On the numerical solution of matrix Bessel equations

Eduardo Kausel¹ | Hauke Gravenkamp²

¹Professor of Civil and Environmental Engineering, Massachusetts Institute of Technology, Massachusetts 02139, USA
²Department of Civil Engineering, University of Duisburg-Essen, 45141 Essen, Germany

Correspondence
Eduardo Kausel, Professor of Civil and Environmental Engineering, Massachusetts Institute of Technology, Massachusetts 02139, USA.
Email: kausel@mit.edu

This article deals with the numerical evaluation of certain systems of second order differential equations that are analogous in structure to the scalar Bessel equations, i.e. \( A(z^2u'' + z'u') + (z^2M - G)u = 0 \). The governing equation includes three distinct coefficient matrices \( A, M, G \) that we assume to be real, symmetric, and positive definite. However, allowance is also made for \( G \) to be semi-definite. We first review the conditions under which rigorous closed-form solutions are possible, and this occurs when the various matrices share a common set of eigenvectors, even if not the eigenvalues. We find in that case that the response consists in a linear combination of conventional Bessel functions. More generally, we propose three alternative methods to accomplish the evaluation:

- An ascending series solution based on the Frobenius method.
- An effective asymptotic solution that provides results in very close agreement with the first method. It is stable and accurate for a wide range of values of the argument more so than the first method.
- A direct numerical integration solution using the fifth order Runge-Kutta-Nyström algorithm. It works well, except for some idiosyncrasies that depend on the presence of eigenvalues with widely different magnitudes, as will be explained.

We determined that all of these techniques produce highly accurate response functions when measured in terms of the residual to the differential equation, and in all cases we found that residual to be negligible.

KEYWORDS
Bessel equation, Frobenius method, matrix differential equations, wave propagation

1 | INTRODUCTION

In certain problems in mechanics on the propagation of waves formulated in cylindrical coordinates and when the material properties are allowed to change with the azimuth, one is led to systems of differential equations of the Bessel type for which the technical literature provides few answers and even fewer solution methods. We examine herein one special case of such matrix Bessel equation, namely one where the three matrices involved are real and symmetric, and they are all positive definite except for one that can be positive semi-definite.

Consider the homogeneous system of differential equations in the dimensionless variable \( z = \omega r/\beta \), in which \( r \) is a radial distance, \( \omega \) is a frequency, and \( \beta \) is a convenient reference wave velocity. Unless \( \omega \to 0 \), it is not necessary to distinguish between \( r \) and \( z \). Written compactly in matrix form, the system of differential equations to be explored herein is
\[
\begin{aligned}
\left[ A \left( z^2 \frac{d^2}{dz^2} + z \frac{d}{dz} \right) + z^2 M - G \right] u &= 0
\end{aligned}
\]

where all matrices are real & symmetric, and they have homogeneous physical dimensions. Of these, \( A, M \) are both positive definite, while \( G \) is either positive definite or positive semi-definite. Except for the use of matrices, the governing equation is analogous to that for Bessel functions, so similar characteristics can be expected.

Except for degenerate cases, this system of \( N \) differential equations of second order will possess \( 2N \) independent solutions, half of which may be singular at \( z = 0 \) and the other half will be regular. In addition, any linear combination of solutions is also an acceptable solution, albeit not an independent one. Furthermore, if \( u \) is any solution, then \( cu \) is also a solution, where \( c \) is any arbitrary constant scalar. This shows that all of the \( 2N \) solutions are defined only up to a multiplicative constant. The classical approach to this type of problem consists in converting the system of second order differential equations into a system of first order differential equations with twice the number of unknowns. This strategy has a number of drawbacks, not the least of which is the increased number of degrees of freedom as well as the loss of symmetry of the matrices. In the ensuing, we abstain from taking that path and explore instead a direct solution to the system of second order differential equations.

The literature on algorithms to solve this type of equations is rather sparse, to the point that we have failed to find more than a handful of references on the numerical tools needed to evaluate the complete solution to our equations. Among these, the work that comes closest to our problem can be found in a set of papers by Navarro et al. Still, these authors considered a matrix differential equation of the Bessel type that was simpler in structure than our Equation (1), inasmuch as their system involved just one matrix and not three, as is the case herein. To solve that case, they elaborated on an ascending series solution that is similar—even if not identical— to one of our own solutions described later on, and which we shall revisit as well. Beyond that, their paper is mainly occupied with demonstrating miscellaneous abstract mathematical properties and proofs such as completeness and convergence tests, but neither delving into the computational details nor showing any specific numerical examples. These details are indeed important because although the ascending series can be shown to be rigorously convergent, that fact alone is not sufficient to guarantee a practical solution. Indeed, for large values of the independent variable \( z \) the series fails to converge due to severe cancellation errors.

In general, the trio of matrices \( A, M, G \) do not share common eigenvectors, so in most cases they cannot be diagonalized simultaneously by any linear transformation. This implies in turn that they cannot readily be separated into a set of uncoupled equations for Bessel functions.

## 2 | PRELIMINARY CONSIDERATIONS

### 2.1 | Eigenvalue problems

The solution methods to be discussed in the ensuing will rely on the solution of two linear eigenvalue problems, namely

\[
A \Psi \Lambda^2 = G \Psi
\]

\[
A \Phi K^2 = M \Phi
\]

with diagonal spectral matrices

\[
\Lambda^2 = \text{diag}\{ \lambda^2_j \}, \quad K^2 = \text{diag}\{ \kappa^2_j \}, \quad j = 1, 2, \ldots, N
\]

and modal matrices

\[
\Psi = \{ \psi_1, \psi_2, \ldots, \psi_N \}, \quad \Phi = \{ \phi_1, \phi_2, \ldots, \phi_N \}.
\]

Because the matrices are both real and symmetric and are positive definite or semi-definite, we know that all eigenvalues are real and non-negative; indeed, they are strictly positive if \( G \) is positive definite. These eigenvalue problems satisfy the orthogonality conditions

\[
\Psi^T A \Psi = I, \quad \Psi^T G \Psi = \Lambda^2,
\]

\[
\Phi^T A \Phi = I, \quad \Phi^T M \Phi = K^2,
\]
where we have also conveniently normalized the eigenmatrices as indicated. We observe in passing that because \( A, M, G \) have the same physical dimensions, the eigenvalues \( \lambda_j^2, \kappa_j^2 \) are all dimensionless.

### 2.2 Scaling of matrices

Prior to solving the system of equations, it is advantageous to scale all matrices by the \( N^{th} \) root of the determinant of the leading, non-singular matrix \( A \). The scaling factors and the scaled matrices are defined as follows

\[
\tilde{A} = \frac{1}{\sqrt[A]{|A|}}, \quad \tilde{G} = \frac{1}{\sqrt[A]{|G|}}, \quad \tilde{M} = \frac{1}{\sqrt[A]{|M|}}, \quad \tilde{z} = z\sqrt{\frac{M}{A}}.
\]

Dividing the matrix differential equation by \( A \), scaling \( M \) as indicated, and redefining the independent variable as \( z \to \tilde{z} \) will then lead us to the differential equation in scaled and dimensionless coordinates

\[
\left[ \tilde{A} \left( \frac{d^2}{d\tilde{z}^2} + \tilde{z} \frac{d}{d\tilde{z}} \right) + \tilde{z}^2 \tilde{M} - \tilde{G} \right] \tilde{u} = 0.
\]

These scaled matrices exhibit the convenient properties

\[
|\tilde{A}| = 1, \quad |\tilde{M}| = 1.
\]

Scaling also affects the eigenvalue problem as follows:

\[
\tilde{A} \tilde{\Psi} \tilde{\Lambda}^2 = \tilde{G} \tilde{\Psi}, \quad \tilde{\Lambda} = \Lambda, \quad \text{and} \quad \tilde{A} \tilde{\Phi} \tilde{K}^2 = \tilde{M} \tilde{\Phi}, \quad \tilde{K} = \sqrt{\frac{A}{M}} K.
\]

These satisfy

\[
|\tilde{\Psi}^T| |\tilde{\Lambda}| |\tilde{\Psi}| = |I| = 1 \quad \Rightarrow \quad |\tilde{\Psi}| = \pm 1,
\]

\[
|\tilde{\Psi}^T| |\tilde{G}| |\tilde{\Psi}| = |\Lambda|^2 = \prod_{j=1}^{N} \lambda_j^2 \quad \Rightarrow \quad |\tilde{G}| = \prod_{j=1}^{N} \lambda_j^2,
\]

\[
|\tilde{\Phi}^T \tilde{\Lambda} \tilde{\Phi}| = |\tilde{\Phi}^T| |\tilde{\Lambda}| |\tilde{\Phi}| = 1 \quad \Rightarrow \quad |\tilde{\Phi}| = \pm 1,
\]

\[
|\tilde{\Phi}^T \tilde{M} \tilde{\Phi}| = |\tilde{\Phi}^T| |\tilde{M}| |\tilde{\Phi}| = \tilde{K}^2 = 1 \quad \Rightarrow \quad \prod_{j=1}^{N} \kappa_j^2 = 1.
\]

We also define

\[
\mathcal{M} = \tilde{\Psi}^T \tilde{M} \tilde{\Psi}.
\]

In general, this is not a diagonal matrix, but it is symmetric. Furthermore, it satisfies the simple property

\[
|\mathcal{M}| = |\tilde{\Psi}^T |\tilde{M}| |\tilde{\Psi}| = 1.
\]

In the ensuing, we shall assume that the matrices have been scaled as indicated, even if we shall omit the overbar to avoid notational overload.

### 2.3 Modal expansion

As a first step in the solution of our matrix Bessel equation, we proceed to make the substitution (with the implicit replacement \( \Psi \to \tilde{\Psi} \), and so forth)

\[
u = \Psi y
\]
which after some straightforward transformations leads us to the simpler differential equation

\[ \frac{d^2 y}{dz^2} + \frac{d y}{dz} \left( z^2 M - \Lambda^2 \right)y = 0 \]

in which all matrices are diagonal, except for \( M = \Psi^T M \Psi \) that is symmetric and satisfies \( |M| = 1 \). Also, all elements of \( \Lambda^2 \) are non-negative real numbers; in addition, if \( G \) is positive definite, then \( \Lambda^2 \) is positive. Before treating the most general case, we examine first two special cases, namely the static case \( \omega = 0 \) and the so-called ‘proportional matrices’ case that takes place when the three matrices \( A, M, G \) do share a common set of eigenvectors, even if not the eigenvalues.

3 | STATIC CASE

The static case \( \omega = 0 \) is the simplest, for it admits an exact, closed form solution without the need for any simplifying assumptions. Recalling that \( z = \omega r / \beta \), we begin by making the replacements

\[ \frac{z}{dz} \rightarrow \frac{r}{dr}, \quad \frac{z^2}{d^2 z} \rightarrow \frac{r^2}{d^2 r}, \quad z^2 \rightarrow 0 \]

in which case the scaled differential equation simplifies into

\[ r^2 \frac{d^2 y}{dr^2} + \frac{dr}{dr} \Lambda^2 y = 0. \]

This homogeneous equation in \( N \) uncoupled differential equations admits solutions of the form

\[ y(r) = r^\Lambda c + r^{-\Lambda} d, \quad r^{\pm \Lambda} = \text{diag} \{ r^{\pm \lambda} \} \]

where \( c, d \) are arbitrary vectors of integration constants. Hence, the complete solution to our original matrix equation is

\[ u(r) = \Psi (r^\Lambda c + r^{-\Lambda} d) = \sum_{j=1}^{N} \psi_j (c_j r^{\lambda_j} + d_j r^{-\lambda_j}). \]

An equivalent useful solution is obtained by redefining the constants of integration such that \( c_j \rightarrow c_j \kappa_j^{\lambda_j}, \quad d_j \rightarrow d_j \kappa_j^{-\lambda_j} \). With these substitutions, the alternative solution becomes

\[ u(r) = \sum_{j=1}^{N} \psi_j \left[ c_j (\kappa_j r)^{\lambda_j} + d_j (\kappa_j r)^{-\lambda_j} \right]. \]

The advantage of this latter form is that it agrees with the leading term of the dynamic solution in terms of Bessel functions.

4 | PROPORTIONAL MATRICES

4.1 | Exact solution with proportional matrices

We now move on to consider another special case, namely that of a trio of matrices \( A, M, G \) which share a set of common eigenvectors. This not only gives us a preliminary peek into the general characteristics of the non-proportional case, but also provides a convenient reference point, and a yardstick against which to measure the quality of the solution tools for the general case that we shall consider later on.

Assume (at least for now) that the structure of \( A, M, G \) is such that they can be diagonalized simultaneously by a modal transformation, i.e. that the eigenvalue problems for the two pairs of matrices \( A, M \) and \( A, G \) share the same eigenvectors \( \phi_j = \psi_j \), even if not the eigenvalues. In that case, \( M \) is a diagonal matrix, namely

\[ M = \Psi^T M \Psi = K^2 = \text{diag} \{ \kappa_j^2 \} \].
Dividing by \( z^2 \), the differential equation thus decouples into the scalar equations

\[
\frac{d^2 y_j}{dr^2} + \frac{1}{z} \frac{dy_j}{dr} + \left( \kappa_j^2 - \frac{1}{z^2} \lambda_j^2 \right) y_j = 0
\]

the solution of which is

\[
y_j = a_j J_{\lambda_j}(\kappa_j z) + b_j Y_{\lambda_j}(\kappa_j z)
\]

or

\[
y_j = c_j H^{(1)}_{\lambda_j}(\kappa_j z) + d_j H^{(2)}_{\lambda_j}(\kappa_j z)
\]

where \( J_{\lambda_j}, Y_{\lambda_j} \) are the Bessel functions of the first and second kind (i.e. the Bessel and Neumann functions) and \( H^{(1)}_{\lambda_j} = J_{\lambda_j} + iY_{\lambda_j}, \ H^{(2)}_{\lambda_j} = J_{\lambda_j} - iY_{\lambda_j} \) are the Bessel functions of the third kind commonly known as the first and second Hankel functions. Also, \( a_j, b_j, c_j, d_j \) are arbitrary constants of integration. Define next the diagonal Bessel matrices

\[
J_{\lambda} = \text{diag} \left\{ J_{\lambda_j}(\kappa_j z) \right\}, \quad N_{\lambda} = \text{diag} \left\{ Y_{\lambda_j}(\kappa_j z) \right\},
\]

\[
H^{(1)}_{\lambda} = \text{diag} \left\{ H^{(1)}_{\lambda_j}(\kappa_j z) \right\}, \quad H^{(2)}_{\lambda} = \text{diag} \left\{ H^{(2)}_{\lambda_j}(\kappa_j z) \right\},
\]

by means of which the solution to the system of homogeneous differential equations can be written in compact form as

\[
u = \Psi [J_{\lambda} a + N_{\lambda} b] \quad a = \{a_j\}, \quad b = \{b_j\}
\]

\[
u = \Psi [H^{(1)}_{\lambda} c + H^{(2)}_{\lambda} d] \quad c = \{c_j\}, \quad d = \{d_j\}
\]

where \( a, b, c, d \) are column vectors of arbitrary constants. This is the exact solution to the system of differential equations with proportional matrices. To round up this section, we now discuss the conditions under which the three matrices share common eigenvectors.

### 4.2 Conditions for proportionality

We can show that the necessary and sufficient condition for the trio of matrices \( \mathbf{A}, \mathbf{M}, \mathbf{G} \) to have the same eigenvectors is for the triple product \( \mathbf{AM}^{-1} \mathbf{G} \) to be a symmetric matrix, which in principle could be established even before computing the modes. To demonstrate this property, assume first that \( \Psi, \mathbf{A} \) are the eigenvectors and spectral matrix for the pair of matrices \( \mathbf{A}, \mathbf{G} \) and that \( \Psi \mathbf{M} \) is not necessarily diagonal, even if it is symmetric because \( \mathbf{M} \) is symmetric. Then

\[
\mathbf{A} = \Psi^{-T} \Psi^{-1}, \quad \mathbf{M}^{-1} = \Psi \Psi^{-1} \Psi^{-T}, \quad \mathbf{G} = \Psi^{-T} \Lambda \Psi^{-1}
\]

so that

\[
\mathbf{AM}^{-1} \mathbf{G} = \Psi^{-T} \Psi^{-1} \Psi \Psi^{-1} \Psi^{-T} \Lambda \Psi^{-1} = \Psi^{-T} \Psi \Lambda \Psi^{-1}.
\]

Transposing this result while considering the symmetry of the matrices, we are led to

\[
\mathbf{GM}^{-1} \mathbf{A} = (\Psi^{-T} \Psi \Lambda \Psi^{-1})^T = \Psi^{-T} \Lambda \Psi^{-1}.
\]

Comparing the two expressions above, we see that for the two products to be equal, i.e. for the product \( \mathbf{AM}^{-1} \mathbf{G} \) to be a symmetric matrix, the equality \( \Psi \Lambda^{-1} \Lambda = \Lambda \Psi^{-1} \) must hold. Since the eigenvalue matrix \( \Lambda \) is diagonal by construction and if all eigenvalues \( \lambda_j^2 \) in \( \Lambda \) are distinct, then this identity can be satisfied if and only if \( \Psi = \{\psi_j^T \mathbf{M} \psi_j\} = \{\mathbf{m}_{jj}\} = \text{diag} \{\kappa_j^2\} \) were to be a diagonal matrix, in which case it can be guaranteed that the matrices are proportional. Still, in some rare cases this symmetry test could fail. This could happen if a subset of the eigenvectors \( \phi_j, \ j = 1, \ldots, n \) of the pair \( \mathbf{A}, \mathbf{M} \) were to coincide with the eigenvectors \( \psi_j \) of the pair \( \mathbf{A}, \mathbf{G} \) such that \( \Psi \mathbf{M} \) would consist of a diagonal block associated with the common eigenvectors \( \phi_j = \psi_j \), and a non-diagonal block associated with the non-repeated eigenvectors \( j = n + 1, \ldots, N \). If in addition the latter were to be associated with identical eigenvalues \( \lambda_j^2 \), then all pairs of elements in the off-diagonal products \( \lambda_j^2 \mathbf{m}_{ij} = \mathbf{m}_{ij} \lambda_j^2 \) would be
equal, in which case the test of proportionality would fail. In that case, the triple product would be symmetrical even though the matrices are non-proportional. Moreover, when two or more identical eigenvalues exist, then the corresponding eigenvectors \( \psi_i, \psi_j \) are not unique either, inasmuch as any two distinct vectors in the subspace spanned by the eigenvectors associated with the repeated eigenvalues is also a valid eigenvector. If so, then we would find that \( \mathcal{M} \) is not unique either. However, if at least one eigenvalue for the distinct eigenvectors participating in the fully-populated (i.e. non-proportional) block of \( \mathcal{M} \) were not identical to the others, then the symmetry test would still hold.

**Example of repeated eigenvalues**

We demonstrate herein a simple example of a non-proportional matrix trio \( \mathbf{A}, \mathbf{M}, \mathbf{G} \) wherein the pair of matrices \( \mathbf{A}, \mathbf{G} \) exhibit repeated eigenvalues. This in turn causes the proportionality test to fail. Consider the following \( 3 \times 3 \) set of matrices

\[
\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 11 & -2 & -8 \\ -2 & 8 & 4 \\ -8 & 4 & 23 \end{bmatrix} = \psi \mathbf{A}^2 \psi^{-1}, \quad \mathcal{M} = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \quad \mathcal{M}^{-1} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}
\]

\[
\mathbf{A}^2 = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 28 \end{bmatrix}, \quad \psi = \begin{bmatrix} 3 & 1 & -2 \\ 2 & -2 & 1 \\ 1 & 1 & 4 \end{bmatrix}, \quad \frac{1}{\sqrt{14}} \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{21}} \end{bmatrix}, \quad \psi^{-1} = \psi^T \text{ (normalized!)}
\]

\[
\mathbf{M}^{-1} = \mathbf{M}^{-1} = \begin{bmatrix} 2.6308 & -0.6745 & 0.9841 \\ -0.6745 & 1.7462 & -0.5238 \\ 0.9841 & -0.5238 & 1.6230 \end{bmatrix}, \quad \mathbf{AM}^{-1} \mathbf{G} = \begin{bmatrix} 22.4159 & -6.7217 & -1.1116 \\ -6.7217 & 13.2232 & 0.3333 \\ -1.1116 & 0.3333 & 27.3609 \end{bmatrix}
\]

Although \( \mathbf{A} \mathbf{M}^{-1} \mathbf{G} \) is symmetric, the matrices are surely not proportional because \( \mathcal{M} \) is not diagonal. This is because the first two eigenvalues are equal and affect the non-diagonal block of \( \mathcal{M}^{-1} \). Moreover, using Matlab to solve the eigenvalue problem for \( \mathbf{A}, \mathbf{G} \), we find a rather different set of eigenvectors \( \psi \) that are linear transformations of the eigenvectors listed above. That is, we find a different \( \mathcal{M} \) when starting with the postulated \( \mathbf{M} \) given above.

**4.3 A general class of proportional matrices**

Although the symmetry of the triple product is simple and theoretically interesting, this condition is rarely useful from a practical point of view, especially because the triple product is computationally expensive. Still, it can be used to demonstrate that a general class of matrices –not necessarily all inclusive– which does satisfy the proportionality condition is of the form

\[
\mathbf{M} = \mathbf{A}^{\frac{1}{2}} \left\{ \sum_{r=0}^{N-1} a_r \left[ \mathbf{A}^{-\frac{1}{2}} \mathbf{G} \mathbf{A}^{-\frac{1}{2}} \right]^r \right\} \mathbf{A}^{\frac{1}{2}}
\]

where the \( a_r \) are arbitrary parameters and \( s \) is a positive number. In particular, if \( s = 1 \), the preceding simplifies into

\[
\mathbf{M} = \mathbf{A} \sum_{r=0}^{N-1} a_r (\mathbf{A}^{-1} \mathbf{G})^r
\]

and since from the eigenvalue problem \( \mathbf{A}^{-1} \mathbf{G} = \psi \mathbf{A}^2 \psi^{-1} \), then

\[
\mathbf{M} = \mathbf{A} \sum_{r=0}^{N-1} a_r (\psi \mathbf{A}^2 \psi^{-1})^r = \mathbf{A} \psi \left\{ \sum_{r=0}^{N-1} a_r \mathbf{A}^2 r \right\} \psi^{-1}
\]

\[
|\mathbf{M}| = |\mathbf{A}| \sum_{r=0}^{N-1} a_r \mathbf{A}^2 r = |\mathbf{A}| \prod_{j=1}^{N} \left( \sum_{r=0}^{N-1} a_r \mathbf{A}^2 r \right)
\]
\( \Psi^T M \Psi = \Psi^T A \Psi \left( \sum_{r=0}^{N-1} a_r A^{2r} \right) = \sum_{r=0}^{N-1} a_r A^{2r} = K^2 = \text{diag}\{ \kappa_j^2 \} \)

\[ \kappa_j^2 = \sum_{r=0}^{N-1} a_r \lambda_j^{2r}. \]

However, observe that after normalization with \( \tilde{A} = A / \sqrt{|A|}, \tilde{M} = M / \sqrt{|M|} \), then

\[ \mathfrak{M} = \Psi^T \tilde{M} \Psi = \tilde{K}^2 = \text{diag}\{ \tilde{\kappa}_j^2 \}, \quad |\mathfrak{M}| = 1, \quad \prod_{j=1}^{N} \tilde{\kappa}_j^2 = 1. \]

A simple case is when all coefficients except the first and the second are non-zero, i.e. when

\[ M = a_0 A + a_1 G \]

with arbitrary coefficients \( a_0, a_1 \). This special case is what has given rise to the moniker of ‘proportional matrices’ in the first place. This expression is analogous to that of the Rayleigh damping matrix in Structural Dynamics. A simpler case still is that of a homogeneous medium with \( M = a_0 A \equiv \kappa^2 A \), in which case \( a_0 = \sqrt{|M|/|A|} \equiv \kappa^2 \) before normalization, and \( \tilde{M} = \tilde{A}, \mathfrak{M} = \tilde{I}, \tilde{\kappa}_j^2 = 1 \) after normalization and for any \( j \). Another interesting case is one where \( G \) is positive definite instead of just semi-definite. If so, then \( G \) is nonsingular and solutions with \( s < 0 \) are also acceptable. For example, with \( s = -1 \) and \( a_1 \neq 0 \) we could write \( M = a_i G^{-1} \), which is also of the proportional type. Finally, any linear combination of proportional matrices as given above is also a proportional matrix.

5 | GENERAL CASE OF NON-PROPORTIONAL MATRICES

Assume next that the trio of matrices \( A, M, G \) does not exhibit any common set of eigenvectors. Hence, the equations cannot be uncoupled by means of a linear transformation, and \( \mathfrak{M} = \Psi^T M \Psi \) is not diagonal, even if it is symmetric. When this is the case, the solution to the system of differential equations is substantially more complicated and involved, as will be seen. In the ensuing, we consider various ways of accomplishing that goal.

The reader may recall that in the numerical evaluation of the classical Bessel functions, two methods are often used, which are the ascending series for low arguments (typically \( |z| \leq 10 \)) and the asymptotic solution for large argument. Although the fully coupled matrix version exhibits strong similarities to the scalar case, a complicating factor here is that the convergence with one or the other method may depend on characteristic numerical parameters that may be unknown ahead of time. Smaller values of these parameters would suggest the use of the ascending series method, while high values might require the use of the asymptotic solution. Although the ascending series solution will be demonstrated to be strictly convergent from a theoretical point of view, for large values of the argument it will fail to converge due to severe cancellation of large magnitude terms with alternating signs. Fortunately, that shortcoming can be overcome by means of the asymptotic solution.

5.1 | Method 1: ascending series solution for ‘low’ arguments \( z \)

In the ensuing, we shall assume that none of the eigenvalues \( \lambda_j \) differ by an integer, i.e. \( \lambda_k \neq \lambda_j + n \) for any \( k > j, n \), with \( j, n \) being integers. In addition, it will be assumed that none of the eigenvalues themselves are integers. The need for these assumptions will be made clear in the ensuing.

We begin by seeking solutions of the Frobenius-type, trying an ascending series of the form

\[ y(z) = \sum_{n=0}^{\infty} z^{n+\lambda} a_n \]

where \( \lambda \) is a parameter to be determined, as are also the vectors \( a_n \). Then

\[ \frac{dy}{dz} = \sum_{n=0}^{\infty} (n + \lambda) z^{n+\lambda-1} a_n, \quad \frac{d^2 y}{dz^2} = \sum_{n=0}^{\infty} (n + \lambda)(n + \lambda - 1) z^{n+\lambda-2} a_n \]
and
\[ z^2 \frac{d^2 y}{dz^2} + z \frac{dy}{dz} = \sum_{n=0}^{\infty} (n + \lambda)^2 z^{n+\lambda} a_n. \]

Defining \(a_{-2} = a_{-1} = 0\), we can write
\[ z^2 y = \sum_{n=0}^{\infty} z^{n+\lambda+2} a_n = \sum_{n=2}^{\infty} z^{n+\lambda} a_{n-2} = \sum_{n=0}^{\infty} z^{n+\lambda} a_{n-2}. \]

Substituting these expressions into the differential equation, we obtain
\[ \sum_{n=0}^{\infty} \left\{ \left( (n + \lambda)^2 I - \Lambda^2 \right) a_n + \mathfrak{M} a_{n-2} \right\} z^{n+\lambda} = 0 \]

which we shall satisfy by setting to zero each of the terms within the braces. For \(n = 0\), this gives
\[ (\lambda^2 I - \Lambda^2) a_0 = 0 \]

which can be satisfied identically if \(\lambda^2\) equals any of the elements \(\lambda_j^2\) in \(\Lambda^2\) and \(a_0 = e_j\) is a unit vector,
\[ \lambda = \lambda_j = \sqrt{\lambda_j^2} \geq 0, \quad a_0 = e_j = \{0, 0, \ldots, 1, 0, \ldots\}^T, \]

with the solitary 1 occupying the \(j^{th}\) row of \(e_j\). For \(n = 1\), we have
\[ \left[ (1 + \lambda_j)^2 - \Lambda^2 \right] a_1 = 0. \quad (2) \]

The term in brackets is a diagonal matrix whose elements \(d_{kk}\) can be factored as
\[ d_{kk} = [1 + (\lambda_k + \lambda_j)] [1 - (\lambda_k - \lambda_j)] \]

None of the two factors in brackets can be zero unless \(|\lambda_k| \pm |\lambda_j| = 1\), and this is when considering both cases of all positive or all negative eigenvalues. Thus, if any two eigenvalues were to differ by 1, then Equation (2) could be satisfied with \(a_1 \neq 0\), in which case the series method about to be described and which is based on \(a_1 = 0\), will fail. Moreover, if all eigenvalues were to be integers, then so would also be their differences \(\lambda_k - \lambda_j\), and chances are excellent that for at least one pair that difference might be 1, which is not acceptable. This is then the reason for ruling out integer eigenvalues, or non-integer eigenvalues that differ by 1. Still, this does not rule out obtaining alternative ascending series of the Frobenius type, but these would be different from those considered herein.

As already mentioned, we shall now assume that none of the eigenvalues differs by an integer, and in particular, that \(\lambda_j + 1\) is not an eigenvalue. If so, then the above matrix equation will only be satisfied if \(a_1 = 0\), and by recursion \(a_3 = a_5 = \cdots = 0\).

For even \(n \geq 2\), we infer that
\[ a_n = -\left( (n + \lambda_j)^2 I - \Lambda^2 \right)^{-1} \mathfrak{M} a_{n-2}, \]

Inasmuch as both \(I\) and \(\Lambda^2\) are diagonal, the inverse of the term in square brackets is trivial. However, the \(\mathfrak{M}\) factor is generally a fully populated, but symmetric matrix. To avoid dealing solely with even indices \(n\), it suffices to re-define \(n \rightarrow 2n\) (i.e. exclude the odd terms). With some obvious changes in the definition of the coefficients \(a_n\) this changes the recursion into
\[ a_n = -\frac{1}{4} \left( \left( n + \frac{1}{2} \lambda_j \right)^2 I - \frac{1}{4} \Lambda^2 \right)^{-1} \mathfrak{M} a_{n-2}, \quad n = 1, 2, 3, \ldots \]

where
\[ D^{(j)}_n = \left( \left( n + \frac{1}{2} \lambda_j \right)^2 I - \frac{1}{4} \Lambda^2 \right)^{-1} \]

(3)
in which case the ascending series solution based on the $j^{th}$ mode would be of the form
\[ y^{(j)} = z^{\lambda_j} \sum_{n=0}^{\infty} \left(-\frac{1}{4}z^2\right)^n a_n^{(j)} \]
where the newly re-defined series coefficients $a_n \rightarrow a_n^{(j)}$ would follow from the recursion
\[ a_n^{(j)} = D_n^{(j)} \mathbf{M} a_{n-1}^{(j)} \]
that is, we absorbed the $(-\frac{1}{4})$ factor within the recursion for $a_n$ into the powers of $z$. Finally, we modify the series $y^{(j)}$ by means of a constant factor $\frac{1}{2}\frac{\lambda_j}{\Gamma(\lambda_j+1)}$ so that the series solution for the $j^{th}$ mode is given by the expression
\[ y^{(j)} = \sum_{n=0}^{\infty} \left(-\frac{1}{4}z^2\right)^n a_n^{(j)}. \] (4)
This is permissible because the solution to the differential equation is defined only up to a multiplicative constant. Furthermore, the reciprocal of the gamma function is known to be an entire (or holomorphic) function for any complex argument,[4] in which case it is well defined. The reasons for this scaling will be explained shortly. Next, we generalize the preceding results to include the contribution for all of the modes and thus obtain a solution of the form
\[ y = \sum_{j=1}^{N} j_j c_j = J_\lambda c \]
where
\[ J_\lambda = \{j_1, j_2 \ldots j_N \} \]
in which the $c_j$ are arbitrary constants of integration. Finally, a nearly identical –yet independent– solution can be constructed with the simple replacement $\lambda_j \rightarrow -\lambda_j$ (i.e. considering the two valid roots of $\lambda_j^2$). This is a series whose first few terms will involve negative powers of $z$, which in turn implies that such a solution will exhibit a singularity at the origin. Thus, the complete solution is of the form
\[ y = J_\lambda c + J_{-\lambda} d \]
\[ u(z) = \Psi (J_\lambda c + J_{-\lambda} d) \]
where the $c, d$ are vectors of arbitrary constants. This assumes that none of the $\lambda$ are integers, because otherwise the Frobenius series, as presented, would break down, as indicated in the line following Equation (2).

**Behavior at low argument $z \to 0$**

Examining the preceding expression for the ascending series, it will be seen that at very low values of the argument $z$, the functions $j_j$ can be approximated by the first two terms
\[ j_j(z) \sim \left(\frac{1}{2}z\right)^{\lambda_j} \frac{1}{\Gamma(\lambda_j+1)} \left[a_0^{(j)} - \frac{1}{4}z^2 a_1^{(j)} \right] \]
\[ = \left[ 1 - \frac{1}{\Gamma(\lambda_j+1)} \left(e\left(\frac{1}{2}z\right)^{\lambda_j} - \left(\frac{1}{2}z\right) \sum_{j=1}^{\infty} \left(\frac{1}{2}z\right)^{\lambda_j} \right) \left[1 - \frac{1}{4}A^2 \right] m_j \right] \]
where $m_j$ is the $j^{th}$ column of $\mathbf{M}$. In the $+\lambda_j$ case (i.e. $J_j$), all eigenvalues are positive, so the functions $j_j$ are dominated by the first term, whose only non-zero element is the $j^{th}$ one and tends to the value $\left(\frac{1}{2}z\right)^{\lambda_j} / \Gamma(\lambda_j+1)$. This equals in turn the leading term in the ascending series for the Bessel functions $J_{\lambda_j}(z)$. Hence, for small arguments $z$, we observe that $\lim_{z \to 0} J_j = \text{diag} \{J_{\lambda_j}(z)\}$, which implies that $J_{\lambda_j}$ is regular and approaches a diagonal matrix when $z \to 0$. 
Contrariwise, the second independent solution with \(-\lambda_j\) (i.e. \(-\Lambda\)) may exhibit a somewhat different behavior, and may not necessarily cause \(J_{\lambda}\) to tend to a purely diagonal matrix as \(z \to 0\). Firstly, at low values of \(z\), the diagonal elements of \(J_{\lambda}\) are again dominated by the leading terms that are in addition singular, because the exponents are negative. That term, however, contributes nothing to the off-diagonal elements, so these coupling terms need not be singular. This is because the leading term in \(e_j\) affects only the diagonal term, and not the coupling terms. Then, examining the second term in the series, we see that it is of the form \((\frac{1}{2}z)^{2-\lambda_j}\), which will not be singular if \(|\lambda_j| < 2\). Therefore, some coupling terms may tend to zero while their symmetrically located elements may tend to \(\pm \infty\).

**Convergence**

Following Navarro et al.,\(^1\) we now go on to show that the terms of the ascending series define a decaying sequence, i.e. that the series is convergent. We accomplish this by observing that

\[
a_{n}^{(j)} = D_n^{(j)} M a_{n-1}^{(j)}
\]

and

\[
|a_n^{(j)}| = |D_n^{(j)} M a_{n-1}^{(j)}| \leq \|D_n^{(j)}\| \|M\| |a_{n-1}^{(j)}|.
\]

Denoting by \(f_n\) the summands in Equation (4), i.e.

\[
f_n^{(j)}(z) = \left(-\frac{1}{4} z^2\right)^n a_n^{(j)}.
\]

and substituting Equation (3), we obtain for a given \(z\) (d’Alembert’s ratio test)

\[
\lim_{n \to \infty} \frac{|f_n^{(j)}|}{|f_{n-1}^{(j)}|} = \lim_{n \to \infty} \frac{\frac{1}{4} z^2 \|M\|}{\left(n + \frac{1}{2} \lambda_j \right)^2 \left(1 - \frac{1}{4} \lambda_j^2\right)} = \lim_{n \to \infty} \left(\frac{z}{2n}\right)^2 = 0.
\]

Hence, the series converges absolutely. Here, \(\|\cdot\|\) denotes the spectral norm while \(\|\cdot\|\) is the Euclidean norm. The corresponding series of functions, Equation (4), converges pointwise. However, as the number of terms required to achieve a given accuracy depends on \(z\), the Frobenius series does not converge uniformly. Furthermore, for low \(n\) the terms in the series grow in magnitude up to some maximum, and only then begin to decrease. According to Equation (5), we can estimate this maximum to occur around \(n \approx \frac{1}{2} z\), while the precise value depends on the eigenvalues \(\Lambda\). For this alternating series, this problem is exacerbated by the fact that the final summation may be small in comparison with the largest terms in the series. The implication is that for large \(z\) there will be many large magnitude terms with alternating signs, in which case the summation will ultimately fail due to severe cancellation errors. This is not a failure of convergence in the mathematical sense, but rather a numerical problem related to the finite precision with which numbers are manipulated in a digital computer, and this is why an asymptotic solution is needed. As shall be seen later on when treating the asymptotic solution method, a transition zone exists, roughly taking place at \(z \approx 15\), where the numerical error of the asymptotic solution is comparable to that of the ascending series solution. Thus, both solutions are needed when the calculations extend over a wide range of \(z\) values.

### 5.2 Method 2: Frobenius method in the style of Navarro et al.

Following the series method proposed by Navarro et al.,\(^1\) and defining \(Y = \{y_1, y_2, \ldots y_N\}\) as the collected solution matrix, the solution to our differential equation

\[
z^2 \frac{d^2 Y}{dz^2} + z \frac{dY}{dz} + \left(z^2 M - \Lambda^2\right) Y = 0
\]

is given by the summation

\[
Y(z) = \left[\sum_{k=0}^{\infty} C_k z^k I\right] z^L
\]
where \( \mathbf{I} \) is the identity matrix and \( \mathbf{C}_k, \mathbf{L} \) are (possibly complex) matrices defined later on. The derivatives are then

\[
\frac{d\mathbf{Y}}{dz} = \left[ \sum_{k=0}^{\infty} \mathbf{C}_k (k\mathbf{I} + \mathbf{L}) z^{(k-1)} \right] \mathbf{z}^L
\]

\[
\frac{d^2\mathbf{Y}}{dz^2} = \left[ \sum_{k=0}^{\infty} \mathbf{C}_k (k\mathbf{I} + \mathbf{L}) (k\mathbf{I} + \mathbf{L} - \mathbf{I}) z^{(k-2)} \right] \mathbf{z}^L
\]

hence,

\[
z^2 \frac{d^2\mathbf{Y}}{dz^2} + z \frac{d\mathbf{Y}}{dz} + (z^2 \mathfrak{m} - \Lambda^2) \mathbf{Y} = \left\{ \sum_{k=2}^{\infty} \left[ \mathbf{C}_k (k\mathbf{I} + \mathbf{L})^2 - \Lambda^2 \mathbf{C}_k \right] z^{2k} + \mathfrak{m} \sum_{k=2}^{\infty} C_{k-2} z^{2k} \right\} \mathbf{z}^L = \mathbf{0}.
\]

Equating to zero the summands for each index, we obtain

\[
\begin{align*}
\mathbf{C}_0 \mathbf{L}^2 - \Lambda^2 \mathbf{C}_0 &= \mathbf{O} \\
\mathbf{C}_1 (\mathbf{I} + \mathbf{L})^2 - \Lambda^2 \mathbf{C}_1 &= \mathbf{O} \\
\mathbf{C}_k (k\mathbf{I} + \mathbf{L})^2 - \Lambda^2 \mathbf{C}_k + \mathfrak{m} \mathbf{C}_{k-2} &= \mathbf{O}
\end{align*}
\]

where \( \mathbf{O} \) is the null matrix. From the first of these, and assuming \( \mathbf{C}_0 \) to be non-singular, we conclude that

\[
\mathbf{L}^2 = \mathbf{C}_0^{-1} \Lambda^2 \mathbf{C}_0
\]

i.e. \( \mathbf{L}^2 \) is a similarity transformation on the spectral matrix \( \Lambda^2 \), which implies in turn

\[
\mathbf{L} = \mathbf{C}_0^{-1} \Lambda \mathbf{C}_0
\]

so the eigenvalues of \( \mathbf{L} \) and \( \Lambda \) are identical. We now recognize our original Frobenius method as the result of arbitrarily choosing \( \mathbf{C}_0 = \mathbf{I} \), which implies that in that case \( \mathbf{L} = \Lambda \).

Next, we assume as before that none of the eigenvalues differ by an integer constant, in which case the term \( \mathbf{I} + \mathbf{L} \) is never singular. This in turn implies that the only solution of the second equations is \( \mathbf{C}_1 = \mathbf{O} \), and by recursion \( \mathbf{C}_{2k+1} = \mathbf{O} \), i.e. all odd coefficient matrices vanish. Finally, the recursion for the even coefficients is

\[
\Lambda^2 \mathbf{C}_k - \mathbf{C}_k (k\mathbf{I} + \mathbf{L})^2 = \mathfrak{m} \mathbf{C}_{k-2}.
\]

Given a known right hand side, this is a Lyapunov equation of the form \( \mathbf{A} \mathbf{X} + \mathbf{X} \mathbf{B} = \mathbf{C} \), for which there exist established solution methods; that solution is particularly simple when \( \mathbf{L} \) is a diagonal matrix. More generally, observe that \( \mathbf{C}_0 \) could even be a complex or Hermitian matrix, although it is not a priori clear to us the advantage of such a choice. In the light of this uncertainty, we digress briefly to consider the effect of choosing \( \mathbf{L} = \mathbf{C}_0^{-1} \Lambda \mathbf{C}_0 \) with \( \mathbf{C}_0 \neq \mathbf{I} \). The common, leading term of the series is then

\[
z^L = \exp(\mathbf{L} \ln z) = \exp \left( \mathbf{C}_0^{-1} (\Lambda \ln z) \mathbf{C}_0 \right) = \mathbf{I} + \mathbf{C}_0^{-1} (\Lambda \ln z) \mathbf{C}_0 + \frac{1}{2!} \left[ \mathbf{C}_0^{-1} (\Lambda \ln z) \mathbf{C}_0 \right]^2 + \cdots
\]

\[
= \mathbf{C}_0^{-1} \left[ \mathbf{I} + (\Lambda \ln z) + \frac{1}{2!} (\Lambda \ln z)^2 + \cdots \right] \mathbf{C}_0 = \mathbf{C}_0^{-1} \exp(\Lambda \ln z) \mathbf{C}_0
\]

\[
= \mathbf{C}_0^{-1} z^\Lambda \mathbf{C}_0
\]

in which case

\[
\mathbf{Y}(z) = \left[ \sum_{k=0}^{\infty} \mathbf{C}_k z^k \right] z^L = \left[ \sum_{k=0}^{\infty} \left( \mathbf{C}_k \mathbf{C}_0^{-1} \right) z^{k+\Lambda} \right] \mathbf{C}_0.
\]

Also, from the recursion, we obtain multiplying from the right by \( \mathbf{C}_0^{-1} \)

\[
\Lambda^2 (\mathbf{C}_k \mathbf{C}_0^{-1}) - (\mathbf{C}_k \mathbf{C}_0^{-1}) (k\mathbf{I} + \mathbf{L}) \mathbf{C}_0^{-1} \mathbf{C}_0 = \mathfrak{m} (\mathbf{C}_{k-2} \mathbf{C}_0^{-1})
\]
\[ \Lambda^2 (C_k C_0^{-1}) - (C_k C_0^{-1}) (kI + \Lambda)^2 = \mathfrak{m} (C_{k-2} C_0^{-1}) . \]

With the definition \( \tilde{C}_k = C_k C_0^{-1} \), \( \tilde{C}_0 = I \), we conclude that the new recursion is

\[ \Lambda^2 \tilde{C}_k - \tilde{C}_k (kI + \Lambda)^2 = \mathfrak{m} \tilde{C}_{k-2} \]

which coincides with the recursion obtained by choosing \( C_0 = I \) in the first place, except that now

\[ Y(z) = \left[ \sum_{k=0}^{\infty} \tilde{C}_k z^{kI + \Lambda} \right] C_0 \]

has a trailing multiplicand \( C_0 \), that is, we now obtain a linear combination of solutions, which are proper (even if not independent) solutions too. Considering this finding, we restate again our previous doubt: What is the technical or numerical advantage gained by choosing \( C_0 \neq I \)? Observe also that the recursion for the \( \tilde{C}_k \) is much simpler when \( C_0 = I \), because it is a Lyapunov problem \( AX + XB = C \) with diagonal matrices \( A, B \).

A second set of solutions is obtained by making the replacement \( L \rightarrow -L \) while assuming that none of the eigenvalues is an integer so that the matrix \( L - kI \) is never singular. The new solution is then of the form

\[ Y(z) = \left[ \sum_{k=0}^{\infty} D_k z^{kI} \right] z^{-L}. \]

These matrices obey the recursion

\[
\begin{align*}
D_0 L^2 - \Lambda^2 D_0 &= 0 \quad \Rightarrow D_0 = C_0 \\
D_1 (I + L^2) - \Lambda^2 D_1 &= 0 \\
D_k (kI - L)^2 - \Lambda^2 D_k + \mathfrak{m} D_{k-2} &= 0
\end{align*}
\]

Navarro et al.\(^{(1)}\) provide a rigorous proof demonstrating that the two sets of solution with \( \pm L \) constitute the full solution to the system of equations. However, they make no suggestions or recommendations as to how to choose the initial, non-singular coefficient matrix \( C_0 \), nor do they provide any numerical examples for a broad range of the independent parameter \( z \).

5.3 | Method 3: asymptotic solution

In the ensuing we proceed to find an approximate solution based on neglecting certain terms in negative powers of \( z \) in the differential equation. Because this solution converges to the exact one as \( z \rightarrow \infty \), we refer it to as an asymptotic solution. To arrive at this solution we begin by making a change of variable

\[
\begin{align*}
y &= \frac{1}{\sqrt{z}} w \\
y' &= \frac{1}{\sqrt{z}} \left( w' - \frac{1}{2z} w \right) \\
y'' &= \frac{1}{\sqrt{z}} \left( w'' - \frac{1}{z} w' + \frac{3}{4z^2} w \right)
\end{align*}
\]

and

\[ z^2 y'' + z y' = \frac{1}{\sqrt{z}} \left( z^2 w'' + \frac{1}{4} w \right) . \]

Omitting the common factor \( \sqrt{z} \), this implies a differential equation of the form

\[ z^2 \frac{d^2 w}{dz^2} + \left[ z^2 \mathfrak{m} + \frac{1}{4} I - \Lambda^2 \right] w = 0 \]
where $\mathcal{M}$ is generally a fully populated matrix. Equivalently

$$\frac{d^2w}{dz^2} + \mathcal{M}w - \frac{1}{z^2} \left[ \Lambda^2 - \frac{1}{4} I \right] w = 0.$$  

This expression no longer contains a term in the first derivative. For reasons of convenience, we now choose to write this differential equation in full matrix form as

$$\frac{d^2W}{dz^2} + \mathcal{M}W - \frac{1}{z^2} \left[ \Lambda^2 - \frac{1}{4} I \right] W = 0$$

where $W = \{w_j\}$ is an $N \times N$ matrix of solutions. To find an approximation in the asymptotic sense, we proceed to make a simplification: For large $z$, it is reasonable to assume that we can neglect the last term in brackets, in which case the differential equation simplifies into

$$\frac{d^2\tilde{W}}{dz^2} + \mathcal{M}\tilde{W} = 0$$

where $\tilde{W}$ is a first order (or principal) asymptotic approximation to $W$. Let $\mathcal{F}$, $\mathcal{K}$ be the modal and spectral matrices of the eigenvalue problem for $\mathcal{M}$, namely

$$\mathcal{M}\Phi = \Phi \mathcal{K}^2, \quad \mathcal{M}^{-1} f_j = \kappa_j^2 f_j, \quad \mathcal{F} = \{f_j\}, \quad \mathcal{K} = \text{diag}\{\kappa_j\}.$$  

In the light of the earlier definition of the normalized matrix $\mathcal{M} = \Psi^T \mathcal{M} \Psi$ using the modes and orthogonality conditions for the pair of matrices $\mathcal{A}$, $\mathcal{G}$, namely $\Psi^T \mathcal{A} \Psi = I$ and $\Psi^T \mathcal{G} \Psi = \Lambda^2$, the above expressions imply

$$\Psi^T \mathcal{M} \Phi = \mathcal{F} \mathcal{K}^2, \quad \mathcal{M}(\Psi \Phi) = \Lambda(\Psi \Phi) \mathcal{K}^2$$

that is,

$$\Phi = \Psi \Phi, \quad \mathcal{F}^T \mathcal{M} \Phi = \mathcal{K}^2, \quad \mathcal{F}^T \mathcal{F} = I.$$  

If $\mathbf{C}$ is a square matrix of arbitrary constants, we see that the simplified differential equation admits solutions of the form

$$\tilde{W} = \mathcal{F} \exp(\pm i \mathcal{K} z) \mathbf{C}, \quad Y_0 = \frac{1}{\sqrt{z}} \tilde{W} = \frac{1}{\sqrt{z}} \mathcal{F} \exp(\pm i \mathcal{K} z) \mathbf{C}$$

which provides the principal asymptotic solution. That this is so can be seen by substituting the last of these expressions into the asymptotic differential equation, after which we obtain

$$(-\mathcal{F} \mathcal{K}^2 + \mathcal{M} \mathcal{F}) \exp(\pm i \mathcal{K} z) \mathbf{C} = \mathbf{0}$$

which is clearly satisfied because the term in parenthesis is the eigenvalue problem. Hence, the solution given above is indeed the principal asymptotic solution, and since we consider both positive and negative exponents, the full range of $2N$ solutions is taken into account. In addition, the solution makes no requirements on the constants of integration $\mathbf{C}$, which up to this point remain arbitrary.

We shall also need an expression for the powers of $\mathcal{M}$. For this purpose, from the definition we have

$$\mathcal{M} = \mathcal{F}^{-1} \mathcal{K}^2 \mathcal{F}^{-1} \mathcal{F}^T.$$  

It follows that

$$\mathcal{M}^{\pm n} = \mathcal{F} K^{\pm n} \mathcal{F}^T.$$
and in particular
\[ \mathcal{M}^{\pm i} = \mathcal{F} K^{\pm i} \mathcal{F}^T. \]

We now proceed to find the parameters and constants needed so that the asymptotic solution will closely approximate the ascending series solution at some adequately large abscissa \( z \). From the above modal expansion, we infer that
\[ \tilde{W}' = \pm i \mathcal{F} K \exp(\pm i K z) C, \quad \tilde{W}'' = -\mathcal{F} K^2 \exp(\pm i K z) C. \]

Let \( y_k(z), y'_k(z) \) be the \( k^{th} \) actual solution at some intermediate or adequately large \( z \) obtained by means of the Frobenius method when \( L = \text{diag} \{ \lambda_j \} > 0 \). Hence
\[ \mathbf{Y} = \{ y_k \}, \quad \mathbf{Y}' = \{ y'_k \}, \quad k = 1, 2, \ldots, N \]

On the other hand, any one of the \( N \) principal asymptotic solutions together with its spatial derivative are of the form
\[ \tilde{y}_m = \frac{1}{\sqrt{z}} f_m \cos(\kappa_m z - \theta_m) d_m, \]
\[ \tilde{y}'_m = -\left[ \frac{1}{\sqrt{z}} \kappa_m f_m \sin(\kappa_m z - \theta_m) d_m + \frac{1}{2z} \tilde{y}_m \right] \]
where \( \theta_m, d_m \) are as yet free parameters. We now proceed to use these \( N \) asymptotic solutions \( \tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_N \) as a vector base to express and match the actual solution \( y_k \) at some fixed \( z = z_0 \):
\[ y_k(z_0) = \frac{1}{\sqrt{z_0}} \sum_{m=1}^{N} f_m \cos(\kappa_m z_0 - \theta_{mk}) d_{mk} \]
\[ y'_k(z_0) = \frac{1}{\sqrt{z_0}} \left\{ \sum_{m=1}^{N} f_m \left[ -\kappa_m \sin(\kappa_m z_0 - \theta_{mk}) - \frac{1}{2z_0} \cos(\kappa_m z_0 - \theta_{mk}) \right] d_{mk} \right\} \]
where \( \theta_{mk}, d_{mk} \) are the appropriate constants that are needed to match the \( k^{th} \) actual solution \( y_k \) at \( z_0 \). Multiplying these expressions by the transposed modal vector \( f_j^T \) weighed by \( \sqrt{z_0} \), considering the orthogonality \( f_j^T f_m = \delta_{jm} \), and then reverting the dummy index \( j \) into \( m \) we obtain
\[ x_{mk} = \sqrt{z_0} f_m^T y_k = \cos(\kappa_m z_0 - \theta_{mk}) d_{mk} \]
\[ x'_{mk} = \sqrt{z_0} f_m^T y'_k = -\left[ \kappa_m \sin(\kappa_m z_0 - \theta_{mk}) + \frac{1}{2z_0} \cos(\kappa_m z_0 - \theta_{mk}) \right] d_{mk}. \]

It follows that
\[ \cos(\kappa_m z_0 - \theta_{mk}) d_{mk} = x_{mk} \]
\[ \sin(\kappa_m z_0 - \theta_{mk}) d_{mk} = -\left( x'_{mk} + \frac{1}{2z_0} x_{mk} \right) / \kappa_m. \]

Dividing the second of these expressions by the first we obtain
\[ \tan(\kappa_m z_0 - \theta_{mk}) = -\left( x'_{mk} + \frac{1}{2z_0} x_{mk} \right) / \kappa_m x_{mk} \]
and by inversion
\[ \theta_{mk}(z_0) = \kappa_m z_0 - \theta_{mk} = \arctan \left[ -\frac{\left( x'_{mk} + \frac{1}{2z_0} x_{mk} \right)}{\kappa_m x_{mk}} \right]. \]
Also,
\[ d_{mk} = \sqrt{(x_{mk})^2 + \left( x_{mk}' + \frac{1}{2z_0} x_{mk} \right)^2 / \kappa_m^2} \]
and clearly,
\[ \theta_{mk}(z) = \theta_{mk}(z_0) + \kappa_m \Delta z, \quad \Delta z = z - z_0 \]

Therefore, the solution for \( z \geq z_0 \) can then be written compactly as
\[
\begin{align*}
y_k(z) &= \frac{1}{\sqrt{z}} \sum_{m=1}^{N} f_m \cos (\theta_{mk} + \kappa_m \Delta z) d_{mk} = \frac{1}{\sqrt{z}} \mathbb{F} c_k \\
y_k'(z) &= -\frac{1}{\sqrt{z}} \mathbb{F} \left( K_s + \frac{1}{2z} c_k \right)
\end{align*}
\]
where \( c_k, s_k \) are the column vectors
\[
c_k = \{ \cos (\theta_{mk} + \kappa_m \Delta z) d_{mk} \}, \quad s_k = \{ \sin (\theta_{mk} + \kappa_m \Delta z) d_{mk} \}
\]
in which \( m \) is the row index. Clearly, we can now write the complete asymptotic solution as
\[
\begin{align*}
\tilde{Y}(z) &= \frac{1}{\sqrt{z}} \mathbb{F} C, \\
\tilde{Y}'(z) &= -\frac{1}{\sqrt{z}} \mathbb{F} \left[ Ks + \frac{1}{2z} C \right]
\end{align*}
\]
\[
C(z) = \{ c_1 \ldots c_N \} = \{ c_{mk} \}, \quad S(z) = \{ s_1 \ldots s_N \} = \{ s_{mk} \}
\]
where
\[
c_{mk} = \cos (\theta_{mk} + \kappa_m \Delta z) d_{mk}, \quad s_{mk} = \sin (\theta_{mk} + \kappa_m \Delta z) d_{mk}.
\]

Combining both terms into a single matrix expression, we obtain
\[
\begin{align*}
\left\{ \begin{array}{l}
\tilde{Y}(z) \\
\tilde{Y}'(z)
\end{array} \right\} &= \frac{1}{\sqrt{z}} \left\{ \begin{array}{c}
\mathbb{F} \\
\mathbb{O}
\end{array} \right\} \left\{ \begin{array}{c}
I \\
2z K
\end{array} \right\} \left\{ \begin{array}{c}
C \\
S
\end{array} \right\} \\
\end{align*}
\]

Observe that we can expand the elements of the \( C, S \) matrices as
\[
c_{mk} = (d_{mk} \cos \theta_{mk}) \cos (\kappa_m \Delta z) - (d_{mk} \sin \theta_{mk}) \sin (\kappa_m \Delta z)
\]
\[
s_{mk} = (\sin \theta_{mk} d_{mk}) \cos (\kappa_m \Delta z) + (\cos \theta_{mk} d_{mk}) \sin (\kappa_m \Delta z)
\]
or in matrix form (do observe that the diagonal matrices \( \cos (K \Delta z), \sin (K \Delta z) \) must trail \( C_0, S_0 \))
\[
C = C_0 \cos (K \Delta z) - S_0 \sin (K \Delta z), \quad \cos (K \Delta z) = \text{diag} \{ \cos (\kappa_m \Delta z) \}
\]
\[
S = S_0 \cos (K \Delta z) + C_0 \sin (K \Delta z), \quad \sin (K \Delta z) = \text{diag} \{ \sin (\kappa_m \Delta z) \}
\]
where
\[
C_0 = C(\Delta z = 0) = \{ d_{mk} \cos \theta_{mk} \}, \quad S_0 = S(\Delta z = 0) = \{ d_{mk} \sin \theta_{mk} \}.
\]

Hence, the asymptotic solution can be written as
\[
\begin{align*}
\left\{ \begin{array}{l}
\tilde{Y}(z) \\
\tilde{Y}'(z)
\end{array} \right\} &= \frac{1}{\sqrt{z}} \left\{ \begin{array}{c}
\mathbb{F} \\
\mathbb{O}
\end{array} \right\} \left\{ \begin{array}{c}
I \\
2z I
\end{array} \right\} \left\{ \begin{array}{c}
C_0 \cos (K \Delta z) - S_0 \sin (K \Delta z) \\
S_0 \cos (K \Delta z) + C_0 \sin (K \Delta z)
\end{array} \right\}.
\end{align*}
\]
In particular, if we match the exact (i.e. Frobenius) solution at \( z = z_0 \) such that \( \tilde{Y}(z_0) = Y(z_0) \equiv Y_0 \) and \( \tilde{Y}'(z_0) = Y'(z_0) \equiv Y'_0 \), we obtain

\[
\begin{align*}
\begin{bmatrix} Y_0 \\ Y'_0 \end{bmatrix} &= \frac{1}{\sqrt{z}} \begin{bmatrix} \mathfrak{F} & O \\ O & -\mathfrak{F} \end{bmatrix} \begin{bmatrix} 1 & O \\ 1 & K \end{bmatrix} \begin{bmatrix} C_0 \\ S_0 \end{bmatrix}
\end{align*}
\]

and after inversion

\[
\begin{align*}
\begin{bmatrix} C_0 \\ S_0 \end{bmatrix} &= \sqrt{z_0} \begin{bmatrix} 1 & K^{-1} \\ -\frac{1}{2z_0} I & -\mathfrak{F} \end{bmatrix} \begin{bmatrix} \mathfrak{F}^T & O \\ O & -\mathfrak{F}^T \end{bmatrix} \begin{bmatrix} Y_0 \\ Y'_0 \end{bmatrix}
\end{align*}
\]

Hence,

\[
\begin{align*}
\begin{bmatrix} \tilde{Y}(z) \\ \tilde{Y}'(z) \end{bmatrix} &= \frac{1}{\sqrt{z}} \begin{bmatrix} \mathfrak{F} & O \\ O & -\mathfrak{F} \end{bmatrix} \begin{bmatrix} 1 & O \\ 1 & K \end{bmatrix} \begin{bmatrix} C_0 \cos(K\Delta z) - S_0 \sin(K\Delta z) \\ S_0 \cos(K\Delta z) + C_0 \sin(K\Delta z) \end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
= \sqrt{\frac{1}{z_0}} \begin{bmatrix} \mathfrak{F} & O \\ O & -\mathfrak{F} \end{bmatrix} \begin{bmatrix} 1 & O \\ 1 & K \end{bmatrix} \begin{bmatrix} \mathfrak{F}^T Y_0 \cos(K\Delta z) + K^{-1} \mathfrak{F}^T \left( \frac{1}{2z_0} Y_0 + Y'_0 \right) \sin(K\Delta z) \\ \mathfrak{F}^T Y_0 \sin(K\Delta z) - K^{-1} \mathfrak{F}^T \left( \frac{1}{2z_0} Y_0 + Y'_0 \right) \cos(K\Delta z) \end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
= \sqrt{\frac{1}{z_0}} \begin{bmatrix} Y_0 \cos(K\Delta z) + \mathfrak{F} K^{-1} \mathfrak{F}^T \left( \frac{1}{2z_0} Y_0 + Y'_0 \right) \sin(K\Delta z) \\ \frac{1}{2z_0} \left( 1 - \frac{z_0}{z} \right) Y_0 + Y'_0 \cos(K\Delta z) - \frac{1}{2z_0} \mathfrak{F} K^{-1} \mathfrak{F}^T \left( \frac{1}{2z_0} Y_0 + Y'_0 \right) + \mathfrak{F} \mathfrak{F}^T Y_0 \sin(K\Delta z) \end{bmatrix}
\end{align*}
\]

With the equivalence \( \mathfrak{M}^{\pm 1} = \mathfrak{F} K^{\pm 1} \mathfrak{F}^T \) this can be written as

\[
\begin{align*}
\begin{bmatrix} \tilde{Y}(z) \\ \tilde{Y}'(z) \end{bmatrix} = \sqrt{\frac{1}{z_0}} \begin{bmatrix} Y_0 \cos(K\Delta z) + \mathfrak{M}^{-\frac{1}{2}} \left( \frac{1}{2z_0} Y_0 + Y'_0 \right) \sin(K\Delta z) \\ \frac{1}{2z_0} \left( 1 - \frac{z_0}{z} \right) Y_0 + Y'_0 \cos(K\Delta z) - \frac{1}{2z_0} \mathfrak{M}^{-\frac{1}{2}} \left( \frac{1}{2z_0} Y_0 + Y'_0 \right) + \mathfrak{M}^{+\frac{1}{2}} Y_0 \sin(K\Delta z) \end{bmatrix}
\end{align*}
\]

which provides an explicit expression for the asymptotic solution in terms of the reference solution at \( z_0 \).

**Example**

Consider the following \( 2 \times 2 \) non-proportional system:

\[
A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}.
\]

Figure 1(a) shows a comparison of the response obtained by the Frobenius method with positive characteristic numbers \( \lambda_j \) and the solution by the asymptotic method just described and when the parameters for the latter are defined by the conditions at \( z_0 = 15 \). The agreement, although not perfect, is still excellent. Of course, the higher the starting value \( z_0 \), the more accurate the asymptotic solution becomes, but by then the Frobenius method may begin to fail for the reasons elaborated earlier on. This is illustrated in Figure 1(b), which depicts the root-mean-square \( R(z) \) (of the four components) of the residual in the differential equation, a topic that will be described in further detail in the ensuing. Observe the systematic rise in the error of the Frobenius solution, which begins exceeding the error of the asymptotic solution when \( z \approx 24 \).
Interpretation in terms of Bessel functions

We saw previously that any one of the asymptotic solutions is of the form

\[ y_k(z) = \frac{1}{\sqrt{z}} \sum_{m=1}^{N} f_m \cos \left( k_m z - \phi_m^{(k)} \right) d_m^{(k)}. \]

But from the principal asymptotic form of the Bessel function, we know that

\[ \frac{1}{\sqrt{z}} \cos \left( k_m z - \phi_m^{(k)} \right) = \frac{1}{\sqrt{z}} \cos \left\{ k_m z - \frac{\pi}{2} \left( \nu_m^{(k)} + \frac{1}{2} \right) \right\} \sim J_{\nu_m^{(k)}}(k_m z) \]

with

\[ \phi_m^{(k)} = \frac{\pi}{2} \left( \nu_m^{(k)} + \frac{1}{2} \right) \]

that is,

\[ \nu_m^{(k)} = \frac{2}{\pi} \phi_m^{(k)} - \frac{1}{2}. \]

Hence, the asymptotic response can just as well be written as

\[ y_k(z) = \sum_{m=1}^{N} f_m J_{\nu_m^{(k)}}(k_m z) d_m^{(k)}, \quad k = 1, 2, \ldots, N \]

which is now given in terms of Bessel functions. In the light of this finding, it is now clear that other valid asymptotic solutions can be of the form

\[ y_k(z) = \sum_{m=1}^{N} f_m Y_{\nu_m^{(k)}}(k_m z) d_m^{(k)}, \quad y_k(z) = \sum_{m=1}^{N} f_m H_{1}^{(1,2)}(k_m z) d_m^{(k)} \]

which are expressed in terms of either Neumann or Hankel functions. Observe, however, that in the asymptotic range there is virtually no difference between the Bessel functions of order shifted by \( \frac{1}{2} \pi \) and the Neumann functions of that same order, so these are merely convenient and not truly different forms.
Error analysis of asymptotic solution

We now proceed to determine the residual that arises when the asymptotic solution is substituted into the actual differential equation. When expressed in transformed coordinates, the exact differential equation and its solution are

\[ \mathcal{L}_1(W) = \frac{d^2W}{dz^2} + \left[ \mathfrak{M} + \frac{1}{z^2} \left( \frac{1}{4} I - \Lambda^2 \right) \right] W = 0, \quad Y = W/\sqrt{z}. \]

By contrast, the asymptotic solution \( \bar{W} \) is obtained by neglecting the last term in \( z^{-2} \). That is, \( \bar{W} \) is a solution to the differential equation

\[ \mathcal{L}_2(\bar{W}) = \frac{d^2\bar{W}}{dz^2} + \mathfrak{M} \bar{W} = 0. \]

Clearly, if we substitute the asymptotic solution into the original differential equation, we satisfy automatically the first part of it and are left with the residual

\[ R_W = \mathcal{L}_1(\bar{W}) = \frac{1}{z^2} \left( \frac{1}{4} I - \Lambda^2 \right) \bar{W}. \]

Now, in our algorithm for the asymptotic solution, we adjusted the constants involved in that solution so as to match the exact functional value and its derivative at some specific \( z_0 \). That is, we started from the expressions

\[ \bar{Y}(z_0) = \frac{1}{\sqrt{z}} \mathfrak{F} C, \quad \bar{Y}'(z_0) = -\frac{1}{\sqrt{z}} \mathfrak{F} \left[ K S + \frac{1}{2z} C \right] \]

which we adjusted so that at \( z = z_0 \) the exact and the asymptotic solutions agreed, i.e.

\[ \{ \bar{Y}(z_0) \} = \{ Y_0 \} = \frac{1}{\sqrt{z_0}} \{ \mathfrak{F} O \quad O \} \left[ \begin{array}{c} I \\ \frac{1}{2z_0} I - \mathfrak{K} \end{array} \right] \{ C_0 \}. \]

where

\[ C_0 = C(\Delta z = 0) = \{ \cos \theta_m^{(k)} a_m^{(k)} \}, \quad S_0 = S(\Delta z = 0) = \{ \sin \theta_m^{(k)} a_m^{(k)} \} \]

so

\[ \{ C_0 \} = \sqrt{z_0} \left[ \frac{1}{\sqrt{z}} \mathfrak{F}^T Y_0 \right] \{ S_0 \} = \frac{1}{\sqrt{z}} \left[ -\frac{1}{2z_0} K^{-1} \right] \left[ \begin{array}{c} \mathfrak{F}^T \\ O - \mathfrak{F}^T \end{array} \right] \left[ \begin{array}{c} I \\ \frac{1}{2z_0} I - \mathfrak{K} \end{array} \right] \left[ \begin{array}{c} Y_0 \\ Y'_0 \end{array} \right] \]

or explicitly

\[ C_0 = \sqrt{z_0} \mathfrak{F}^T Y_0, \quad S_0 = -K^{-1} \mathfrak{F}^T \left[ \frac{1}{2z_0} Y_0 + Y'_0 \right]. \]

Hence,

\[ \bar{Y}(z) = \frac{1}{\sqrt{z}} \mathfrak{F} \left[ C_0 \cos(K\Delta z) - S_0 \sin(K\Delta z) \right]. \]

It follows that the residual \( R(z) = \frac{1}{\sqrt{z}} R_W \) of the original differential equation at each component is

\[ R(z) = \frac{1}{z^2} \left( \frac{1}{4} I - \Lambda^2 \right) Y_A(z) \]

\[ = \frac{1}{z^{5/2}} \left( \frac{1}{4} I - \Lambda^2 \right) \mathfrak{F} \left[ C_0 \cos(K\Delta z) - S_0 \sin(K\Delta z) \right]. \]
We can also express the previous result as

\[
R(z) = \frac{1}{z^2} \sqrt{\frac{z_0}{z_0}} \left( A^2 - \frac{1}{4} I \right) \left\{ Y_0 \cos(K \Delta z) + \frac{1}{2z_0} Y_0 + Y'_0 \right\} \sin(K \Delta z) \]  

(6)

where \( \mathbf{M}^{-1} = \mathbf{F} K^{-1} \mathbf{F}^T \). Observe that at \( z = z_0 \), the residual is not zero but instead

\[
R(z_0) = \frac{1}{z_0^2} \left( A^2 - \frac{1}{4} I \right) Y_0
\]

(7)

that is, even though we match the actual functional values at \( z_0 \), we still fail to satisfy exactly the actual differential equation. At values \( z > z_0 \) i.e. \( \Delta z > 0 \), the residual both oscillates due to the \( \sin(K \Delta z) \) and \( \cos(K \Delta z) \) terms and decays as \( z^{-5/2} \).

Figure 2 illustrates a logarithmic plot of the residual \( R(z) \) (evaluated as the root-mean square, or RMS, of the four residuals of the differential equation) for the example presented earlier, and plotted for very wide set of values of the independent variable \( z \). The plot also includes a best-fit reference line with a slope equal to -2.5, as Equation (6) predicts. This line was obtained by fitting via linear regression a straight line to the oscillating residual within the interval \([15 \leq z \leq 800] \).

Observe further that the residual starts with values on the order of \( 10^{-2} \), which is also the order of magnitude of error of the function itself, inasmuch as the norm of the leading element in Equation (7) is of order 1. Thus, the asymptotic solution is reasonably accurate.

**Completeness and linear independence of the asymptotic solution**

In the foregoing we saw that a general form of the asymptotic solution together with its derivative can be expressed as

\[
\tilde{\mathbf{W}} = \mathbf{F} \left[ \exp(iKz) \mathbf{C}_1 + \exp(-iKz) \mathbf{C}_2 \right]
\]

\[
\tilde{\mathbf{W}}' = iK \mathbf{F} \left[ \exp(iKz) \mathbf{C}_1 + \exp(-iKz) \mathbf{C}_2 \right]
\]

where \( \mathbf{C}_1, \mathbf{C}_2 \) are matrices of arbitrary (possibly complex) constants. In matrix form, this is given by

\[
\begin{bmatrix}
\tilde{\mathbf{W}} \\
\tilde{\mathbf{W}}'
\end{bmatrix} =
\begin{bmatrix}
\mathbf{F} & \mathbf{O} \\
\mathbf{O} & \mathbf{F}
\end{bmatrix}
\begin{bmatrix}
\exp(iKz) & \exp(-iKz) \\
iK \exp(iKz) & -iK \exp(-iKz)
\end{bmatrix}
\begin{bmatrix}
\mathbf{C}_1 \\
\mathbf{C}_2
\end{bmatrix}.
\]

Suppose next that we prescribe the solution at \( z = z_0 \) to be the identity matrix. In other words, the asymptotic solution for \( z > z_0 \) will define a complete set of Green’s functions or fundamental solutions. In that case

\[
\begin{bmatrix}
\mathbf{F} & \mathbf{O} \\
\mathbf{O} & \mathbf{F}
\end{bmatrix}
\begin{bmatrix}
\exp(iKz) & \exp(-iKz) \\
iK \exp(iKz) & -iK \exp(-iKz)
\end{bmatrix}
\begin{bmatrix}
\mathbf{C}_1 \\
\mathbf{C}_2
\end{bmatrix} =
\begin{bmatrix}
\mathbf{I} & \mathbf{O} \\
\mathbf{O} & \mathbf{I}
\end{bmatrix}
\]

so

\[
\begin{bmatrix}
\mathbf{C}_1 \\
\mathbf{C}_2
\end{bmatrix} =
\begin{bmatrix}
\exp(iKz) & \exp(-iKz) \\
iK \exp(iKz) & -iK \exp(-iKz)
\end{bmatrix}^{-1}
\begin{bmatrix}
\mathbf{F}^T & \mathbf{O} \\
\mathbf{O} & \mathbf{F}^T
\end{bmatrix}.
\]
Clearly, $C_1, C_2$ are each of size $N \times 2N$, which is also the size of the fundamental solution $\tilde{W}$. The solution for $z > z_0$ is then

$$
\begin{bmatrix}
\tilde{W} \\
\tilde{W}'
\end{bmatrix} = \begin{bmatrix} \mathfrak{g} & 0 \\ 0 & \mathfrak{g} \end{bmatrix} \begin{bmatrix} \exp(iKz) & \exp(-iKz) \\ iK\exp(iKz) & -iK\exp(-iKz) \end{bmatrix} \begin{bmatrix} \exp(iKz) & \exp(-iKz) \\ iK\exp(iKz) & -iK\exp(-iKz) \end{bmatrix}^{-1} \begin{bmatrix} \mathfrak{g}^T & 0 \\ 0 & \mathfrak{g}^T \end{bmatrix}.
$$

Inasmuch as all submatrices are diagonal, the inversion is trivially composed of the inverses of the $2 \times 2$ submatrices

$$
\begin{bmatrix}
\exp(ik_jz_0) & \exp(-ik_jz_0) \\
ik_j\exp(ik_jz_0) & -ik_j\exp(-ik_jz_0)
\end{bmatrix}^{-1} = \frac{1}{2ik_j} \begin{bmatrix} ik_j\exp(-ik_jz_0) & \exp(-ik_jz_0) \\ ik_j\exp(ik_jz_0) & -\exp(ik_jz_0) \end{bmatrix}
$$

the determinant of which is

$$
\Delta_j = -\frac{1}{2ik_j}.
$$

Also, the determinant of the modal matrix is $|\mathfrak{g}| = \pm 1$. It follows that the determinant of the complete solution is

$$
W = \begin{bmatrix} \tilde{W} \\ \tilde{W}' \end{bmatrix} = \begin{bmatrix} \exp(iKz) & \exp(-iKz) \\ iK\exp(iKz) & -iK\exp(-iKz) \end{bmatrix} \begin{bmatrix} \exp(iKz) & \exp(-iKz) \\ iK\exp(iKz) & -iK\exp(-iKz) \end{bmatrix}^{-1} = \prod_{j=1}^N \frac{2ik_jz_0}{2ik_j} = 1
$$

which is referred to as the Wronskian of the solution. Inasmuch as it is nowhere zero, that shows that the asymptotic solution thus constructed constitutes a linearly independent set that is complete and proper. Following a similar development, one can also show that the physical solution $\tilde{Y} = \frac{1}{\sqrt{z}}\tilde{W}$ also forms a linearly independent set satisfying

$$
W = \begin{bmatrix} \tilde{Y} \\ \tilde{Y}' \end{bmatrix} = \prod_{j=1}^N \frac{2ik_jz_0}{2ik_j} = \left(\frac{z_0}{z}\right)^N.
$$

### 5.4 Method 4: numerical integration via Runge-Kutta

As already mentioned, numerical experiments demonstrate that the Frobenius solution works well for small to intermediate values of the argument, but that it fails at large values. Besides the asymptotic solution, another alternative to overcome this problem is by using a hybrid method, wherein the solution to both $u$ and $\frac{d}{dz}u$ is found using the Frobenius method with small arguments. Thereafter, the starting solution is extended via numerical integration using the highly accurate Runge-Kutta method in its vectorial version,[5] or better and more accurately still, the Runge-Kutta-Nyström version of that algorithm.[6] We begin with a concise summary of the latter method.

Let $y_0, v_0 = \frac{d}{dz}y_0$ be the complete solution at some initial, fixed (low magnitude) value $z_0$ found via the Frobenius method of the previous sections, say one column in the matrix $Y_0 = J_j(z_0)$. Consider also a generic point $z_n$ (with $n > 0$) whose solution is $y_n = y(z_n)$ and the solution at a neighboring point $z_{n+1} = z_n + \Delta z$ is $y_{n+1}$. We begin by defining the differential operator

$$
f_n = f(y_n, v_n, z_n) = \frac{d^2y_n}{dz^2} = -\frac{1}{z_n}v_n - \left(m - \frac{1}{z_n^2}\lambda^2\right)y_n.
$$

### Runge-Kutta-Nyström algorithm

The Nyström version of the Runge-Kutta algorithm in matrix form proceeds along four steps that differ slightly from those in the conventional Runge-Kutta method:

$$
\begin{align*}
d_1 &= y_n \\
d_2 &= y_n + \frac{1}{2}\Delta z g_1 + \frac{1}{8}\Delta z^2 h_1 \\
d_3 &= d_2 \\
d_4 &= y_n + \Delta z g_1 + \frac{1}{2}\Delta z^2 h_3
\end{align*}
\begin{align*}
g_1 &= v_n \\
g_2 &= v_n + \frac{1}{2}\Delta z h_1 \\
g_3 &= v_n + \frac{1}{2}\Delta z h_3 \\
g_4 &= v_n + \Delta z h_3
\end{align*}
\begin{align*}
h_1 &= f(d_1, g_1, z_n) \\
h_2 &= f(d_2, g_2, z_n + \frac{1}{2}\Delta z) \\
h_3 &= f(d_3, g_3, z_n + \frac{1}{2}\Delta z) \\
h_4 &= f(d_4, g_4, z_n + \Delta z)
\end{align*}
and the solution at step $z_{n+1}$ is

$$y_{n+1} = y_n + \Delta z \, g_1 + \frac{1}{6} \Delta z^2 \left( h_1 + h_2 + h_3 \right)$$

$$v_{n+1} = v_n + \frac{1}{6} \Delta z \left( h_1 + 2h_2 + 2h_3 + h_4 \right).$$

**Example**

Consider the $2 \times 2$ matrices in the $N = 2$ space

$$A = \frac{1}{6} \left\{ \begin{array}{cc} 2 & 1 \\ 1 & 2 \end{array} \right\}, \quad M = \frac{1}{2} \left\{ \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right\}, \quad G = \left\{ \begin{array}{cc} -1 & -1 \\ -1 & 2 \end{array} \right\}$$

which satisfy

$$\sqrt{|A|} = \frac{1}{6} \sqrt{3}, \quad \sqrt{|M|} = \frac{1}{2}, \quad \sqrt{|G|} = 1, \quad AM^{-1}G = \frac{1}{3} \left\{ \begin{array}{cc} 1 & 0 \\ -1 & 3 \end{array} \right\}.$$ 

Inasmuch as the triple product is not symmetric, we know for sure that the matrices are not proportional. Hence, the solution will be fully coupled. The square root of the eigenvalues of the pair of matrices $A, G$ and the coupling matrix $\mathcal{M}$ are respectively

$$\mathcal{A} = \left\{ \begin{array}{cc} \sqrt{8 - 2\sqrt{13}} & 0 \\ 0 & \sqrt{8 + 2\sqrt{13}} \end{array} \right\} = \left\{ \begin{array}{cc} 0.8882 & 0 \\ 0 & 3.9001 \end{array} \right\}$$

$$\mathcal{M} = \left\{ \begin{array}{cc} 0.594252 & 0.138675 \\ 0.138675 & 1.715149 \end{array} \right\}, \quad |\mathcal{M}| = 1.$$

To establish the derivatives at an initial point near $z_0 \approx 0.10$, we estimated numerically the derivatives of all matrix elements at the third point in the Frobenius series method using the 5-point central differences method:

$$y_3' = \left. \frac{d}{dz} y \right|_{z_3} \approx \frac{1}{24\Delta z} \left[ 2(y_1 - y_3) - 16(y_2 - y_4) \right]$$

$$y_3'' = \left. \frac{d^2}{dz^2} y \right|_{z_3} \approx \frac{1}{24(\Delta z)^2} \left[ -2(y_1 + y_5) - 60y_3 + 32(y_2 + y_4) \right].$$

This allowed us to avoid the use of an ascending series for the derivatives. We also verified thereafter the satisfaction of the differential equation by evaluating the residue with the above differences, and we found that residue to be negligible. Because of the estimation of the first derivative via the 5-point central differences method, the Runge-Kutta solution now starts actually at the third discrete point, which is inconsequential.

Figure 3(a) shows the numerical solution obtained using the Frobenius method as dashed lines, and the numerical solution using Runge-Kutta as solid lines in the interval $[z] = 0.50 : 0.01 : 30.00$. As can be seen, the solid lines cover the dashed lines up to $z \approx 27$, and disagree strongly thereafter. Whereas the Runge-Kutta solution continues returning a negligible residual (Figure 3(b)), that is not the case with the ascending series, which fails in that high range. Indeed, using Runge-Kutta we have extended the solution to very high values such as $z = 50$, and the residual continued being negligible, i.e. the solution continued being accurate. This is because the largest term in the ascending series at $z = 30$ is $\approx 10^{14}$ (this is for the 15th term in the summation) whereas the converged solution is of order 1. Hence, there is total loss of accuracy in the computation.

More generally and in the context of other numerical experiments, we have observed that the Frobenius method generally fails above $z = 25$ when using double precision arithmetic. While we have also managed to extend further the practical range of $z$ by resorting to quadruple precision, we achieved that extension at a significant computational cost.

We should mention in passing that the Runge-Kutta method is ‘reversible’ or bidirectional, i.e. it works just as well with increasing or decreasing values of $z$, or equivalently, with positive or negative values of the increment $\Delta z$. This can prove useful when starting with asymptotic solutions for high values of $z$ and then working downwards from there. More generally still, the Runge-Kutta numerical integration need not be based on any special starting values at $z = z_0$ obtained via the Frobenius method. Instead, it would suffice to prescribe any arbitrary set of initial conditions, of which we could define a total $2N$ of these.
FIGURE 3  Solution of the $2 \times 2$ system with the Frobenius method compared against the numerical integration via Runge-Kutta (a) and RMS of the residual error (b). The deviation of the first of these after $z \approx 27$ results from severe cancellation errors in the series.

The simplest such initial conditions for the response matrix $Y$ and its derivative, when written in matrix form, could be written succinctly as

$$\begin{bmatrix} Y & 0 \\ 0 & Y' \end{bmatrix} \bigg|_{z=z_0} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

with the right-hand side being an identity matrix. This is a $2N \times 2N$ matrix, each of whose columns define the initial conditions for one of the $2N$ independent solutions at $z = z_0$. That is, in one of these we would impose one component of the response to be unity, and all of the other components together with all of the derivatives to be zero. Or vice-versa, all response components can be defined as being zero and one of the derivatives is defined as unity. Of course, arbitrary initial conditions could be still imposed later on for which the solution would follow by linear superposition of the fundamental solutions so obtained. The reader will recognize this strategy as the solution to an initial value problem for which the unit response functions define a set of fundamental solutions or Green’s functions.

Finally, we should mention that inasmuch as the Runge-Kutta Method is strictly a numerical procedure that does not rely on solving any eigenvalue problem, we could just as well use it for general, non-symmetric, non-proportional matrices $A, G, M$, i.e. we could directly solve Equation (1) without further transformations, and merely prescribe appropriate initial conditions.

**Potential pitfalls with numerical integration**

Extensive testing carried out with the Runge-Kutta method for large systems of differential equations of the Bessel type seemed to confirm the general robustness of this integration method, yet at the same time it also revealed some fundamental problems—which at first proved puzzling to us—that might arise. The problems had to do with the fact that the response functions for the non-proportional case constitute a complicated superposition of Bessel functions $J_\nu, Y_\nu$ of multiple orders $\nu = \lambda_j$, namely the eigenvalues of the pair of matrices $A, G$, and that the largest of these eigenvalues can control the response—at least numerically. In addition, for sufficiently high argument $z$ (i.e. in the asymptotic range), there is hardly any difference between $J_\nu(z)$ and $Y_{\nu+\pi/2}(z)$.

The problems encountered can be explained by means of the simple example illustrated in Figure 4 showing $J_{15}(z)$ and $Y_{15+\pi/2}(z)$. Both of these functions are of a relatively high order $\nu$. The first detail that should be observed is that the Bessel function is flat all the way up to $z = 10$, and is practically zero below that value. Indeed, at the not-so-small value of $z = 5$, we have $J_{15}(5) = 0.48 \cdot 10^{-6}$, which is in fact still very small. Suppose then that we are integrating numerically beginning at $z_0 = 1$ (a point at which $J_{15}(1) = 2 \times 10^{-17}$) and impose there initial conditions that we know will not contain contributions of high order. As the numerical integration then progresses from left to right, tiny round off errors get introduced into the response functions and these are ‘interpreted’ by the integration engine as participation of the high mode to the response function. Although at first negligible, these diminutive errors become dramatically manifest as $z$ increases and take over the response with a very high, even if false, modal participation. In effect, the response ‘blows up’ and the computation is rendered invalid. Thus,
the numerical integration becomes unreliable whenever it is started at a $z_0$ such that $J_{\max(\lambda_j)}(z_0)$ is still a very small number, i.e. within this function’s flat domain. Apparently, such flat domain might be characterized by a simple rule such as $z_0 < \frac{2}{3} \max(\lambda_j)$.

The opposite problem becomes apparent when starting from the asymptotic solution and integrating in reverse order from high to low $z$. It will then be seen how ‘exploding’ branches appear in the low $z$ range. This is because in the asymptotic range there is hardly any difference between a Bessel function of any given order and a Neumann function of that same order shifted by $\frac{1}{2} \pi$, and vice-versa. Therefore, as the integration progresses down, the computation will be taken over by the appropriate Neumann function. The end-result is that the computation diverges dramatically near the origin, and this gets worse as the order $\lambda_j$ of the functions is increased. As a practical rule, the low limit at which reverse-direction integration becomes unreliable is defined by the first zero of $Y_{\max(\lambda_j)}(z_0)$, which roughly takes place at $z_0 = \frac{6}{5} \nu_{\max}$. Conceivably, the first of these problems might be avoided with some kind of cleaning or rinsing, by means of which the participation of high modes is filtered out as the computation progresses, either at each step or after a finite number of such steps. We have, however, not explored the feasibility or practicality of such an alternative.

The reader should also understand that the numerical integration cannot be started from the starting point $z_0 = 0$, and this for several reasons. First, $z_0 = 0$ is the same as $\omega = 0$, which is the simpler ‘static’ problem considered in the earlier sections of the paper. Therefore, once we were to start integrating, we would be doing so along values of increasing $r$ and not $z$, and thus remain in the static domain. That is, we cannot infer the dynamic solution starting from the static solution. Secondly, it should be observed that in the differential equation used in the context of the RKN method, $z_n^2$ appears as a divisor and so the equation has an essential singularity at $z_n = 0$. In the light of these considerations, it seems best to start the numerical integration at values of $z_0$ for which one knows (and can verify) that the solution is accurate, using the Frobenius method as the starting tool.

6 | CONCLUSIONS

In this article, we have dealt with the numerical solution of a matrix set of differential equations whose structure is similar to that of the scalar differential equation for the Bessel functions, namely

$$A\left(u'' + \frac{1}{z} u'\right) + \left(M - \frac{1}{z^2} G\right) u = 0$$

where $A, M, G$ are positive definite, symmetric matrices. This applied mathematical task arose in the context of a wave propagation problem cast in cylindrical coordinates when the media exhibits material inhomogeneity in the azimuthal direction. If the matrices share a common set of eigenvectors, i.e. if they are ‘proportional’ in the general sense that they are simultaneously diagonalizable, then the exact solution follows from solving the common eigenvalue problem for these matrices, and the solution consists of combinations of classical Bessel functions.

More generally, when the three matrices are not proportional and the solution remains fully coupled, we have proposed three effective methods of finding the solution:
1. By application of the Frobenius method that leads to the so-called ascending series for the solution. This method works very well, but is restricted to relatively low values of the independent variable $z$, say $z \leq 15$ (but after the matrices have been appropriately scaled and normalized). At large $z$ the series solution fails because of severe cancellation errors.

2. By using asymptotic expressions that provide very good results when $z > 15$ or so. Hence, when combined with the first method, one can cover virtually the entire range of interest.

3. By direct numerical integration via the accurate Runge-Kutta-Nyström algorithm, which works very well for most intervals in the independent variable, and is valid whatever the structure of the matrices, real or complex or non-symmetric. However, it exhibits some idiosyncrasies related to spurious (or more precisely, improperly scaled) solutions that get introduced by the accumulation of numerical error and which are associated with large eigenvalues. These become manifest in the form of rapidly rising branches that are immediately apparent to the analyst.

4. By appropriate combinations of all of the above.

ACKNOWLEDGEMENTS

The research described in this article received no funding from any source, and represents curiosity-driven investigations. Still, the second author wishes to acknowledge and thank the travel grant received from the Mercator Research Center Ruhr that allowed him to visit the Massachusetts Institute of Technology during the summer of 2018, and which made this cooperation with the University of Duisburg-Essen possible.

ORCID

Eduardo Kausel https://orcid.org/0000-0002-7196-1402
Hauke Gravenkamp https://orcid.org/0000-0002-2641-6384

REFERENCES

[1] E. Navarro, R. Company, L. Jódar, *Applicationes Mathematicae* 1993, 22, 11.
[2] E. Navarro, M. V. Ferrer, L. Jodar, *Analysis* 1993, 13, 259.
[3] L. Jodar, E. A. Navarro, M. V. Ferrer, *Appl. Math.* 1995, 40, 391.
[4] M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions*, 10th ed., National Bureau of Standards 1972.
[5] E. Kausel, *Advanced Structural Dynamics*, 1st ed., Cambridge University Press 2017.
[6] E. J. Nyström, *Über die numerische Integration von Differentialgleichungen*, vol. 50, Acta Societatis Scientiarium Fennicae 1925.

How to cite this article: E. Kausel, H. Gravenkamp. On the numerical solution of matrix Bessel equations. *Z Angew Math Mech.* 2019;99:e201800288. https://doi.org/10.1002/zamm.201800288