Optimization and Pole Assignment in Control System Design

Eric K. Chu
Department of Mathematics & Statistics,
Monash University, Clayton,
Victoria 3168, Australia
eric.chu@sci.monash.edu.au

Keywords: condition number, linear feedback, pole assignment, robustness.

Abstract
Some elementary optimization techniques, together with some not so well-known robustness measures and condition numbers, will be utilized in pole assignment. In particular, “Method 0” by Kautsky et al (1985) for optimal selection of vectors is shown to be convergent to a local minima, with respect to the condition number \( \frac{1}{2} \| X \|_F^2 - \ln | \det X | \). This contrasts with the misconception by Kautsky et al that the method diverges, or the recent discovery by Yang and Tits (1995) that the method converges to stationary points.

1 Introduction
Consider a controllable time-invariant linear multivariable system controlled by output feedback

\[
\frac{dx}{dt} = Ax + Bu, \quad y = Cx
\]

with the given system matrix \( A \in \mathbb{R}^{n \times n} \), input matrix \( B \in \mathbb{R}^{n \times m} \) and output matrix \( C \in \mathbb{R}^{l \times n} \). We shall assume the matrices \( B \) and \( C \) to be of full rank. The system (1) is controlled by output feedback \( u = Ky = K\dot{x} \), giving rise to the closed-loop system

\[
\frac{dx}{dt} = A_c x = (A + BKC)x
\]

For state feedback, all the states in \( x \) can be observed or measured and we have \( C = I_n \) and \( A_c = (A + BK) \).

In the pole assignment problem (PAP), a feedback matrix \( K \in \mathbb{R}^{m \times l} \) is sought so that the closed-loop system matrix \( A_c \equiv (A + BKC) \) has prescribed spectrum \( \Omega = \{ \lambda_1, \ldots, \lambda_n \} \). Note that \( \Omega \) is closed under complex conjugation, so that \( \rho \in \Omega \Rightarrow \overline{\rho} \in \Omega \), because \( A_c \in \mathbb{R}^{n \times n} \).

The PAP with state feedback is solvable for arbitrary closed-loop poles in \( \Omega \) if and only if the system (1), characterized by \( \{A, B\} \), is completely controllable [9, 42, 46, 56]. That is, either of the matrices

\[
[B, AB, A^2B, \ldots, A^{n-1}B], \quad [sI_n - A, B]
\]

has full rank (for the latter, for all eigenvalues \( s \in \lambda(A) \)). With output feedback, we also require that the system (1) is completely observable, i.e. \( \{A^T, C^T\} \) is completely controllable, and the Kimura condition [34, 35] \( m + l \geq n \) holds. Note that for a particular \( \Omega \), the pole assignment problem may be solvable without \( \{A, B, C\} \) being controllable or observable, if \( \lambda(A + BKC) = \Omega \) for some \( K \). Also, it is common to assume that the algebraic multiplicity of any closed-loop pole is less than or equal to \( \min\{m, l\} \) so as to avoid defective eigenvalues (for exceptions, see, e.g., [11, 21, 22, 23, 36]).

1.1 PAP with State Feedback
For the history of the PAP, see [41, 43, 44, 45, 48]. Most papers in pole assignment contain some survey of the field; see the references in the bibliography and the references therein.

Classification is a subjective exercise. There are different interpretations of words like “methods”, “algorithms”, “solution”, etc. In particular, some hold the view that “algorithms” have to be numerically stable, with numerically sound components and, preferably, supported by the corresponding backward error analysis. Ideally, convergence, independent of starting values, for any iterative processes may need to be investigated theoretically and numerically. Efficiency may have to be supported by operation counts and extensive numerical experiments. All in all, and as a result of our ignorance and limited resources, we shall attempt a brief summary of existing methods for the state feedback pole assignment problem (SPAP) in this subsection.

There are essential four types of methods for the SPAP:

Classical Methods Transform system (1) into one or several SISO systems or canonical forms (Frobenius, Lu-nenberger, Jordan), or involve the controllability matrix (e.g. [26, 61]). Numerically unsound and inefficient techniques involving determinants or characteristic polynomials were often used.

Direct Methods Transform the system into canonical form using stable unitary matrices (e.g. the Schur form [43, 44, 45, 55]).

Matrix Equation Methods Solving Sylvester-like matrix
Eigenvector Methods

Selecting the closed-loop eigenvectors \( \mathbf{x}_i \), the columns of the matrix \( \mathbf{X} \), from some admissible subspaces (see, e.g., [32, 36] and §2).

Other Methods

The Kautsky-Nichols-Van Dooren (KNV) algorithms [32] attracted much attention. Based on similar formulation, others proposed modified or generalized methods in [6, 65, 66] (more details in §3 and §4 later).

Classical methods are usually inefficient or numerically unstable [32, 45]. Matrix equation methods usually require the solution of Sylvester equations so that open-loop poles of \( \mathbf{A} \) cannot be re-assigned, and the parameterisation of the feedback matrix \( \mathbf{K} \) in terms of \( \mathbf{G} \) in (2) is unnatural. The most efficient and numerically stable methods to date are direct methods, with the methods by Petkov et al [47], Minnis and Paige [43, 44, 45] and Varga [55] considered to be the “state of the art” methods for the SPAP. However, these methods do not take into account the under-determined nature of the SPAP. In the eigenvector methods, available degrees of freedom (when the number of inputs \( m \) is greater than one) will be utilized to optimize the condition of the closed-loop spectrum.

The PAP is an important problem in control system design, although its practical usefulness has been continuously under dispute. Many systems were controlled via pole assignment and related techniques. Others preferred other techniques in state space (such as LQR) or the frequency domain (like \( H^\infty \)). A satisfactory general-purpose method has not yet been found. It has never been clear on how the closed-loop poles in \( \Omega \) could be selected. There was doubt about the condition of the PAP for large scale systems Numerically stable algorithms were not available, until recently). How do we choose the closed-loop eigenvectors? In spite of the questions raised here and elsewhere, the search for numerically stable algorithms for the PAP has never ceased. Possibly, it is partly the consequence of the challenge presented by such a simply stated problem. Other reasons are:

1. The general OPAP, as well as the ROPAP and approximate PAP (APAP), are still open.
2. New applications have been found, ranging from stabilizing the starting point for Newton’s iteration for the solution of the algebraic Riccati equation [9], to the design of neural networks [10].
3. Related to (1) above, many problems in control system design give rise to the PAP, as a component of their solution process. For example, the partial pole assignment problem [17, 49] for large scale systems requires pole assignment for small subsystems.
4. More general systems, such as second-order systems [8, 28] and descriptor systems [14], have been investigated and generalizations of the PAP are obvious candidates for their control.

1.4 Robust Pole Assignment

The PAP is one of the most researched problems in control system design (see the long list of references in the bibliography). However, it was not until the 1980’s that modern numerical practices began to impact on control system design. Numerically stable algorithms were proposed for testing controllability [9, 42, 46, 56] and the state feedback pole [32, 43, 44, 45, 47, 55]. Available degrees of freedom were not utilized to improve robustness of the closed-loop system until the mid-1980’s [6, 36, 32]. Earlier attempts in applying optimization to the PAP involved naive over-powering approaches, like minimizing a weighted sum of the squares of the distances between the closed-loop poles from of \((\mathbf{A} + \mathbf{BKC})\) and their desired
positions in \( \Omega \) [1]. With robustness, various optimization techniques can obviously be applied to the robust PAP, minimizing some robustness measure or condition number with the pole assignment requirement providing the constraints. This has been the approaches in [6, 36, 32]. However, curiously and unfortunately, even basic, rudimentary theory in optimization was seldom applied to the related optimization problems. For instance, iterative processes (like Methods 0, 1, 2 and 3 in [32]) were proposed for the state feedback pole assignment problem, in which a robustness measure or condition number was minimized. However, as far as we can recall, no one has written down the corresponding optimality conditions! This paper investigates some of these optimization problems. Some elementary optimization techniques, together with some not so well-known robustness measures and condition numbers, borrowed from sizing and stability analysis and Nash [6], and Yang and Tits [63, 64, 65, 66] is important to the development of our work and will be described in more details in later Sections.

Finally, there has been some interesting work recently in [27, 37], on applying gradient flow techniques [16] to the PAP. These techniques require the solution of a matrix differential equation, which will only be competitive on custom-built neural networks. However, these new developments illustrate again the vast possibility and unfulfilled potential of applying optimization (and neural network computing) to control system design.

2 KNV Algorithms

For the PAP for the system (1), we are seeking a feedback matrix \( K \) which satisfies the closed-loop eigenvalue problems, for \( j = 1, \cdots, n \),

\[
(A + BKC)x_j = \lambda_j x_j
\]

and, for \( i = 1, \cdots, n \),

\[
y_i^H(A + BKC) = \bar{\lambda}y_i^H
\]

for \( \lambda_j \in \Omega \) and \( x_j \), \( y_i \neq 0 \).

Let

\[
B = Q_{B1}R_B = [Q_{B1}, Q_{B2}]
\]

be, respectively, the QR decompositions [25] of \( B \) and \( C \). The matrix \( Q_B (Q_C) \) is orthogonal and \( R_B (R_C) \) is \( m \times m \) \((l \times l)\) and nonsingular. Premultiply (3) by \( Q_{B2}^T \) to eliminate the dependence on \( B \), we have, \( \forall j \),

\[
Q_{B2}^T(A - \lambda_j I)x_j = 0
\]

The conditions in (5) can be interpreted as constraints on the closed-loop eigenvectors \( x_j \), which have to be selected from the null-spaces defined. Let \( \mathcal{U}_j \) be the corresponding invariant subspace, i.e., \( \forall j \),

\[
\mathcal{U}_j \equiv \text{Null} \{Q_{B2}^T(A - \lambda_j I)\}
\]

It can be proven from the controllability of \( \{A, B \} \) that \( \mathcal{U}_j \) is \( m \)-dimensional. Let \( S_j \in \mathbb{C}^{n \times m} \) be unitary, with its columns spanning \( \mathcal{U}_j \). Then we need to choose the right-eigenvectors \( x_j \) such that, \( \forall j \),

\[
x_j = S_j u_j
\]

Similarly, postmultiply (4) by \( Q_{C2} \), we have the corresponding constraints on the left-eigenvectors \( y_i \), \( \forall j \),

\[
y_i^H(A - \lambda_j I)Q_{C2} = 0
\]

and \( \forall i, y_i = T_i v_i \) with

\[
\mathcal{V}_j \equiv \text{Null} \{Q_{C2}^T(A^T - \bar{\lambda}_j I)\}
\]

Here \( T_i \in \mathbb{C}^{n \times l} \) is unitary and \( \mathcal{V}_i = \text{span} T_i \) is \( l \)-dimensional. Note that there are now \( n(m + l) \) unknowns in \( u_j \) and \( \mathcal{V}_i \), \( (i,j = 1, \cdots, n) \) with the biorthogonality condition \( y_i^H x_j = \delta_{ij} \) providing \( n^2 \) equations (c.f. with the Kimura condition \( m + l \geq n \)).

After selecting the eigenvectors such that \( X = [x_1, \cdots, x_n] \) and \( Y = [y_1, \cdots, y_n] \) are nonsingular, we can retrieve the feedback matrix \( K \) by

\[
K = B^T(XAY^H - A)C^T
\]

\[
= R_B^{-1}Q_{B1}^T(XAY^H - A)Q_{C1}R_C^{-1}
\]

with \( Y^H X = I_n \) and \( A = \text{diag} \{\lambda_1, \cdots, \lambda_n\} \).

For the SPAP, recall that \( C = C^T = C^{-1} = I_n \), \( \mathcal{V}_i = \mathbb{C}^n \) and there is no constraint on \( Y \).

As \( \kappa(X) \equiv \|X\| \|Y\| \) represents a condition number of the closed-loop eigenvalue problems (3) and (4), the RSPAP has been reduced to the selection of \( X \) so as to minimize a robustness measure, such as \( \kappa \). (For more general results on condition numbers and perturbation analysis of ordinary and generalized eigenvalue problems, see [12, 13, 50].) In addition, it has been shown in [32, 15] that \( \kappa \) and other related robustness measures can be linked to measures in the frequency domain, which are popular in engineering circles.

Kautsky et al [32] proposed three different algorithms for the selection of right-eigenvectors \( x_j \) from \( \mathcal{U}_j \), which will be described later. See [32] for numerical experiments comparing various algorithms.

2.1 Method 0

Method 0 is based on the heuristic that the perfectly conditioned eigenvalues for a symmetric matrix has identical left and right-eigenvectors. Assume that all the eigenvalues are real and let \( X_+ \) be constructed from \( X \) by deleting the eigenvector \( x_j \); i.e.,

\[
X_+ = [x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n]
\]
In an iteration in Method 0, all the eigenvectors \( \mathbf{x}_j \) are updated one after another, by selecting \( \mathbf{x}_j \) to be the projection of the null space of \( X^T \) onto \( \mathcal{U}_j \). When \( X \) is nonsingular, \( X^T \) has a one dimensional null space identical to the closed-loop left-eigenvector corresponding to \( \lambda_j \). The selection process thus optimizes the degree of orthogonality of the right-eigenvectors by rank-1 updates. No objective function was optimized and the method was thought in [32] to possess a “convergence” problem, with respect to \( \kappa(X) \). While this observation was factually correct, Yang and Tits [65, 66] have proven that the method converges with respect to (maximizing) another robustness measure \( |\det X| \).

Method 0 cannot handle complex eigenvalues in its original form, due to the need to update two complex conjugate eigenvectors at the same time. However, this can be remedied, as in the generalization by Yang and Tits in [65, 66]. Similar generalizations were considered by Kautsky et al but were abandoned, because of the “convergence” problems perceived at the time.

One more comment on Method 0 — the order in which the eigenvectors are updated can affect the efficiency of the iteration tremendously. One strategy is to arrange \( \{\|y_i\|\} \) in ascending order and update \( \mathbf{x}_j \) in the same order [15]. Note that changing \( \mathbf{x}_j \) affects all the left-eigenvectors \( \mathbf{y}_i \) (\( i \neq j \)), thus the strategy duals with the better conditioned eigenvalues last. The strategy has proven to be effective in numerical experiments in [15] and can be applied to all rank-1 update methods. Similar strategies can easily be applied to methods which update more than one vector.

In [65, 66], the iterative process is linked to the maximization of \( |\det X| \). This will be discussed in §4 and §5.

### 2.2 Method 1

Method 1 was the method of choice in [32] and was implemented in the MATLAB command \texttt{PLACCE} in the Control Toolbox [40]. Only the case when all the eigenvalues are real was discussed in [32], although any simple test example shows that the complex case can be handled by the MATLAB command \texttt{PLACCE}. It is unclear how the complex case was treated in the MATLAB Control Toolbox.

The method can be summarized as follows:

\[
\min_{\mathbf{x}_j} \| Y \|_{F}^2 \quad \text{such that} \quad \mathbf{x}_j \in \mathcal{U}_j, \quad \| \mathbf{x}_j \| = 1 \quad (7)
\]

Note that the above process is equivalent to the minimization of \( \kappa_F(X) \) with the same constraints. The approach also avoids the problem associated with the nonuniqueness of solutions when minimizing \( \kappa(X) \) or \( \kappa_F(X) \), as \( \kappa_j(\alpha X) = \kappa_j(X) \) (\( J = 2, F \)) (see [6] and §3 for more details).

For each iteration which updates the eigenvectors \( \mathbf{x}_j \) (\( j = 1, \ldots, n \)) in some order, we have

\[
X = [X_-, \mathbf{x}_j] = Q \begin{bmatrix} R & Q_1^T \mathbf{x}_j \\ 0 & q_2^T \mathbf{x}_j \end{bmatrix} \quad (8)
\]

where

\[
Q_1 R = [Q_1, q_2] \begin{bmatrix} R \\ 0 \end{bmatrix} = X_-
\]

denotes the QR decomposition of \( X_- \). Note that some permutation matrix has been ignored in (8), without loss of generality. As the matrix \( Y \) satisfies

\[
Y^T = X^{-1} = \begin{bmatrix} R^{-1} & -\beta^{-1} R^{-1} Q_1^T \mathbf{x}_j \\ 0 & \beta^{-1} \end{bmatrix}, \quad \beta = q_2^T \mathbf{x}_j
\]

the minimization problem in (7) reduces to

\[
\min_{\mathbf{u}_j} \| \beta^{-1} \begin{bmatrix} R^{-1} Q_1^T S_j \\ I_m \end{bmatrix} \mathbf{u}_j \| \quad (9)
\]

using the fact that \( \mathbf{x}_j^T \mathbf{x}_j = \mathbf{u}_j^T \mathbf{u}_j = 1 \). From the QR decomposition

\[
S_j^T \mathbf{q}_2 = \gamma \mathbf{e}_1
\]

we have

\[
\beta = \mathbf{q}_2^T S_j \mathbf{u}_j = \gamma \mathbf{e}_1^T Z^T \mathbf{u}_j = \gamma z_1, \quad \mathbf{z} = Z^T \mathbf{u}_j = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}
\]

From (9), the original minimization in (7) is further reduced to

\[
\min_{\mathbf{z}} \left\| \begin{bmatrix} R^{-1} Q_1^T S_j \\ I_m \end{bmatrix} Z \begin{bmatrix} 1 \\ \bar{z} \end{bmatrix} \right\| / |\gamma z_1| \quad (11)
\]

or the simpler

\[
\min_{\bar{\mathbf{z}}} \left\| \begin{bmatrix} R^{-1} Q_1^T S_j \\ I_m \end{bmatrix} Z \begin{bmatrix} 1 \\ \bar{z} \end{bmatrix} \right\|, \quad \bar{\mathbf{z}} = \mathbf{z}_2 / z_1 \quad (11)
\]

It is easy to see that the minimization problem in (11) is a standard linear least squares problem, which can be solved by using, again, the QR decomposition [25].

In [15], another interpretation of Method 1 in terms of eigenvectors was presented. It can be proven that the rank-1 update process is equivalent to minimizing a quotient of quadratics in \( \mathbf{u}_j \). Thus the process is equivalent to solving a symmetric generalized eigenvalue problem \( M_1 \mathbf{x} = \lambda M_2 \mathbf{x} \), which can be solved in general by the symmetric QZ algorithm [25]. The transformation of the minimization problem in (7) to the simpler (11) is done via the QR decomposition (10), which is equivalent to the first step of the QZ algorithm, when the rank-1 \( M_2 \) is transformed.

For rank-2 updates are required, as in the case when some eigenvalues \( \lambda_j \) are complex, the minimization in (11) will then involve quotients of quartics. At each of the rank-2 update steps, these quotients of quartics may be minimized by established routines in unconstrained optimization, such as BFGS Quasi-Newton method with line search or a trust-region Newton-like method [2, 24].

### 2.3 Method 2 and 3

Method 2 involves the heuristic that a well-conditioned set of eigenvectors should be nearly orthogonal. The method starts off with an arbitrary orthogonal set of reference vectors \( \{ \mathbf{z}_j \} \). For small values of \( m \), a weighted sum of the squares of the angles or distances between \( \mathbf{z}_j \) and the subspace \( \mathcal{U}_j \) is minimized. It is realized through changing a particular pair
of eigenvectors by Jacobi rotations [25]. Details can be found in [32].

Method 2 can handle complex eigenvalues with ease, as it updates a pair of vectors simultaneously, although the generalization has not been included in [32]. When the number of inputs \( m \) is greater than \( n/2 \), it will be more efficient to deal with the orthogonal complements of \( U_j \), in which case we have Method 3.

Methods 2 and 3 were found to be satisfactory in terms of efficiency and convergence. However, Method 1 was preferred in [32], mainly due to its direct minimization of a condition number.

### 3 Byers-Nash Approaches

The work by Byers and Nash [6] extended the approach by Kautsky et al. As one of the main contributions of the paper, Byers and Nash pointed out that minimizing condition numbers such as \( \kappa(X) \) and \( \kappa_F(X) \) does not yield a unique solution, because \( \kappa_k(X) = \kappa_k(A_X) \) \( (k = 2, F) \). This will give rise to singular Hessian matrices and hence severe difficulties when standard optimization routines are applied. As a result, an extra constraint like \( \|X\|_F = 1 \) has to be imposed, or modified robustness measures such as

\[
  f_3(X) = \|X\|_F^2 + \|X^{-1}\|_F^2
\]

have to be used [6]. The measure \( f_3 \) can be shown to be related to an upper bound of \( \kappa_F(X) \), when \( X \) is nonsingular [6, Eq. 6], where

\[
  \kappa_F(X) = \|X\|_F \|X^{-1}\|_F \leq \frac{1}{2} f_3
\]

Other measures considered in [6], coupled with the constraint \( \|X\|_F = 1 \), were

\[
  f_1 \equiv \kappa_F(X), \quad f_2 \equiv -\frac{1}{f_1}, \quad f_4 \equiv -\frac{1}{f_3}, \quad f_5 \equiv \log f_3
\]

There is still some scaling flexibility in the solution \( X \) but this can be eliminated by insisting on choosing the first component of eigenvectors \( X_j \) to be real and positive.

The truncated Newton method (with line search), the conjugate gradient method and Newton’s method [2, 24] were then applied to the robustness measures, and some numerical results were reported in [6, §4–6]. Various derivatives of the robustness measures were then derived [6, §5].

Interestingly, Quasi-Newton methods (like BFGS with inexact line-search) were not applied, fearful of their “ineffectiveness” for “large-scale” problems. Also, optimality conditions were not investigated. Nevertheless, the work by Byers and Nash represents one of the few serious attempts in applying optimization techniques to the PAP.

### 4 Yang-Tits Algorithm

In [65, 65], Yang and Tits investigated Method 0 in [32] and proved its convergence with respect to maximizing the robustness measure \( |\det X| \). Similar measures have been used in [63, 64] and were linked to other condition numbers and robustness measures in [30, 38].

The main contributions of the work by Yang and Tits are as follows:

1. They resurrected the simple and powerful Method 0 in [32], by proving its convergence with respect to maximizing \( |\det X| \). Only convergence to stationary points can be proven, although numerical experiments illustrated that these stationary points were most likely to be local minima.

2. They generalized the rank-1 update approach for Method 0 in [32] to a rank-2 update method, which is more efficient than its original version for real eigenvalues. Note that rank-2 updates have to be used for complex eigenvalues.

3. With the new insight into Method 0, various algorithms were re-assessed in [66].

We next describe the generalized Method 0 (with rank-2 updates). Similar to the original Method 0 in [32], the matrix \( X_m = [x_1, \ldots, x_{2n-2}, x_{2n+1}, \ldots, x_n] \). The null space of \( X_m^H \) is then projected, respectively, onto the appropriate subspaces \( U_{2i-2} \) and \( U_{2i+1} \). To enforce convergence, the new eigenvectors are selected to be those which are closest to the ones they replaced.

The work by Yang and Tits made important advances towards understanding the optimization problems related to the selection of eigenvectors in the KNV algorithms, an area where little work has been done. Convergence to stationary points was proven. Again interestingly, no optimality conditions were written down for the optimization problems involved.

In §5–7, some of the results by Yang and Tits will be extended. The rank-2 update generalization is actually a special case of a rank-\( n \) update method, which can be derived from the necessary optimality conditions. In addition, many other eigenvector selection algorithms can be shown to be closely related to Method 0.

### 5 Robustness Measures

In [67, Ch. 4] as well as [17, 19, 20, 29], the following condition numbers were used to investigate sizing and least-change secant methods: (for s.p.d. matrices \( A \))

1. \( \kappa(A) = \frac{\lambda_n}{\lambda_1} = \|A\| \|A^{-1}\|, \) where \( \lambda_1 \) and \( \lambda_n \) are respectively the largest and smallest eigenvalues of \( A \).

2. \( \omega(A) = \frac{\text{tr}(A)/n}{\text{det}(A)^{1/n}}, \) the ratio between the arithmetic mean and geometric mean of the eigenvalues of \( A \).

3. \( \sigma(A) = \frac{\lambda_1}{\text{det}(A)^{1/n}}. \)

4. \( \tau(A) = \frac{\text{tr}(A)}{\lambda_n}. \)

5. \( \max \text{det}(A) \text{ such that } \lambda_1(A) \leq 1, \) the restricted maximum determinant measure [60].
In the study of least-change secant methods [67], the above condition numbers were applied to measure the distance between $A$ and $\alpha I_n$, a multiple of the identity matrix. As a result, the redundancy issues raised by Byers and Nash [6] have no significance here. Note that the measure in (V) is obviously similar to (II) as well as the determinant measure used by Yang and Tits [65, 66]. The measures in (II)–(V) were described as “uniform”, as they involved all the eigenvalues of $A$, as compared to only $\lambda_1$ and $\lambda_n$ for $\kappa$ in (I).

Some important results from [67, §4] are quoted here: (with $\lambda_i$ arranged in descending order)

**Lemma 5.1** ([67, Proposition 4.1])

The uniform condition number $\omega(A)$ satisfies

(i) $1 \leq \omega(A) \leq \kappa(A) < \frac{(\kappa(A)+1)^2}{\kappa(A)} \leq 4\omega(A)^n$, with equality in the first and second inequalities if and only if $A$ is a (nontrivial) multiple of the identity and equality in the last if and only if

$$\lambda_2 = \cdots = \lambda_{n-1} = \frac{\lambda_1 + \lambda_n}{2}$$

(ii) $\omega(\alpha A) = \omega(A)$, for all $\alpha > 0$.

(iii) If $n = 2$, $\omega(A)$ is isotonic with $\kappa(A)$.

(iv) The measure $\omega$ is pseudoconvex on the set of s.p.d. matrices, and thus any stationary point is a global minimizer of $\omega$.

(v) Let $V$ be a full rank $m \times n$ matrix, $n \leq m$. Then the optimal column scaling that minimizes the measure $\omega$, i.e.,

$$\min \omega([VD]^TVD)$$

over all positive diagonal matrices $D$ is given by

$$D_{ii} = \frac{1}{\|V_i\|}, \quad i = 1, \cdots, n$$

where $V_i$ is the $i$th column of $V$.

The proof of Lemma 5.1 can be found in [20]. For a definition of pseudoconvexity, see [18].

**Lemma 5.2** ([67, Proposition 4.2])

The measure $\sigma(A)$ satisfies

(i) $1 \leq \sigma(A) \leq n\kappa(A) \leq n\kappa(A)^n \leq 4n\sigma(A)^n$.

(ii) $\sigma(\alpha A) = \sigma(A)$, for all $\alpha > 0$.

(iii) The measure $\sigma$ is a pseudoconvex function on the set of s.p.d. matrices, and thus any stationary point is a global minimizer.

The proof of Lemma 5.2 can be found in [60].

**Lemma 5.3** ([67, Proposition 4.3])

The measure $\tau(A)$ satisfies

(i) $1 \leq \omega(A) \leq \tau(A) \leq \kappa(A) \leq 4\omega(A)^n$.

(ii) $\tau(\alpha A) = \tau(A)$, for all $\alpha > 0$.

(iii) The measure $\tau$ is pseudoconvex on the set of all s.p.d. matrices, and thus any stationary point is a global minimizer.

The proof of Lemma 5.3 can be found in [67, p. 32].

The lemmas in this section on the condition numbers $\omega$, $\sigma$ and $\tau$, contain powerful results on the relation between various condition numbers and their convexity. However, the convergence results are only valid for unconstrained optimization of the condition numbers in the set of s.p.d. matrices. For the PAP, the optimality conditions derived from the robustness measures are of great interest. These conditions may form the basis of future algorithms. The number of possible combinations of robustness measures and various generalizations of the PAP is enormous and we shall only present some selected results of our preliminary study.

### 6 Method 0 Revisited

In this section, we assume that the $S_j \in C^{n \times m}$ is extended to an $n \times n$ unitary matrix $[S_j, \hat{S}_j]$. That is, the columns of $S_j$ form a unitary basis of $U_j^\perp$, the orthogonal complement of $U_j$.

#### 6.1 The Yang and Tits Measure

In this subsection, we shall consider the robustness measure $-|\det X|$ used in [65, 66]. Note that the measure is equivalent to $\sigma(X^H X)$ or the restricted maximum determinant measure in the previous section. It is easy to modify the condition number $-\det A$ for s.p.d. matrices, by replacing $A$ with $X^H X$. Consequently, we have to minimize the robustness measure $-\det(X^H X) = -\det(X)^2$ or $-|\det X|$, exactly the measure used by Yang and Tits.

Consider the eigenvector selection problem for the RSPAP:

$$\min_{\|x_j\|=1} \{-|\det X|\} \quad \text{s.t.} \quad \hat{S}_j^H x_j = 0 \quad \forall j \quad (12)$$

The minimization problem in (12) can actually be stated as one without constraints, with the variables $x_j$ replaced by $S_j u_j$, as in (6). The formulation in (12) gives rise to a simpler exposition and is retained. Similar comments hold for similar minimization problems in the following sections.

The Lagrangian of the minimization problem in (12) equals, with $\alpha_j$ and $\beta_j$ denoting the Lagrange multipliers,

$$\mathcal{L} = -|\det X| + \sum_j \frac{1}{2} \alpha_j (x_j^H x_j - 1) - \beta_j \hat{S}_j^H x_j$$

The Karush-Kuhn-Tucker (KKT) conditions implies that, $\forall j$,

$$-y_j |\det X| + \sum_j \alpha_j x_j - \hat{S}_j \beta_j = 0 \quad (13)$$
Premultiply (13) by $x_j^H$, we obtain $\alpha_j = -|\det X|$. Premultiply (13) by $\hat{S}_j^H$, we obtain $\beta_j = -|\det X|\hat{S}_j^H y_j$. Substitute $\alpha_j$ and $\beta_j$ back into (13), we have the optimality condition: \( \forall j, \)
\[
x_j = \left( I_n - \hat{S}_j^H \right) y_j = S_j S_j^H y_j \tag{14}\]

If we update only a particular $j$, (14) represents exactly Method 0 in [32]! Note that the constraints $||x_j|| = 1$ destroy the convexity of the feasible region.

Note that (14) is a set of $n$ nonlinear simultaneous equations in $x_j$ and its solution by any iterative process represents a rank-$n$ update eigenvector selection method.

### 6.2 Entropy

The convexity of the feasible region can be restored by considering the following minimization problem:

\[
\min_{x_j} \frac{1}{2} ||X||_F^2 - \ln |\det X| \quad \text{s.t.} \quad \hat{S}_j^H x_j = 0 \quad \forall j \tag{15}\]

The condition number used in (15) is the “entropy measure” which can be considered to be a generalization of the condition number $\omega$ in the previous section.

Instead of the ratio between the arithmetic and geometric means, let us replace the numerator in the ratio by its exponential, i.e. the ration
\[
\exp\left\{ (\text{tr} A)/n \right\} / (\text{det} A)^{1/n}
\]

With $X^H X$ substituted into $A$ and logarithm taken, we have
\[
\frac{2}{n} E_n(X) = \frac{2}{n} \left\{ \frac{1}{2} ||X||_F^2 - \ln |\det X| \right\}
\]

which leads to the entropy measure in (15).

Note that $E(X)$ does not involve the inverse $X^{-1}$ and so is convenient to differentiate and manipulate. More importantly, the optimization problem in (15) does not have any constraint on the sizes of $x_j$ and the feasible region for $x_i$ now consists of the invariant subspace $U_j$. The combined feasible region for $X$ is obviously convex. However, the optimization problem in (15) is not convex in general, as the objective function is not convex (due to the term $-\ln |\det X|$; it is locally convex from Lemma 6.1).

Consider the Lagrangian of the minimization in (15):
\[
\mathcal{L} \equiv \frac{1}{2} ||X||_F^2 - \ln |\det X| - \sum_j \beta_j^H S_j^H x_j \tag{16}\]

The KKT conditions then yield
\[
x_j - y_j - \hat{S}_j \beta_j = 0
\]
as in the steps of the previous subsection, premultiplications of (16) by $x_j^H$ and $S_j^H$ yield, after some simple rearrangement, $\forall j$,
\[
\beta_j = S_j^H y_j \tag{17}\]

and
\[
x_j = S_j S_j^H y_j \tag{18}\]

Notice that the scaling of $x_j$ is somehow done implicitly, due to the inclusion of the term $\frac{1}{2} ||X||_F^2$ in $E(X)$, so that
\[
||x_j||^2 = x_j^H (S_j S_j^H y_j) = x_j^H y_j = 1 \quad \forall j
\]

Also, there is no redundancy issue as mentioned in [6], since, in general $E(\alpha X) \neq E(X)$.

Not surprisingly, (18) and (14) contain an identical system of nonlinear equations in $x_j$, which gives rise to Method 0 and its generalizations.

Recall the convergence results (to stationary points) by Yang and Tits [65, 66]. We shall show that the correct condition number used in analysing Method 0 should be the entropy measure $E(X)$ in (15). This is based on the following Lemma:

**Lemma 6.1** The Hessian $\nabla^2 \mathcal{L}$ of the Lagrangian function (differentiating with respect to elements in $X$ columnwise, and then the elements in $[\beta_1, \ldots, \beta_n]$)
\[
\mathcal{L}(x_1, \ldots, x_n; \beta_1, \ldots, \beta_n) \equiv \frac{1}{2} ||X||_F^2 - \ln |\det X| - \sum_j \beta_j^H S_j^H x_j
\]
is given by
\[
\nabla^2 \mathcal{L} = \left[ \begin{array}{cc} I_{n^2} + \hat{Y} \hat{S} & \hat{S} \\ \hat{S}^H & 0 \end{array} \right], \quad \hat{S} = \left[ \hat{S}_1, \ldots, \hat{S}_n \right]
\]

and $\hat{Y}$ has $n^2 n \times n$ submatrices, with the $(i, j)$th subblock being $y_i y_j^H$.

The proof of Lemma 6.1 is elementary, involving differentiation of the Lagrangian function $\mathcal{L}$, and will be left as an exercise.

From Lemma 6.1, it is easy to show that the second order sufficient optimality condition [24] is satisfied by a solution $x_j^*$ of (18). For the minimizer $x_j^*$ and the corresponding $\beta_j^* = \hat{S}_j^H y_j^*$ (from (17)), with $x_j^*$ $(j = 1, \ldots, n)$ and $\beta_j^*$ $(j = 1, \ldots, n)$ stacked into $z$, we have
\[
z^H \nabla^2 \mathcal{L}(x_j^*, \beta_j^*) z = ||X||_F^2 + n > 0
\]

Consequently, the generalized version of Method 0, i.e. the solutions of (18), converge to local minima.

### 7 Method 1

Let us return to Method 1 in [32] and consider the minimization problem
\[
\min_{||x_j|| = 1} \frac{1}{2} ||X^{-1}||_F^2 \quad \text{s.t.} \quad \hat{S}_j^H x_j = 0 \quad (\forall j)
\]

The Lagrangian for the minimization problem equals
\[
\frac{1}{2} \text{tr} \left( X^{-1} X^{-H} \right) + \frac{1}{2} \alpha_j (x_j^H x_j - 1) - \sum_j \beta_j^H S_j^H x_j
\]
The KKT conditions are
\[-Y^H y_j + \alpha_j x_j - \hat{S}_j \beta_j = 0\]

With the notation \(\lambda_{ij} = \langle y_i^H y_j \rangle / \|y_j\|^2\), the following optimality conditions can be obtained: \(\forall j\),
\[x_j = \frac{1}{\|y_j\|^2} S_j S_j^H Y Y^H y_j = S_j S_j^H \sum_i \lambda_{ij} y_i\]  

(19)

As in other generalizations of the PAP, similar nonlinear equations as in (19) are obtained. Note that (19) can be considered as a generalization of Method 0, which projects the left-eigenvectors onto span \(S_j\) as right-eigenvectors. The method indicated in (19) obviously involves a similar but more sophisticated process. Ideally when \(\{y_i\}\) are mutually orthogonal, we have \(\lambda_{ij}\) equals a nontrivial multiple of the Kronecker delta \(\delta_{ij}\). Consequently, the equations in (19) are identical to those in (18) or (14).

When only a particular \(j\) is updated, as in the original Method 1 by Kautsky el al, we have the necessary optimality condition
\[x_j = \frac{1}{\|y_j\|^2} S_j S_j^H Y Y^H y_j\]  

(20)

Solving (20) will thus be equivalent to the process described in [32]. Note that \(Y^H = X^{-1}\) can be updated easily, using the well-known rank-1 update of inverse results. Obviously, there is no reason to stop at updating one or two eigenvectors, and (20) could be solved simultaneously for all \(j\), as in (19). Thus, \(x_j\) is projected from \(y_j\), first obliquely by the projection \(Y Y^H\) then orthogonally by \(S_j S_j^H\), and finally scaled by \(\|y_j\|^2\).

References


