BIT Numerical Mathematics **44:** 181–188, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

STABILITY OF THE MATRIX FACTORIZATION FOR SOLVING BLOCK TRIDIAGONAL SYMMETRIC INDEFINITE LINEAR SYSTEMS*

JINXI ZHAO[1], WEIGUO WANG[2] and WEIQING REN[3,**]

[1] *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, P.R.China. email: jxzhao@nju.edu.cn*

[2] *Department of Mathematics, Ocean University of China, Qingdao, 266071, P.R.China*

[3] *Department of Mathematics, Nanjing University, Nanjing 210008, P.R.China*

Abstract.

In this paper, we propose a new factorization method for block tridiagonal symmetric indefinite matrices. We also discuss the stability of the factorization method. As a measurement of stability, an effective condition number is derived by using backward error analysis and perturbation analysis. It shows that under some suitable assumptions, the solution obtained by this factorization method is acceptable. Numerical results demonstrate that the factorization is stable if its condition number is not too large.

AMS subject classification (2000): 65F05, 65F35.

Key words: indefinite systems, backward stability, condition number, rounding-error analysis, quasidefinite matrix.

1 Introduction.

We are concerned in this paper with the stability of a factorization for the equation

$$Bx = b, \tag{1.1}$$

where B is a block tridiagonal symmetric indefinite matrix. The block tridiagonal matrix is of the form

$$B = \begin{pmatrix} K & -A & 0 \\ -A^\top & -C & G \\ 0 & G^\top & D \end{pmatrix}, \tag{1.2}$$

where K is a symmetric positive definite matrix of order m and A and G have dimensions $m \times n$ and $n \times l$ with full column rank, C and D are symmetric positive semidefinite. The matrix (1.2) is symmetric indefinite and nonsingular.

The special linear system (1.1) has many applications, e.g., the remaining (linearized) Euler–Lagrange equations can be obtained in the matrix form (1.2) (see [11]).

* Received January 2002. Revised October 2003. Communicated by Axel Ruhe.

** The work is partly supported by China NSF grant (♯60073029).

In [7], with a coupled DEM-FEM formulation, the necessity to use Lagrange multipliers will induce a three-field mixed system that, in the case of an impervious porous material with an incompressible pore fluid, has the form (1.2).

Direct methods with partial pivoting strategies for symmetric indefinite matrices producing a block diagonal matrix consisting of blocks of order 1 or 2 are given in [1, 2]. The stability analysis of the Cholesky factorization for quasidefinite systems is developed in [4]. Using the stability analysis for the factorization $L\widetilde{D}M^{\mathrm{T}}$ of nonsymmetric positive definite matrix, the conditions are derived under which the Cholesky factorization is stable for quasidefinite systems. Forsgren, Gill and Shinnerl give a rounding-error analysis of the symmetric indefinite factorization when applied to t-diagonally dominant systems (see [3]).

The factorization method was presented in [10, 12]. This method inherited the advantage of Cholesky factorization with small storage and low computation costs. For this we have the following theorem [10].

Theorem 1. *Given any symmetric indefinite matrix as* (1.2), *then we have*

$$B = LJL^{\mathrm{T}}, \tag{1.3}$$

$$L = \begin{pmatrix} L_{11} & & 0 \\ L_{21} & L_{22} & \\ 0 & L_{32} & L_{33} \end{pmatrix}, \qquad J = \begin{pmatrix} I_m & & \\ & -I_n & \\ & & I_l \end{pmatrix}, \tag{1.4}$$

where $L_{11} \in R^{m\times m}, L_{22} \in R^{n\times n}, L_{33} \in R^{l\times l}$ *are lower triangular matrix,* $L_{21} \in R^{n\times m}, L_{32} \in R^{l\times n}$. I_m, I_n *and* I_l *are identity matrix.*

L_{ij} can be easily calculated from the following matrix equations:

$$K = L_{11}L_{11}^{\mathrm{T}}, \tag{1.5}$$

$$-A^{\mathrm{T}} = L_{21}L_{11}^{\mathrm{T}}, \tag{1.6}$$

$$C + L_{21}L_{21}^{\mathrm{T}} = L_{22}L_{22}^{\mathrm{T}}, \tag{1.7}$$

$$-G^{\mathrm{T}} = L_{32}L_{22}^{\mathrm{T}}, \tag{1.8}$$

$$D + L_{32}L_{32}^{\mathrm{T}} = L_{33}L_{33}^{\mathrm{T}}. \tag{1.9}$$

In this paper, we discuss the backward rounding error and the stability of the factorization method (1.3) considering the special structures B and the perturbation E sufficiently. We will examine conditions under which the factorization method may be used reliably. In Section 2 we give a backward rounding-error analysis of (1.1) by means of the factorization method. The stability condition of the factorization is described in Section 3. In Section 4 we describe the sensitivity of the solution of this class of systems when the matrix is changed by perturbation. The analysis suggests an effective condition number for these equations and indicates that under suitable assumptions, the solution can be computed accurately. Finally, some numerical results are given to demonstrate the prediction.

2 Backward rounding error analysis.

In this section we give a backward error analysis of the solution of (1.1) by means of the factorization method. Throughout, we use the "standard model" of floating-point arithmetic in which the evaluation of an expression in floating-point arithmetic is denoted by $fl(\cdot)$, with

$$fl(a \circ b) = (a \circ b)(1+\delta), \quad |\delta| \le u, \circ = +, -, *, /$$

and

$$fl(x^{1/2}) = x^{1/2}(1+\varepsilon), \quad |\varepsilon| \le 1.00001u$$

(see, for example, Higham [6]). Here u is the unit round-off associated with the particular machine being used.

In the next lemma, the backward error analysis is presented.

LEMMA 2. *Let B be the same as that in* (1.2) *and assume* $\max\{m+3, n+3\}u \le 0.01$. *$\hat{L}$ is the computed factorization factor of B. Then we have*

$$B + \Delta B = \hat{L}J\hat{L}^{\mathrm{T}},$$

where ΔB satisfies

$$|\Delta B| \le \frac{(m+7)1.01u}{1-3.00002u}|\hat{L}||\hat{L}^{\mathrm{T}}|. \tag{2.1}$$

PROOF. Let $\hat{L}_{ij}$ denote the subblock of $\hat{L}$ in analogy with that in (1.4), which are computed from Equations (1.5)$\sim$(1.9) respectively. The backward error of the factorization is accumulated from each of the five steps. First, let

$$K + \Delta K = \hat{L}_{11}\hat{L}_{11}^{\mathrm{T}}$$

then we have the following inequality as it in [8],

$$|\Delta K| \le \frac{1.01(m+2)u}{1-3.00002u}|\hat{L}_{11}||\hat{L}_{11}^{\mathrm{T}}|. \tag{2.2}$$

Similar results can be established for the computable factors $\hat{L}_{21}$ and $\hat{L}_{22}$:

$$\begin{aligned} &-(A^{\mathrm{T}} + \Delta A^{\mathrm{T}}) = \hat{L}_{21}\hat{L}_{11}^{\mathrm{T}}, \\ &|\Delta A^{\mathrm{T}}| \le \frac{1.01(m+3)u}{1-2.02u}|\hat{L}_{21}||\hat{L}_{11}^{\mathrm{T}}| \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} &C + \Delta C + \hat{L}_{21}\hat{L}_{21}^{\mathrm{T}} = \hat{L}_{22}\hat{L}_{22}^{\mathrm{T}}, \\ &|\Delta C| \le \frac{\max\{m+7, n+6\}1.01u}{1-3.00002u}(|\hat{L}_{21}||\hat{L}_{21}^{\mathrm{T}}| + |\hat{L}_{22}||\hat{L}_{22}^{\mathrm{T}}|). \end{aligned} \tag{2.4}$$

Finally, we establish for the computed factors $\hat{L}_{32}$ and $\hat{L}_{33}$. Let $L_{22} = [\nu_{ij}], L_{32} = [\mu_{ij}], L_{33} = [h_{ij}]$, from (1.8) and (1.9), the μ_{ij} and h_{ij} are computed from

$$\mu_{ij} = -\left(g_{ji} - \sum_{p=1}^{j-1} \mu_{ip}\nu_{jp}\right)/\nu_{jj}, \quad i = 1, 2, \ldots, l, j = 1, 2, \ldots, n. \tag{2.5}$$

and

$$h_{ij} = \begin{cases} (D_{ii} + \sum_{k=1}^{n} \mu_{ik}^2 - \sum_{p=1}^{j-1} h_{ip}^2)^{1/2}, & i = j, \\ (D_{ij} + \sum_{k=1}^{n} \mu_{ik}\mu_{jk} - \sum_{p=1}^{j-i} h_{ip}h_{jp})/h_{jj}, & i > j, \end{cases} \tag{2.6}$$
$$i = 1, 2, \ldots, l, j = 1, 2, \ldots, i.$$

We obtain

$$|\Delta G^{\mathrm{T}}| \le \frac{1.01(m+3)u}{1 - 2.02u} |\hat{L}_{32}||\hat{L}_{22}^{\mathrm{T}}| \tag{2.7}$$

and

$$|\Delta D| \le \frac{\max\{n+7, l+6\}1.01u}{1 - 3.00002u} (|\hat{L}_{32}||\hat{L}_{32}^{\mathrm{T}}| + |\hat{L}_{33}||\hat{L}_{33}^{\mathrm{T}}|). \tag{2.8}$$

Then,

$$B + \Delta B = \hat{L} J \hat{L}^{\mathrm{T}}.$$

From (2.2)∼(2.8), we have

$$|\Delta B| \le \frac{\max\{n+7, l+7, l+6\}1.01u}{1 - 3.00002u} (|\hat{L}||\hat{L}^{\mathrm{T}}|). \tag{2.9}$$

For A and G are full column rank, i.e., $m \ge n \ge l$. (2.1) is derived immediately from (2.9). □

We consider the backward error resulting from solution of the triangular systems,

$$Lx = f \quad \text{and} \quad Uy = g,$$

where $L \in R^{n\times n}$ is lower triangular and $U \in R^{n\times n}$ upper triangular. It follows from [9] that the intermediate vectors $\hat{x}$ and $\hat{y}$ satisfy $(\hat{L} + \Delta L)\hat{x} = f$ and $(\hat{U} + \Delta U)\hat{y} = g$, where ΔL and ΔU have the same element-wise bound

$$|\Delta L| \le \frac{nu}{1 - nu}|\hat{L}| \quad \text{and} \quad |\Delta U| \le \frac{nu}{1 - nu}|\hat{U}|. \tag{2.10}$$

In the next theorem we show that the computed solution $\hat{x}$ of (1.1) is the exact solution of $(B+E)\hat{x} = b$, where E is an error matrix. The theorem is established by accumulating the backward error.

Theorem 3. *Let B be as in* (1.2)*, and $\hat{x}$ is the computed solution of $Bx = b$ by the factorization. Then $\hat{x}$ is the exact solution of $(B + E)\hat{x} = b$ with*

$$|E| \le \frac{3(m+n+l)u}{1 - (m+n+l)u} |\hat{L}||\hat{L}^{\mathrm{T}}|.$$

PROOF. Let $\hat{L}J\hat{L}^{\mathrm{T}}$ be the computed factorization of B. In the following we denote $J\hat{L}^{\mathrm{T}}$ as $\hat{U}$, an upper triangular matrix. To solve $Bx = b$, we must solve two triangular systems. Let the computed solution $\hat{Y}$ and $\hat{x}$ satisfy $(\hat{L} + \Delta L)\hat{y} = b$ and $(\hat{U} + \Delta U)\hat{x} = y$, with bounds for ΔL and ΔU given in (2.10). Hence $\hat{x}$ satisfies

$$(\hat{L} + \Delta L)(\hat{U} + \Delta U)\hat{x} = b,$$

i.e.,

$$(B + E)\hat{x} = b,$$

where

$$E = \Delta B + \Delta L\hat{U} + \hat{L}\Delta U + \Delta L\Delta U.$$

Ignoring elements of order u^2, from (2.1) and (2.10), we have

$$\begin{aligned} |E| &\le |\Delta B| + |\Delta L||\hat{U}| + |\hat{L}||\Delta U| + |\Delta L||\Delta U| \\ &\le \frac{(m+7)1.01u}{1 - 3.00002u}|\hat{L}||\hat{U}| + \frac{2(m+n+l)u}{1-(m+n+l)u}|\hat{L}||\hat{U}| \\ &\le \frac{3(m+n+l)u}{1-(m+n+l)u}|\hat{L}||\hat{L}^{\mathrm{T}}|. \end{aligned}$$

In the last inequality, we assume

$$(m + n + l) \ge 1.01(m + 7). \qquad \square$$

In the preceding Theorem 3, the error E is a symmetric error matrix with the special structure as B.

3 Stability of the factorization.

In the backward error analysis of solving $Bx = b$, it is shown that the computed solution $\hat{x}$ is the exact solution of the perturbed system $(B+E)\hat{x} = b$, where the size of E is bounded by an expression involving the size of the computed factor $\hat{L}$. Algorithms that produce $\hat{L}$ of sufficiently bounded size are therefore considered stable. Based on the factorization $L\widetilde{D}M^{\mathrm{T}}$ for nonsymmetric positive definite matrix $\bar{H}$, the stability analysis of the Cholesky factorization for symmetric quasidefinite matrix is given (see [4]). Here the nonsymmetric positive definite matrix is

$$\bar{H} = \begin{pmatrix} K & A \\ -A^{\top} & C \end{pmatrix}$$

and the symmetric quasidefinite matrix is

$$H = \begin{pmatrix} K & -A \\ -A^{\top} & -C \end{pmatrix} = \begin{pmatrix} K & A \\ -A^{\top} & C \end{pmatrix} \begin{pmatrix} I_n & \\ & -I_m \end{pmatrix}.$$

In the following, we consider directly the stability of the factorization as (1.2) without using the factorization $L\widetilde{D}M^{\mathrm{T}}$.

ASSUMPTION 4. *For some scalar γ of moderate size,*

$$\||\hat{L}||\hat{L}^{\mathrm{T}}|\|_F \le \gamma \||L||L^{\mathrm{T}}|\|_F.$$

From Equation (1.5)∼(1.9)*, we have*

$$\begin{aligned}
&\|L_{11}\|_F^2 = \mathrm{tr}(K),\\
&\|L_{21}\|_F^2 = \mathrm{tr}(A^{\mathrm{T}}K^{-1}A),\\
&\|L_{22}\|_F^2 = \mathrm{tr}(A^{\mathrm{T}}K^{-1}A + C),\\
&\|L_{32}\|_F^2 = \mathrm{tr}(G^{\mathrm{T}}(A^{\mathrm{T}}K^{-1}A + C)^{-1}G),\\
&\|L_{33}\|_F^2 = \mathrm{tr}(G^{\mathrm{T}}(A^{\mathrm{T}}K^{-1}A + C)^{-1}G + D).
\end{aligned}$$

Hence

$$\begin{aligned}
\||L||L^{\mathrm{T}}|\|_F &\le \|L\|_F\|L^{\mathrm{T}}\|_F\\
&= \mathrm{tr}(K) + \mathrm{tr}(C) + \mathrm{tr}(D) + 2\mathrm{tr}(A^{\mathrm{T}}K^{-1}A)\\
&\quad + 2\mathrm{tr}(G^{\mathrm{T}}(A^{\mathrm{T}}K^{-1}A + C)^{-1}G)\\
&= (1 + \omega(B))(\mathrm{tr}(K) + \mathrm{tr}(C) + \mathrm{tr}(D)),
\end{aligned}$$

where

$$\omega(B) = \frac{2\mathrm{tr}(A^{\mathrm{T}}K^{-1}A) + 2\mathrm{tr}(G^{\mathrm{T}}(A^{\mathrm{T}}K^{-1}A + C)^{-1}G)}{\mathrm{tr}(K) + \mathrm{tr}(C) + \mathrm{tr}(D)}. \tag{3.1}$$

Combining Theorem 3 *and Assumption* 4*, we have*

$$\|E\|_2 \le \|E\|_F \le \frac{3\gamma(m+n+l)u}{1-(m+n+l)}(1 + \omega(B))(\mathrm{tr}(K) + \mathrm{tr}(C) + \mathrm{tr}(D)). \tag{3.2}$$

THEOREM 5. *If B is as in* (1.2)*, the factorization $B = LJL^{\mathrm{T}}$ is stable if $\omega(B)$ is not too large.*

4 The condition number of the augmented system.

Let the computed solution $\hat{x}$ of $Bx = b$ satisfy the perturbed system

$$(B + E)x = b.$$

Then the usual sensitivity bound takes the form

$$\frac{\|x - \hat{x}\|}{\|x\|_2} \le \frac{\alpha}{1-\alpha}, \quad \text{where } \alpha = \frac{\|E\|_2}{\|B\|_2}\kappa_2(B).$$

From Assumption 4 and (3.2), we have

$$\alpha \le \frac{3\gamma(m+n+l)u}{1-(m+n+l)u}\,\frac{\mathrm{tr}(K) + \mathrm{tr}(C) + \mathrm{tr}(D)}{\|B\|_2}(1 + \omega(B))\kappa_2(B). \tag{4.1}$$

A simple calculation shows that

$$\mathrm{tr}(K) + \mathrm{tr}(C) + \mathrm{tr}(D) \le (m+n+l)\|B\|_2. \tag{4.2}$$

From (4.1) and (4.2), we obtain the following result showing that the relative error is bounded by quantity involving $(1 + \omega(B))\kappa_2(B)$.

Theorem 6. *Let B be symmetric indefinite as* (1.2)*, and $\hat{x}$ is the computed solution of $Bx = b$, then*

$$\frac{\|x - \hat{x}\|}{\|x\|_2} \leq \frac{\alpha}{1 - \alpha},$$

with

$$\alpha \leq \frac{3\gamma(m+n+l)^2 u}{1 - (m+n+l)u}\phi(B),$$

where $\phi(B) = (1 + \omega(B))\kappa_2(B)$, $\omega(B)$ is defined in (3.1)*.*

The theorem tells us that under Assumption 4, system (1.1) can be solved accurately as long as $\phi(B)$ is not too large. We therefore interpret $\phi(B)$ to be the effective condition number of generalized Cholesky factorization for solving the system (1.1).

5 Numerical experiments.

We have used MATLAB in PC to implement the factorization of $Bx = b$, where B has the form (1.2). The results confirm our prediction.

Example 1. Let $m = n = 10, l = 5; K = \text{diag}(k_{ii}), k_{11} = \varepsilon, k_{22}, \ldots, k_{mm}$ be positive and randomly generated by MATLAB, A and G have full column rank, C and D be zero matrices. We list some results corresponding to the example in Table 5.1.

Example 2. Let K, A, G be as in Example 1; let C and D be symmetric positive semidefinite random matrices. The results listed in Table 5.2.

The results show that for various ε, though the spectral condition number of B is almost invariant, the precision of the computed solution is different. This confirms our predictions: (i) for small $\omega(B)$, the factorization is stable; (ii) for small $\phi(B)$, the computed solution $\hat{x}$ is reliable.

Note that small $\omega(B)$ and $\phi(B)$ are both sufficient, but not necessary, for ensuring an accurate factorization factor and the reliable solution of $Bx = b$. For large $\omega(B)$ and $\phi(B)$, it is difficult to draw any conclusion.

Table 5.1: Numerical results for the Example 1.

ε	$\kappa_2(B)$	$\omega(B)$	$\phi(B)$	$\|B - LJL^{\mathrm{T}}\|_F$	$\frac{\|x-\hat{x}\|_2}{\|x\|_2}$
10^2	7.1860e+02	6.4539e+00	5.3564e+03	1.6834e−14	9.0382e−14
10^0	4.2444e+01	3.3010e+02	1.4435e+03	4.7234e−15	5.0320e−15
10^{-2}	4.2236e+01	2.3321e+02	9.8921e+03	3.3934e−14	3.8136e−14
10^{-4}	4.2241e+01	1.9735e+04	8.3365e+05	3.0106e−12	1.9367e−12
10^{-6}	4.2241e+01	1.9699e+06	8.3209e+07	2.8257e−10	1.8954e−10
10^{-8}	4.2241e+01	1.9698e+08	8.3207e+09	2.7447e−08	2.2862e−08

Table 5.2: Numerical results for the Example 2.

ε	$\kappa_2(B)$	$\omega(B)$	$\phi(B)$	$\|B - LJL^{\mathrm{T}}\|_F$	$\frac{\|x-\hat{x}\|_2}{\|x\|_2}$
10	1.6485e+02	2.2222e+00	5.3116e+02	5.0286e−15	4.7348e−15
10^0	1.6504e+02	2.8013e+00	6.2736e+02	6.3152e−15	4.0532e−15
10^{-2}	1.6515e+02	2.3192e+01	3.9954e+03	4.5329e−14	3.1294e−14
10^{-4}	1.6516e+02	2.0582e+03	3.4009e+05	5.7003e−12	2.3596e−12
10^{-6}	1.6516e+02	2.0556e+05	3.3939e+07	4.5493e−10	2.7224e−10
10^{-8}	1.6516e+02	2.0556e+07	3.3949e+09	3.8521e−08	3.3042e−08

Acknowledgement.

The authors are indebted to Prof. Åke Björck and Prof. Axel Ruhe for their comments which improved the formulation and interpretation of several results. Also the authors wish to thank the anonymous referees for their valuable suggestions that improved the paper.

REFERENCES

1. J. R. Bunch and B. N. Parlett, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.
2. J. R. Bunch, *Partial pivoting strategies for symmetric matrices*, SIAM J. Numer. Anal., 11 (1974), pp. 521–528.
3. A. Forsgren, P. E. Gill, and J. R. Shinnerl, *Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 187–211.
4. P. E. Gill, M. A. Saunders, and J. R. Shinnerl, *On the stability of Cholesky factorization for symmetric quasidefinite systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 35–46.
5. G. H. Golub and C. F. Van Loan, *Matrix Computation*, 3rd ed., The John Hopkins University Press, Baltimore, 1996.
6. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM Publications, Philadelphia, PA, 1996.
7. H. Modaressi, *A diffuse element-finite element technique for transient coupled analysis*, Int. J. Numer. Methods Eng., 39 (1996), pp. 3809–3838.
8. Wei-qing Ren, *The solution of augmented systems*, manuscript, 1996.
9. J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 2nd ed., Springer-Verlag, New York, 1993.
10. Weiguo Wang, *A matrix factorization method for solving block tridiagonal symmetric indefinite linear systems*, J. Nanjing Univ. Math. Biqu., 14 (1997).
11. S. L. Weissman, *High-accuraty low-order three-dimensional brick elements*, Int. J. Numer. Methods Eng., 39 (1996), pp. 2337–2361.
12. Jinxi Zhao, *The generalized Cholesky factorization method for saddle point problem*, Appl. Math. Comput., 92 (1998), pp. 49–58.