{A24}$$

$$q_{21}(r^2) = r^4 - (|m_{11}|^2 + |m_{22}|^2)\ r^2 + |\det(M)|^2$$

$$q_{12}(r^2) = q_{21}(x)$$

$$q_{22}(r^2) = -2\ |m_{21}|^2\ r^2$$

There is a nonzero vector t satisfying (A13) if and only if

$$\det(Q(r^2)) = q_{11}(r^2)q_{22}(r^2) - q_{21}(r^2)q_{12}(r^2) = 0. \tag{A24}$$

But it is easy to check that:

$$\det(Q(r^2)) = p(r) = -\left[r^4 - (|m_{11}|^2 + |m_{22}|^2)\ r^2 + |\det(M)|^2\right]^2 + 4\ r^4\ |m_{12}|^2\ |m_{21}|^2 \tag{A25}$$

If r is real, the matrix $Q(r^2)$ is real, symmetric and the kernel is real. The demonstration is complete.