

***Filtering Frequencies in a Shift-and-invert  
Lanczos Algorithm for the Dynamic Analysis of  
Structures***

Zemaite, Mante and Tisseur, Francoise and Kannan,  
Ramaseshan

2018

MIMS EPrint: **2018.17**

Manchester Institute for Mathematical Sciences  
School of Mathematics

The University of Manchester

Reports available from: <http://eprints.maths.manchester.ac.uk/>

And by contacting: The MIMS Secretary  
School of Mathematics  
The University of Manchester  
Manchester, M13 9PL, UK

ISSN 1749-9097

# FILTERING FREQUENCIES IN A SHIFT-AND-INVERT LANCZOS ALGORITHM FOR THE DYNAMIC ANALYSIS OF STRUCTURES \*

MANTE ZEMAITE<sup>†</sup>, FRANÇOISE TISSEUR<sup>‡</sup>, AND RAMASESHAN KANNAN<sup>§</sup>

**Abstract.** The shift-and-invert Lanczos algorithm is a commonly used solution procedure to compute the eigenpairs of large, sparse eigenvalue problems that arise when approximating the elastic dynamic response of large structures under the influence of seismic forces. Not all eigenvectors are equally important to the response when the structure is exposed to a mass-dependent external force of the form  $g(t)Mb$ , where  $M$  is the mass matrix of the system and  $b$  the rigid body vector. Structural engineers select eigenvectors  $x_j$ ,  $j = 1, \dots, \ell$ , such that their cumulative mass participation, measured as  $\sum_{j=1}^{\ell} (x_j^T Mb)^2 / (b^T Mb)$ , is above a target threshold  $\xi$ . We show that when the starting vector for the unshifted Lanczos algorithm is the spatial distribution vector  $b$ , the Lanczos procedure can be used to provide an estimate of the cumulative mass participation. This allows us to identify intervals containing eigenvalues whose eigenvectors have a large contribution to the cumulative mass participation and filter out intervals containing eigenvalues whose eigenvectors have a negligible contribution. We use this information to devise a sequence of shifts  $\sigma_1, \dots, \sigma_p$  for the shift-and-invert Lanczos algorithm as well as a stopping criterion for the iteration with shift  $\sigma_i$  so that the cumulative mass participation of the computed eigenvectors reaches the required level  $\xi$ . Numerical experiments on real engineering problems show that our approach computes up to 80% fewer eigenvectors and requires fewer shifts, on average, than the more general shifting strategy proposed by Ericsson and Ruhe (Math. Comp., 35 (1980)).

**Key words.** Shifting strategy, shift-and-invert Lanczos algorithm, orthogonal polynomials, symmetric generalised eigenvalue problem, structural analysis.

**1. Introduction.** A structural dynamics problem consists of finding the response of a structure, for instance, a building or a bridge, given some dynamic loading. Such problems may be written in the form of a system of second order differential equations

$$(1.1) \quad M\ddot{u}(t) + D\dot{u}(t) + Ku(t) = f(t)$$

that results from the finite element discretization of the equation of motion, together with some initial conditions. The mass matrix  $M \in \mathbb{R}^{n \times n}$  is usually symmetric positive semidefinite (denoted by  $M \geq 0$ ), the stiffness matrix  $K \in \mathbb{R}^{n \times n}$  is symmetric positive definite ( $K > 0$ ), the damping matrix  $D$  is symmetric and often positive definite,  $u(t)$  is the displacement, and  $f(t)$  is the time-dependent external load on the given structure. Here, we concentrate on external forces of the form

$$(1.2) \quad f(t) = g(t)Mb,$$

where  $g(t)$  is a scalar function and  $0 \neq b \in \mathbb{R}^n$  is the spatial distribution vector (also called rigid body vector or spatial vector of loading patterns). External forces of the form (1.2) are particular to earthquake loading, where  $g(t)$  is the input earthquake acceleration.

Projection methods are usually employed to reduce the dimension  $n$  of the system (1.1). These methods consist of constructing a matrix  $X_\ell \in \mathbb{R}^{n \times \ell}$  of full rank and

---

\*Version of May 25, 2018.

<sup>†</sup>School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK (mante.zemaityte@manchester.ac.uk).

<sup>‡</sup>School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK (francoise.tisseur@manchester.ac.uk). This work was supported by a Royal Society-Wolfson Research Merit Award.

<sup>§</sup>Arup, 3 Piccadilly Place, Manchester, M1 3BN, UK (Ramaseshan.Kannan@arup.com).

transforming (1.1) into the reduced system

$$(1.3) \quad X_\ell^T M X_\ell \ddot{v}(t) + X_\ell^T D X_\ell \dot{v}(t) + X_\ell^T K X_\ell v(t) = g(t) X_\ell^T M b,$$

which is then solved for  $v(t)$ . An approximate solution to (1.1) is obtained as  $u(t) \approx X_\ell v(t)$ . In the mode superposition method, the columns  $x_1, \dots, x_\ell$  of  $X_\ell$  are eigenvectors corresponding to finite eigenvalues of the associated generalized eigenvalue problem (GEP)

$$(1.4) \quad (K - \lambda M)x = 0,$$

where we assume generalized proportional damping, so that the damping matrix  $D$  in (1.1) is diagonalizable by the matrix of eigenvectors of (1.4). If the columns of  $X_\ell$  are  $M$ -orthonormal, i.e.,  $x_i^T M x_j = \delta_{ij}$  with  $\delta_{ij}$  the Kronecker delta, then the reduced system (1.3) can be rewritten as  $\ell$  decoupled second order differential equations,

$$(1.5) \quad \ddot{v}_j(t) + 2\zeta_j \omega_j \dot{v}_j(t) + \omega_j^2 v_j(t) = g(t) x_j^T M b, \quad j = 1, \dots, \ell,$$

where  $\omega_j = \sqrt{\lambda_j}$ ,  $\lambda_j > 0$  is an eigenvalue of (1.4) with corresponding eigenvector  $x_j$  and  $x_j^T D x_j = 2\zeta_j \omega_j \delta_{ij}$  for some  $\zeta_j \geq 0$  [16, Chap. 18].

Not all eigenvectors are equally important to a given system of differential equations and it is clear from (1.5) that the response  $v_j$  depends on both the frequencies  $\omega_j$  and the magnitude of  $x_j^T M b$ , which is called the *mass participation factor* of  $x_j$ . They satisfy

$$(1.6) \quad \sum_{j=1}^n \frac{(x_j^T M b)^2}{b^T M b} = 1$$

(see Section 2.1). Within the context of seismic analysis and design, structural engineers aim to achieve 80 to 90% of mass participation in (1.6) using only a subset of the eigenvectors (see the justification in section 2.1). This leads to the following problem.

**PROBLEM 1.1.** *For a given proportion  $\xi \in (0, 1)$  and a spatial distribution vector  $b$ , find the smallest number of  $M$ -orthonormal eigenvectors  $x_{i_k}$ ,  $k = 1, \dots, \ell$ , of (1.4) such that*

$$(1.7) \quad \sum_{k=1}^{\ell} \frac{(x_{i_k}^T M b)^2}{b^T M b} \geq \xi,$$

where  $\{i_1, \dots, i_\ell\} \subseteq \{1, \dots, n\}$ .

Problem 1.1 is easy to solve if we can compute all the eigenvectors of the GEP (1.4), but this is not feasible for problems of large dimensions. It is usually the eigenvectors corresponding to small eigenvalues that contribute the most to the total mass participation (see for example [2] or Section 2.1), so in previous work Problem 1.1 was relaxed to the following problem.

**PROBLEM 1.2.** *For a given proportion  $\xi \in (0, 1)$  and a spatial distribution vector  $b$ , find the  $M$ -orthonormal eigenvectors  $x_i$ ,  $i = 1, \dots, \ell$  associated with the  $\ell$  smallest eigenvalues of (1.4), where  $\ell$  is the smallest integer such that*

$$(1.8) \quad \sum_{i=1}^{\ell} \frac{(x_i^T M b)^2}{b^T M b} \geq \xi.$$

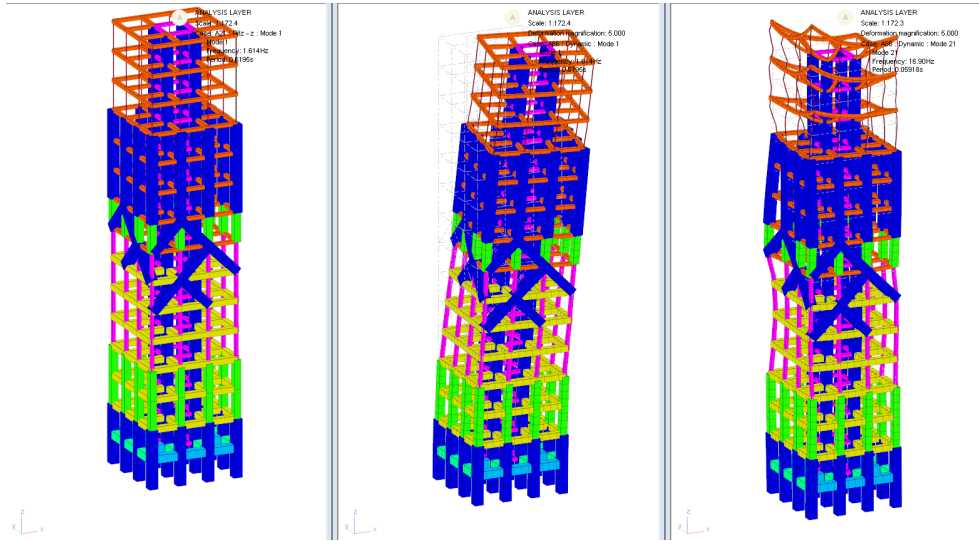


FIG. 1. The eigenvectors (modes of vibration) of a tall building structure. From left to right: the undeformed shape of the structure; its first eigenvector,  $x_1$  corresponding to the smallest eigenvalue, which is predominantly a “sway mode” (i.e., large mass participation factor in the cartesian  $x$ -direction); and another eigenvector,  $x_{20}$ , which is a “bouncing mode” (large participation along cartesian  $z$ -direction). Deformations are exaggerated.

TABLE 1  
Mass participation factors for  $M$ -orthonormal eigenvectors  $x_1$  and  $x_{20}$  from Figure 1.

Eigenvector $x_j$	Mass participation factor $(x_j^T M b)^2 / (b^T M b)$	
	$b$ along cartesian $x$	$b$ along cartesian $z$
$x_1$	0.199	$10^{-6}$
$x_{20}$	$10^{-8}$	0.209

Recovering sufficient mass participation can be a challenge depending on the geometry of the structure and orientation of  $b$ . The vector  $b$  represents the “rigid body” deformation of the structure in cartesian  $x$ -,  $y$ -, and  $z$ -directions. For instance a given structure may have a geometry that makes it easy to solve Problem 1.2 with only a handful of smallest eigenvectors (i.e.,  $\ell$  is small) when  $b$  is along the  $x$ -direction but might require a large value of  $\ell$  when  $b$  is along  $z$ . This is graphically illustrated in Figure 1 which shows the first and twentieth eigenvectors,  $x_1$  and  $x_{20}$ , overlaid as displacements on the structure’s geometry. As can be seen from the diagram in the middle,  $x_1$  is a sway mode wherein large parts of the structure deform in the cartesian  $x$ -direction. By contrast,  $x_{20}$  (right of middle) is a bouncing mode and therefore the structure deforms significantly in the  $z$ -direction. The mass participations of these eigenvectors are tabulated in Table 1 for  $b$  in the  $x$ - and  $z$ -directions.

A natural approach to solve Problem 1.2 is to apply the shift-and-invert Lanczos algorithm (see [1], [17] and references therein) to  $K - \lambda M$  with or without shifts depending on how many eigenvectors will be needed to satisfy (1.8). Instead of using eigenvectors in (1.8), Wilson et al. [20] proposed to use the Ritz vectors resulting from a variant of the Arnoldi algorithm later known as the WYD algorithm [9], [19]. The use of the Lanczos vectors was later suggested by Nour-Omid and Clough [13].

But as noted in [2], there is no analysis to support the use of Lanczos vectors and Ritz vectors (unless they have converged to eigenvectors of  $K - \lambda M$ ). In particular, we show in section 2 that under suitable conditions on  $g(t)$  and the initial conditions for (1.1), the quantity  $(1 - \xi)^2$  provides an upper bound on the relative error between the response  $u(t)$  to the ODE (1.1) and its approximation by  $X_\ell v(t)$  when the columns of  $X_\ell$  are eigenvectors chosen such that (1.7) or (1.8) holds. So we concentrate here on the computation of eigenvectors for Problems 1.1–1.2.

Given an initial shift  $\sigma_1$ , a common strategy to determine the sequence of shifts  $\sigma_2, \sigma_3, \dots$  to be employed in a shift-and-invert Lanczos process for symmetric GEPs is to choose the new shift  $\sigma_i$ ,  $i > 1$ , such that the largest converged eigenvalue  $\lambda_{\max}$  is halfway between the old shift  $\sigma_{i-1}$  and the new shift  $\sigma_i$ , namely  $\sigma_i = 2\lambda_{\max} - \sigma_{i-1}$  [5]. This way, if the eigenvalues are roughly evenly distributed across the spectrum then the same number of eigenvalues is expected to converge to the left of the new shift as the number of eigenvalues that have converged to the right of the old shift. The Lanczos iteration with shift  $\sigma_i$  is then stopped when the smallest converged eigenvalue in that iteration coincides with the largest converged eigenvalue with shift  $\sigma_{i-1}$ . One issue with this shifting strategy for Problem 1.2 is that the shift-and-invert Lanczos algorithm may return too many eigenvectors with small normalized mass participation. As a result, the number  $\ell$  of returned eigenpairs may be unnecessarily large, whereas eigenvectors with large normalized mass participations corresponding to eigenvalues that are not in the lower end of the spectrum may not be detected.

Our main contribution is the presentation of a new shifting strategy for the shift-and-invert Lanczos algorithm together with a stopping criterion for the iteration with shift  $\sigma_i$  that are specifically designed to approximate the solution to Problem 1.1. For this, we use the theory of orthogonal polynomials to show that a few steps of the unshifted inverse Lanczos algorithm applied to  $K^{-1}M$  with starting vector  $b$  provides, at no additional cost, information about the location of the eigenvalues whose corresponding eigenvectors have non-negligible mass participation and also helps to identify intervals where eigenvalues have negligible mass participation. We use this information to devise a shifting strategy for the shift-and-invert Lanczos process so that condition (1.7) holds, albeit perhaps not for the smallest number  $\ell$  of eigenvectors. This shifting strategy performs especially well in the cases where the eigenvectors with non-negligible mass participation correspond to larger eigenvalues while there are intervals of smaller eigenvalues whose corresponding eigenvectors have negligible contribution to the overall response of the structure. These are the cases that are the most problematic for available methods.

Numerical experiments performed on real structural engineering problems show an often large reduction in the number  $\ell$  of eigenvectors computed to approximate the solution to Problem 1.1 using our new shifting strategy as opposed to the number of eigenvectors computed to approximate the solution to Problem 1.2 using the more general shifting strategy of Ericsson and Ruhe. The use of a new shift in the shift-and-invert Lanczos process has a cost since it leads to a new matrix factorization. For our set of test problems, our numerical experiments show that the number of shifts used with our new shifting strategy is, on average, smaller than with Ericsson and Ruhe’s strategy.

Unlike previous attempts to solve Problems 1.1–1.2, we pay special attention to issues that can occur when the mass matrix  $M$  is singular. In this case, the Lanczos process proceeds with a quasi-inner product. In finite precision arithmetic, the computed Lanczos vectors and Ritz vectors have components in the null space of  $M$  and the magnitude of these unwanted components grows during the iterations.

This can either delay or prevent convergence of the Ritz vectors [4]. These issues are overcome with the use of an appropriate starting vector and implicit filtering [12].

In the next section we give some preliminary material that includes a justification of Problem 1.1 and a discussion of issues that arise when using a quasi-inner product in a shift-and-invert Lanczos process, as well as possible remedies. We describe and justify our new shifting strategy in section 3, and illustrate its performance on a number of real structural engineering problems in section 4.

**2. Preliminaries.** Following [2], we show in section 2.1 that, under some assumptions on the initial conditions and the input function  $g(t)$ , choosing  $\xi$  close to 1 in (1.7) guarantees a small error between the exact response  $u(t)$  and its approximation as a linear combination of the eigenvectors  $x_{i_k}$ ,  $k = 1, \dots, \ell$  satisfying (1.7).

We recall in section 2.2 the shift-and-invert Lanczos process for the GEP (1.4) and discuss issues related to the use of a quasi-inner product defined by the symmetric semidefinite matrix  $M$ .

**2.1. Upper bound on the response error.** The  $n \times n$  mass matrix  $M$  in (1.1) is often singular in applications. As a result, the GEP (1.4) has the eigendecomposition

$$X^T(K - \lambda M)X = \Lambda - \lambda \begin{bmatrix} I_r & \\ & 0 \end{bmatrix},$$

where  $I_r$  is the  $r \times r$  identity matrix with

$$r := \text{rank}(M) \leq n,$$

$\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  has real positive diagonal entries displaying the  $r$  finite eigenvalues as  $\lambda_j$ ,  $j = 1, \dots, r$  (the remaining  $n - r$  eigenvalues being at infinity), and  $X$  is a nonsingular matrix containing the corresponding eigenvectors  $x_1, \dots, x_n$ . Note that the eigenvectors  $x_j$ ,  $j = r + 1, \dots, n$  associated with the eigenvalues at infinity belong to the null space of  $M$ , i.e.,

$$(2.1) \quad Mx_j = 0, \quad j = r + 1, \dots, n.$$

Let

$$(2.2) \quad u(t) = Xv(t) = \sum_{j=1}^n v_j(t)x_j,$$

where  $v_j(t)$  is the  $j$ th entry of the vector  $v(t)$ . We assume generalized damping so that  $X^TDX = 2\text{diag}(\zeta_1\omega_1, \dots, \zeta_n\omega_n)$  for some  $\zeta_j \geq 0$  and  $\omega_j = \sqrt{\lambda_j}$ ,  $j = 1, \dots, n$ , and rewrite (1.1) as

$$(2.3) \quad \begin{aligned} \ddot{v}_j(t) + 2\zeta_j\omega_j\dot{v}_j(t) + \omega_j^2v_j(t) &= g(t)x_j^T Mb, & j = 1, \dots, r, \\ v_j(t) &= 0, & j = r + 1, \dots, n. \end{aligned}$$

The system of uncoupled equations (2.3) can then be solved by direct integration, yielding

$$(2.4) \quad \begin{aligned} v_j(t) &= (x_j^T Mb) \int_0^t \frac{e^{-\zeta_j\omega_j(t-s)}}{\tilde{\omega}_j} \sin(\tilde{\omega}_j(t-s))g(s)ds, & j = 1, \dots, r, \\ v_j(t) &= 0, & j = r + 1, \dots, n, \end{aligned}$$

where we let  $\tilde{\omega}_j = \omega_j(1 - \zeta_j)^{1/2}$  and assume for simplicity that  $u(0) = \dot{u}(0) = 0$  so that  $v(0) = \dot{v}(0) = 0$ .

As mentioned in the introduction, in practical applications,  $n$  is large and it is infeasible to compute all  $n$  eigenpairs of (1.4) so the solution  $u(t)$  in (2.2) is approximated instead by

$$(2.5) \quad \tilde{u}(t) = \sum_{j=1}^{\ell} v_j(t) x_j, \quad \ell \leq r.$$

We use the  $M$ -quasi vector norm,

$$\|y\|_M = \langle y, y \rangle_M^{1/2} = (y^T M y)^{1/2},$$

to measure the relative error between the exact solution and its approximation,

$$(2.6) \quad \frac{\|u(t) - \tilde{u}(t)\|_M}{\|u(t)\|_M} = \frac{(\sum_{j=\ell+1}^r v_j^2(t))^{1/2}}{(\sum_{j=1}^r v_j^2(t))^{1/2}}.$$

We rewrite  $v_j(t)$  in (2.4) as  $v_j(t) = h_j(t) x_j^T M b$ . If the spatial distribution vector  $b$  has no components in the null space of  $M$  (this is usually the case), i.e.,  $b = \sum_{j=1}^r b_j x_j$ , then

$$(2.7) \quad \sum_{j=1}^r (x_j^T M b)^2 = \sum_{j=1}^r b_j^2 = b^T M b,$$

which on using (2.1) is equivalent to (1.6). Now if we assume that  $h_{\min} \leq |h_j(t)| \leq h_{\max}$  for  $t > 0$ ,  $j = 1, \dots, r$ , and some positive scalars  $h_{\min}, h_{\max}$ , then it follows from (2.6) and (2.7) that

$$\begin{aligned} \frac{\|u(t) - \tilde{u}(t)\|_M}{\|u(t)\|_M} &\leq \frac{h_{\max}}{h_{\min}} \frac{(\sum_{j=\ell+1}^r (x_j^T M b)^2)^{1/2}}{(\sum_{j=1}^r (x_j^T M b)^2)^{1/2}} = \frac{h_{\max}}{h_{\min}} \left( 1 - \frac{\sum_{j=1}^{\ell} (x_j^T M b)^2}{b^T M b} \right)^{1/2} \\ &\leq \frac{h_{\max}}{h_{\min}} (1 - \xi)^{1/2}, \end{aligned}$$

where  $\xi$  is as in (1.7). Thus, under the above assumptions on the initial conditions  $u(0)$ ,  $\dot{u}(0)$ , and the functions  $h_j$ , choosing a  $\xi$  close to 1 guarantees a small relative error between the exact solution  $u$  and its approximation  $\tilde{u}$ . We refer to section 4 for an illustration of the above analysis.

Finally, we note that the factor  $1/\omega_j$  in (2.4) suggests that the eigenvectors corresponding to smaller eigenvalues  $\omega_j^2$  (i.e., lower frequencies) are more likely to have a larger contribution to the response  $u(t)$  in (2.2) than those corresponding to the higher frequencies.

## 2.2. Shift-and-invert Lanczos process with semi-definite inner product.

Applying the Lanczos algorithm with an  $M$ -quasi inner product to compute approximate eigenpairs of the definite pencil  $K - \lambda M$  with  $M \geq 0$  was first suggested by Scott [18]. Given a shift  $\sigma$  near the eigenvalues of interest and a starting vector  $w$ , this Lanczos procedure constructs a matrix  $Q_k$  whose columns, called the Lanczos vectors, form an  $M$ -orthonormal basis for the  $k$ th order Krylov subspace

$$\mathcal{K}_k(A, w) = \text{span}\{w, Aw, \dots, A^{k-1}w\} = \text{span}\{q_1, q_2, \dots, q_k\},$$

where  $A = (K - \sigma M)^{-1}M$ . Given the first  $k \geq 1$   $M$ -orthonormal Lanczos vectors  $q_1 = w/\|w\|_M, \dots, q_k$  and  $q_0 = 0$ , the next Lanczos vector  $q_{k+1}$  is obtained from the three-term recurrence

$$(2.8) \quad \beta_{k+1}q_{k+1} = (K - \sigma M)^{-1}Mq_k - \alpha_k q_k - \beta_k q_{k-1},$$

where  $\alpha_k = q_k^T M(K - \sigma M)^{-1}Mq_k$ ,  $\beta_1 = \|w\|_M = w^T M w$ , and  $\beta_k$  for  $k > 1$  is such that  $\|q_k\|_M = q_k^T M q_k = 1$ . The three-term recurrence (2.8) can be rewritten in matrix form as

$$(2.9) \quad (K - \sigma M)^{-1}M Q_k = Q_k T_k + \beta_{k+1} q_{k+1} e_k^T,$$

where  $Q_k = [q_1, \dots, q_k] \in \mathbb{R}^{n \times k}$  and  $T_k \in \mathbb{R}^{k \times k}$  is symmetric tridiagonal with

$$(T_k)_{jj} = \alpha_j, \quad j = 1, \dots, k, \quad (T_k)_{j,j+1} = \beta_{j+1}, \quad j = 1, \dots, k-1.$$

Since, by construction,  $q_{k+1}$  is  $M$ -orthogonal to the columns of  $Q_k$  and  $Q_k^T M Q_k = I_k$ , it follows from (2.9) that

$$(2.10) \quad Q_k^T M (K - \sigma M)^{-1} M Q_k = T_k.$$

Now, for an eigenpair  $(\theta_j^{(k)}, s_j^{(k)})$  of  $T_k$ , the pair

$$(2.11) \quad (\lambda_j, x_j) := (1/\theta_j^{(k)} + \sigma, Q_k s_j^{(k)})$$

is called a Ritz pair for  $K - \lambda M$  and is considered as an approximate eigenpair of  $K - \lambda M$  if the scaled residual (which approximates the backward error for  $(\lambda_j, x_j)$  [10, Thm. 2.1])

$$(2.12) \quad \eta(\lambda_j, x_j) = \frac{\|(K - \lambda_j M)x_j\|_2}{(\|K\|_1 + |\lambda_j| \|M\|_1) \|x_j\|_2}$$

is below a given tolerance.

An algorithm implementing the three-term recurrence (2.8) is provided in Algorithm 2.1. This is essentially the algorithm provided in [14]. Breakdown occurs when  $\beta_{k+1} \leq 0$  in step 10. The number  $\ell$  of eigenpairs returned will depend of the convergence criterion in step 3. We discuss the latter in section 3.3. Some more comments follow.

- (a) **Choice of starting vector and implicit filtering.** When  $M$  is singular, Nour-Omid et al [14, Sec. 2.2] recommend choosing a starting vector  $w$  in the range of  $(K - \sigma M)^{-1}M$ . Indeed, since  $\mathbb{R}^n = \text{range}((K - \sigma M)^{-1}M) \oplus \text{null}((K - \sigma M)^{-1}M)$  and  $\text{null}((K - \sigma M)^{-1}M) = \text{null}(M)$ , we have that

$$\text{range}((K - \sigma M)^{-1}M) \cap \text{null}(M) = \{0\}.$$

For an eigenvector  $x_i$  of  $K - \lambda M$  with finite eigenvalue  $\lambda_i$ , we have that

$$(2.13) \quad x_i \in \text{range}(K^{-1}M) = \text{range}((K - \sigma M)^{-1}M).$$

If the starting vector  $w$  for the shift-and-invert Lanczos procedure is in the range of  $(K - \sigma M)^{-1}M$  then so is the first Lanczos vector  $q_1$  and the subsequent Lanczos vectors  $q_2, \dots, q_k$  if operations are performed in exact arithmetic. As a result, the Ritz vectors lie in  $\text{range}((K - \sigma M)^{-1}M)$ . Now if



---

**Algorithm 2.1** Shift-and-invert Lanczos algorithm

---

This algorithm takes as input two  $n \times n$  symmetric matrices  $M \geq 0$  and  $K > 0$ , a shift  $\sigma \geq 0$  such that  $K - \sigma M$  is nonsingular, and a starting vector  $0 \neq w \in \text{range}((K - \sigma M)^{-1}M)$ . It returns  $\ell$  eigenpairs (converged Ritz pairs)  $(\lambda_{i_j}, x_{i_j})$ ,  $j = 1, \dots, \ell$ .

- 1  $q_0 = 0$ ,  $z = Mw$ ,  $\beta_1 = \sqrt{w^T z}$
  - 2 Factor  $K - \sigma M$ .
  - 3 for  $k = 1, 2, \dots$ , until convergence
  - 4  $z = z/\beta_k$ ,  $q_k = w/\beta_k$
  - 5  $v = (K - \sigma M)^{-1}z - \beta_k q_{k-1}$
  - 6  $\alpha_k = v^T z$
  - 7  $w = v - \alpha_k q_k$
  - 8 Reorthogonalize  $w$  against  $q_1, \dots, q_k$  with respect to  $\langle \cdot, \cdot \rangle_M$  if necessary.
  - 9  $z = Mw$
  - 10  $\beta_{k+1} = \sqrt{w^T z}$
  - 11 Compute the  $k$  eigenpairs  $(\theta_j^{(k)}, s_j^{(k)})$ ,  $j = 1, \dots, k$  of  $T_k = \text{tridiag}(\beta, \alpha, \beta)$  with  $\alpha = [\alpha_1, \dots, \alpha_k]$  and if  $k > 1$ ,  $\beta = [\beta_2, \dots, \beta_k]$ .
  - 12 Check for convergence of the Ritz pairs  $(\lambda_j, x_j) = (\frac{1}{\theta_j^{(k)}} + \sigma, Q_k s_j^{(k)})$ ,  
 $j = 1, \dots, k$ , where  $Q_k = [q_1, \dots, q_k]$ .
  - 13 end
- 

$w \notin \text{range}((K - \sigma M)^{-1}M)$  then the Ritz vectors may have unwanted components in the null space of  $M$ , which in turn slows down their convergence. Note that in finite precision arithmetic, even if  $w \in \text{range}((K - \sigma M)^{-1}M)$ , rounding errors prevent the computed Lanczos vectors from staying in the range of  $(K - \sigma M)^{-1}M$ . The unwanted components in  $\text{null}(M)$  are mainly introduced when solving the linear systems with  $K - \sigma M$ . Since this operation is performed at each iteration, the accumulation can be rapid. The latter is set off at the beginning when the starting vector  $w$  is constructed as the product of  $(K - \sigma M)^{-1}M$  with some  $y \in \mathbb{R}^n$ , since this operation itself can introduce a null space component. The starting vector can be put in the right space explicitly by forming a particular projection matrix [4, Section 2.3], or solving the linear system to higher precision, however both methods come at a substantial cost. Instead, as suggested by Meerbergen [12], we apply an implicit filter that alters the starting vector implicitly, producing Lanczos vectors lying in  $\text{range}((K - \sigma M)^{-1}M)$ . This approach is briefly discussed below.

For  $q = q_R + q_N \in \mathbb{R}^n$  with  $q_R \in \text{range}(M)$  and  $q_N \in \text{null}(M)$ , we have that  $\|q\|_M = \|q_R\|_M$ , i.e., the components in the null space of  $M$  are undetectable by the  $M$ -norm. So  $\|q_N\|_2$  can be arbitrarily large even when  $\|q\|_M = 1$ . Hence, if at step  $k$  of the Lanczos process we have

$$(2.14) \quad \|q_k\|_2 > \text{tol} \|q_1\|_2,$$

for some tolerance  $\text{tol} \gg 1$  (we choose  $\text{tol} = 10^4$ ) then we apply implicit filtering. For this let

$$(2.15) \quad \underline{T}_k = \underline{V}_k R_k$$

be the QR factorization of  $\underline{T}_k = [T_k, \beta_{k+1} e_k]^T \in \mathbb{R}^{k+1 \times k}$  with  $R_k \in \mathbb{R}^{k \times k}$  upper triangular and  $\underline{V}_k \in \mathbb{R}^{k+1 \times k}$  with orthonormal columns. Then  $\tilde{Q}_k =$

$Q_{k+1}\mathbf{V}_k$  has  $M$ -orthonormal columns, and if we let

$$\tilde{\mathbf{T}}_{k-1} = R_k \mathbf{V}_{k-1}$$

then it is not difficult to show that the matrix  $\tilde{T}_{k-1}$  obtained from  $\tilde{\mathbf{T}}_{k-1}$  by deleting its last row is tridiagonal and satisfies the Lanczos recurrence

$$(2.16) \quad (K - \sigma M)^{-1} M \tilde{Q}_{k-1} = \tilde{Q}_{k-1} \tilde{T}_{k-1} + \tilde{\beta}_k \tilde{q}_k e_{k-1}^T,$$

where  $\tilde{\beta}_k$  is such that  $\|\tilde{q}_k\|_M = 1$ . It is shown in [12, Theorem 3.1] that  $\text{range}(\tilde{Q}_k) = \text{range}((K - \sigma M)^{-1} M Q_k)$ . The implicit filter implicitly pre-multiplies the Lanczos vectors by  $K^{-1}M$ , thereby removing any components in  $\text{null}(M)$ . Since the dimension of the projection space is reduced by one, we continue the Lanczos algorithm by forming the next Lanczos vector  $\tilde{q}_{k+1}$  from  $\tilde{q}_k$  as well as updating  $T_k$  in step 11 with  $\tilde{T}_{k-1}$ . Note that the implicit filtering described above is essentially one step of the unshifted QR algorithm applied to  $T_k$ . A Lanczos vector  $q_k$  with a large 2-norm means that  $q_k$  has large components in the null space of  $M$ . As a result, the smallest eigenvalue of  $T_k$  becomes close to zero and the unshifted QR step pushes that very small eigenvalue to the bottom of the tridiagonal matrix. Indeed, if  $T_k$  had a zero eigenvalue then after one step of unshifted QR, the last row of  $T_k$  would be zero. The construction of  $\tilde{T}_{k-1}$  corresponds to a deflation of the eigenpair of  $T_k$  with smallest eigenvalue, thereby removing the Ritz pair in (2.11) with a large (possibly infinite) eigenvalue.

- (b) **Testing for convergence of eigenpairs.** As explained earlier, we consider the Ritz pair  $(\lambda_j, x_j) := (1/\theta_j^{(k)} + \sigma, Q_k s_j^{(k)})$ , where  $(\theta_j^{(k)}, s_j^{(k)})$  is an eigenpair of  $T_k$ , to have converged if its backward error  $\eta(\lambda_j, x_j)$  in (2.12) is below a given tolerance.

We suggest to only compute  $\eta(\lambda_j, x_j)$  when the Ritz pair  $(\lambda_j, x_j)$  is likely to have converged. This can be checked as follows. On using (2.9), we have that

$$\begin{aligned} (K - \lambda_j M)x_j &= (K - \sigma M)x_j - \frac{1}{\theta_j^{(k)}} Mx_j \\ &= \frac{1}{\theta_j^{(k)}} (K - \sigma M)(\theta_j^{(k)} x_j - (K - \sigma M)^{-1} Mx_j) \\ &= \frac{1}{\theta_j^{(k)}} (K - \sigma M)(Q_k T_k - (K - \sigma M)^{-1} M Q_k) s_j^{(k)} \\ &= -\frac{1}{\theta_j^{(k)}} (K - \sigma M) \beta_{k+1} q_{k+1} (e_k^T s_j^{(k)}) \end{aligned}$$

so that

$$\|(K - \lambda_j M)x_j\|_2 \leq \frac{\|K\|_2 + |\sigma| \|M\|_2}{|\theta_j^{(k)}|} |e_k^T s_j^{(k)}| |\beta_{k+1}| \|q_{k+1}\|_2.$$

If  $s_j^{(k)}$  is normalized such that  $\|s_j^{(k)}\|_2 = 1$  then

$$\|x_j\|_M^2 = x_j^T M x_j = s_j^{(k)T} Q_k^T M Q_k s_j^{(k)} = s_j^{(k)T} s_j^{(k)} = 1$$

and

$$1 = \|x_j\|_M^2 = |x_j^T M x_j| \leq \|x_j\|_2 \|M x_j\|_2 \leq \|x_j\|_2^2 \|M\|_2 \leq \sqrt{n} \|x_j\|_2^2 \|M\|_1,$$

where we used Cauchy-Schwarz for the first inequality. Without loss of generality, we can assume that  $\|M\|_1 = \|K\|_1 = 1$  (if not replace  $K - \lambda M$  by  $K/\|K\|_1 - \tilde{\lambda}M/\|M\|_1$  with  $\tilde{\lambda} = \lambda\|M\|_1/\|K\|_1$ ). Hence,  $\|x_j\|_2 \geq n^{-1/2}$  and  $\|K\|_2 + |\sigma|\|M\|_2 \leq n^{1/2}(\|K\|_1 + |\sigma|\|M\|_1)$ . Now if  $|\lambda_j| \approx |\sigma|$  then

$$(2.17) \quad \eta(\lambda_j, x_j) \lesssim n^{1/4} |e_k^T s_j^{(k)}| |\beta_{k+1}| \|q_{k+1}\|_2 / |\theta_j^{(k)}|.$$

All the quantities in the above approximate upper bound are readily available during the Lanczos steps including  $\|q_{k+1}\|_2$  (see point (a) and (2.14)). Hence, we suggest to only compute  $\eta(\lambda_j, x_j)$  when the upper bound in (2.17) is below, say,  $10 \times \text{tol}$ .

### 3. Shifting strategies for an approximate solution to problems 1.1–1.2.

A considerable number of eigenvectors may be required to solve Problem 1.2 for problems where there are large normalized mass participation factors

$$(3.1) \quad m(x_j) := \frac{(x_j^T M b)^2}{b^T M b}$$

corresponding to eigenvalues  $\lambda_j$  that lie further away from the lower end of the spectrum. In this case, the Lanczos algorithm (Algorithm 2.1 with  $\sigma = 0$  as we are aiming for the small eigenvalues first) becomes increasingly slow and memory intensive. So if condition (1.8) is not satisfied by the converged eigenvectors after a given number, say  $k_{\max}$ , steps of Algorithm 2.1 with zero shift, we then restart Algorithm 2.1 with a sequence of shifts. As mentioned in the introduction, Ericsson and Ruhe [5] propose to use the following sequence:

$$(3.2) \quad \sigma_1 = \lambda_{\max} + \frac{\lambda_{\max}}{2}, \quad \sigma_i = 2\lambda_{\max} - \sigma_{i-1}, \quad i \geq 2,$$

where  $\lambda_{\max}$  is the largest converged eigenvalue. Algorithm 2.1 with shift  $\sigma_i$  is stopped when the smallest converged eigenvalue coincides with the largest converged eigenvalue with the shift  $\sigma_{i-1}$ , or once condition (1.8) is satisfied, in which case we terminate the computation. The shift-and-invert Lanczos algorithm with this shifting strategy, which we refer to as SIL, approximates the solution to Problem 1.2.

An issue with Problem 1.2 is that its solution may include many eigenvectors with very small or negligible normalized mass participation factors. Also, the shifting strategy (3.2) is not as effective when  $K - \lambda M$  has clustered eigenvalues, as is the case for the real structural engineering problems we consider in section 4.

In what follows, we show that Algorithm 2.1 with  $\sigma = 0$  and starting vector  $w$  equal to the spatial distribution vector  $b$  provides at almost no additional cost information about where the eigenvalues associated with eigenvectors of non negligible normalized mass participation lie, while at the same time identifies the parts of the spectrum that do not contribute much to the total mass participation. We use this information to devise a new shifting strategy.

**3.1. Estimating the cumulative mass participation.** Let  $q_1, q_2, \dots$  be the Lanczos vectors generated by the three-term recurrence (2.8) with  $\sigma = 0$  and  $0 \neq q_1 \in \text{range}(K^{-1}M)$ . Each Lanczos vector  $q_i$  can be written as

$$(3.3) \quad q_i = p_i(K^{-1}M)q_1,$$

where  $p_i$  is called the  $i$ th Lanczos polynomial. It is well known from the theory of orthogonal polynomials [6, Chap. 2], [8, Chap. 4] that these polynomials are orthogonal with respect to the inner product defined in terms of the Riemann-Stieltjes integral

$$(3.4) \quad \langle p_i, p_j \rangle_\phi := \int_a^b p_i(\mu) p_j(\mu) d\phi(\mu),$$

where  $a \leq \mu_{\min}$ , and  $b \geq \mu_{\max}$ , with  $\mu_{\min}$  and  $\mu_{\max}$  the smallest and the largest eigenvalues of  $K^{-1}M$ , respectively (i.e.,  $\mu_{\min} = 1/\lambda_{\max}$  and  $\mu_{\max} = 1/\lambda_{\min}$  where  $\lambda_{\max}$  and  $\lambda_{\min}$  are the largest and smallest eigenvalues of  $K - \lambda M$ ). The distribution function  $\phi(\mu)$  is a step function with jumps at the eigenvalues  $\mu_i$  of  $K^{-1}M$ , and is given by

$$(3.5) \quad \phi(\mu) = \sum_{i=1}^n \phi_i^2 h(\mu - \mu_i),$$

where

$$h(t) = \begin{cases} 1 & t \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

is the Heaviside function and the  $\phi_i$ 's are the coefficients of the first Lanczos vector  $q_1$  when expressed in the  $M$ -orthonormal basis  $x_1, \dots, x_r$  for  $\text{range}(K^{-1}M)$ , namely

$$(3.6) \quad q_1 = \sum_{i=1}^r \phi_i x_i,$$

(and  $\phi_i = 0$  for  $i = r+1, \dots, n$ ).

It turns out that the step function  $\phi(\mu)$  coincides with the cumulative mass participation sum when the starting vector for the Lanczos algorithm with zero shift is the spatial distribution vector  $b$  in (1.2) with  $b \in \text{range}(K^{-1}M)$ , as we now show.

**PROPOSITION 3.1.** *Let  $0 \neq b$  be the spatial distribution vector in (1.2) and assume that  $b \in \text{range}(K^{-1}M)$ . If  $w = b$  in Algorithm 2.1 with  $\sigma = 0$ , then for the step function  $\phi(\lambda)$  in (3.5) we have that*

$$\phi_i^2 = m(x_i), \quad i = 1, \dots, n,$$

where  $m(x_i)$  is the normalized mass participation of the eigenvector  $x_i$  in (3.1).

*Proof.* The first Lanczos vector is given by  $q_1 = b/\|b\|_M$  and, for  $i = 1, \dots, r$ ,

$$m(x_i) = \frac{(x_i^T M b)^2}{\|b\|_M^2} = \frac{(x_i^T M q_1 \|b\|_M)^2}{\|b\|_M^2} = (x_i^T M q_1)^2 = (x_i^T M \sum_{j=1}^n \phi_j x_j)^2 = \phi_i^2.$$

For  $i = r+1, \dots, n$ ,  $x_i \in \text{null}(M)$  so that  $Mx_i = 0$  and hence  $m(x_i) = 0 = \phi_i^2$ .  $\square$

If the nonzero eigenvalues  $\mu_j$ ,  $j = 1, \dots, r$ , of  $K^{-1}M$  are ordered by decreasing values then it follows from (3.5) and Proposition 3.1 that

$$\phi(\mu_j) = \sum_{i=1}^j m(x_i), \quad j \leq r.$$

Let  $(\theta_i^{(k)}, s_i^{(k)})$ ,  $i = 1, \dots, k$ , be the eigenpairs of the tridiagonal matrix  $T_k$  resulting from  $k$  steps of the unshifted Lanczos algorithm (Algorithm 2.1 with  $\sigma = 0$ ) and ordered such that

$$\theta_1^{(k)} \geq \theta_2^{(k)} \geq \dots \geq \theta_k^{(k)}.$$

The Lanczos polynomials  $p_i$  in (3.3) are not only orthogonal with respect to the inner product (3.4), they are also orthogonal with respect to the inner product

$$(3.7) \quad \langle p_i, p_j \rangle_{\tau_k} = \int_a^b p_i(\mu) p_j(\mu) d\tau_k(\mu), \quad 1 \leq i, j \leq k$$

induced by the step function

$$(3.8) \quad \tau_k(\mu) = \sum_{i=1}^k \tau_{k,i}^2 h(\mu - \theta_i^{(k)}),$$

where  $\tau_{k,i}$  is the first entry of the eigenvector  $s_i^{(k)}$ , namely

$$\tau_{k,i} = e_1^T s_i^{(k)}$$

(see [6, Chap. 2], [8]). As a result,  $\tau_k(\mu)$  and the distribution function  $\phi(\mu)$  in (3.5) have the same modified moments up to degree  $2k - 1$ , namely

$$(3.9) \quad \langle 1, p_i \rangle_\phi = \langle 1, p_i \rangle_{\tau_k}, \quad i = 1, \dots, 2k - 1.$$

In turn, by the following theorem due to Karlin and Shapley [11, Thm. 22.2], it follows that  $\tau_k(\mu)$  serves as a good approximation to  $\phi(\mu)$ .

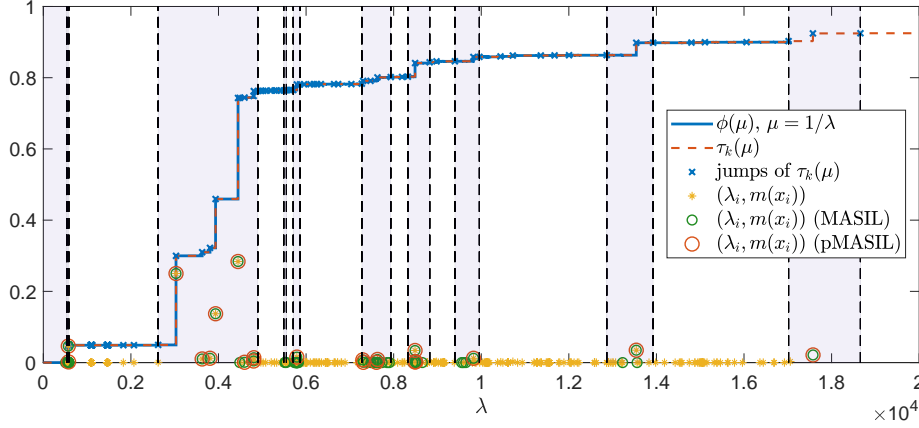
**THEOREM 3.2.** *If  $\phi(\mu)$  in (3.5) and  $\tau_k(\mu)$  in (3.8) have the same modified moments up to degree  $2k - 1$  then, if not identically zero, the difference function  $\phi(\mu) - \tau_k(\mu)$  has  $2k - 1$  sign changes.*

Thus if  $\tau_k$  and  $\phi$  do not coincide, the vertical and horizontal steps of  $\tau_k$  will intersect  $\phi$  exactly  $2k - 1$  times. This theory has been used, for instance, in estimating eigenvalue distribution [6], and in constructing polynomial preconditioners [7]. In our case, we use the step function  $\tau_k$ , obtained after  $k$  steps of the Lanczos algorithm, as an approximation to the cumulative mass participation sum, i.e.,

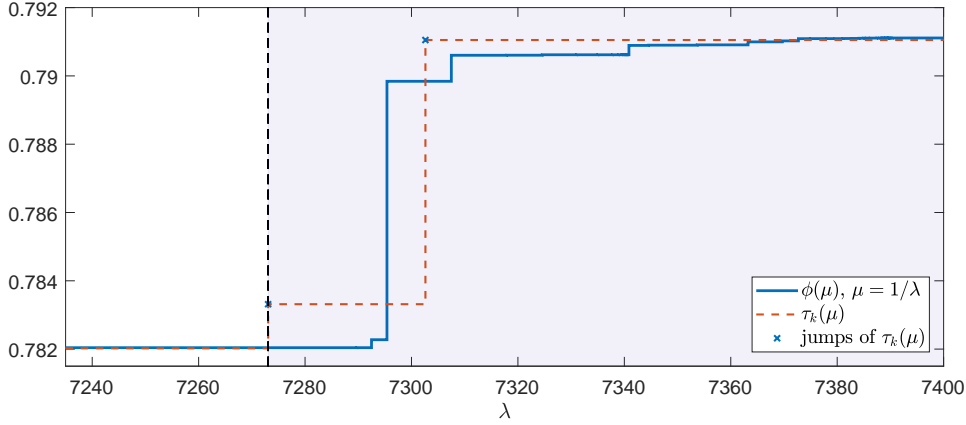
$$\tau_k(\mu_\ell) \approx \phi(\mu_\ell) = \sum_{i=1}^{\ell} m(x_i), \quad 1 \leq \ell \leq n.$$

This is illustrated in Figure 2(a)–(b) for a real structural engineering problem called **chilled** problem. On plot (a), the step functions  $\tau_k$  and  $\phi$  appear to coincide at least to the eye but plot (b), which is a closeup of plot (a) around  $\lambda = 7.3 \times 10^3$  shows that the step functions indeed intersect. We will return to Figure 2 at the end of section 3.2.

**3.2. A new shifting strategy.** The step function  $\tau_k(\mu)$  in (3.8) is readily available after  $k$  steps of Algorithm 2.1 with  $\sigma = 0$  and starting vector  $w = b$ . Intuitively, the eigenvalues corresponding to the eigenvectors with largest normalized mass participation should lie under tall and narrow steps of  $\phi(\lambda)$ , whereas short and wide steps would indicate intervals of eigenvalues corresponding to eigenvectors of negligible mass participation (see Figure 2 for illustration). To make this formal, denote by



(a) chilled problem in the  $x$ -direction,  $n = 93445$ ,  $k_{\max} = 200$ .



(b) Closeup the above stair graph.

FIG. 2. Step functions  $\phi(\mu)$  and  $\tau_k(\mu)$  for the **chilled** problem in the  $x$ -direction. The small yellow stars in (a) correspond to the smallest eigenvalues  $\lambda_i$  and their normalized mass participation factors  $m(x_i)$  such that  $\sum_i m(x_i) \geq 0.9$ . The green and red circles correspond to the pairs  $(\lambda_i, m(x_i))$  obtained by MASIL and pMASIL, respectively, and such that  $\sum_i m(x_i) \geq 0.9$ . The shading shows the intervals selected by the new shifting strategy.

$\Psi(c, d)$  the total mass participation of eigenvectors whose corresponding eigenvalues lie in the interval  $[c, d]$ , that is,

$$(3.10) \quad \Psi(c, d) = \sum_{j \in \mathcal{J}} m(x_j), \quad \mathcal{J} = \{j \in \{1, \dots, n\} : \lambda_j \in [c, d]\},$$

where  $\lambda_j = 1/\mu_j$  is the  $j$ th largest eigenvalue of  $K - \lambda M$ . Since by Theorem 3.2 the function  $\phi(\mu) - \tau_k(\mu)$  has exactly  $2k - 1$  sign changes, we have that for  $1 \leq i < j \leq k$ ,

$$(3.11) \quad \sum_{\ell=i+1}^{j-1} \tau_{k,\ell}^2 \leq \Psi\left(\frac{1}{\theta_i^{(k)}}, \frac{1}{\theta_j^{(k)}}\right) \leq \sum_{\ell=i-1}^{j+1} \tau_{k,\ell}^2.$$

We can now make use of the bounds in (3.11) to identify the union of intervals of smallest total length over which we are guaranteed to satisfy

$$(3.12) \quad \sum_{j=1}^{\ell} m(x_{i_j}) \geq \xi$$

for some  $\xi \in (0, 1)$ . Since we require that the total mass participation over that union of intervals is at least  $\xi$ , we must look at the lower bounds of (3.11). Let us denote by  $\gamma_{ij}$  the following ratios

$$\gamma_{ij} := \frac{\sum_{\ell=i+1}^{j-1} \tau_{k,\ell}^2}{\frac{1}{\theta_j^{(k)}} - \frac{1}{\theta_i^{(k)}}},$$

so that, by (3.11),

$$\gamma_{ij} \leq \frac{\Psi\left(\frac{1}{\theta_i^{(k)}}, \frac{1}{\theta_j^{(k)}}\right)}{\frac{1}{\theta_j^{(k)}} - \frac{1}{\theta_i^{(k)}}}.$$

Assuming that the eigenvalues are distinct and roughly evenly distributed, large  $\gamma_{ij}$  indicate high relative mass participation in the interval  $[1/\theta_i^{(k)}, 1/\theta_j^{(k)}]$ , and small  $\gamma_{ij}$  indicate low relative mass participation. Since the lower bound for  $\Psi(1/\theta_i^{(k)}, 1/\theta_{i+1}^{(k)})$  is 0, we look at intervals of the form  $[1/\theta_{i-1}^{(k)}, 1/\theta_{i+1}^{(k)}]$ . Thus to simplify the following discussion, define  $\gamma_i$  to be

$$\gamma_i := \gamma_{i-1, i+1} = \frac{\tau_{k,i}^2}{\frac{1}{\theta_{i+1}^{(k)}} - \frac{1}{\theta_{i-1}^{(k)}}}.$$

Suppose that after  $k$  steps of the Lanczos algorithm the first  $\ell$  Ritz pairs  $(\lambda_\nu, x_\nu)$ ,  $\nu = 1, \dots, \ell$  of  $K - \lambda M$  have converged. Let us denote by  $\xi_\ell$  the sum of their mass participation factors, that is,

$$\xi_\ell = \sum_{\nu=1}^{\ell} m(x_\nu).$$

To construct a union of intervals of smallest total length over which we are guaranteed to attain the remaining mass participation  $\xi - \xi_\ell$ , we pick the  $s$  largest  $\gamma_i$ , say  $\gamma_{i_1}, \dots, \gamma_{i_s}$ ,  $i_\nu \in \{\ell + 1, \dots, k - 1\}$  such that

$$(3.13) \quad \sum_{\nu=1}^s \tau_{k, i_\nu}^2 \geq \xi - \xi_\ell.$$

The wanted union of intervals is then  $\bigcup_{\nu=1}^s [1/\theta_{i_\nu-1}^{(k)}, 1/\theta_{i_\nu+1}^{(k)}]$ . By merging the neighbouring and the overlapping intervals we can construct a set of  $s' \leq s$  disjoint intervals

$$(3.14) \quad [1/\theta_{j_\nu}^{(k)}, 1/\theta_{j_\nu}^{(k)}], \quad \nu = 1, \dots, s'.$$

We then choose the shifts  $\sigma_\nu$  to be the midpoints of those intervals, namely,

$$\sigma_\nu = \frac{1/\theta_{j_\nu}^{(k)} + 1/\theta_{j_\nu}^{(k)}}{2}, \quad \nu = 1, \dots, s'.$$

We end the shift-and-invert Lanczos iteration with shift  $\sigma_\nu$  whenever the sum of the mass participation of the converged Ritz vectors with eigenvalues in  $[1/\theta_{i_\nu}^{(k)}, 1/\theta_{j_\nu}^{(k)}]$  attains the minimum in (3.11), i.e.,  $\sum_{\ell=i_\nu+1}^{j_\nu-1} \tau_{k,\ell}^2$ .

REMARK 3.3. *The inequality (3.13) may not hold if even  $s = k - \ell + 1$  although in practice, it is usually satisfied after a small number of steps  $k$ . If (3.13) does not hold then we can increase  $k$ , or reduce  $\xi$ .*

As an illustration, let us look at Figure 2(a). The jumps of  $\tau_k(\mu)$  correspond to the points  $(1/\theta_i^{(k)}, \tau_k(\theta_i^{(k)}))$ ,  $i = 1, \dots, k$  with  $k = k_{\max}$ . There are 34 more jumps outside of plot (a) corresponding to those  $\theta_i^{(k)}$  such that  $1/\theta_i^{(k)} > 2 \times 10^4$ . The plot shows as small yellow stars the normalized mass participation factors  $m(x_i)$  of the eigenvectors  $x_i$  corresponding to the  $\ell$  smallest eigenvalues  $\lambda_i$ , computed by the unshifted Lanczos algorithm and thus solving Problem 1.2. Although not visible on plot (a), the eigenvalues of the `chilled` problem are clustered: for example, the smallest eigenvalue is around 507, there are 47 eigenvalues in the interval [535, 538] and 25 in [570, 581]. The first shaded region from the left corresponds to the interval  $[1/\theta_1^{(k)}, 1/\theta_\ell^{(k)}]$  containing the converged Ritz values computed by the unshifted Lanczos algorithm. The remaining shaded regions correspond to the  $s' = 9$  disjoint intervals in (3.14). They define nine shifts for the `chilled` problem (there is a small interval just after the first interval that corresponds to the shift  $\sigma = 0$ ). The non shaded areas correspond to intervals containing eigenvalues associated with eigenvectors of negligible mass participation. These intervals are ignored by our approach.

**3.3. An algorithm for the approximate solution to Problem 1.1.** Given two  $n \times n$  matrices  $M \geq 0$  and  $K > 0$ , a spatial distribution vector  $b \in \text{range}(K^{-1}M)$ , a proportion  $\xi \in (0, 1)$ , and a maximum number of iterations  $k_{\max}$  our algorithm for the approximate solution to Problem 1.1 goes through the following steps.

**step 1** Call Algorithm 2.1 with  $\sigma = 0$ ,  $w = b$ , and the implicit filtering turned off.

Stop the Lanczos iterations at step  $k$  when either

- (a) the converged Ritz vectors  $x_j$ ,  $j = 1, \dots, \ell_0$  are such that  $\sum_{j=1}^{\ell_0} m(x_j) \geq \xi$ , or
- (b)  $k$  has reached a number of  $k_{\max}$  iterations (or larger if  $\sum_{j=1}^k \tau_{k,j}^2 \geq \xi$  is not satisfied).

If (a) holds then return the converged Ritz vectors  $x_j$ ,  $j = 1, \dots, \ell_0$  as an approximate solution to Problem 1.1. End the algorithm.

If (b) holds then

- if  $\|q_k\|_2 \leq \text{tol}\|q_1\|_2$  then save the converged Ritz vectors  $x_j$ ,  $j = 1, \dots, \ell_0$ , let  $\xi_{\ell_0} = \sum_{j=1}^{\ell_0} m(x_j)$  and proceed to **step 2** with the converged and unconverged Ritz pairs  $(1/\theta_i^{(k)}, x_i)$ ,  $i = 1, \dots, k$ .
- if  $\|q_k\|_2 > \text{tol}\|q_1\|_2$  then keep the computed Ritz pairs  $(1/\theta_i^{(k)}, \tilde{x}_i)$ ,  $i = 1, \dots, k$  for **step 2**. Apply implicit filtering as described in section 2.2(a) and continue the Lanczos iterations (Algorithm 2.1). Save the converged Ritz vectors  $x_j$ ,  $j = 1, \dots, \ell_0$ , let  $\xi_{\ell_0} = \sum_{j=1}^{\ell_0} m(x_j)$  and proceed to **step 2**.

**step 2** Construct a sequence of disjoint intervals  $[1/\theta_{i_\nu}^{(k)}, 1/\theta_{j_\nu}^{(k)}]$ ,  $\nu = 1, \dots, s'$ , as discussed in section 3.2, using the  $k$  converged and unconverged Ritz pairs from **step 1**. Compute the sequence of shifts  $\sigma_\nu = (1/\theta_{j_\nu}^{(k)} + 1/\theta_{i_\nu}^{(k)})/2$ ,  $\nu = 1, \dots, s'$ .

**step 3** For  $\nu = 1, \dots, s'$ , call Algorithm 2.1 with  $\sigma = \sigma_\nu$  and  $w = (K - \sigma_\nu M)^{-1} M \tilde{w}$ ,



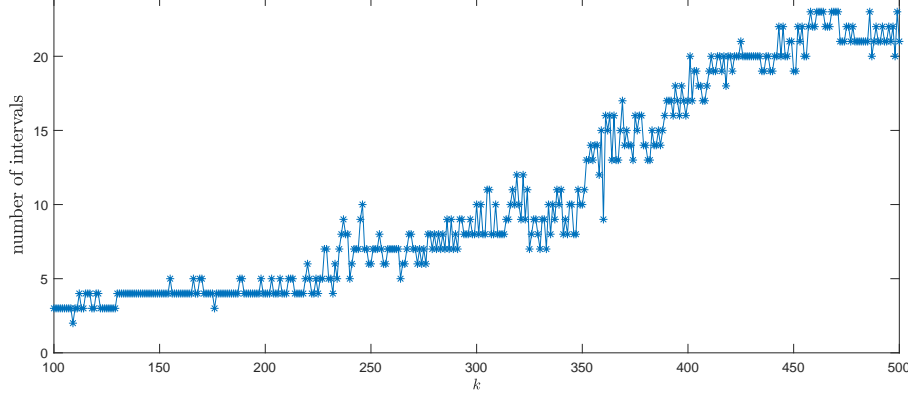


FIG. 3. Number of intervals  $s'$  in (3.14) as a function of the number  $k$  of unshifted Lanczos steps for the `local_modes` problem in the  $z$ -direction.

where  $\tilde{w}$  is the sum of the unconverged Ritz vectors from the previous step. Apply implicit filtering when  $\|q_k\|_2 > \text{tol}\|q_1\|_2$ . Stop the Lanczos iterations when either

- (i) the converged Ritz vectors  $x_j$ ,  $j = p+1, \dots, p+\ell_\nu$  with  $p = \ell_0 + \sum_{i=1}^{\nu-1} \ell_i$ , are such that

$$\xi_{\ell_\nu} := \sum_{j=p+1}^{p+\ell_\nu} m(x_j) \geq \sum_{\ell=i_\nu+1}^{j_\nu-1} \tau_{k,\ell}^2,$$

or

- (ii)  $\bar{\xi} := \xi_{\ell_0} + \sum_{j=1}^{\nu} \xi_{\ell_j} \geq \xi$ .

Stop the for loop when (ii) holds. Return the converged Ritz vectors  $x_j$ ,  $j = 1, \dots, p$  with  $p = \ell_0 + \sum_{i=1}^{\nu} \ell_i$  as an approximate solution to Problem 1.1. End the algorithm.

Implicit filtering cannot be used in the first call to Algorithm 2.1 in **step 1** since this would alter the starting vector and we could not apply our shifting strategy. In **step 1(b)** the condition  $\sum_{j=1}^k \tau_{k,j}^2 \geq \xi$  is usually satisfied after a small number of steps  $k$  (see remark 3.3). The choice of the maximum number of iterations  $k_{\max}$  is important. If  $k_{\max}$  is too small then the approximation  $\tau_k(\lambda)$  of  $\phi(\lambda)$  is too rough and leads to large intervals and shifts that are not close enough to eigenvalues with eigenvectors that have large mass participation. On the other hand, a too large  $k_{\max}$  can lead to unnecessary computations, in particular of converged eigenvectors with negligible mass participation but can also result in too many shifts being identified by our shifting strategy. This last point is illustrated in Figure 3 for a real engineering problem. The plot shows that the number of shifts (or intervals in (3.14)) determined by our shifting strategy increases as the number  $k$  of unshifted Lanczos steps increases. In practice, we found that taking  $k_{\max} = 200$  is a reasonable choice. For `tol` we choose `tol` =  $10^4$ , and we consider that the eigenpair  $(\lambda_j, x_j)$  has converged if  $\eta(\lambda_j, x_j) \leq nu$ , where  $u$  is the machine precision and  $\eta(\lambda_j, x_j)$  is defined in (2.12).

**4. Numerical experiments.** For our numerical experiments we used matrices  $M$  and  $K$ , and spatial distribution vectors  $b$  provided by **Arup Group Limited** that were constructed by the finite element software package Oasys GSA [15] from models

TABLE 2

List of test problems together with their size  $n$  and the number  $k_{\max}$  of unshifted Lanczos steps for SIL and the new method MASIL. The last four columns list the number of eigenvectors needed to satisfy the MPF condition (3.12) with  $\xi = 0.9$  for SIL and MASIL, and their purged versions pSIL and pMASIL. The number of shifts used by SIL and MASIL is provided inside brackets.

Problem (direction)	$n$	$k_{\max}$	SIL (shifts)	MASIL (shifts)	pSIL	pMASIL
local_modes (z)	51,348	200	977 (20)	367 (3)	463	275
ccnb (x)	57,152	200	115 (1)	94 (6)	61	56
chilled (x)	93,445	200	449 (2)	70 (9)	32	19
chilled (y)		100	107 (1)	14 (4)	6	5
chilled (y)		200	49 (0)	49 (0)	6	6
TT (y)	131,835	200	488 (6)	404 (9)	211	140
TT (z)		200	1173 (18)	994 (6)	566	413

of real structural engineering problems. These are listed in Table 2. Our numerical experiments are performed with MATLAB. As mentioned in the introduction,  $b$  is always associated with a direction, either  $x$ ,  $y$ , or  $z$ . So for a given problem, we will have different solutions depending on the chosen direction since they correspond to different spatial distribution vectors.

We compare the following approaches to solve our problem:

- SIL: shift-and-invert Lanczos algorithm with the Ericsson and Ruhe shifting strategy (3.2),
- MASIL: mass accumulating shift-and-invert Lanczos algorithm with the new shifting strategy described in section (3.3).

When no shifts are used, SIL and MASIL are identical, except for their starting vectors: SIL uses  $w = K^{-1}Mb$  as suggested in [20] and MASIL uses  $w = b$ .

In practice, the cumulative sum of the mass participation factors of the eigenvectors returned by SIL and MASIL is always slightly larger than the wanted proportion  $\xi$ . So we have the possibility to remove from the list of computed eigenvectors those with smallest mass participation so that the cumulative sum of the mass participation factors of the remaining eigenvectors is exactly  $\xi$  or just above. We refer to this small modification of SIL and MASIL as pSIL and pMASIL, respectively, where the “p” stands for extra purging step. An illustration of the purging step can be seen in Figure 2(a), where the pairs  $(\lambda_i, m(x_i))$  from MASIL are shown as small green circles and those kept by pMASIL are shown as large red circles.

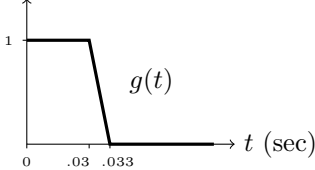
In Table 2, we compare the number of computed eigenvectors required to satisfy the mass participation condition (3.12) with  $\xi = 0.9$ . Some directions for the problems are excluded from the table. This is either because shifts were not employed (e.g. ccnb in the  $y$ - or the  $z$ -direction, where the 90% mass participation was reached in fewer than  $k_{\max}$  steps and no shift), or too many eigenvectors were required to achieve 90% mass participation, exceeding time and memory constraints (e.g., for the chilled problem in  $z$ -direction and  $\xi = 0.75 \ll 0.9$ , SIL returns 6959 eigenvectors whereas MASIL returns 4201 eigenvectors).

The new shifting strategy allows us to exclude intervals containing eigenvalues whose eigenvectors have a negligible mass participation and shift in the middle of intervals containing eigenvalues with eigenvectors of large mass participation. As a result, the number of eigenvectors returned by MASIL can be much smaller than that

returned by SIL. This is, for example, the case for the `local_modes` problem in the  $z$ -direction and the `chilled` problem in the  $x$ -direction. The number of shifts used by SIL or MASIL depends on the problem and the shifting strategy employed. Notice the large number of shifts employed by SIL for the `local_modes` problem in the  $z$ -direction and for the `TT` problem in the  $z$ -direction. For MASIL, the number of shifts needed is known before hand so if this number is too high then there is always the possibility to increase the value of  $k_{\max}$  or to see if there is a  $k < k_{\max}$  that leads to larger search intervals but fewer of them. We reported results for two different values of  $k_{\max}$  for the `chilled` problem in the  $y$ -direction. For  $k_{\max} = 100$ , SIL uses only one shift but returns 107 eigenvectors whereas MASIL uses 4 shifts but returns only 14 eigenvectors. No shifts are needed if we increase  $k_{\max}$  to 200.

Now comparing the number of eigenvectors returned by pSIL to that of SIL shows that a large proportion of the eigenvectors computed by SIL have a negligible mass participation and can be purged away. The last column in Table 2 shows that MASIL still returns many eigenvectors with negligible mass participation that can be removed while still maintaining the condition (3.12) but the reduction is not as drastic as for SIL. Note that the number of eigenvectors returned by pMASIL is lower or equal than those returned by pSIL.

Let us now look at the quality of the approximation  $\tilde{u}(t) = \sum_{j=1}^p v_j(t)x_j$  to the response  $u(t)$  of (1.1) when the  $p$  eigenvectors  $x_j$  are computed by SIL, MASIL, pSIL or pMASIL. For the time dependent external load  $f(t) = g(t)Mb$  we use

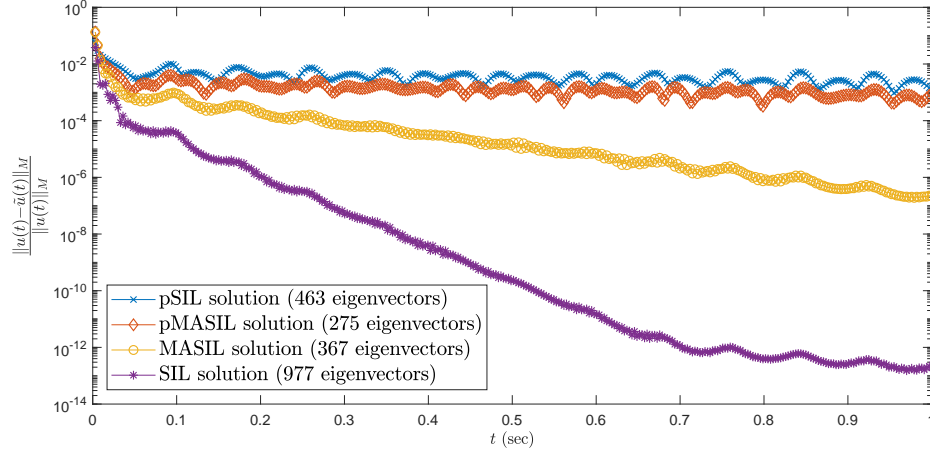
$$g(t) = \begin{cases} 1, & 0 \leq t \leq 0.03 \\ 11 - \frac{10^3}{3}t, & 0.03 \leq t \leq 0.033 \\ 0, & 0.033 \leq t \leq 1 \end{cases}$$


for the loading function  $g$  as suggested in [3, Fig. 2]. We employ 2% damping which consists of setting  $\zeta_j = 0.02$  in (2.4) for all  $j$ . For the relative error between the exact solution  $u(t)$  and its approximation  $\tilde{u}(t)$ , the analysis in section 2.1 tells us to expect

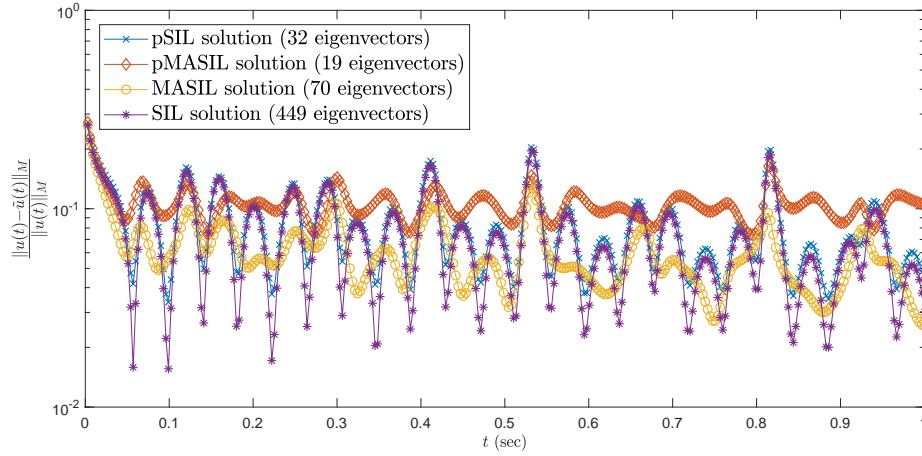
$$(4.1) \quad \frac{\|u(t) - \tilde{u}(t)\|_M}{\|u(t)\|_M} \lesssim \sqrt{1 - \xi} \approx 0.3.$$

We do not have access to  $u(t)$ . Therefore as a reference we use the solution obtained from SIL with  $\xi = 0.99$ . The relative errors for the approximate responses  $\tilde{u}(t)$  returned by SIL, pSIL, MASIL and pMASIL with  $\xi = 0.9$ , are shown in Figure 4 for the `local_modes` problem in the  $z$ -direction in (a) and for the `chilled` problem in the  $x$ -direction in (b). The relative errors agree with (4.1). Not surprisingly, the relative errors for the pSIL and pMASIL solutions are close to  $\sqrt{1 - \xi} \approx 0.3$ . We note that for the `chilled` problem, the MASIL solution with 14 eigenvectors is almost as good as the SIL solution which requires 101 eigenvectors, whereas for the `local_modes` problem pMASIL algorithm outperforms pSIL algorithm with significantly fewer eigenvectors.

Finally, we plot in Figure 5(a)-(b) the  $i$ th entry of the reference solution  $u(t)$  and its approximations obtained by SIL, pSIL, MASIL, and pMASIL for the `ccnb` and `TT` problems. We chose the index  $i$  for which the reference solution has the largest amplitude. For the `TT` problem in the  $y$ -direction in plot (a), all solutions agree with the reference solution even the pMASIL solution which uses significantly fewer



(a) `local_modes` problem in the  $z$ -direction,  $k_{\max} = 200$ .

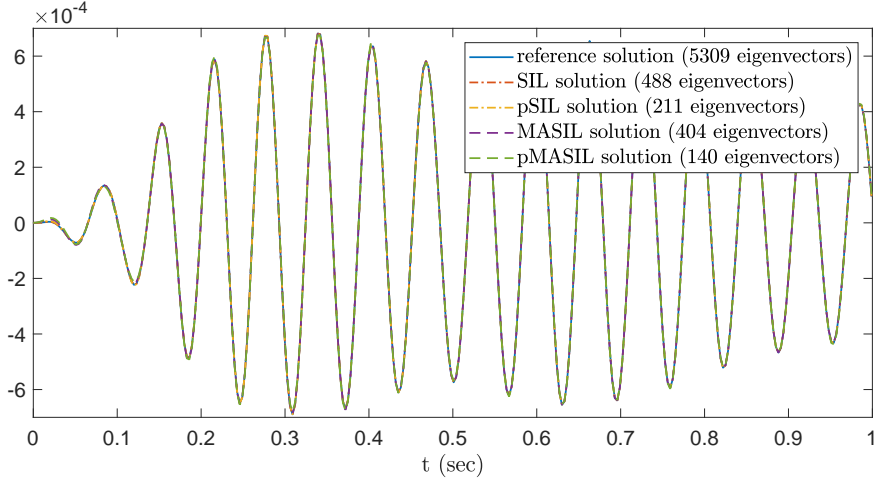


(b) `chilled` problem in the  $x$ -direction,  $k_{\max} = 200$ .

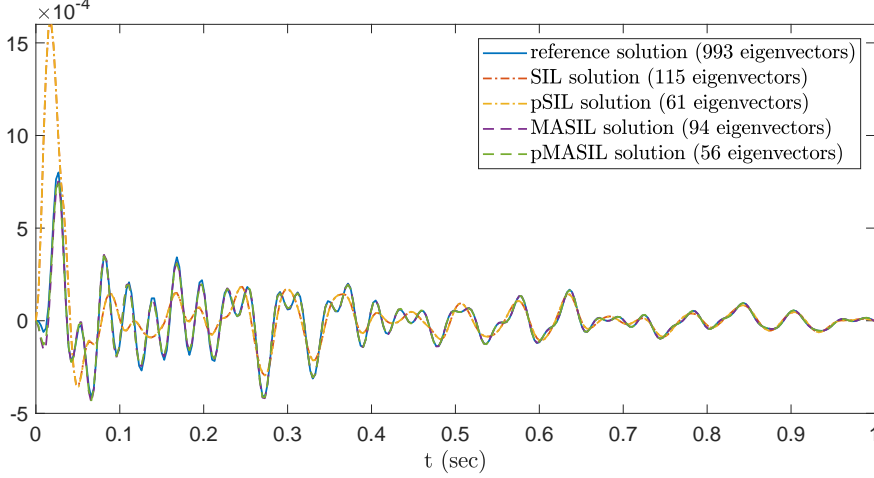
FIG. 4. Relative error between the reference solution  $u(t)$  and its approximation  $\tilde{u}(t)$  obtained from SIL, MASIL, pSIL, and pMASIL with a proportion of  $\xi = 90\%$  for the total mass participation.

eigenvectors. Although (4.1) holds for all the approximate solutions, the SIL and pSIL solutions do not agree as well with the reference solution for the `ccnb` problem in the  $x$ -direction in plot (b), whereas the MASIL and pMASIL solutions do, thereby suggesting a better selection of eigenvectors representing the solution.

**5. Conclusion.** We have shown that if the Lanczos process is applied to  $K^{-1}M$  with starting vector equal to the spatial distribution vector  $b$ , then the Lanczos polynomials are orthogonal with respect to the inner product induced by a step function  $\phi$  that coincides with the cumulative mass participation sum, that is, the quantity we are interested in to solve Problems 1.1–1.2. The Lanczos polynomials are also orthogonal with respect to an inner product induced by a step function  $\tau_k$  which, unlike  $\phi$ , is readily available at step  $k$  of the Lanczos process. The step function  $\tau_k$



(a) TT problem in the  $y$ -direction,  $i = 21943$ .



(b) **ccnb** problem in the  $x$ -direction,  $i = 14802$ .

FIG. 5. Plot of the  $i$ th entry of the response vector  $u(t)$  (reference solution) and its approximations obtained by SIL, pSIL, MASIL, and pMASIL.

offers an approximation to  $\phi$  and hence to the cumulative mass participation sum. The eigenvalues of  $K - \lambda M$  lie on the positive real line and we use  $\tau_k$  to identify intervals containing eigenvalues whose corresponding eigenvectors have non negligible mass participation as well as intervals containing eigenvalues whose eigenvectors have negligible contribution to the cumulative mass participation. We use this information to construct a sequence of shifts  $\sigma_1, \sigma_2, \dots, \sigma_p$  for the shift-and-invert Lanczos algorithm as well as a stopping criterion for the shift-and-invert Lanczos steps with shift  $\sigma_i$ ,  $i = 1, \dots, p$  so that (1.7) holds. The numerical experiments we performed on real engineering problems show that our approach computes up to 80% fewer eigenvectors and requires fewer shifts, on average, compared with the shifting strategy proposed by Ericsson and Ruhe.

**Acknowledgments.** The authors wish to thank Stefan Güttel for pointing out the connection between the cumulative mass participation and the measure  $\phi$  for which the Lanczos polynomials are orthogonal.

#### REFERENCES

- [1] ZHAOJUN BAI, JAMES W. DEMMEL, JACK J. DONGARRA, AXEL RUHE, AND HENK A. VAN DER VORST, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.
- [2] HARN C. CHEN AND ROBERT L. TAYLOR, *Using Lanczos vectors and Ritz vectors for computing dynamic responses*, Eng. Comput., 6 (1989), pp. 151–157.
- [3] J. M. DICKENS, J. M. NAKAGAWA, AND M. J. WITTBRODT, *A critique of mode acceleration and modal truncation augmentation methods for modal response analysis*, Computers & Structures, 62 (1997), pp. 985–998.
- [4] THOMAS ERICSSON, *A generalised eigenvalue problem and the Lanczos algorithm*, North-Holland Mathematics Studies, 127 (1986), pp. 95–119.
- [5] THOMAS ERICSSON AND AXEL RUHE, *The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems*, Math. Comp., 35 (1980), pp. 1251–1268.
- [6] BERND FISCHER, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2011.
- [7] BERND FISCHER AND ROLAND W. FREUND, *On adaptive weighted polynomial preconditioning for Hermitian positive definite matrices*, SIAM J. Sci. Comput., 15 (1994), pp. 408–426.
- [8] GENE H. GOLUB AND GERARD MEURANT, *Matrices, Moments and Quadrature with Applications*, Princeton University Press, Princeton, NJ, USA, 2010.
- [9] JIANMIN GU, ZHENG-DONG MA, AND GREGORY M. HULBERT, *A new load dependent Ritz vector method for structural dynamics analyses: quasi-static Ritz vectors*, Finite Elements in Analysis and Design, 36 (2000), pp. 261–278.
- [10] DESMOND J. HIGHAM AND NICHOLAS J. HIGHAM, *Structured backward error and condition of generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 493–512.
- [11] SAMUEL KARLIN AND LLOYD S. SHAPLEY, *Geometry of moment spaces*, American Mathematical Society, 1953.
- [12] KARL MEERBERGEN, *The Lanczos method with semi-definite inner product*, BIT Numerical Mathematics, 41 (2001), pp. 1069–1078.
- [13] BAHRAM NOUR-OMID AND RAY W. CLOUGH, *Dynamic analysis of structures using lanczos co-ordinates*, Earthquake Engrg. Struct. Dyn., 12 (1984), pp. 565–577.
- [14] BAHRAM NOUR-OMID, BERESFORD N. PARLETT, THOMAS ERICSSON, AND PAUL S. JENSEN, *How to implement the spectral transformation*, Math. Comp., 48 (1987), pp. 663–673.
- [15] OASYS LIMITED, *Oasys GSA*, available from <http://www.oasys-software.com/gsa>. Retrieved on May 10, 2018.
- [16] JR. ROY R. CRAIG, *Structural dynamics: an introduction to computer methods*, Wiley, New York, 1981.
- [17] YOUSEF SAAD, *Numerical Methods for Large Eigenvalue Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, revised ed., 2003. Updated edition of the work first published by Manchester University Press in 1992.
- [18] D. S. SCOTT, *The advantages of inverted operators in Rayleigh–Ritz approximations*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 68–75.
- [19] EDWARD L. WILSON, *A new method of dynamic analysis for linear and nonlinear systems*, Finite Elements in Analysis and Design, 1 (1985), pp. 21–23.
- [20] EDWARD L. WILSON, MING-WU YUAN, AND JOHN M. DICKENS, *Dynamic analysis by direct superposition of Ritz vectors*, Earthquake Engineering & Structural Dynamics, 10 (1982), pp. 813–821.