On the singular value decomposition of matrices generated by finite elements

Z. Drmač and K. Veselić

July 14, 2004

Abstract

We propose a method computing the eigenfrequences and eigenforms of a system consisting of multiply connected chains. The method has high relative accuracy and is based on a completely pivoted LR decomposition of the natural factor of the stiffness matrix. The decomposition preserves the structure of the matrix in spite of complete pivoting and is then followed by a one-sided Jacobi type eigenreduction.

To find the singular values σ_i of a matrix G with a *high relative accuracy* means to get the relative errors $\delta\sigma_i/\sigma_i$ small even then when σ_i differ greatly in their magnitude. The most important singular values are often the smallest ones; they are most menaced. In order to get this accuracy the matrix itself must determine well its singular values with high relative accuracy e.g. in the sense that the perturbation

$$G \mapsto G + \delta G \text{ with } |\delta G_{ij}| \le \varepsilon |G_{ij}| \tag{1}$$

implies the estimate

$$|\delta\sigma_i| \le C\varepsilon|\sigma_i| \tag{2}$$

for the corresponding singular values σ_i, $\sigma_i + \delta\sigma_i$, respectively. Here C is a reasonable constant. Algorithms of qr-type with this kind of accuracy exist for bidiagonal matrices ([3, 8]). In the more general case of totally acyclic matrices an algorithm with this type of accuracy is bisection ([4]). There exist classes of dense matrices for which (1) implies (2). Those are characterised by a "scaled" condition number. In fact, if

$$G = DB \quad , \quad B \text{ has full column rank } \quad , D \text{ diagonal} \tag{3}$$

(1) implies (2) with

$$C = \| |B| \|_2 \; \|B^{-1}\| \quad , \tag{4}$$

([1, 5, 14, 12]) A nearly optimal D carries the Euclidian columns of G on its diagonal ([13]). The matrices for which $cond(B)$ is a moderate are called well-scaled from right. Since the transposed (square) matrix G^T has the same singular values as G the right scaling $G = B_1 D_1$ has the same consequence. Here Jacobi methods are shown to have high relative accuracy ([5, 11, 6, 10]).

In this paper we are interested in matrices which are generated by differential equations e.g. in finite element formulation of vibrational systems.[1] A vibrational system is governed by the differential equation

$$K\ddot{x} + Mx = 0 \quad , \tag{5}$$

where K (stiffness matrix) and M (mass matrix) are real symmetric and positive definite. For an oscillator chain fixed at one end and free at the other K is tridiagonal:

$$K = \begin{pmatrix} k_1 & -k_1 & & & \\ -k_1 & k_1 + k_2 & -k_2 & & \\ & -k_2 & \ddots & \ddots & \\ & & \ddots & \ddots & -k_{n-1} \\ & & & -k_{n-1} & k_{n-1} + k_n \end{pmatrix} \quad , \tag{6}$$

[1] Some of the results in this paper were announced in [7].

where $k_1, \dots, k_n > 0$ are the stiffness parameters while

$$M = diag(m_1, \dots, m_n) \tag{7}$$

contains the masses. The eigenfrequencies ω_i of this system are the square roots of the eigenvalues of the problem

$$Kx = \lambda Mx \quad ,$$

i.e. the singular values of the matrix

$$G = DA_0\hat{D} \quad , \tag{8}$$

with $D = diag(\kappa_1, \dots, \kappa_n)$, $\kappa_i = \sqrt{k_i}$, $\hat{D} = M^{-1/2} = diag(\mu_1, \dots, \mu_n)$ and

$$A_0 = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{pmatrix} \quad , \tag{9}$$

an incidence matrix. Note that DA_0 is just the right Cholesky factor of K. Here G is bidiagonal and the accurate SVD is possible. Now (6) is but a very special case of matrices which describe vibrational systems and are of the form

$$K = \sum_{i,j} k_{ij} z_{ij} z_{ij}^T \quad , \quad M = diag(m_1, \dots, m_n) \quad , \tag{10}$$

where k_{ij} are positive parameters and z_{ij} are vectors, defined as follows.

$$z_{ij} = \begin{pmatrix} 0 & \cdots & 0 & \varepsilon & 0 & \cdots & 0 & -\varepsilon & 0 & \cdots & 0 \end{pmatrix} = (e_i^T - e_j^T)\varepsilon \, , \varepsilon = \pm 1 \tag{11}$$

$$z_{ii} = \begin{pmatrix} 0 & \cdots & 0 & \varepsilon & 0 & \cdots & 0 \end{pmatrix} = \varepsilon_i^T \, , \varepsilon = \pm 1 \, , \tag{12}$$

where e_i^T is the i-th canonical basis vector in R^m. These matrices are the subject of the present paper.

Obviously $K = G^T G$ and this leads to the singular value problem for the matrix G in (8) with

$$A_0 = \begin{pmatrix} z_{11}^T \\ z_{12}^T \\ \vdots \end{pmatrix} \tag{13}$$

which we call a *net incidence matrix*. An example of such a matrix is

$$A_0 = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 \end{pmatrix} \tag{14}$$

Typically our matrices will have full column rank, but this is not necessary for what follows here. To every such matrix we can associate an undirected graph which may contain links to the earth like in the 6. and 8. row above. Thus, the number of n of rows equals the number of links while the number m of columns equals the number of vortices in the graph.

We thus have to compute the SVD of G in (8) with $D = diag(d_1, \dots, d_n)$, $\hat{D} = diag(\hat{d}_1, \dots, \hat{d}_m)$, $d_i, \hat{d}_j > 0$ and A_0 is a net incidence matrix. Quite generally, it is advisable to compute the eigenvalues of a positive definite matrix K as the singular values of its factor, especially, if that factor can be assembled directly as in (13). The reason is that in this way we avoid squaring which may heavily harm small eigenvalues. The particular structure of a net matrix makes it possible to compute the eigensolution with much more accuracy. For a net matrix the singular values depend on the parameters κ_i, μ_i "with the condition one". Indeed, as an immediate consequence of the minimax principle the perturbation $d_i + \delta d_i, \hat{d}_i + \delta\hat{d}_i$ with $|\delta d_i| \le \varepsilon d_i, |\delta\hat{d}_i| \le \varepsilon\hat{d}_i$ implies (2) with $C \approx 1/2$. [2]. Thus, the 'well-behavedness' of a net matrix is not due to a scaled condition or to a sparsity pattern, but to a particular way of housekeeping its elements. An accurate algorithm must therefore work with this structure either the whole time or at least until the matrix is safely transformed to another more common well-behaved type. We present here an algorithm which performs this task. It begins by an LR decomposition with complete pivoting and then switches to a Jacobi like algorithm. The key tehnical result is that an LR decomposition with complete pivoting of a net matrix is fast and accurate and leads to suitably conditioned matrices L and R. The efficiency of this decomposition (it needs some n^2 operations, n the number of rows) is due to the fact that the net structure *is not destroyed during the Gaussian eliminations even in the presence of pivoting.* For the same reason the only arithmetic operations which are really performed are multiplications and divisions.[3] Once we have such a decomposition we can apply e.g. the Jacobi type algorithm proposed in [2] which is based on the eigensolver for indefinite symmetric matrices devised in [15, 11]. We now come to our central result.

Theorem 1 *Let G be a net matrix of the form (8) with A_0 from (13). Then any LR decomposition using Gaussian elimination with complete pivoting yields*

$$P_1 G P_2^T = LR\,, \tag{15}$$

where $P_{1,2}$ are permutation matrices, R is an upper triangular net matrix such that every $R_{i,i}$ absolutely dominates all elements in $R_{i:m,i:m}$ and L is a unit lower triangular matrix with all off-diagonals absolutely ≤ 1. Moreover, the net matrix property is preserved during the whole elimination. The only real-arithmetic operations of the algorithm are multiplications and divisions.

Proof. We consider the first elimination step. After the first pivoting permutation we have $G_{11} = \varepsilon_1 d_1 \hat{d}_1$ (note that permutations of rows or columns do not destroy the net property). If the rest of the first row or column vanishes everything is proved. Else there is just one non-vanishing $G_{1j} = -\varepsilon d_1 \hat{d}_j$ in the first row. The updating takes place for each i where the first column does not vanish and reads $\varepsilon_i d_i \hat{d}_1$. Now we have two possibilities:

(i) the original i, j element vanishes. Then the updated element reads

$$-\frac{-\varepsilon d_1 \hat{d}_j \varepsilon_i d_i \hat{d}_1}{\varepsilon d_1 \hat{d}_1} = \varepsilon_i d_i \hat{d}_j \tag{16}$$

Now, if the i-th row has one more non-vanishing element then it must read $-\varepsilon_i d_i \hat{d}_{j_1}$, for $j_1 \ne j$ and the net property is preserved. The same is true if there is no further non vanishing element in the i-th row. Moreover, by complete pivoting we have $d_1\hat{d}_1 \ge d_i\hat{d}_1, d_1\hat{d}_j$ and thus,

$$d_i \hat{d}_j \le d_1 \hat{d}_1\,, \tag{17}$$

(ii) the original i, j-element does not vanish, it then reads $-\varepsilon_i d_i \hat{d}_j$. Thus, the i-th row is proportional to the 1-st and after elimination the i-th row completely vanishes.

Thus, we have proved that the net structure goes over to the updated matrix $(G_{i,j})_{i,j>1}$. By iterating this we obtain a net matrix $R = D' B_1 \hat{D}'$ where D' is some part of D and $\hat{D}'$ a permutation of $\hat{D}$. By complete pivoting the number $\hat{d}_1 d_1$ dominates all other elements of G and thus the first column of L does not exceed one. In the same way by (17) the pivot element dominates all

[2]Note that the estimate for a general bidiagonal matrix is obtained just in transforming it to a form (8), the larger constant obtained there (it is linear in n) is due to the loss upon this transformation ([1, 3])

[3]A standard way to find the SVD of a matrix G would, of course, be to first find its QR decomposition and then continue by a bidiagonalisation or, if high relative accuracy is required, by a one-sided Jacobi as in [5]. This does not seem to be generally viable, as we show below.

elements of the updated matrix, which shows that R has the required properties as well. From the construction it is obvious that the only really used operations are multiplications and divisions. All cancellations (they occur whenever closed loops are present) are handled symbolically. QED.

The fact that there are no proper subtractions or additions is favorable for error estimates. The whole computation consists of combinatorial manipulations on the incidence matrix A_0 and updatings of the diagonals of D, $\hat{D}$.

Since the matrix R inherits the net property of G it must be of the form

$$R = D(I - S_n) \ , \quad D = diag(R_{11}, \ldots, R_{nn}) \ . \tag{18}$$

Here S_n is a strictly upper triangular non-negative matrix of order n with the row sums not exceeding one and D is diagonal.

Theorem 2 *Let S_n be as above. Then*

$$0 \le A_n = (I - S_n)^{-1} \le T_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ 0 & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

If, in addition, all row sums in the strict upper triangle of S_n are one, then the last column of A_n has all ones.

Proof. By $S_n^n = 0$ we have

$$A_n = I_n + S_n + \cdots + S_n^{n-1} \ . \tag{19}$$

We proceed by induction. Obviously

$$S_n = \begin{pmatrix} 0 & s \\ 0 & S_{n-1} \end{pmatrix} \ ,$$

where S_{n-1} is a matrix of the same type as S_n. From the obvious recursive formula

$$S_n^k = \begin{pmatrix} 0 & sS_{n-1}^{k-1} \\ 0 & S_{n-1}^k \end{pmatrix}$$

we obtain another

$$A_n = \begin{pmatrix} 1 & sA_{n-1} \\ 0 & A_{n-1} \end{pmatrix} \ , \quad n = 2, 3, \ldots \ , \tag{20}$$

For $n = 2$ we have

$$S_2 = \begin{pmatrix} 0 & \alpha \\ 0 & 0 \end{pmatrix} \ , \quad 0 \le \alpha \le 1$$

and

$$A_2 = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix} \ .$$

Thus, for $n = 2$ the theorem is true. Let the theorem hold for an integer $n - 1 \ge 2$. By the induction assumption we already have $A_{n-1} \le T_{n-1}$. By $s \ge 0$ and $A_{n-1} \ge 0$ we have

$$0 \le sA_{n-1} \le sT_{n-1} \le \left[\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}\right]$$

and thus $A_n \le T_n$. Let, finally, the sum of all rows in the strict upper triangle of S_n be one. Then the same is true of S_{n-1} so, again by the induction assumption, all last-column elements of A_n except possibly the first, are one. Since the sum of s is one the $1,n$-element of A_n as a scalar product of s and the last column of A_{n-1} is one as well. Q.E.D.

Since the spectral norm is a pointwise monotone on positive matrices we have

$$\|A_n\|_2 \le \|T_n\|_2 = \frac{1}{2\sin\frac{\pi}{4n+2}} \approx \frac{2n}{\pi} \tag{21}$$

Here the first inequality is attainable for $I - S_n = A_0$ from (9). We note a rather sharp estimate

$$\|T_n\|_2 \le \sqrt{\frac{n(n+1)}{2}} \ .$$

Thus, we have the following estimates for the extremal singular values of $I - S_n$

$$\sigma_{min}(I - S_n) \ge \sqrt{\frac{2}{n(n+1)}} \ ,$$

$$\sigma_{max} \le \sqrt{n} \ .$$

The matrix R from (18) is well-scaled from the left. The same estimates follow for matrices of type $n \times m$:

$$R = D \left[\begin{pmatrix} I - S_n & * \end{pmatrix} \right] \ , \ D = diag(R_{11}, \ldots, R_{nn}) \ . \tag{22}$$

A similar condition estimate is expected to hold for L (we have no proof as yet). There is, of course, an easy uniform estimate $condL \le const.2^n$, based on (19), but the experimental evidence strongly indicates a better, polynomial behaviour in n. This makes it possible to use the algorithm proposed in [2] upon the matrix LDR and to get at a highly accurate SVD.

The above result is related to another decomposition of G namely to the QR decomposition with column pivoting, a natural step towards an SVD of G.

First note that K is weakly diagonally dominant i.e it has the following properties

1. the diagonal is positive
2. the off-diagonal is non-positive
3. the sum of each column is non-negative

From this it follows that K is positive semidefinite. Without loss of generality we can assume that K is irreducible. An example of such K is (10). Conversely, it is immediately seen that all symmetric matrices with the properties 1. 2. 3. above are of the form (10).[4]Obviously, for an irreducible K the positive definiteness means that one of k_{ii} is positive.

It is important to note that the *Cholesky factor of K coincides with the QR factor of the matrix G from (13).*

We quote a well-known simple result (cf. e.g. [9]).

Theorem 3 *If K from (10) is positive definite then its Cholesky factor has properties 1. 2. 3. and is a monotonously increasing function of K (with respect to the elementwise ordering). The inverse K^{-1} is a positive matrix and is a monotonously decreasing function of K.*

Corollary 4 *Any K of the form (10) has a Cholesky factor or, equivalently, any corresponding*

$$G = diag(k_{11}, k_{12}, \ldots) \left[\begin{pmatrix} z_{11} \\ z_{12} \\ \vdots \end{pmatrix} \right]$$

has the QR-factor which is well scaled.

Note also that all assertions above are absolutely independent of the permutation of rows and columns in K or, equivalently, of columns in G. This additional freedom can be used, if the matrix K is substituted by $M^{-1/2}KM^{-1/2}$, M diagonal then instead of R above we obtain $RM^{-1/2}$ which will inherit the well-scaledness of R, if the diagonals of M are increasing. Otherwise the left scaled condition of $RM^{-1/2}$ may be spoiled. This necessitates the coresponding prepermutation of K,

[4] One should distignuish between the matrices (k_{ij}) and K, the latter having elements K_{ij}. Of course, $k_{ij} = -K_{ij}$ for $i \ne j$.

M or equivalently, G, M. (This permutation may be different from the one in standard column pivoting, applied to $GM^{-1/2}$.)

Thus, taking a properly pivoted QR decomposition of the net matrix G always leads to a well scaled matrix R with the same singular values. The problem is the stability of this QR decomposition - in the sense of our type of accuracy, of course. This stability can indeed be established for some very simple graphs like simple loop or simple chain. But even then the standard QR decomposition needs some modifications and it is not clear how to do it in general ([7]). Another case in which the QR decomposition is stable (see [6]) is the one in which M has a small condition number (e.g. if M is scalar). In the general case the well behavedness of G heavily relies on its net property which is, as a rule, *destroyed during the QR decomposition* in contrast to the aforementioned LR decomposition which preserves it. This is one more evidence that looking for mere unitarity does not help much, if high relative accuracy is required. Note that our LR factorisation makes it possible to solve, accurately as well as efficiently, the corresponding equilibrium equation

$$Kx = b \tag{23}$$

which can be written as

$$R^*L^*LRx = b \tag{24}$$

which essentially reduces to four times backward substitution. To see this it is enough to show how to solve the factorised system

$$L^*Ly = c\ , \tag{25}$$

where

$$L = \begin{pmatrix} L_1 \\ L_2 \end{pmatrix}$$

and L_1 is lower triangular and non-singular. The solution is obtained in two steps. The first step is the underdetermined system

$$L = \begin{pmatrix} L_1^* & L_2^* \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = c\,,$$

whose solution is $z_1 = -L_1^{-*}L_2^*z_2 + L^{-*}c$, z_2 free. The second step is the overdetermined system

$$L = \begin{pmatrix} L_1 \\ L_2 \end{pmatrix} y = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} -L_1^{-*}L_2^*z_2 + L_1^*b \\ z_2 \end{pmatrix}$$

The solution reads

$$y = -L_1^{-1}L_1^{-*}(-L_2^*L_2y + c) \tag{26}$$

This takes some $2m(n-m)+m^2$ operations, adding the count for solving with R and R^* we obtain total operational count of some

$$2m(n-m) + 2m^2 \tag{27}$$

operations. Since the number n of the links is usually at most twice or thrice the number m of the knots we have a count with quadratic growth in m, the number of unknowns. Since the count for the pivoted LR decomposition is of the same order this is very favorable.

Since all matrices entering (24) are well scaled we also expect a sort of high relative accuracy for the solution.

References

[1] Barlow, J., Demmel, J., Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices, SIAM J. Num. Anal. **27** 1988 762-791.

[2] Demmel J. Eisenstat S., Gu M., Slapničar I., Veselić K., Notes on computing the SVD with high relative accuracy preprint 1995

[3] Demmel, J., Kahan, W., Accurate Singular Values of Bidiagonal Matrices, SIAM J. Sci. Stat. Comp., **11** 1990 873-912.

[4] Demmel J., Gragg W., On computing accurate singular values and eigenvalues of acyclic matrices, Linear Algebra Appl. **85** (1993) 203-218.

[5] Demmel, J., Veselić, K., Jacobi's method is more accurate than QR, SIAM J. Matr. Anal. Appl. **13** 1992 1204-1245.

[6] Drmač, Z., Computing the singular value decomposition, Ph. D. dissertation Fernuniversität Hagen 1994

[7] Drmač, Z., Veselić, K., Note on the accuracy of the eigensolution of matrices generated by finite elements in SVD and signal processing (eds. M. Moonen and B. de Moor) Elsevier Amsterdam 1995 pp. 115-121.

[8] Fernando K., V., Parlett B., Accurate singular values and differential qd algorithms, Numer. Math. **67** (1994) 191-229.

[9] Horn R., J., Johnson Ch. R., Matrix Analysis, Cambridge University Press 1985.

[10] Mathias, R., Accurate Eigensystem Computations by Jacobi Methods IMA preprint, 1993.

[11] Slapničar, I., Accurate symmetric eigenreduction by a Jacobi method, Fernuniversität Hagen, Ph. D. Thesis 1992.

[12] Slapničar I., Veselić K., Perturbations of the eigenprojections of factorised Hermitian matrices, Linear Algebra Appl 218 (1995)

[13] van der Sluis, A., Condition numbers and equilibration of matrices, Numer. Math. **14** (1969) 14-23.

[14] Veselić, K., Slapničar, I., Floating-point perturbations of Hermitian matrices, Linear Algebra Appl **195** (1993) 81 - 116.

[15] Veselić, K., A Jacobi eigenreduction algorithm for definite matrix pairs, Numer. Math. **64** (1993) 241-269.

[16] Veselić, K., Slapničar, I., Floating-point perturbations of Hermitian matrices, Fernuniversität Hagen preprint 1992, to appear in Linear Algebra Appl.

[17] Barlow, J., Demmel, J., Computing accurate eigensystems of scaled diagonally dominant matrices, SIAM J. Numer Anal. **27** (1988) 762-791.

[18] Demmel, J., Veselić, K., Jacobi's method is more accurate than QR, Courant Institute preprint 1989, to appear in SIAM J. Matr. Anal. Appl.

[19] Golub, G. H., van Loan, Ch. F., Matrix computations, J. Hopkins University Press, Baltimore 1989.

[20] K. Veselić, Note on the accuracy of symmetric eigenreduction algorithms, Fernuniversität Hagen preprint 1993.