Lawrence Berkeley National Laboratory

Applied Math & Comp Sci

Title

An algebraic sub-structuring method for large-scale eigenvalue calculation

Permalink

https://escholarship.org/uc/item/8fw31896

Journal

Lawrence Berkeley National Laboratory

Authors

Yang, C.
Gao, W.
Bai, Z.
et al.

Publication Date

2004-05-26

eScholarship.org Powered by the California Digital Library
University of California

An Algebraic Sub-structuring Method for Large-scale Eigenvalue Calculation

C. Yang*, W. Gao*†, Z. Bai‡, X. Li*, L. Lee§, P. Husbands*, E. Ng*

May 25, 2004

Abstract

We examine sub-structuring methods for solving large-scale generalized eigenvalue problems from a purely algebraic point of view. We use the term *algebraic sub-structuring* to refer to the process of applying matrix reordering and partitioning algorithms to divide a large sparse matrix into smaller submatrices from which a subset of spectral components are extracted and combined to provide approximate solutions to the original problem. We are interested in the question of which spectral components one should extract from each sub-structure in order to produce an approximate solution to the original problem with a desired level of accuracy. Error estimate for the approximation to the smallest eigenpair is developed. The estimate leads to a simple heuristic for choosing spectral components (modes) from each sub-structure. The effectiveness of such a heuristic is demonstrated with numerical examples. We show that algebraic sub-structuring can be effectively used to solve a generalized eigenvalue problem arising from the simulation of an accelerator structure. One interesting characteristic of this application is that the stiffness matrix produced by a hierarchical vector finite elements scheme contains a null space of large dimension. We present an efficient scheme to deflate this null space in the algebraic sub-structuring process.

1 Introduction

Sub-structuring is a commonly used technique for studying the static or dynamic properties of large engineering structures [3, 8, 12, 13, 17, 18, 19, 20, 21, 28, 32, 34]. The basic idea of sub-structuring is analogous to the concept of domain-decomposition widely used in the numerical solution of partial differential equations [30]. By dividing a large structure model or computational domain into a few smaller components (sub-structures), one can often obtain an approximate solution to the original problem from a linear combination (or direct sum) of solutions to similar problems defined on the sub-structures. Because solving problems on

*Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. {CYang@lbl.gov, WGGao@lbl.gov, XSLi@lbl.gov, PJRHusbands@lbl.gov, EGNg@lbl.gov} This work was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC03-76SF00098.

†Department of Mathematics, Fudan University, Shanghai, P. R. China 200433. The research of this author is carried out while he is visiting the Computational Research Division at Lawrence Berkeley National Laboratory.

‡Department of Computer Science, University of California, Davis, CA 95616, {bai@cs.ucdavis.edu}

§Stanford Linear Accelerator Center, Menlo Park, CA 94025. {liequan@slac.stanford.edu} The research of this author was supported by U.S. Department of Energy under contract number DE-AC03-76SF00515.

each sub-structure requires far less computational power than what would be required to solve the entire problem as a whole, sub-structuring can lead to a significant reduction in the computational time required to carry out a large-scale simulation and analysis.

The automated multi-level sub-structuring (AMLS) method [5, 6, 7, 22] is an extension of a simple sub-structuring method called *component mode synthesis* (CMS) [12, 20] originally developed in the 1960s. Detailed analysis is provided in [9, 10] to assess the level of accuracy attainable by the CMS method. The analysis is based on the standard finite element theory and properties of the partial differential equation governing the evolution of the structure. The recent work by Bennihof and Lehoucq [7] provides a high level mathematical description of the AMLS in a continuous variational setting. It also provides a framework for describing AMLS in matrix algebra notation.

Recent studies have shown that sub-structuring can be used successfully in the vibration and acoustic analysis of large-scale finite element models of automobile bodies [22, 25]. The frequency response analysis performed in these studies requires computing several thousand eigenvalues and eigenvectors associated with a large-scale symmetric generalized eigenvalue problem. The timing results reported in [22, 25] indicate that AMLS is significantly faster than conventional Lanczos-based approaches [26, 16].

It is important to note that the accuracy achieved by a sub-structuring method such as AMLS is typically lower than that achieved by the standard Lanczos algorithm. However, in many applications, the level of accuracy required from an approximate solution to an algebraic problem is no more than what is provided by the finite element scheme used to discretize the original continuous problem. Thus, the use of algebraic sub-structuring is easily justified as long as the error associated with the sub-structuring approximation does not exceed that produced by the finite element discretization.

Our focus in this paper is to examine sub-structuring methods for solving large-scale eigenvalue problems from a purely algebraic point of view. We use the term *algebraic sub-structuring* to refer to the process of applying matrix reordering and partitioning algorithms (such as the *nested dissection* algorithm [15]) to divide a large sparse matrix into smaller submatrices from which a subset of spectral components are extracted and combined to form an approximate solution to the original eigenvalue problem. We are interested in the question of which spectral components one should choose from each sub-structure and what level of accuracy one can expect from the approximate solution associated with a particular choice of sub-structure components. Our analysis is related to but different from the recent work by Berkas and Saad [4] who view algebraic sub-structuring as an approximation to the spectral Schur complement method [1, 2, 11].

Our interest in algebraic sub-structuring is motivated in part by an application arising from the simulation of the electromagnetic field associated with the next generation particle accelerator design [24]. We will show through a numerical example that algebraic sub-structuring can be used effectively to compute the cavity resonance frequencies and the electromagnetic field generated by a linear particle accelerator model. One interesting characteristic of this application is that the stiffness matrix produced by a hierarchical vector finite elements scheme contains a null space of large dimension. We will show how to effectively deflate this null space in the sub-structuring calculation.

Our presentation is organized as follows. In Section 2, we give a brief overview of the algorithmic ingredients of a simple algebraic sub-structuring method. The accuracy of the approximate eigenpairs returned from sub-structuring is analyzed in Section 3 by examining

the elements of the eigenvectors associated with a canonical matrix pencil congruent to the original matrix pencil. The analysis leads to a simple heuristic for choosing spectral components from each sub-structure. In Section 4, we discuss how to deflate the null space introduced by the stiffness matrix in algebraic sub-structuring. Our analysis of algebraic sub-structuring is confirmed by numerical examples presented in Section 5. These examples also suggests that algebraic sub-structuring can be used effectively to solve generalized eigenvalue problems arising from electromagnetic field simulations.

Throughout this paper, capital and lower case Latin letters denote matrices and vectors respectively, while lower case Greek letters denote scalars. An $n \times n$ identity matrix will be denoted by I_n. The j-th column of the identity matrix is denoted by e_j. The transpose of a matrix A is denoted by A^T. We use $\|x\|$ to denote the standard 2-norm of x, and use $\|x\|_M$ to denote the M-norm defined by $\|x\|_M = \sqrt{x^T M x}$. We will use $\angle_M(x, y)$ to denote the M-inner product induced acute angle (M-angle for short) between x and y. This angle can be computed from

$$\cos \angle_M(x, y) = \frac{x^T M y}{\|x\|_M \|y\|_M}.$$

Similarly, we use $\angle_M(x, \mathcal{S})$ to denote the M-angle between a vector x and a subspace $\mathcal{S}$. This angle can be computed from

$$\cos \angle_M(x, \mathcal{S}) = \frac{\|Q^T M x\|_2}{\|x\|_M}, \tag{1}$$

where Q is an M-orthonormal basis of the subspace $\mathcal{S}$, i.e., $\mathcal{S} = \text{span}\{S\}$ and $Q^T M Q = I$.

A matrix pencil (K, M) is said to be *symmetric definite* if both K and M are symmetric and M is positive definite. A matrix pencil (K, M) is said to be *congruent* to another pencil (A, B) if there exits a nonsingular matrix P, such that $A = P^T K P$ and $B = P^T M P$.

2 Algebraic Sub-structuring

In this section, we will briefly describe a single-level algebraic sub-structuring algorithm. Our description does not make use of any information regarding the geometry or the physical structure on which the original problem is defined.

We are concerned with solving the following generalized algebraic eigenvalue problem

$$Kx = \lambda M x, \tag{2}$$

where K is symmetric and M is symmetric positive definite. We assume K and M are both sparse. They may or may not have the same sparsity pattern. Suppose the rows and columns of K and M have been permuted so that these matrices can be partitioned as

$$K = \begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{pmatrix} K_{11} & & K_{13} \\ & K_{22} & K_{23} \\ K_{13}^T & K_{23}^T & K_{33} \end{pmatrix} \end{array} \quad \text{and} \quad M = \begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{pmatrix} M_{11} & & M_{13} \\ & M_{22} & M_{23} \\ M_{13}^T & M_{23}^T & M_{33} \end{pmatrix} \end{array}, \tag{3}$$

where the labels n_1, n_2 and n_3 are inserted to the top and left borders of the partitioned matrices to indicate the dimension of each sub-matrix block. The permutation can be accomplished by applying a matrix ordering and partitioning algorithm such as the nested dissection algorithm [15] to the matrix $K + M$.

The pencils (K_{11}, M_{11}) and (K_{22}, M_{22}) now define two algebraic sub-structures that are connected by the third block rows and columns of K and M which we will refer to as the *interface* block. We assume that n_1 and n_2 are about the same size, and n_3 is much smaller than n_1 and n_2.

A single-level algebraic sub-structuring algorithm proceeds by performing a block factorization $K = LDL^T$, where

$$L = \begin{pmatrix} I_{n_1} & & \\ & I_{n_2} & \\ K_{13}^T K_{11}^{-1} & K_{23}^T K_{22}^{-1} & I_{n_3} \end{pmatrix}, \tag{4}$$

and D is a block diagonal matrix

$$D = \begin{pmatrix} K_{11} & & \\ & K_{22} & \\ & & \widehat{K}_{33} \end{pmatrix}.$$

The last diagonal block of D, often known as the *Schur complement*, is defined by

$$\widehat{K}_{33} = K_{33} - K_{13}^T K_{11}^{-1} K_{13} - K_{23}^T K_{22}^{-1} K_{23}. \tag{5}$$

The inverse of the lower triangular factor L defines a congruent transformation that, when applied to the matrix pencil (K, M), yields a new matrix pencil $(\widehat{K}, \widehat{M})$, where

$$\widehat{K} = L^{-1} K L^{-T} = D = \begin{pmatrix} K_{11} & & \\ & K_{22} & \\ & & \widehat{K}_{33} \end{pmatrix}, \tag{6}$$

and

$$\widehat{M} = L^{-1} M L^{-T} = \begin{pmatrix} M_{11} & & \widehat{M}_{13} \\ & M_{22} & \widehat{M}_{23} \\ \widehat{M}_{13}^T & \widehat{M}_{13}^T & \widehat{M}_{33} \end{pmatrix}. \tag{7}$$

Note that the eigenvalues of $(\widehat{K}, \widehat{M})$ are identical to those of (K, M), and the corresponding eigenvectors are related to the eigenvectors of the original problem (2) through $\widehat{x} = L^T x$.

Our sub-structuring algorithm constructs a subspace in the form of

$$S = \begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccc} k_1 & k_2 & n_3 \end{array} \\ \begin{pmatrix} S_1 & & \\ & S_2 & \\ & & I_{n_3} \end{pmatrix} \end{array} \tag{8}$$

where S_1 and S_2 consist of k_1 and k_2 selected eigenvectors of (K_{11}, M_{11}) and (K_{22}, M_{22}) respectively. These eigenvectors will be referred to as *sub-structure modes* in the discussion that follows. Note that k_1 and k_2 are typically much smaller than n_1 and n_2, respectively.

The approximation to the desired eigenvalues and eigenvectors of the pencil $(\widehat{K}, \widehat{M})$ are obtained by projecting the pencil $(\widehat{K}, \widehat{M})$ onto the subspace spanned by S, i.e., we seek θ and $q \in \mathbb{R}^{k_1+k_2+n_3}$ such that

$$(S^T \widehat{K} S) q = \theta (S^T \widehat{M} S) q. \tag{9}$$

It follows from the standard Rayleigh-Ritz theory [29, page 213] that θ serves as an approximation to an eigenvalue of (K, M), and the vector formed by $z = L^{-T}Sq$ is the approximation to the corresponding eigenvector.

A summary of the single-level algebraic sub-structuring algorithm described in this section is provided below.

Algorithm: Single-level Algebraic Sub-structuring

Input: A matrix pencil (K, M), where K is symmetric and M is symmetric positive definite;

Output: $\theta_j \in \mathbb{R}$ and $z_j \in \mathbb{R}^n$, $(j = 1, 2, ..., k)$ such that $Kz_j \approx \theta_j M z_j$.

1. Order K and M to be in the form of (3)

2. Perform block factorization $K = LDL^T$;

3. Apply the congruent transformation defined by L^{-1} and L^{-T} to (K, M) to obtain $(\widehat{K}, \widehat{M})$ defined by (6) and (7);

4. Compute a subset of eigenpairs of the sub-structures (K_{11}, M_{11}) and (K_{22}, M_{22}). The eigenvectors of each sub-structure form the columns of S_1 and S_2 respectively;

5. Project the matrix pencil (K, M) into subspace spanned by columns of $Z = L^{-T}S$ where S is defined by (8);

6. Compute k desired eigenpairs (θ_j, q_j) from $(Z^T K Z, Z^T M Z)$, and set $z_j = Zq_j$ $(j = 1, 2, ..., k)$;

A few remarks are in order.

- Note that the most expensive computational task associated with this algorithm is the block factorization $K = LDL^T$ and the congruent transformation of M required for projecting M into the subspace spanned by $Z = L^{-T}S$. These computational tasks must be carried out with great care in order to reduce memory requirements and floating point operations. However, it is beyond the scope of this paper to discuss these important implementation issues.

- Since $k_1 \ll n_1$ and $k_2 \ll n_2$, Step 4 of the algorithm can be carried out by using a shift-invert Lanczos algorithm to obtain a small number of desired eigenpairs from each sub-structure. The cost of this computation is generally small compared to the rest of the computation, especially when this algorithm is extended to a multi-level scheme.

- Similarly, because n_3 is typically much smaller than n_1 and n_2, the dimension of the projected problem (9) is significantly smaller than that of the original problem. Thus, the cost of solving (9) is also small compared to Steps 2 and 3 of the algorithm.

- Decisions must be made on how to select eigenvectors from each sub-structure. The selection should made in such a way that the subspace spanned by the columns of Z retains enough spectral information from (K, M). The process of choosing appropriate eigenvectors from each sub-structure will be referred to as *mode selection*. We will postpone the discussion on this key aspect of the algebraic sub-structuring algorithm until the next section.

The algebraic sub-structuring algorithm presented in this section can be extended in two ways. First, the matrix reordering and partitioning scheme used to create the block structure of (3) can be applied recursively to (K_{11}, M_{11}) and (K_{22}, M_{22}) respectively to produce a multi-level division of (K, M) into smaller sub-matrices. The reduced computational cost associated with finding selected eigenpairs from these even smaller sub-matrices further improves the efficiency of the algorithm. Second, one may replace I_{n_3} in (8) with a subset of eigenvectors of the interface pencil $(\widehat{K}_{33}, \widehat{M}_{33})$. This modification will further reduce the computational cost associated with solving the projected eigenvalue problem (9). A combination of these two extensions yields the AMLS algorithm presented in [22, 7]. However, we will limit the scope of our presentation to a single level sub-structuring algorithm in this paper.

3 Accuracy and Error Estimation

Algebraic sub-structuring allows one to break a large-scale eigenvalue problem into a set of smaller sub-problems that are easier to solve. The algorithm would be less attractive to use if one has to compute all eigenvalues and eigenvectors of each sub-problem. Fortunately, such a calculation is typically not necessary as we will show in this section. In practice, only a small subset of the eigenvectors of (K_{11}, M_{11}) and (K_{22}, M_{22}) are needed to assemble the projection subspace spanned by the columns of the matrix S in (8). To simplify the analysis, we will work with the matrix pencil $(\widehat{K}, \widehat{M})$, where $\widehat{K}$ and $\widehat{M}$ are defined by (6) and (7) respectively. As we noted earlier, $(\widehat{K}, \widehat{M})$ and (K, M) have the same set of eigenvalues. If $\hat{x}$ is an eigenvector of $(\widehat{K}, \widehat{M})$, then $x = L^{-T}\hat{x}$ is an eigenvector of (K, M), where L is the transformation defined in (4).

Let $(\mu_j^{(i)}, v_j^{(i)})$ be the j-th eigenpair of the i-th sub-problem

$$K_{ii}v^{(i)} = \mu^{(i)} M_{ii} v^{(i)},$$

where $v_j^{(i)}$ is M_{ii}-orthonormal, i.e., $(v_j^{(i)})^T M_{ii} v_k^{(i)} = \delta_{j,k}$. To simplify our discussion, we assume that $\mu_j^{(i)}$ has been ordered such that

$$\mu_1^{(i)} \leq \mu_2^{(i)} \leq \ldots \leq \mu_{n_i}^{(i)}. \tag{10}$$

Let us define $V_i = (v_1^{(i)}\ v_2^{(i)}\ \ldots\ v_{n_i}^{(i)})$, $V = \mathrm{diag}(V_1, V_2, I_{n_3})$ and $\Lambda_i = \mathrm{diag}(\mu_1^{(i)}, \mu_2^{(i)}, \ldots, \mu_{n_i}^{(i)})$. An eigenvector of $(\widehat{K}, \widehat{M})$, say $\widehat{x}$, can be expressed as a linear combination of columns of V. That is,

$$\widehat{x} = Vy = \begin{pmatrix} V_1 & & \\ & V_2 & \\ & & I_{n_3} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \tag{11}$$

where $y = (y_1^T, y_2^T, y_3^T)^T$ is an eigenvector of the following generalized eigenvalue problem

$$\begin{pmatrix} \Lambda_1 & & \\ & \Lambda_2 & \\ & & \widehat{K}_{33} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \lambda \begin{pmatrix} I_{n_1} & & G_{13} \\ & I_{n_2} & G_{23} \\ G_{13}^T & G_{13}^T & \widehat{M}_{33} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \tag{12}$$

where $G_{13} = V_1^T \widehat{M}_{13}$ and $G_{23} = V_2^T \widehat{M}_{23}$. Note the matrix pencil defined by (12) can be obtained by applying V^T from the left to $\widehat{K}\widehat{x} = \lambda \widehat{M}\widehat{x}$ and expressing $\widehat{x}$ by $\widehat{x} = Vy$. This

pencil is clearly congruent to the pencils $(\widehat{K}, \widehat{M})$ and (K, M). Thus they share the same set of eigenvalues. We will refer to (12) as a *canonical form* of the generalized eigenvalue problem (2).

If $\widehat{x}$ can be well approximated by a linear combination of the columns of S, as suggested by our description of the the algorithm in the previous section, then the vectors y_1 and y_2 must contain only a few large entries. All other components of y_1 and y_2 are likely to be small and negligible. In this section, we seek to formalize this key concept by developing *a priori* error bounds for the approximations to the smallest eigenvalue of $(\widehat{K}, \widehat{M})$ and the corresponding eigenvector. As we will see below, these bounds can be expressed in terms of the small components of y_1 and y_2.

Suppose $\Lambda_i - \lambda I_{n_i}$ is nonsingular, for $i = 1, 2$. It follows from the first two block rows of the canonical eigenproblem (12) that

$$y_i = (\Lambda_i - \lambda I)^{-1} \lambda G_{i3} y_3, \tag{13}$$

Consequently, we can express the j-th element of y_i by

$$e_j^T y_i = \frac{\lambda}{\mu_j^{(i)} - \lambda} g_j^{(i)} = \frac{1}{\mu_j^{(i)}/\lambda - 1} g_j^{(i)}, \tag{14}$$

where $g_j^{(i)} = e_j^T G_{i3} y_3$. It is easy to see from (14) that, when $|\mu_j^{(i)}/\lambda| \approx 1$, $|e_j^T y_i|$ will be relatively large provided $|g_j^{(i)}|$ is bounded from below. On the other hand, if $\mu_j^{(i)}$ is far away from λ, and if $|g_j^{(i)}|$ is bounded from above, $|e_j^T y_i|$ will be relatively small. Thus, if λ is surrounded by a few eigenvalues of (K_{ii}, M_{ii}), and if S_i contains only the eigenvectors associated with these eigenvalues, one would expect to obtain an accurate approximation to λ by solving the projected problem (9).

To make the above statements more precise, we introduce some additional notation. Let us define

$$\rho_k(\omega) = \left| \frac{\lambda_k}{\omega - \lambda_k} \right|, \tag{15}$$

where λ_k is the k-th eigenvalue of (K, M). If $|g_j^{(i)}| \in [\gamma_1, \gamma_2]$ for some modest sized constants $\gamma_1 < \gamma_2$, then $\rho_k(\mu_j^{(i)})$ provides a reliable measure for $|e_j^T y_i|$.

It is easy to verify that

$$\rho_k(\mu_{j+1}^{(i)}) \leq \rho_k(\mu_j^{(i)}) \quad \text{for } \mu_j^{(i)} > \lambda_k$$

and

$$\rho_k(\mu_j^{(i)}) \leq \rho_k(\mu_{j+1}^{(i)}) \quad \text{for } \mu_j^{(i)} < \lambda_k.$$

These inequalities suggest that $\rho_k(\mu_j^{(i)})$, and therefore $|e_j^T y_i|$, is relatively large when $\mu_j^{(i)}$ is sufficiently close to λ_k.

Let us now focus on the special case in which $k = 1$, i.e., the case associated with the smallest eigenvalue. Because (K_{ii}, M_{ii}) represents the restriction of the pencil $(\widehat{K}, \widehat{M})$ to a subspace, all of its eigenvalues satisfy

$$\lambda_1 < \mu_j^{(i)} \leq \lambda_n.$$

Consequently, the inequality

$$\rho_1(\mu_1^{(i)}) \geq \rho_1(\mu_2^{(i)}) \geq \cdots \geq \rho_1(\mu_{n_1}^{(i)}) \tag{16}$$

holds. Suppose $k_i < n_i$ is the smallest integer such that $\rho_1(\mu_{k_i+1}^{(i)}) \leq \tau$ for some $\tau \ll 1$, then we can assert, under the assumption

$$|g_j^{(i)}| \leq \gamma, \quad \text{for some small constant } \gamma,$$

that $|e_j^T y_i|$ is relatively small for all $j > k_i$. This assertion follows directly from (16) and the observation made in (14). Hence, if our goal is to seek an accurate approximation to λ_1 by projecting $(\widehat{K}, \widehat{M})$ into a subspace $\mathcal{S}$ spanned by the columns of

$$S = \begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \end{array} \begin{array}{c} \begin{array}{ccc} k_1 & k_2 & n_3 \end{array} \\ \begin{pmatrix} S_1 & & \\ & S_2 & \\ & & I_{n_3} \end{pmatrix} \end{array}, \tag{17}$$

it is natural to set S_i to include only the leading k_i columns of V_i.

Given this choice of subspace, it remains to be shown how much accuracy one can expect from the approximate eigenvalue and eigenvector obtained by applying the Rayleigh-Ritz procedure to S. To simplify our discussion, let us assume that λ_1 is simple. Suppose θ_1 is the smallest eigenvalue of the projected problem

$$(S^T \widehat{K} S)q = \theta(S^T \widehat{M} S)q,$$

and q_1 is the corresponding eigenvector. We will now quantify the accuracy of the *Ritz pair* (θ_1, u_1), where $u_1 = Sq_1$, by providing *a priori* error bounds for both $\theta_1 - \lambda_1$ and $\angle_M(\widehat{x}_1, u_1)$ in terms of small elements of y_1 and y_2. Note that $\angle_M(\widehat{x}_1, u_1)$ is the M-inner product induced angle (between $\widehat{x}$ and u_1) defined in Section 1.

To develop these error bounds, we use the following theorem, which is a generalization of a similar theorem associated with a standard symmetric eigenvalue problem [31, 33].

Theorem 1 *Let $K, M \in \mathbb{R}^{n\times n}$ be symmetric matrices and M be positive definite. Suppose the eigenpairs of (K, M), (λ_i, x_i), have been ordered so that*

$$\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n.$$

If (θ_i, u_i), $i = 1, 2, ..., k$, are Ritz pairs associated with a k-dimensional subspace $\mathcal{S}$ ordered so that

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k,$$

then

$$\theta_1 - \lambda_1 \leq (\lambda_n - \lambda_1) \sin^2 \angle_M(x_1, \mathcal{S}), \tag{18}$$

$$\sin \angle_M(u_1, x_1) \leq \sqrt{\frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1}} \sin \angle_M(x_1, \mathcal{S}), \tag{19}$$

where $\angle_M(u_1, x_1)$ denotes the M-angle between the vectors x_1 and u_1, and $\angle_M(x_1, \mathcal{S})$ denotes the M-angle between x_1 and the subspace $\mathcal{S}$.

A proof of Theorem 1 is included in the appendix. The theorem suggests that the accuracy of the desired Ritz pair is determined largely by the M-angle between the exact eigenvector $\widehat{x}_1$ and the subspace $\mathcal{S}$ from which the Ritz pair is extracted. We now focus on seeking a bound for $\sin\angle_M(\widehat{x}_1, \mathcal{S})$. The following theorem, which is a generalization of a similar theorem in [35, page 250], provides a useful characterization for $\sin\angle_M(\widehat{x}_1, \mathcal{S})$.

Theorem 2 *Let x be a vector with $\|x\|_M = 1$ and let $\mathcal{S}$ be a subspace. Then*

$$\sin\angle_M(x, \mathcal{S}) = \min_{w\in\mathcal{S}} \|x - w\|_M.$$

Theorem 2 suggests that we can provide a bound for $\sin\angle_M(\widehat{x}_1, \mathcal{S})$ by measuring the distance between $\widehat{x}_1$ and a particular choice of a vector $w \in \mathcal{S}$ that is "close" to $\widehat{x}_1$ in M-norm.

Our choice of such a vector $w \in \mathcal{S}$ is made as follows. We define $\hat{y}_i$ $(i = 1, 2)$ by

$$e_j^T \widehat{y}_i = \begin{cases} e_j^T y_i & \text{for } \; j \le k_i, \\ 0 & \text{for } \; k_i < j \le n_i, \end{cases} \tag{20}$$

where y_i satisfies

$$\begin{pmatrix} \Lambda_1 & & \\ & \Lambda_2 & \\ & & \widehat{K}_{33} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \lambda_1 \begin{pmatrix} I_{n_1} & & G_{13} \\ & I_{n_2} & G_{23} \\ G_{13}^T & G_{13}^T & \widehat{M}_{33} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}. \tag{21}$$

It is easy to verify that

$$w = \begin{pmatrix} V_1 & & \\ & V_2 & \\ & & I \end{pmatrix} \begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ y_3 \end{pmatrix} \in \mathcal{S} = \text{span}\{S\}.$$

For this particular choice of w, we can easily show that

$$\widehat{x}_1 - w = \begin{pmatrix} V_1 & & \\ & V_2 & \\ & & I \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ 0 \end{pmatrix},$$

where $h_i = y_i - \widehat{y}_i$ for $i = 1, 2$. Consequently, we have

$$\|\widehat{x}_1 - w\|_M^2 = \begin{pmatrix} h_1^T & h_2^T & 0 \end{pmatrix} \begin{pmatrix} I & & G_{13} \\ & I & G_{23} \\ G_{13}^T & G_{23}^T & \widehat{M}_{33} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ 0 \end{pmatrix} = h_1^T h_1 + h_2^T h_2.$$

Hence, we can now conclude that

$$\sin\angle_M(\widehat{x}_1, \mathcal{S}) = \min_{w\in\mathcal{S}} \|\widehat{x}_1 - w\|_M \le \sqrt{h_1^T h_1 + h_2^T h_2}. \tag{22}$$

Note that the vector w is essentially obtained from (11) by truncating components associated with the trailing $n_i - k_i$ elements of y_i. These elements are typically small, and they form the non-zero entries of h_i.

Combining (18) and (19) with (22), we obtain the following result.

Theorem 3 *Let $\widehat{K}$ and $\widehat{M}$ be matrices defined in (6) and (7) respectively. Let $(\lambda_i, \widehat{x}_i)$ $(i = 1, 2, ...n)$ be eigenpairs of the pencil $(\widehat{K}, \widehat{M})$, ordered so that $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$. Let (θ_i, u_i) $(i = 1, 2, ..., k)$ be the Ritz pairs associated with a k-dimensional subspace $\mathcal{S}$ spanned by the columns of S defined in (17), ordered so that $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$. Then*

$$\theta_1 - \lambda_1 \leq (\lambda_n - \lambda_1)(h_1^T h_1 + h_2^T h_2), \tag{23}$$

$$\sin \angle_M(u_1, x_1) \leq \sqrt{\frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1}} \sqrt{h_1^T h_1 + h_2^T h_2}, \tag{24}$$

where $h_i = y_i - \widehat{y}_i$, and y_i, $\widehat{y}_i$ $(i = 1, 2)$ are defined by (21) and (20) respectively.

Theorem 3 indicates that the accuracy of (θ_1, u_1) is proportional to the size of $h_1^T h_1 + h_2^T h_2$, a quantity that provides a cumulative measure of the "truncated" components in (11). Similar *a priori* error estimates can be made for other Ritz pairs by utilizing a generalization of Theorems 4.5 and 4.6 in [31, pp 135-136] which are developed for standard eigenvalue problems. However, to keep our presentation concise, we will not pursue this type of error estimate in this paper.

To end this section, we provide an estimate for $h_1^T h_1 + h_2^T h_2$ that is independent of the number of non-zero elements in h_1 and h_2. Note that the nonzero elements of h_i are those elements of y_i associated with

$$\rho_1(\mu_j^{(i)}) < \tau < 1.$$

If $|g_j^{(i)}| \leq \gamma$ for some moderate sized constant γ, then it follows from (14) that each individual element of h_i is either zero or tiny. Moreover, since $\rho_1(\mu_j^{(i)})$ decreases rapidly as $\mu_j^{(i)}$ increases, we can establish a bound for $h_i^T h_i$ in terms of τ under some mild conditions.

To simplify our notation, we will drop the superscript of $\mu_j^{(i)}$ in the following. Suppose the eigenvalues μ_j of (K_{ii}, M_{ii}) are distinct and

$$\min_{j \geq k_i} (\mu_{j+1} - \mu_j) \geq \delta,$$

for some constant $\delta > 0$. Then it is easy to show that

$$\begin{aligned}
h_i^T h_i &= \sum_{j=k_i+1}^{n_i} (e_j^T h_i)(e_j^T h_i) \\
&= \sum_{j=k_i+1}^{n_i} \rho_1^2(\mu_j)(e_j^T G_{i3} y_3)^2 \\
&\leq \left[\sum_{j=k_i+1}^{n_i} \rho_1^2(\mu_j) \right] \gamma^2 \\
&\leq \frac{\gamma^2}{\delta} \int_{\mu_{k_i+1}}^{\mu_{n_i}} \rho_1^2(\omega) d\omega \\
&= \frac{(\lambda_1 \gamma)^2}{\delta} \left(\frac{1}{\mu_{k_i+1} - \lambda_1} - \frac{1}{\mu_{n_i} - \lambda_1} \right).
\end{aligned}$$

Note that μ_{n_i} is typically much larger than λ_1. Thus the term $1/(\mu_{n_i} - \lambda_1)$ in the above

expression is negligible. Hence,

$$\begin{aligned} h_i^T h_i &\leq \frac{(\lambda_1\gamma)^2}{\delta}\left(\frac{1}{\mu_{k_i+1}-\lambda_1}\right) \\ &= \frac{\lambda_1\gamma^2}{\delta}\rho_1(\mu_{k_i+1}) \\ &\leq \frac{\lambda_1\gamma^2}{\delta}\tau. \end{aligned} \tag{25}$$

Combining (25) with inequalities (23) and (24) stated in Theorem 1, we obtain

$$\frac{\theta_1-\lambda_1}{\lambda_1} \leq (\lambda_n-\lambda_1)(2\alpha\tau), \tag{26}$$

$$\sin\angle_M(\widehat{x}_1,u_1) \leq \sqrt{\lambda_1\left(\frac{\lambda_n-\lambda_1}{\lambda_2-\lambda_1}\right)\sqrt{2\alpha\tau}}, \tag{27}$$

where $\alpha=\gamma^2/\delta$.

We should mention that the presence of multiple (or tightly clustered) eigenvalues of (K_{ii},M_{ii}) does not alter the qualitative measure of the bounds established in (26) and (27). In that case, we can simply replace the definition of δ with the minimum distances between two adjacent eigenvalue clusters and multiply the bounds by the largest multiplicity of the eigenvalues of (K_{ii},M_{ii}).

We should also emphasize that (26) and (27) merely provide a qualitative estimate of the error in the Ritz pair (θ_1,u_1) in terms of the threshold τ that may be used as a heuristic in practice to determine which spectral components of a substructure should be included in the subspace $\mathcal{S}$ defined in (17). It is clear from these inequalities that a smaller τ, which typically corresponds to a selection of more spectral components from each substructure, leads to a more accurate Ritz pair (θ_1,u_1).

4 Null Space Deflation

The LDL^T factorization performed in Step 2 of the algebraic sub-structuring algorithm shown in section 2 relies on the assumption that K is non-singular. When K is singular, one may choose to work with the shifted eigenvalue problem

$$(K-\sigma M)x=(\lambda-\sigma)Mx,$$

where σ is a non-zero shift, if the null space of (K,M) cannot be easily identified.

However, if the null space of (K,M) has a special structure and a basis of the null space can be constructed easily, it may be advantageous to deflate this null space in the sub-structure calculation.

In this section, we discuss how to modify the sub-structuring algorithm presented earlier to handle a special case in which K is singular and the null space of K can be easily identified. This special case arises when we apply a special hierarchical vector finite element discretization scheme [36] to the homogeneous vector wave equation

$$\begin{aligned} \nabla\times(\nabla\times\mathbf{E})-\lambda\mathbf{E} &= 0 \quad \text{in } \Omega, \\ \mathbf{n}\times\mathbf{E} &= 0 \quad \text{on electric boundary } \Gamma_E, \\ \mathbf{n}\times(\nabla\times\mathbf{E}) &= 0 \quad \text{on magnetic boundary } \Gamma_B, \end{aligned} \tag{28}$$

where $\mathbf{E}$ is a vector function representing the electric field, and $\mathbf{n}$ is the unit normal vector. Equation (28) is derived from the Maxwell equation by eliminating the magnetic intensity variables. This equation plays a fundamental role in the simulation of the electromagnetic field associated with next generation accelerator structures [24].

Once (28) is discretized, the stiffness matrix K corresponding to the curl-curl operator often contains many zero rows and columns. If we reorder K to move all the non-zero rows and columns into the leading block of the matrix and apply the same permutation to M, the generalized eigenvalue problem (2) can be partitioned as

$$\begin{pmatrix} K_{11} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{29}$$

The hierarchical vector finite element discretization scheme developed in [36] ensures that the non-zero block K_{11} is symmetric positive definite. The dimension of the (2,2) blocks of K and M, which we will denote by m, is typically much smaller than that of the $(1,1)$ block. But it can be as large as $n/4$, where n is the dimension of K. The null space of K does not contain any physically interesting information. It is purely an artifact of the discretization. We are only interested in the non-zero eigenvalues and the corresponding eigenvectors.

It is easy to show that the non-zero eigenvalues of (K, M) can be obtained by solving the reduced problem

$$K_{11}x_1 = \lambda \widehat{M}_{11}x_1, \tag{30}$$

where $\widehat{M}_{11} = M_{11} - M_{12}M_{22}^{-1}M_{12}^T$. The x_2 component of the eigenvector associated with (K, M) can be recovered by

$$x_2 = M_{22}^{-1}M_{12}^T x_1. \tag{31}$$

The null space of K is automatically deflated in such a scheme.

However, because $\widehat{M}_{11}$ is generally dense, we cannot apply algebraic sub-structuring to $(K_{11}, \widehat{M}_{11})$. Instead, we choose to work with K and M directly. If we simply apply the nested dissection ordering to $K + M$ to obtain the block structure shown in (3), both K_{11} and K_{22} may contain zero rows and columns.

Since we cannot form $K_{i3}^T K_{ii}^{-1}$ when K_{ii} is singular, we replace K_{ii}^{-1} with the pseudo-inverse of K_{ii} in (4). If we reorder K by moving all nonzero rows and columns of K_{ii} to the leading portion of this submatrix, i.e.,

$$K_{ii} = \begin{pmatrix} A_i & 0 \\ 0 & 0 \end{pmatrix},$$

where A_i is non-singular, then the pseudo-inverse of the reordered K_{ii} is simply

$$K_{ii}^{\dagger} = \begin{pmatrix} A_i^{-1} & 0 \\ 0 & 0 \end{pmatrix},$$

Applying the congruent transformation defined by

$$L^{-1} = \begin{pmatrix} I_{n_1} & & \\ & I_{n_2} & \\ -K_{13}^T K_{11}^{\dagger} & -K_{23}^T K_{22}^{\dagger} & I_{n_3} \end{pmatrix}$$

to the reordered K yields a block diagonal matrix $\widehat{K}$ in the form of (6). The diagonal blocks of $\widehat{M} = L^{-1}ML^{-T}$ can be partitioned conformally to give

$$M_{ii} = \begin{pmatrix} B_i & C_i \\ C_i & D_i \end{pmatrix}.$$

The nonzero eigenvalues and corresponding eigenvectors of the i-th subproblem

$$\begin{pmatrix} A_i & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} p_i \\ q_i \end{pmatrix} = \mu^{(i)} \begin{pmatrix} B_i & C_i \\ C_i^T & D_i \end{pmatrix} \begin{pmatrix} p_i \\ q_i \end{pmatrix}$$

can be computed by solving the following reduced problem

$$A_i p_i = \mu^{(i)} (B_i - C_i D_i^{-1} C_i^T) p_i. \tag{32}$$

When a shift-invert Lanczos algorithm with a zero shift is applied to (32), one does not need to form the Schur complement $B_i - C_i D_i^{-1} C_i^T$ explicitly. If A_i and D_i can be easily factored by a sparse direct method, then the matrix vector operations $w \leftarrow A_i^{-1} v$ and $w \leftarrow (B_i - C_i D_i^{-1} C_i^T) v$, which are required at each step of the Lanczos algorithm, can be carried out efficiently with a few sparse matrix vector multiplications and sparse triangular solves.

Deflation may also be necessary for computing the non-zero eigenvalues and the corresponding eigenvectors of the projected problem (9) when the null vectors of (K_{ii}, M_{ii}) are included in S_i. Suppose $\widetilde{K} = S^T K S$ and $\widetilde{M} = S^T M S$. If the null space of (K_{ii}, M_{ii}), which can be easily constructed in this case, is included in S_i, we can permute the rows and columns of $(\widetilde{K}, \widetilde{M})$ to obtain the projected eigenvalue problem

$$\begin{pmatrix} \widetilde{K}_{11} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \theta \begin{pmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ \widetilde{M}_{12}^T & \widetilde{M}_{22} \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix}, \tag{33}$$

where $\widetilde{K}_{11}$ contains only the non-zero rows and columns of $\widetilde{K}$.

Again, the non-zero eigenvalues of (33) can be computed more efficiently by working with the reduced problem

$$\widetilde{K}_{11} p = \theta (\widetilde{M}_{11} - \widetilde{M}_{12} \widetilde{M}_{22}^{-1} \widetilde{M}_{12}^T) p.$$

The q-component of the eigenvector can be recovered from $q = \widetilde{M}_{22}^{-1} \widetilde{M}_{12}^T p$.

5 Numerical Experiments

We present a few numerical examples in this section to illustrate the effectiveness of the single-level algebraic sub-structuring algorithm presented in Section 2. These examples also confirm the analysis carried out in Section 3. All experiments are performed in MATLAB. The desired eigenpairs of all pencils are computed by using the MATLAB `eigs()` function. For illustration purposes, we computed more eigenvalues and eigenvectors of each subproblem than we actually need in the following experiments. In practice, one would only need to compute a selected number eigenpairs of (K_{ii}, M_{ii}) incrementally.

5.1 Example 1 - BCS structural dynamics

The matrices used in this example, BCSSTK09 and BCSSTM09, are part of the Harwell-Boeing BCS sparse matrix collection [14]. These matrices originated from a dynamic analysis of a clamped plate. The dimensions of these matrices are $n = 1083$. The non-zero pattern of the stiffness matrix K is shown in Figure 1. The pattern on the left corresponds to the original K. The one on the right corresponds to a reordered K obtained after the METIS [23] software is used to dissect the matrix into two main sub-structures coupled by a small

separator (interface block). The two sub-structures of the reordered K are nearly identical. The dimension of each sub-structure is $n_1 = n_2 = 513$. The separator contains only 57 rows and columns. The mass matrix M is diagonal in this example. Applying the same reordering to M does not change its structure.

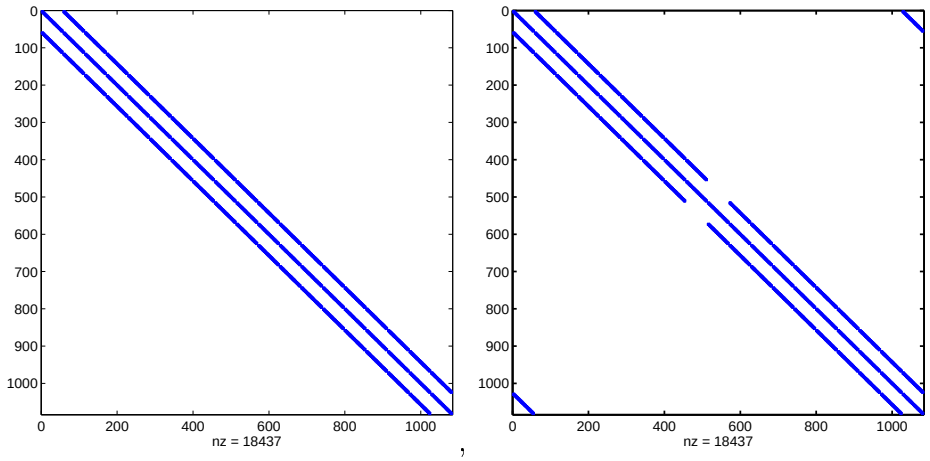

Figure 1: The non-zero pattern of $BCSSTK09$ before and after it is ordered by METIS.

The spectra of the original matrix pencil (K, M) and the sub-structure pencils (K_{ii}, M_{ii}) $(i = 1, 2)$ are shown in Figure 2. There is a large gap between the 361st and the 362nd eigenvalues of (K, M). Similar gaps are present in the spectra of (K_{ii}, M_{ii}). In this example,

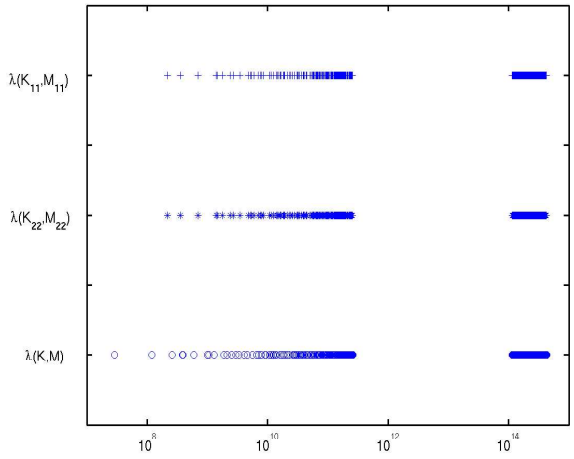

Figure 2: The spectra of the pencils (K_{11}, M_{11}), (K_{22}, M_{22}) and (K, M) associated with the BCS example.

the eigenvalues of interest are the ones at the left end of the spectrum. Naturally, we would select the eigenvectors associated with the smallest eigenvalues of (K_{ii}, M_{ii}) to construct the subspace (8) required in Step 5 of the single-level algebraic sub-structuring algorithm.

To determine how many eigenvectors of (K_{ii}, M_{ii}) we should include in the subspace represented by (8), we examine the ρ-factor defined in (15). It follows from the discussion in Section 3 that one may develop a selection scheme by setting a threshold value τ for ρ_1, i.e.,

one can choose sub-structure modes that satisfy

$$\rho_1(\mu_j^{(i)}) > \tau,$$

for some small τ. However, since the computation of ρ_1 requires the knowledge of exact λ_1 which we do not have in advance, a more practical scheme is perhaps to compute an approximate ρ-factor by replacing λ_1 in (15) with an approximate eigenvalue σ.

We use $\sigma = \min(\mu_1^{(1)}, \mu_1^{(2)})/2$ in all of our experiments, and define

$$\widehat{\rho}_1(\omega) = \left| \frac{\sigma}{\omega - \sigma} \right|. \tag{34}$$

In Figure 3, we plot both $\rho_1(\mu_j^{(i)})$ and $\widehat{\rho}_1(\mu_j^{(i)})$, for $i = 1, 2$. The figure clearly shows that there is essentially no qualitative difference between $\rho_1(\mu_j^{(i)})$ and $\widehat{\rho}_1(\mu_j^{(i)})$. Both decrease rapidly as $\mu_j^{(i)}$ increases. There is a clear gap between $\rho_1(\mu_{171}^{(i)})$ and $\rho_1(\mu_{172}^{(i)})$. A similar gap is observed between $\widehat{\rho}_1(\mu_{171}^{(i)})$ and $\widehat{\rho}_1(\mu_{172}^{(i)})$. These gaps reflect the gaps observed in the spectrum of (K_{ii}, M_{ii}).

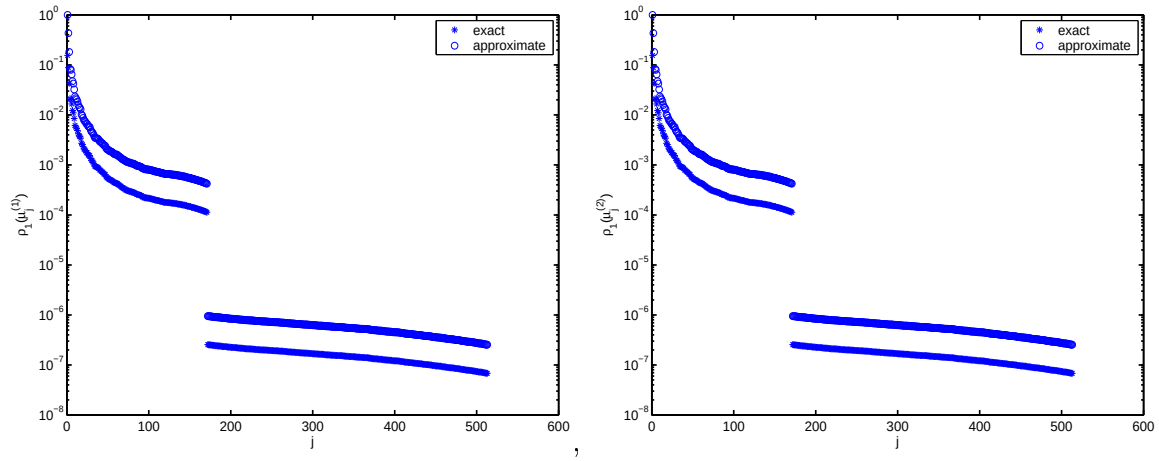

Figure 3: The exact (marked by asterisks) and the approximate (marked by circles) ρ-factors associated with both sub-structures of the BCS problem. The exact ρ-factor is defined by (15), and the approximate ρ-factor is defined by (34).

Several choices of τ values (listed in Table 1) have been tried. The analysis performed in Section 3 indicates that the smaller the value of τ the more accurate the smallest Ritz pair should be. This prediction is confirmed in Figure 4 where we plot the relative errors of the smallest 50 Ritz values extracted from three subspaces constructed by using these different choices of τ values. Notice that with the choice of $\tau = 10^{-4}$, which corresponds to selecting the leading 171 eigenvectors from each sub-structure to form the matrix S_i required in (17), θ_1 exhibits roughly 10 digits of accuracy.

Even though our error estimates presented in Section 3 is targeted only at (θ_1, u_1), Figure 4 shows that the improvement in the accuracy of other Ritz values is also proportional to the decrease of τ.

In this example, the least upper bound for the elements of $g^{(i)}$ used in (14) is roughly $\gamma = 0.28$. Hence, $\rho_1(\mu_j^{(i)})$ provides a reliable upper bound for the magnitude of $e_j^T y_i$ ($i =$

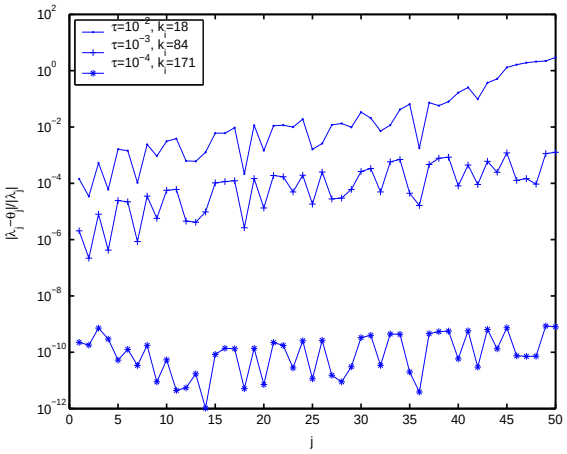

Figure 4: The relative error of the smallest 50 Ritz values extracted from three subspaces constructed by using different choices of the ρ-factor thresholds (τ values) for the BCS problem.

$1, 2$), where $(y_1^T, y_2^T, y_3^T)^T$ is the eigenvector associated with the smallest eigenvalue (λ_1) of the canonical eigenvalue problem (12).

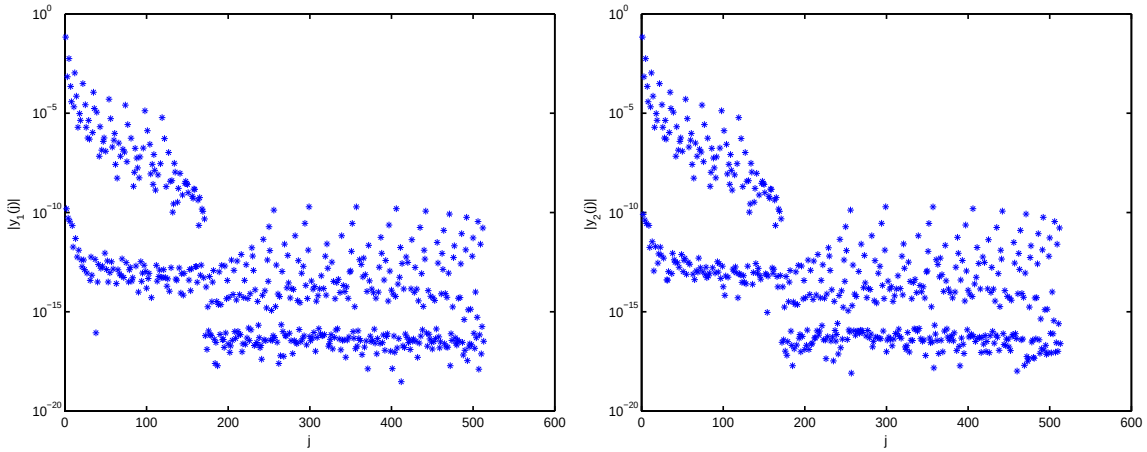

Figure 5: The magnitude of $e_j^T y_1$ (left) and $e_j^T y_2$ (right), where $(y_1^T, y_2^T, y_3^T)^T$ is the eigenvector corresponding to the smallest eigenvalue of the canonical problem (12) associated with the BCS example.

Judging from the small magnitude of $\rho_1(\mu_j^{(i)})$ for $j > 171$, which is less than 10^{-6}, we predict the magnitude of $e_j^T y_i$, $i = 1, 2$, to be tiny for $j > 171$. This is indeed the case as is demonstrated in Figure 5 where we plot $|e_j^T y_i|$ for $i = 1, 2$. We observe that $|e_j^T y_i| < 2 \times 10^{-10}$ for all $j > k_i = 171$ and $i = 1, 2$. This observation, when used in conjunction with Theorem 3, provides a clear explanation for the high accuracy of θ_1 displayed in Figure 4.

Table 1 further illustrates the connections between the mode selection threshold τ, the number of modes selected from each substructure (k_i), the relative accuracy of θ_1 and the error estimates established in Theorem 3. Note that the relative error bound listed in the last column of Table 1, which is calculated directly from the right hand side of (23), tends to be

somewhat pessimistic. However, it does provide a qualitative estimate for the relative accuracy of θ_1.

τ	k_i	$(\theta_1 - \lambda_1)/\lambda_1$	relative error bound
10^{-2}	18	1.4×10^{-4}	3.4×10^{0}
10^{-3}	84	2.0×10^{-6}	6.4×10^{-3}
10^{-4}	171	1.2×10^{-12}	4.2×10^{-12}

Table 1: The effect of τ on the number of selected modes associated with the BCS problem, the relative accuracy of the smallest Ritz value and the relative error bound defined by (23).

It is interesting to see from Figure 5 that among the first 171 elements of both y_1 and y_2, many have magnitudes less than 10^{-10}. This observation suggests that one may potentially reduce the dimension of the subspace (8) by excluding eigenvectors of (K_{ii}, M_{ii}) that are associated with these small entries from S_i. We will pursue this idea further in a follow up paper on mode selection strategies.

We will end this example by pointing out that the large gap between the leading 361 eigenvalues of (K, M) and the rest of the spectrum is a highly favorable feature of this problem. This gap, which also manifests itself in the ρ-factor plots displayed in Figure 3, allows an algebraic sub-structuring algorithm to easily construct a subspace that contains accurate approximations to the leading 361 eigenvalues of (K, M). Figure 6 shows that by setting $k_i = 171$, the leading 361 Ritz values extracted from the subspace $\mathcal{S}$ spanned by columns of (17) all have at least 7 digits of accuracy.

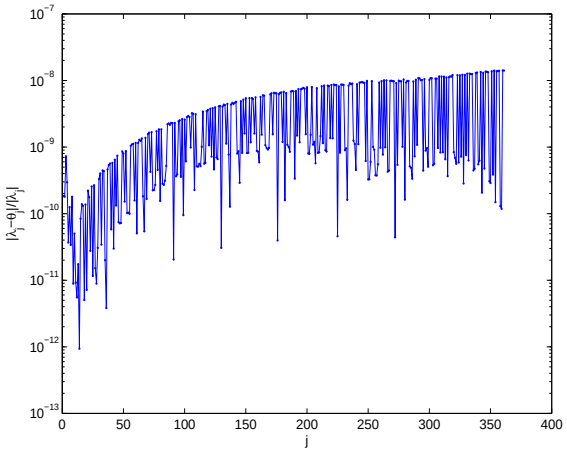

Figure 6: The relative error of the smallest 361 approximate eigenvalues associated with the BCS problem.

5.2 Example 2 - Disk brake squeal prediction

In this example, we consider a generalized eigenvalue problem $Kx = \lambda Mx$ arising from the simulation of disk brake squeal [27]. The matrices K and M are generated by a finite element discretization of a disk brake model. The dimensions of these matrices are $n = 3954$. The mesh used in the finite element discretization has been ordered to yield a stiffness matrix K

that has the sparsity pattern shown in Figure 7. The leading diagonal block forms the first sub-structure. Its dimension is $n_1 = 2853$. The second diagonal block, which is much smaller in size but denser in terms of the non-zero pattern, forms the second sub-structure. Its dimension is $n_2 = 975$. These two sub-structures are connected by a separator that contains 126 rows and columns. The mass matrix M is diagonal in this example.

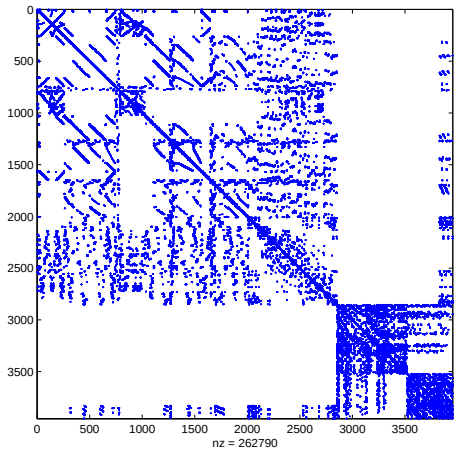

Figure 7: The non-zero pattern of the stiffness matrix K associated with the disk brake structure.

We will illustrate that the algorithm presented in Section 2 works equally well on the sub-structures produced directly from the finite element mesh partition.

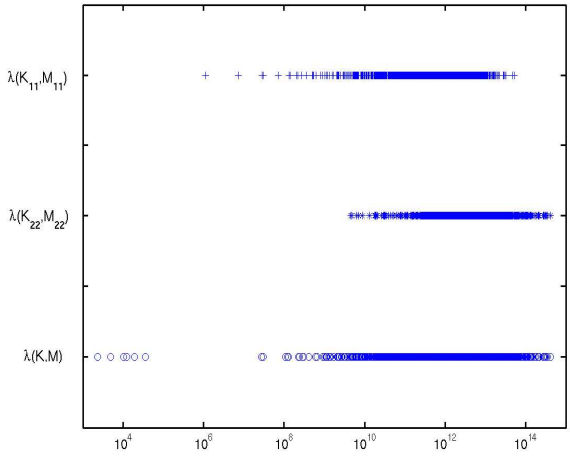

Figure 8: The spectra of the pencils (K, M), (K_{11}, M_{11}) and (K_{22}, M_{22}) associated with the brake structure.

The spectra of (K, M) and (K_{ii}, M_{ii}), $i = 1, 2$, are shown in Figure 8. We are interested in the smallest eigenvalues of (K, M). The smallest 6 eigenvalues of (K, M), which are considerably smaller than the largest eigenvalue, correspond to rigid body motions. These rigid body motions are not deflated in advance in the following calculation.

We observe from the $\widehat{\rho}$-factor plot in Figure 9 that $\widehat{\rho}_1(\mu_j^{(1)}) > \widehat{\rho}_2(\mu_j^{(2)})$, for small j's. This observation suggests that eigenvectors associated with the smallest eigenvalues of (K_{11}, M_{11})

may play a more important role than those of (K_{22}, M_{22}) for constructing the subspace $\mathcal{S}$ defined in (17). This speculation is confirmed in Figure 10 where we plot the magnitude of

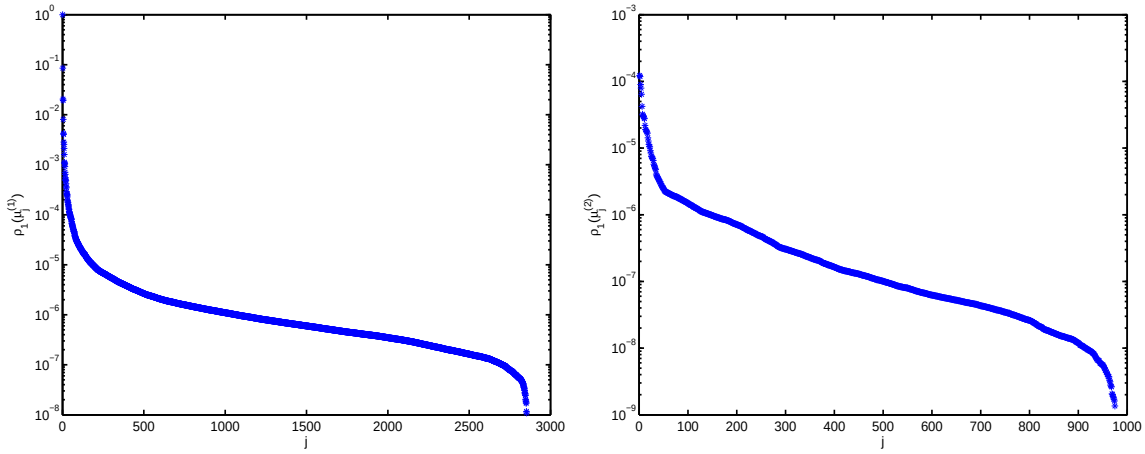

Figure 9: The approximate ρ-factors associated with each sub-structure of the disk brake structure.

each element of y_i, $i = 1, 2$. The first few elements of y_1 are much larger than those of y_2 in magnitude. But $|e_j^T y_1|$ eventually decreases rapidly as j increases.

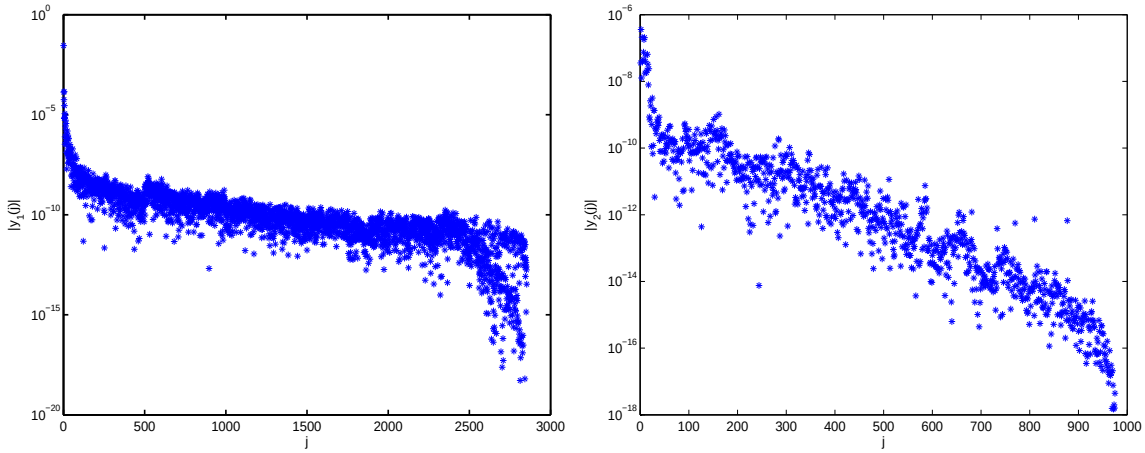

Figure 10: The magnitude of $e_j^T y_1$ (left) and $e_j^T y_2$ (right), where $(y_1^T, y_2^T, y_3^T)^T$ is the eigenvector corresponding to the smallest eigenvalue of the canonical problem (12) associated with the disk brake structure.

We experimented with using three different choices of τ values (listed in Table 2) for selecting sub-structure modes that satisfy $\widehat{\rho}_1(\mu_j^{(1)}) > \tau$. Table 2 shows that, with $\tau = 10^{-3}$, only 14 eigenvectors are selected from the first sub-structure, and none are selected from the second structure. The smallest Ritz value obtained from the subspace (17) constructed by these eigenvectors has roughly 6 digits of accuracy. Although the smallest 6 Ritz values and the corresponding Ritz vectors are not physically interesting, our experiment demonstrates that they can be computed accurately by our sub-structuring algorithm using a sub-space containing only 140 basis vectors.

Figure 11 shows that decreasing the value of τ does not further improve the accuracy of the Ritz values associated with the rigid body motion. However, a smaller τ value does lead to significant improvement in the accuracy of the Ritz values associated with non-rigid vibrations. When τ is set to 10^{-5}, which corresponds to selecting 185 modes from the first sub-structure and 22 modes from the second sub-structure, the relative errors of the smallest 50 Ritz values all have at least 3 digits of accuracy.

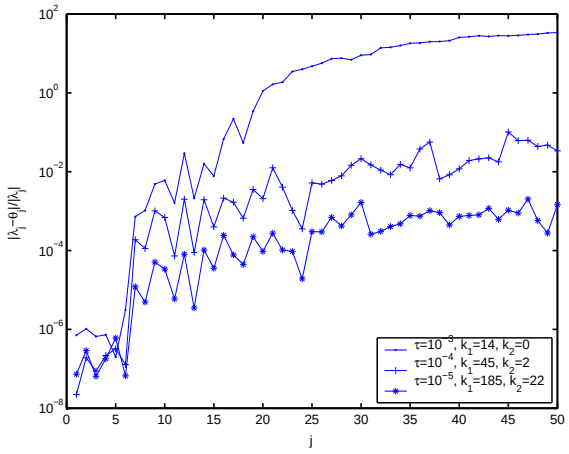

Figure 11: The relative error of the smallest 50 Ritz values extracted from three subspaces constructed by using different choices of the ρ-factor thresholds (τ values) for the disk brake problem.

τ	k_1	k_2	$(\theta_1 - \lambda_1)/\lambda_1$	relative error bound
10^{-3}	14	0	7.2×10^{-7}	4.2×10^{-1}
10^{-4}	45	2	2.2×10^{-8}	1.4×10^{-3}
10^{-5}	185	22	7.3×10^{-8}	2.3×10^{-5}

Table 2: The effect of τ on the number of selected modes associated with the disk brake problem, the relative accuracy of the smallest Ritz value and the relative error bound defined by (23).

5.3 Example 3 - Short traveling wave accelerating structure

We show in this example that algebraic sub-structuring can be used to compute approximate cavity resonance frequencies and the electromagnetic field associated with a small accelerator structure. The matrix pencil used in this example is obtained from a finite element model of a five-cell traveling wave accelerating structure. The three dimensional geometry of the model is shown in Figure 12. The model contains three cavity cells and two couplers. The dimension of the pencil (K, M) is $n = 1898$. The stiffness matrix K has 336 zero rows and columns. As we mentioned in Section 4, these zero rows and columns are produced by a particular hierarchical vector finite element discretization scheme. In order to deflate the null space of (K, M) associated with these zero rows and columns, which are of no physical interest, we perform the following two-stage matrix reordering:

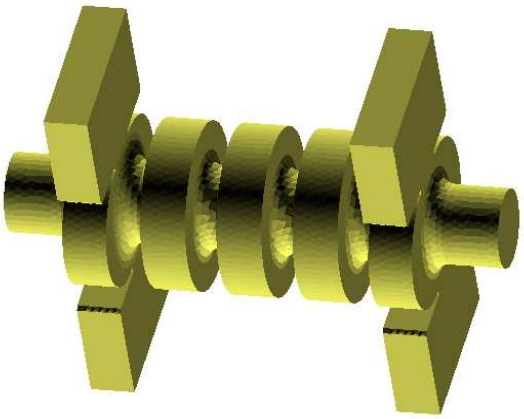

Figure 12: The finite element model corresponding to a 5-cell traveling wave accelerating structure.

- A single-level dissection is applied to $K + M$ first using the METIS [23] software. The dissection produces two sub-structures of sizes $n_1 = 995$ and $n_2 = 887$ respectively. These sub-structures are connected by a small separator (an interface block) which contains only 16 rows and columns. The K_{11} block of the permuted K contains 175 zero rows and columns, the K_{22} block contains 157 zero rows and columns, and K_{33} block contains 6 zero rows and columns.
- The non-zero rows and columns of K_{11}, K_{22} and K_{33} are permuted to the leading blocks of these submatrices. The matrix M is permuted accordingly.

The non-zero patterns of the permuted K and M are displayed in Figure 13.

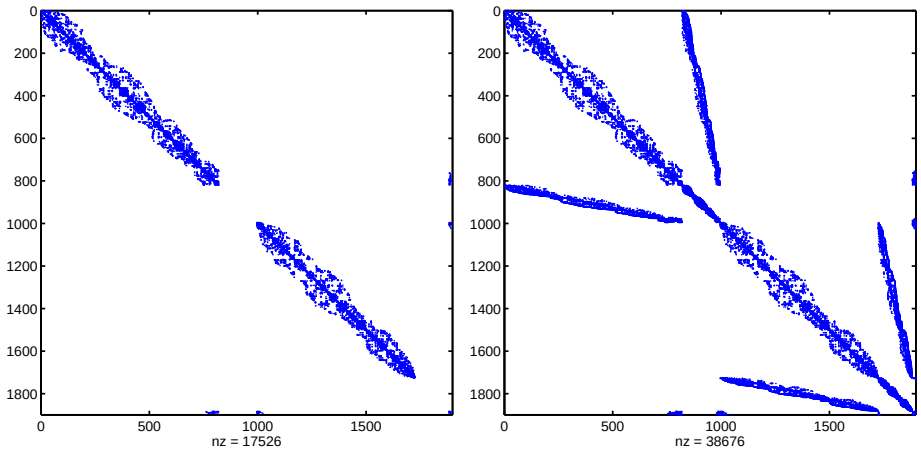

Figure 13: The non-zero pattern of the permuted stiffness matrix K (left) and the mass matrix M (right) associated with the traveling wave accelerating structure.

The distribution of the non-zero eigenvalues of (K, M) is shown in Figure 14. We are interested in the smallest non-zero eigenvalues which appear to be relatively well separated from the large end of the spectrum. In addition to the spectrum of (K, M), we also plot the

spectra of (K_{ii}, M_{ii}) $(i = 1, 2)$ in the Figure 14. Notice that the spectra of both sub-structures show a similar distribution pattern to that of (K, M).

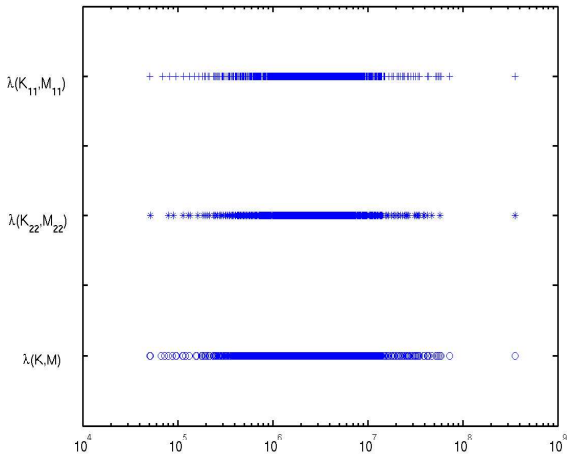

Figure 14: The spectra of the pencils (K_{11}, M_{11}), (K_{22}, M_{22}) and (K, M) associated with the traveling wave accelerating structure.

We plot the $\widehat{\rho}$-factors associated with smallest eigenvalue of the deflated problem in Figure 15. We observe that the $\widehat{\rho}$-factors associated with this example decrease at a somewhat slower rate. Three different choices of τ values were used as the thresholds $(\tau = 0.1, 0.05, 0.01)$

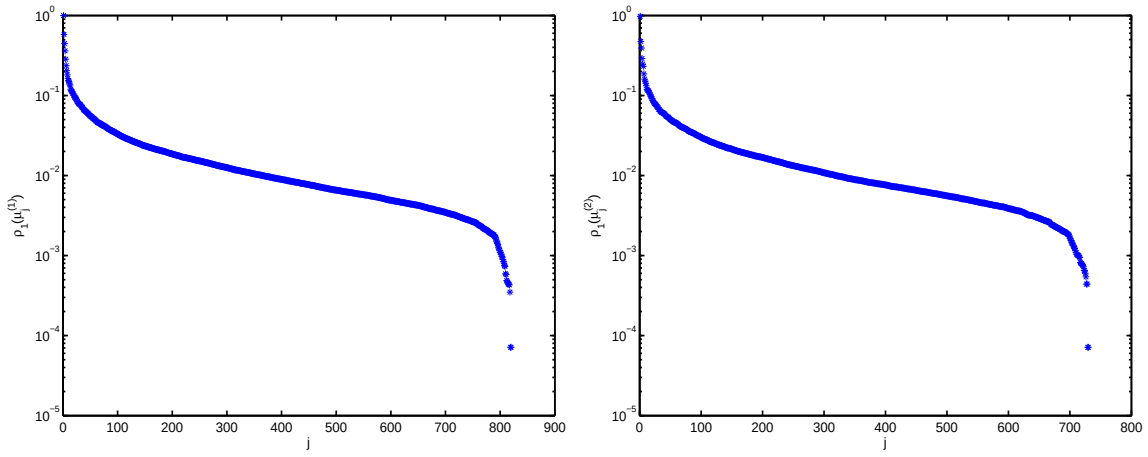

Figure 15: The approximate ρ-factors associated with each sub-structure of the traveling wave accelerating structure.

for selecting sub-structure modes. The relative accuracy of the 50 smallest non-zero Ritz values extracted from the subspaces constructed with these choices of τ values is displayed in Figure 16.

We observe that with $\tau = 0.1$, θ_1 has roughly four digits of accuracy, which is quite sufficient for this particular discretized model. If we decrease τ down to 0.01, most of the smallest 50 non-zero Ritz values have at least 8 digits of accuracy.

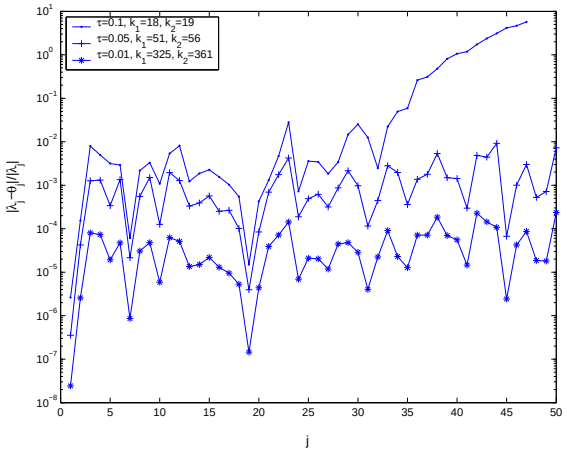

Figure 16: The relative error of the smallest 50 Ritz values extracted from three subspaces constructed by using different choices of the ρ-factor thresholds (τ values) for the traveling wave accelerating problem.

The least upper bound for $g_j^{(i)}$ used in (14) is $\gamma = 0.02$. Thus the ρ-factor gives an overestimate of $|e_j^T y_i|$ in this case. In Figure 17, we plot $|e_j^T y_1|$ and $|e_j^T y_2|$, where $(y_1^T, y_2^T, y_3^T)^T$ is the eigenvector associated with the smallest non-zero eigenvalue of (12). For simplicity, we excluded the values of $|e_j^T y_1|$ and $|e_j^T y_2|$ corresponding to the null space of (K_{11}, M_{11}) and (K_{22}, M_{22}), which have been deflated from our calculations (See Section 4). We observe that $|e_j^T y_i|$ is much smaller compared to $\widehat{\rho}_1(\mu_j^{(i)})$, and it decays much faster than the $\widehat{\rho}$-factor also.

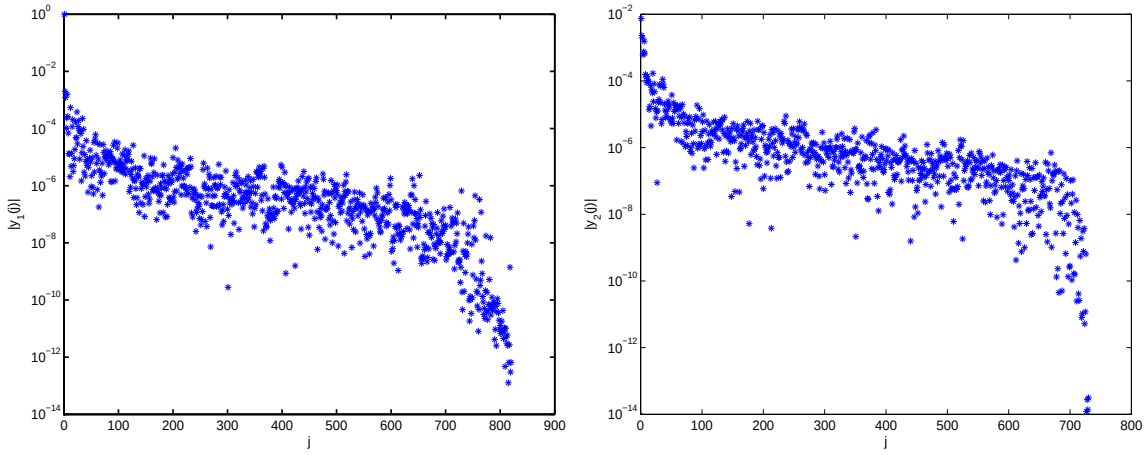

Figure 17: The magnitude of $e_j^T y_1$ (left) and $e_j^T y_2$ (right), where $(y_1^T, y_2^T, y_3^T)^T$ is the eigenvector corresponding to the smallest eigenvalue of the canonical problem (12) associated with the traveling wave accelerating structure.

We conclude this example by listing the mode selection threshold τ, the number of modes selected from each sub-structure (k_i), the relative accuracy of θ_1 and the error estimate computed directly from the right-hand side of (23) in Table 3.

τ	k_1	k_2	$(\theta_1 - \lambda_1)/\lambda_1$	relative error bound
0.1	18	19	1.4×10^{-4}	1.7×10^{-3}
0.05	51	56	1.2×10^{-5}	2.6×10^{-4}
0.01	325	361	2.4×10^{-8}	2.5×10^{-6}

Table 3: The effect of τ on the number of selected modes associated with the traveling wave accelerating structure, the relative accuracy of the smallest Ritz value and the relative error bound defined by (23).

6 Concluding Remarks

A purely algebraic analysis of a single-level sub-structuring algorithm for large-scale eigenvalue calculation is developed in this paper. By applying a sequence of special congruent transformations to (K, M), we turn the original generalized eigenvalue (2) into a canonical problem (12) with a simpler structure. We observed that the desired eigenvector y of the canonical problem (12) often contains only a few large entries. The magnitude of these entries ultimately determines which eigenvectors (modes) of each sub-structure should be included in the subspace (8) from which approximations to the eigenpairs of (K, M) are extracted. All other sub-structure modes can essentially be truncated from (11) without sacrificing the required level of accuracy in our approximation. We provided an explicit *a priori* error estimate for the smallest Ritz pair in terms of the small components of y that are truncated from (11). We also suggested a practical way to estimate the magnitude of each component of y by exploiting its relationship with the "ρ-factor" defined in (15). This estimation leads to a practical way for selecting sub-structure modes by specifying a threshold value τ for the ρ-factor. We showed that the accuracy of smallest Ritz pair is proportional to the size of τ under some mild conditions. A number of numerical examples are provided to confirm our theoretical analysis. Moreover, we demonstrated that an algebraic sub-structuring algorithm can be an effective tool for computing cavity resonance frequencies and the electromagnetic field generated by a linear accelerator structure.

Our analysis of a simple algebraic sub-structuring algorithm can be extended to a multi-level setting. Our error estimate can be made for non-extreme Ritz pairs as well. These topics will be pursued in our future research. Another interesting area that would require further research is the development of a better strategy for selecting sub-structuring modes.

Our presentation has focused on the theoratical aspect of the algebraic sub-structuring algorithm. In a follow up paper, we will discuss the implementation detail and compare the computational cost of a multi-level algebraic sub-structuring algorithm with other methods for large-scale eigenvalue computation.

Appendix

For completeness purpose, we provide a proof for Theorem 1 in this section. Theorem 1 is an extension of Theorem 2.1 in [33], which we will restate below.

Theorem 4 *Let A be a symmetric matrix with eigenpairs (λ_i, x_i), ordered so that*

$$\lambda_1 < \lambda_2 \leq \ldots \leq \lambda_{n-1} < \lambda_n.$$

Suppose (θ_i, u_i) are Ritz pairs formed by applying the Rayleigh-Ritz procedure to a subspace $\mathcal{V}$

spanned by $V \in \mathbb{R}^{n\times k}$. *If these Ritz pairs are ordered so that*

$$\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k,$$

then

$$\theta_1 - \lambda_1 \leq (\lambda_n - \lambda_1)\sin\angle(x_1, \mathcal{V}), \tag{35}$$

$$\sin^2\angle(u_1, x_1) \leq \frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1}\sin\angle(x_1, \mathcal{V}). \tag{36}$$

Proof of Theorem 1: The Ritz values θ_i $(i = 1, 2, ..., k)$ are the eigenvalues of the projected problem

$$(S^T K S)q = \theta(S^T M S)q \tag{37}$$

where the columns of S forms a basis for $\mathcal{S}$. Let q_i be the i-th generalized eigenvector of (37) associated with θ_i. Then the i-th Ritz vector u_i is defined by $u_i = Sq_i$.

Suppose $M = R^T R$ is the Cholesky factorzation of M, where R is upper triangular. If we define $Z = RSQ$, where

$$Q = (q_1\ q_2\ \ldots\ q_k),$$

then it follows from the $S^T MS$-orthogonality of q_i that

$$Z^T Z = I_k.$$

Solving the generalized eigenvalue problem (2) is equivalent to solving the following standard eigenvalue problem

$$R^{-T} K R^{-1} y = \lambda y, \tag{38}$$

where $y = Rx$.

Applying the standard Rayleigh-Ritz procedure to (38) from the subspace spanned by Z yields the following projected eigenvalue problem

$$Z^T R^{-T} K R^{-1} Z g = \theta g. \tag{39}$$

It is easy to show that

$$Z^T R^{-T} K R^{-1} Z = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_k).$$

Moreoever, since the eigenvectors of (39) are simply $g_i = e_i$, for $i = 1, 2, ..., k$, the Ritz vectors associated with (38) are $z_i = Zg_i = RSQe_i = Ru_i$, for $i = 1, 2, ..., k$.

It follows from Theorem 4 that

$$\theta_1 - \lambda_1 \leq (\lambda_n - \lambda_1)\sin\angle(y_1, \mathcal{Z}) \tag{40}$$

$$\sin\angle(z_1, y_1) \leq \sqrt{\frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1}}\sin\angle(y_1, \mathcal{Z}) \tag{41}$$

where y_1 is the eigenvector of (38) corresponding to the smallest eigenvalue λ_1 and $\mathcal{Z} = \mathrm{span}\{Z\} = \mathrm{span}\{RSQ\} = \mathrm{span}\{RS\}$.

Note that $y_1 = Rx_1$, where x_1 is the eigenvector of (2) corresponding to the smallest eigenvalue λ_1. Thus,

$$\cos\angle(z_1, y_1) = z_1^T y_1 = (Ru_1)^T Rx_1 = u_1^T M x_1 = \cos\angle_M(u_1, x_1).$$

Furthermore, it is easy to verify that $W = SQ$ is M-orthonormal, i.e.,

$$W^T M W = Q^T S^T M S Q = I_k.$$

Hence,

$$\cos \angle(y_1, \mathcal{Z}) = \|y_1^T Z\| = \|(Rx_1)^T RSQ\| = \|x_1^T M W\| = \cos \angle_M(x_1, \mathcal{S}).$$

where $\mathcal{S} = \text{span}\{S\}$. Thus, we can now replace $\angle(z_1, y_1)$ and $\angle(y_1, \mathcal{Z})$ with $\angle_M(u_1, x_1)$ and $\angle_M(u_1, \mathcal{S})$ respectively in (40) and (41) to reach the conclusions stated in Theorem 1.

References

[1] A. Abramov. On the separation of the principal part of some algebraic problems. *Zh. Vych. Mat.*, 2:141–145, 1962.

[2] A. Abramov. Remarks on finding the eigenvalues and eigenvectors of matrices which arise in the application of Ritz's method or in the difference method. *Zh. Vych. Mat. Fiz.*, 7:644–647, 1962.

[3] J. S. Arora and D. T. Nguyen. Eigensolution for large structure systems with substrutures. *Int. J. Num. Meth. Eng.*, 15:333–341, 1980.

[4] C. Bekas and Y. Saad. Computation of smallest eigenvalues using spectral schur complements. Technical Report UMSI-2004-6, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 2004.

[5] J. K. Bennighof. Adaptive multi-level substructuring method for acoustic radiation and scattering from complex structures. In A. J. Kalinowski, editor, *Computational methods for Fluid/Structure Interaction*, volume 178, pages 25–38, AMSE, New York, November, 1993.

[6] J. K. Bennighof and C. K. Kim. An adaptive multi-level substructuring method for efficient modeling of complex structures. In *Proceedings of the AIAA 33st SDM Conference*, Dallas, Texas, 1992.

[7] J. K. Bennighof and R. B. Lehoucq. An automated multileve substructuring method for eigenspace computation in linear elastodynamics. *SIAM Journal on Scientific Computing*, to appear, 2003.

[8] R. E. Bishop. The analysis and synthesis of vibration systems. *Journal of Royal Aeronautical Society*, 58, 1954.

[9] F. Bourquin. Analysis and comparison of several component mode synthesis methods on one-dimensional domains. *Numer. Math*, 58:11–34, 1990.

[10] F. Bourquin. Component mode synthesis and eigenvalues of second order operators: Discretization and algorithm. *Mathematical Modeling and Numerical Analysis*, 26:385–423, 1992.

[11] V. S. Chichov. A method for partitioning a high order matrix into blocks in order its eigenvalues. *Zh. Vych. Mat.*, 1:169–173, 1961.

[12] R. R. Craig and M. C. C. Bampton. Coupling of substructures for dynamic analysis. *AIAA Journal*, 6:1313–1319, 1968.

[13] R. R. Craig and C-J. Chang. A review of substructure coupling methods for dynamic analysis. *Advances in Engineering Science, NASA CP-2001*, 2:393–408, 1976.

[14] I.S. Duff, R.G. Grimes, and J.G. Lewis. Users' guide for the Harwell-Boeing sparse matrix collection (release 1). Technical Report RAL-92-086, Rutherford Appleton Laboratory, December 1992.

[15] A. George. Nested dissection of a regular finite element mesh. *SIAM J. Num. Anal.*, 10:345–363, 1973.

[16] R. G. Grimes, J. G. Lewis, and H. D. Simon. A shifted block Lanczos algorithm for solving sparse symmetric generalized eigenproblems. *SIAM Journal on Matrix Analysis and Applications*, 15(1):228–272, January 1994.

[17] T. K. Hasselman. Damping synthesis from substructure tests. *AIAA Journal*, 14:1409–1418, 1976.

[18] B. A. Hunn. A method of calculation the space free resonant modes of an aircraft. *Journal of the Royal Aeronautical Society*, 57:420–422, 1953.

[19] B. A. Hunn. A method of calculating the normal modes of an aircraft. *Quarterly Journal of Mechanics*, 8:38–58, 1955.

[20] W. C. Hurty. Vibrations of structure systems by component-mode synthesis. *Journal of the Engineering Mechanics Dvision, ASCE*, 86:51–69, 1960.

[21] D. D. Kana and S. Huzar. Synthesis of shuttle vehicle damping using substructure test results. *Journal of Spacecraft and Rockets*, 10:790–797, 1973.

[22] M. F. Kaplan. *Implementation of Automated Multilevel Substructuring for Frequency Response Analysis of Structures.* PhD thesis, University of Texas at Austin, Austin, TX, December 2001.

[23] G. Karypis and V. Kumar. METIS – *A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices – Version 4.0.* University of Minnesota, September 1998.

[24] K. Ko, N. Folwell, L. Ge, A. Guetz, V. Ivanov, L. Lee, Z. Li, I. Malik, W. Mi, C. Ng, and M. Wolf. Electromagnetic systems simulation - from simulation to fabrication. SciDAC report, Stanford Linear Accelerator Center, Menlo Park, CA, 2003.

[25] A. Kropp and D. Heiserer. Efficient broadband vibro-accoutic analysis of passenger car bodies using an FE-based component mode synthesis approach. In *Proceedings of the fifth World Congress on Computational Mechanics (WCCM V)*, Vienna University of Technology, 2002.

[26] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide – Solution of Large-scale eigenvalue problems with implicitly restarted Arnoldi Methods.* SIAM, Philadelphia, PA., 1999.

[27] G. Lou, T. W. Wu, and Z. Bai. Disk brake squeal prediction using the ABLE algorithm. *J. of Sound and Vibration*, 272:731–748, 2004.

[28] R. H. MacNeal. Vibrations of composite systems. Technical Report OSRTN-55-120, Office of Scientific Research, Air Research of Scientific Research and Development Command, 1954.

[29] B. N. Parlett. *The Symmetric Eigenvalue Problem.* Prentice-Hall, 1980.

[30] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations.* Oxford University Press, Oxford, UK, 1999.

[31] Y. Saad. *Numerical Methods for Large Eigenvalue Problems.* Halsted Press, 1992.

[32] H. Serbin. Vibration of composite structures. *Journal of the Aeronautical Sciences*, 12(1):108, 1945.

[33] G. L. G. Sleijpen, J. V. D. Eshof, and P. Smit. Optimal a priori error bounds for the Rayleigh Ritz method. *Math. Comp.*, 72(242):667–684, 2002.

[34] T. C. Sofrin. The combination of dynamical systems. *Journal of the Aeronautical Sciences*, 13(6):281–288, 1946.

[35] G. W. Stewart. *Matrix Algorithms, Volume II: Eigensystems.* SIAM, 2001.

[36] Din-Kow Sun, Jin-Fa Lee, and Zoltan Cendes. Construction of nearly orthogonal nedelec bases for rapid convergence with multilevel preconditioned solvers. *SIAM Journal on Scientific Computing*, 23(4):1053–1076, 2001.