Modified Cholesky Algorithms: A Catalog with New Approaches

Haw-ren Fang* Dianne P. O'Leary†

August 8, 2006

Abstract

Given an $n \times n$ symmetric possibly indefinite matrix A, a modified Cholesky algorithm computes a factorization of the positive definite matrix $A+E$, where E is a correction matrix. Since the factorization is often used to compute a Newton-like downhill search direction for an optimization problem, the goals are to compute the modification without much additional cost and to keep $A + E$ well-conditioned and close to A.

Gill, Murray and Wright introduced a stable algorithm, with a bound of $\|E\|_2 = O(n^2)$. An algorithm of Schnabel and Eskow further guarantees $\|E\|_2 = O(n)$. We present variants that also ensure $\|E\|_2 = O(n)$.

Moré and Sorensen and Cheng and Higham used the block LBL^T factorization with blocks of order 1 or 2. Algorithms in this class have a worst-case cost $O(n^3)$ higher than the standard Cholesky factorization. We present a new approach using an LTL^T factorization, with T tridiagonal, that guarantees a modification cost of at most $O(n^2)$.

1 Introduction

Modified Cholesky algorithms are widely used in nonlinear optimization to compute Newton-like directions. Given a symmetric possibly indefinite $n \times n$ matrix A approximating the Hessian of a function to be minimized, the goal is to find a positive definite matrix $\hat{A} = A + E$, where E is small. The search direction Δx is then computed by solving the linear system $(A + E)\Delta x = -g(x)$ where $g(x)$ is the gradient of the function to be minimized. This direction is Newton-like and guaranteed to be downhill. Four objectives to be achieved when computing E are listed below [5, 15, 16].

*Department of Computer Science, University of Maryland, A.V. Williams Building, College Park, Maryland 20742, USA. The work of this author was supported by National Science Foundation Grant CCF 0514213.

†Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, A.V. Williams Building, College Park, Maryland 20742, USA. The work of this author was supported by National Science Foundation Grant CCF 0514213 and Department of Energy Grant DEFG0204ER25655.

Objective 1. If A is sufficiently positive definite, $E = 0$.

Objective 2. If A is not positive-definite, $\|E\|$ is not much larger than $\inf\{\|\Delta A\| : A + \Delta A \text{ is positive definite}\}$ for some reasonable norm.

Objective 3. The matrix $A + E$ is reasonably well-conditioned.

Objective 4. The cost of the algorithm is only a small multiple of n^2 higher than that of the standard Cholesky factorization, which takes $\frac{1}{3}n^3 + O(n^2)$ flops ($\frac{1}{6}n^3 + O(n^2)$ multiplications and $\frac{1}{6}n^3 + O(n^2)$ additions).

Objective 1 ensures that the fast convergence of Newton-like methods on convex programming problems is retained by the modified Cholesky algorithms. Objective 2 keeps the search direction close to Newton's direction, while Objective 3 implies numerical stability when computing the search direction. Objective 4 makes the work in computing the modification small relative to the work in factoring a dense matrix.

There are two classes of algorithms, motivated by the simple case when A is diagonal: $A = \mathrm{diag}(d_1, d_2, \ldots, d_n)$. In this case we can make $A + E = \mathrm{diag}(\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_n)$ positive definite by choosing $\hat{d}_k := \max\{|d_k|, \delta\}$ for $k = 1, \ldots, n$, where $\delta > 0$ is a preset small tolerance. We call such a modification algorithm a Type-I algorithm. Alternatively, we can choose $\hat{d}_k := \max\{d_k, \delta\}$. We call modified Cholesky algorithms of this kind Type-II algorithms. In both types of algorithms, δ must be kept small to satisfy Objective 2, but large enough to satisfy Objective 3. Early approaches were of Type-I [10, Chapter 4][13], whereas more recently Type-II algorithms have prevailed [5, 15, 16].

There are three useful factorizations of a symmetric matrix A as $PAP^T = LXL^T$, where P is a permutation matrix for pivoting, and L is unit lower triangular:

1. the LDL^T factorization, where X is diagonal[1].
2. the LBL^T factorization, where X is block diagonal with block order 1 or 2, [2, 3, 4].
3. the LTL^T factorization, where X is a tridiagonal matrix, and the off-diagonal elements in the first column are all zero [1, 14].

Existing modified Cholesky algorithms use either the LDL^T factorization [10, Chapter 4][15, 16] or the LBL^T factorization [5, 13]. We present new modified LDL^T factorizations and an approach via the LTL^T factorization.

In all we review five modified Cholesky algorithms in the literature and give five new ones, each of which depends on a modification tolerance parameter $\delta > 0$. Satisfaction of Objectives 1–3 is measured by bounds, discussed in detail as the algorithms are introduced, and referenced in Table 1, where the new algorithms are in boldface.

Table 2 lists some notation used in this paper. We use $\mathrm{diag}(a_1, \ldots, a_n)$ to denote the diagonal matrix formed by $a_1, \ldots, a_n$, and $\mathrm{Diag}(A)$ to denote the diagonal matrix formed by the diagonal of matrix A.

[1] If D is nonnegative, it is the Cholesky factorization in the LDL^T form.

Table 1: Satisfaction of the four objectives for Modified Cholesky algorithms.

Algorithm	*Type*	δ	*Obj. 1*	*Obj. 2*	*Obj. 3*	*Obj. 4*
LDL^T:						
GMW81	I	ϵ_M	(4)	(3)	(25)	$O(n^2)$
SE90	II	$\tau\eta$	(32)	(12) (13)	(33)	$O(n^2)$
SE99	II	$\bar{\tau}\eta$	(32)	(12) (19)	(33)	$O(n^2)$
GMW-I	I	ϵ_M	(32)	(22)	(25)	$O(n^2)$
GMW-II	II	$\bar{\tau}\eta$	(32)	(24)	(25)	$O(n^2)$
SE-I	I	$\bar{\tau}\eta$	(32)	(29) (19)	(33)	$O(n^2)$
LBL^T:						
MS79	I	ϵ_M	(34)	(35)	(41)	$\leq O(n^3)$
CH98	II	$\sqrt{u}\|A\|_\infty$	(34)	(36)	(42)	$\leq O(n^3)$
LTL^T:						
LTL^T**-MS79**	I	ϵ_M	(45)	(46)	(48)	$O(n^2)$
LTL^T**-CH98**	II	$\bar{\tau}\eta$	(45)	(47)	(49)	$O(n^2)$

The organization of this paper is as follows. Section 2 presents the modified LDL^T factorizations in the literature and Section 3 presents our new variants. Section 4 describes modified LBL^T factorizations in the literature. Section 5 gives our new LTL^T algorithms. Section 6 summarizes the results of our computational tests. Concluding remarks are given in Section 7.

2 Modified LDL^T Algorithms

Given a LDL^T factorization of a symmetric matrix A, a naïve way to modify A to be positive definite is by making nonpositive elements in the diagonal matrix D positive. However, the factorization of A may fail to exist (e.g., $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$), and even if it does, this method fails to meet Objective 2. For example, in the LDL^T factorization

$$A = \begin{bmatrix} \epsilon & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{1}{\epsilon} & 1 \end{bmatrix} \begin{bmatrix} \epsilon & 0 \\ 0 & -\frac{1}{\epsilon} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{\epsilon} \\ 0 & 1 \end{bmatrix} = LDL^T,$$

the modification is unbounded when $\epsilon \to 0^+$. A 3×3 example is given in [10, Chapter 4].

In a *modified LDL^T algorithm* for a positive definite $\hat{A} = A + E$, $E = \mathrm{diag}(\delta_1, \delta_2, \ldots, \delta_n)$ is typically diagonal, with $\delta_k \geq 0$ computed at the kth step of the factorization, for $k = 1, \ldots, n$ Denote the Schur complement at the kth step by $A_k = \begin{bmatrix} a_k & c_k^T \\ c_k & \bar{A}_k \end{bmatrix}$ for $k = 1, \ldots, n$, where $a_k \in R$ and c_k is a column vector of $n-k$ elements. Initially $A_1 := A$. The factorization can be computed

Table 2: Notation.

Symbol	*Description*
ϵ_M	machine epsilon
u	unit roundoff, $\epsilon_M/2$
A	an $n \times n$ symmetric matrix
ξ	maximum magnitude of off-diagonal elements of A
η	maximum magnitude of diagonal elements of A
$\lambda_i(A)$	ith smallest eigenvalue of A
$\lambda_{\min}(A)$	smallest eigenvalue of A
$\lambda_{\max}(A)$	largest eigenvalue of A
τ	$\sqrt[3]{\epsilon_M}$
$\bar{\tau}$	$\sqrt[3]{\epsilon_M^2}$

by setting

$$L(k+1:n, k:k) := \frac{c_k}{a_k + \delta_k},\ D(k,k) := a_k + \delta_k \text{ and } A_{k+1} := \bar{A}_k - \frac{c_k c_k^T}{a_k + \delta_k} \quad (1)$$

for $k = 1, \ldots, n-1$. The challenge is to determine δ_k to satisfy the four objectives. All the algorithms in Sections 2 and 3 follow this model, although Schnabel and Eskow [15, 16] originally formulated their algorithm in LL^T form.

We may incorporate a diagonal pivoting strategy in our algorithms, symmetrically interchanging rows and columns at the kth step to ensure that $|a_k| \geq |A_k(j,j)|$ (pivoting on the diagonal element of maximum magnitude) or $a_k \geq A_k(j,j)$ (pivoting on the element of maximum value) for $j = 1, \ldots, n-k$. The resulting modified LDL^T factorization is in the form

$$P(A+E)P^T = LDL^T = \bar{L}\bar{L}^T, \quad (2)$$

where P is a permutation matrix.

Gill and Murray introduced a stable algorithm in 1974 [9]. It was subsequently refined by Gill, Murray, and Wright in 1981 [10, Chapter 4]; we call it GMW81 hereafter. Schnabel and Eskow introduced another modified LDL^T algorithm in 1990 [15]. It was subsequently revised in 1999 [16]. We call these algorithms SE90 and SE99, respectively.

2.1 The GMW81 Algorithm

GMW81 determines δ_k in (1) by setting

$$a_k + \delta_k = \max\{\delta, |a_k|, \frac{\|c_k\|_\infty^2}{\beta^2}\}$$

for $k = 1, \ldots, n$, where $\beta > 0$ and the small tolerance $\delta > 0$ are preset. We set $\delta := \epsilon_M$ (machine epsilon) as is common in the literature [5, 16].

The rationale behind GMW81 is that β becomes a bound on the magnitude of the off-diagonal elements in the lower triangular matrix $\bar{L}$ of the Cholesky factorization in (2). The challenge is to choose β such that $\|E\|_2$ is well-controlled and Objective 1 is satisfied. The correction E is bounded by

$$\|E\|_2 \leq (\frac{\xi}{\beta} + (n-1)\beta)^2 + 2(\eta + (n-1)\beta^2) + \delta =: f(\beta), \tag{3}$$

where η and ξ are the maximum magnitudes of the diagonal and off-diagonal elements of A, respectively. Note that since E is diagonal, its 1-norm, 2-norm and ∞-norm are the same.

The overall extra cost of GMW81 relative to the standard Cholesky factorization is $O(n^2)$, so Objective 4 is satisfied. Now we consider Objective 2. The minimum of (3) is

$$\min_{\beta} f(\beta) = 2\xi(\sqrt{n^2-1} + n - 1) + 2\eta + \delta \leq 4n\xi + 2\eta + \delta,$$

which is attained with $\beta^2 = \frac{\xi}{\sqrt{n^2-1}}$ for $n > 1$.

A diagonal pivoting strategy is used in GMW81. The pivot is chosen as the maximum magnitude diagonal element[2].

To satisfy Objective 1, we let $\beta^2 \geq \eta$, so that $E = 0$ if A is sufficiently positive definite [9]. More precisely, $E = 0$ if $\beta^2 \geq \eta$ and

$$\lambda_{\min}(A) \geq \delta. \tag{4}$$

Therefore, β is chosen by

$$\beta^2 := \max\{\eta, \frac{\xi}{\sqrt{n^2-1}}, \epsilon_M\} \tag{5}$$

for $n > 1$. Substituting this into (3), we obtain $\|E\|_2 = O(n^2)$.

2.2 The SE90 Algorithm

SE90 was inspired by a lemma related to the Gerschgorin circle theorem [12, page 344]. We define the i-th Gerschgorin radius and circle by

$$R_i(A) := \sum_{j=1, j\neq i}^{n} |a_{ij}| \text{ and } C_i(A) := \{z : |z - a_{ii}| \leq R_i(A)\}$$

for $1 \leq i \leq n$ and recall that Gerschgorin showed that the eigenvalues of A are contained in the union of the circles $C_i(A)$. Therefore, we could perturb A to be positive semidefinite by setting $\delta_k := \max\{0, -a_{kk} + R_k(A)\}$ for $k = 1, \ldots, n$ in (1). The modification δ_k can be reduced by the following lemma.

[2]Alternatively, we could pivot on the maximum diagonal element, but pivoting on the maximum magnitude usually gives a smaller $\|E\|_2$ in our experiments.

Lemma 1 *Given a symmetric matrix* $A = \begin{bmatrix} a & c^T \\ c & \bar{A} \end{bmatrix} \in R^{n\times n}$*, suppose we add a perturbation* $\delta \geq \{0, -a + \|c\|_1\}$ *to* a*, so that* $a + \delta \geq \|c\|_1$*. The resulting Schur complement*[3] *is* $\hat{A} := \bar{A} - \frac{cc^T}{a+\delta}$*. Then* $C_i(\hat{A}) \subseteq C_{i+1}(A)$ *for* $i = 1, \ldots, n-1$*.*

Proof This proof is a condensed version of that in [15]. Let $\bar{a}_{ij}$ and $\hat{a}_{ij}$ denote the (i, j) entries of $\bar{A}$ and $\hat{A}$ respectively for $1 \leq i, j < n$. Also denote $c = [(c)_1, (c)_2, \ldots, (c)_{n-1}]^T$. For $1 \leq i < n$,

$$R_{i+1}(A) - R_i(\hat{A}) = (R_{i+1}(A) - R_i(\bar{A})) + (R_i(\bar{A}) - R_i(\hat{A})).$$

The difference between $R_{i+1}(A)$ and $R_i(\bar{A})$ is $|(c)_i|$. In addition, the ith column of $\bar{A} - \hat{A}$ is $\frac{(c)_i c}{a+\delta}$, whose 1-norm minus $\frac{(c)_i^2}{a+\delta}$ is the upper bound for $|(R_i(\bar{A}) - R_i(\hat{A}))|$. Therefore,

$$\begin{aligned} R_{i+1}(A) - R_i(\hat{A}) &\geq |(c)_i| - \frac{|(c)_i|(\|c\|_1 - |(c)_i|)}{a+\delta} \\ &= |(c)_i|(1 - \frac{\|c\|_1}{a+\delta}) + \frac{(c)_i^2}{a+\delta} \\ &\geq \frac{(c)_i^2}{a+\delta} = \bar{a}_{ii} - \hat{a}_{ii} \geq 0. \end{aligned}$$

This means that the Gerschgorin circles contract, and the contraction of each circle is no less than the perturbation of the circle center. Therefore, $C_i(\hat{A}) \subseteq C_{i+1}(A)$ for $i = 1, \ldots, n-1$. ■

Following this result, we can make $A + E$ positive semidefinite by setting $\delta_k := \max\{0, -a_k + \|c_k\|_1\}$ in (1) for $k = 1, \ldots, n$. Note that $a_k - \|c_k\|_1$ is the lower endpoint of the Gerschgorin circle $C_1(A_k)$. Repeatedly applying Lemma 1, we obtain $\delta_k \leq \max\{0, -a_{kk} + R_k(A)\}$ for $k = 1, \ldots, n$. Taking the maximum of these values and zero, we define

$$\bar{G} := \max\{0, \max\{-a_{kk} + R_k(A) : k = 1, \ldots, n\}\}.$$

Then $\|E\|_2 \leq \bar{G} \leq \eta + (n-1)\xi$, where η and ξ are the maximum magnitudes of the diagonal and off-diagonal elements of A, respectively. However, this naïve method may fail to satisfy Objective 1.

To satisfy Objective 1, SE90 consists of two phases. The *2-phase strategy* was also presented in [8]. Phase 1 performs steps of the standard Cholesky factorization (i.e., without perturbation, $\delta_k := 0$), as long as all diagonal elements of the next Schur complement are sufficiently positive. The pseudo-code is given in Algorithm 1.

SE90 uses the tolerance $\delta := \tau\eta$, where η is the maximum magnitude of the diagonal elements of A, and $\tau = \sqrt[3]{\epsilon_M}$. Therefore, in Phase 1,

$$\text{Diag}(A_k) \geq \tau\eta I_{n-k+1} \tag{6}$$

[3]Note that the ith row/column of $\bar{A}$ corresponds to the $(i+1)$st row/column of A.

Algorithm 1 Phase 1 of a 2-Phase Strategy.

{Given a symmetric $A \in R^{n\times n}$ and a small tolerance $\delta > 0$.}
$A_1 := A$, $k := 1$
Pivot on the maximum diagonal element of A_1.
{Denote $A_k = \begin{bmatrix} a_k & c_k^T \\ c_k & \bar{A}_k \end{bmatrix}$, then $\mathrm{Diag}(\bar{A}_k) \le a_k I_{n-k}$ after pivoting.}
if $a_1 \ge \delta$ **then**
  while $\mathrm{Diag}(\bar{A}_k - \frac{c_k c_k^T}{a_k}) \ge \delta I_{n-k}$ and $k < n$ **do**
    $A_{k+1} := \bar{A}_k - \frac{c_k c_k^T}{a_k}$
    $k := k + 1$
    Pivot on maximum diagonal of A_k.
  end while
end if

for $k = 1, \ldots, \min\{n, K+1\}$, where K is the number of steps in Phase 1. If A is sufficiently positive definite, then $K = n$ and the factorization completes without using Phase 2. Otherwise, Phase 1 ends when setting $\delta_{K+1} := 0$ results in A_{K+2} having a diagonal element less than δ. It is not hard to see that

$$\hat{\eta} \le \eta \text{ and } \hat{\xi} \le \xi + \eta, \tag{7}$$

where $\hat{\eta}$ and $\hat{\xi}$ (and η and ξ) are the maximum magnitudes of the diagonal and off-diagonal elements of A_{K+1} (and A), respectively [15].

In Phase 2, δ_k is determined by

$$\delta_k := \max\{\delta_{k-1}, -a_k + \max\{\|c_k\|_1, \tau\eta\}\} \le G + \tau\eta, \tag{8}$$

for $k = K+1, \ldots, n-2$, where G is the maximum of zero and the negative of the lowest Gerschgorin endpoint of A_{K+1}. For the case $K = 0$, we set $\delta_0 := 0$. The rationale for $\delta_k \ge \delta_{k-1}$ is because increasing δ_k up to δ_{k-1} does not increase $\|E\|_2$ at this point and may possibly reduce the subsequent δ_i for $k < i \le n$. This *nondecreasing strategy* can be applied to virtually all modified Cholesky algorithms with modifications confined to the diagonal.

In experiments, Schnabel and Eskow [15] obtained a smaller value of $\|E\|_2$ when using special treatment for the final 2×2 Schur complement A_{n-1}, setting

$$\delta_{n-1} = \delta_n := \max\{\delta_{n-2}, -\lambda_1(A_{n-1}) + \max\{\frac{\tau(\lambda_2(A_{n-1}) - \lambda_1(A_{n-1}))}{1-\tau}, \tau\eta\}\} \tag{9}$$

$$\le G + \frac{2\tau}{1-\tau}(G + \eta), \tag{10}$$

where $\lambda_1(A_{n-2})$ and $\lambda_2(A_{n-2})$ are the smaller and larger eigenvalues of A_{n-2}, respectively. The last inequality holds because

$$-\lambda_1(A_{n-1}) \le G \text{ and } \lambda_2(A_{n-1}) - \lambda_1(A_{n-1}) \le 2(G + \eta).$$

In (9), δ_{n-1} and δ_n are chosen to obtain the bound

$$\kappa_2(A_{n-1} + \delta_n I_2) \leq \frac{1 + (\tau/(1-\tau))}{\tau/(1-\tau)} = \frac{1}{\tau}, \tag{11}$$

where $I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Finally, by (8) and (10),

$$\|E\|_2 \leq G + \frac{2\tau}{1-\tau}(G + \eta). \tag{12}$$

If $K = 0$, then $G \leq \eta + (n-1)\xi$. By (7), if $K > 0$, then

$$G \leq (n - K - 1)(\xi + \eta). \tag{13}$$

In either case, $\|E\|_2 = O(n)$. Recall that with GMW81, $\|E\|_2 = O(n^2)$.

Diagonal pivoting is also used in SE90, as well as the later SE99 algorithm. The analysis above does not rely on the pivoting, but pivoting reduces $\|E\|_2$ empirically. In Phase 1, the pivot is chosen as the largest diagonal entry as shown in Algorithm 1.

In Phase 2, one may choose the pivot with the largest lower endpoint of the Gerschgorin circle in the current Schur complement. This provides the least modification at the current step. In other words, after diagonally interchanging rows and columns, $G_1(A_k) \geq G_i(A_k)$ for $k = K{+}1, \ldots, n{-}2$ and $i = 1, \ldots, n{-}k{+}1$, where $G_i(A_k) = a_{ii} - R_i(A_k)$ is the lower endpoint of the ith Gerschgorin circle $C_i(A_k)$. However, computing all $G_i(A_k)$ in Phase 2 takes $\frac{(n-K)^3}{3}$ additions and fails to satisfy Objective 4. The proof of Lemma 1 shows

$$\hat{a}_{ii} - R_i(\hat{A}) \geq \bar{a}_{ii} - R_{i+1}(A) + |(c)_i|(1 - \frac{\|c\|_1}{a + \delta})$$

for $i = 1, \ldots, n{-}1$. Therefore,

$$G_i(A_{k+1}) \geq G_{i+1}(A_k) + |(c_k)_i|(1 - \frac{\|c_k\|_1}{a_k + \delta_k})$$

for $k = 1, \ldots, n{-}1$ and $i = 1, \ldots, n{-}k$. Using this fact, we recursively compute the lower bounds of these Gerschgorin intervals by

$$\hat{G}_i(A_{k+1}) := \hat{G}_{i+1}(A_k) + |(c_k)_i|(1 - \frac{\|c_k\|_1}{a_k + \delta_k})$$

for $k = 1, \ldots, n{-}1$ and $i = 1, \ldots, n{-}k$. The base cases are $\hat{G}_i(A_1) := G_i(A)$ for $i = 1, \ldots, n$. Computing these estimated lower endpoints $\hat{G}_i(A_{k+1})$ for pivoting takes $2(n{-}K)^2$ additions and $\frac{1}{2}(n{-}K)^2$ multiplications. Hence Objective 4 is satisfied.

2.3 The SE99 Algorithm

Although SE90 has a better *a priori* bound on $\|E\|_2$ than GMW81, there are matrices for which SE90 gives an inordinately large $\|E\|_2$. These matrices are generally close to being positive definite. The SE99 algorithm [16], was developed to remedy this. In SE99, condition (6) is *relaxed* into the following two conditions that possibly increase the number of Phase 1 pivots:

$$\mathrm{Diag}(\bar{A}_k - \frac{c_k c_k^T}{a_k}) \geq -\mu\eta I_{n-k}$$

for some $0 < \mu \leq 1$, and

$$\mathrm{Diag}(A_k) \geq -\mu a_k I_{n-k+1}.$$

Schnabel and Eskow suggested $\mu = 0.1$ for SE99 [16]. The pseudo-code of the *relaxed 2-phase strategy* is given in Algorithm 2.

Algorithm 2 Relaxed Phase 1 of a 2-Phase Strategy.

{Given a symmetric $A \in R^{n\times n}$, $\delta > 0$ and $0 < \mu \leq 1$.}
$\eta := \max_{1\leq i\leq n} |A_{ii}|$
if $\mathrm{Diag}(A) \geq -\mu\eta I_n$ **then**
  $A_1 := A$, $k := 1$
  Pivot on the maximum diagonal element of A_1.
  {Denote $A_k = \begin{bmatrix} a_k & c_k^T \\ c_k & \bar{A}_k \end{bmatrix}$, then $\mathrm{Diag}(\bar{A}_k) \leq a_k I_{n-k}$ after pivoting.}
  while $a_k \geq \delta$ and $\mathrm{Diag}(A_k) \geq -\mu a_k I_{n-k+1}$ and $\mathrm{Diag}(\bar{A}_k - \frac{c_k c_k^T}{a_k}) \geq -\mu\eta I_{n-k}$ and $k < n$ **do**
    $A_{k+1} := \bar{A}_k - \frac{c_k c_k^T}{a_k}$
    $k := k + 1$
    Pivot on maximum diagonal of A_k.
  end while
end if

In SE99, $\delta := \bar{\tau}\eta$, where $\bar{\tau} = \sqrt[3]{\epsilon_M^2}$, is smaller than $\tau = \sqrt[3]{\epsilon_M}$ in SE90, potentially keeping $\|E\|$ smaller. In Phase 1, there is no perturbation, so $\delta_k = 0$ for $k = 1, 2, \ldots, K$, with K the number of steps in Phase 1. The modification in Phase 2 turns out to be

$$\delta_k := \max\{\delta_{k-1}, -a_k + \max\{\|c_k\|_1, \bar{\tau}\eta\}\} \leq G + \bar{\tau}\eta, \tag{14}$$

where G is the negative of the lowest Gerschgorin endpoint of A_{K+1}. Recall that we set $\delta_0 := 0$ and δ_k is nondecreasing, so that δ_k is nonnegative.

Since small negative numbers are allowed on the diagonal in Phase 1, two changes have to be made. First, we need to check whether $a_k \geq \delta$ at each step, as shown in Algorithm 2, whereas it is not required in Algorithm 1. Second, it

is possible that SE99 moves into Phase 2 at the last step (i.e., the number of steps in Phase 1 is $K = n - 1$). In such a case,

$$\delta_n := \max\{0, -a_n + \max\{\frac{-\tau}{1-\tau}a_n, \bar{\tau}\eta\}\} \leq G + \frac{\tau}{1-\tau}G + \bar{\tau}\eta. \tag{15}$$

Similar to (9) in SE90, the special treatment in SE99 for the final 2×2 Schur complement in Phase 2 is

$$\begin{aligned} \delta_{n-1}=\delta_n &:= \max\{\delta_{n-2}, -\lambda_1(A_{n-1}) + \max\{\frac{\tau(\lambda_2(A_{n-1}) - \lambda_1(A_{n-1}))}{1-\tau}, \bar{\tau}\eta\}\} \\ &\leq G + \frac{2\tau}{1-\tau}(G+\eta). \end{aligned} \tag{16}$$

By (14), (15) and (16), we obtain

$$\|E\|_2 \leq G + \frac{2\tau}{1-\tau}(G+\eta). \tag{17}$$

Although (17) for SE99 looks the same as (12) for SE90, the bound on G in (17) is different for $0 < K < n$. Due to relaxing, the bounds (7) on $\hat{\eta}$ and $\hat{\xi}$ are replaced by

$$\hat{\eta} \leq \eta \text{ and } \hat{\xi} \leq \xi + (1+\mu)\eta, \tag{18}$$

where $\hat{\eta}$ and $\hat{\xi}$ (and η and ξ) are the maximum magnitudes of the diagonal and off-diagonal elements of A_{K+1} (and A), respectively. Therefore, if $0 < K < n$,

$$G \leq (n-K-1)(\xi + (1+\mu)\eta) + \mu\eta. \tag{19}$$

Recall that K is the number of steps in Phase 1, and SE99 potentially has more steps staying in Phase 1 than SE90.

The pivoting strategy used in SE99 is the same as that in SE90. Note that the bound on $\|E\|_2$ in (17) for SE99 is independent of the pivoting strategy applied, and so is (12) for SE90.

3 New Modified LDL^T Algorithms

This section presents three variants of the LDL^T algorithms, GMW-I, GMW-II and SE-I, and illustrates their performance.

Experiments used a laptop with an Intel Celeron 2.8GHz CPU using IEEE standard arithmetic with machine epsilon $\epsilon_M = 2^{-52} \approx 2.22 \times 10^{-16}$. We measure the size of E by the ratios

$$r_2 = \frac{\|E\|_2}{|\lambda_{\min}(A)|} \text{ and } r_F = \frac{\|E\|_F}{(\sum_{\lambda_i(A)<0} \lambda_i(A)^2)^{1/2}}. \tag{20}$$

Note that assuming $\lambda_{\min}(A) < 0$, the denominators are the norms of the least modification to make the matrix positive semidefinite.

The random matrices in our experiments are of the form $Q\Lambda Q^T$, where $Q \in R^{n\times n}$ is a random orthogonal matrix computed by the method of G. W. Stewart [17], and $\Lambda \in R^{n\times n}$ is diagonal with uniformly distributed random eigenvalues in $[-1, 10000]$, $[-1, 1]$ or $[-10000, -1]$. For the matrices with eigenvalues in $[-1, 10000]$, we impose the condition that there is at least one negative eigenvalue.

3.1 The GMW-I Algorithm

The GMW81 algorithm, a Type-I algorithm, satisfies $\|E\|_2 = O(n^2)$, whereas SE90 and SE99 further guarantee $\|E\|_2 = O(n)$, as shown in (12) and (17), respectively. Schnabel and Eskow [15] pointed out that the 2-phase strategy can drop the bound on $\|E\|_2$ of GMW81 to be $O(n)$. In our experiments, we note that incorporating the 2-phase strategy into GMW81 introduces difficulties similar to those for SE90, and again relaxing provides the rescue.

We denote by GMW-I the algorithm that uses the Relaxed Phase 1 of SE99 with Phase 2 defined by GMW81. Denote the number of steps in Phase 1 by K. Then $\delta_k = 0$ for $k = 1, 2, \ldots, K$. Instead of (3), the bound on $\|E\|_2$ is

$$\|E\|_2 \le (\frac{\hat{\xi}}{\beta} + (n-K-1)\beta)^2 + 2(\hat{\eta} + (n-K-1)\beta^2) + \delta, \tag{21}$$

where $\hat{\eta}$ and $\hat{\xi}$ are the maximum magnitudes of the diagonal and off-diagonal elements of A_{K+1}, respectively. Now we do not need $\beta^2 \ge \hat{\eta}$ to satisfy Objective 1, so β is chosen as the minimizer of (21),

$$\beta^2 = \max\{\frac{\hat{\xi}}{\sqrt{(n-K)^2 - 1}}, \epsilon_M\}$$

for $n - K > 1$. Substituting this into (21) and invoking (18), we obtain

$$\|E\|_2 \le 4(n-K)\hat{\xi} + 2\hat{\eta} + \delta \le 4(n-K)(\xi + (1+\mu)\eta) + 2\eta + \delta = O(n), \tag{22}$$

where we ignore the extreme case $\beta^2 = \epsilon_M$.

We still use $\delta := \epsilon_M$ as in GMW81 and set $\mu = 0.75$ in the relaxed 2-phase strategy since it is an empirically good value for the GMW algorithms. (Recall that $\mu = 0.1$ for SE99.) Pivoting reduces $\|E\|_2$ in the original GMW81 algorithm; we pivot on the maximum element instead of the maximum magnitude element in Phase 2, because on average the resulting $\kappa_2(A + E)$ is smaller in our experiments. We call our variant GMW-I.

Figure 1 shows our experimental result. The GMW-I algorithm performed well, but for the random matrices with eigenvalues in $[-10000, -1]$, the $\|E\|_2$ was a few times larger than in the original GMW81. Nevertheless, in practical optimization problems, negative definite Hessian matrices rarely occur, and indefinite Hessian matrices are usually close to being positive definite. The nondecreasing strategy was also tried. For the random matrices with eigenvalues in $[-1, 1]$ and $[-10000, -1]$, the nondecreasing strategy substantially reduced $\kappa_2(A + E)$ but roughly doubled $\|E\|_F$ (though with $\|E\|_2$ comparable). Note that the bound on $\|E\|_2$ in (3) is preserved with the nondecreasing strategy.

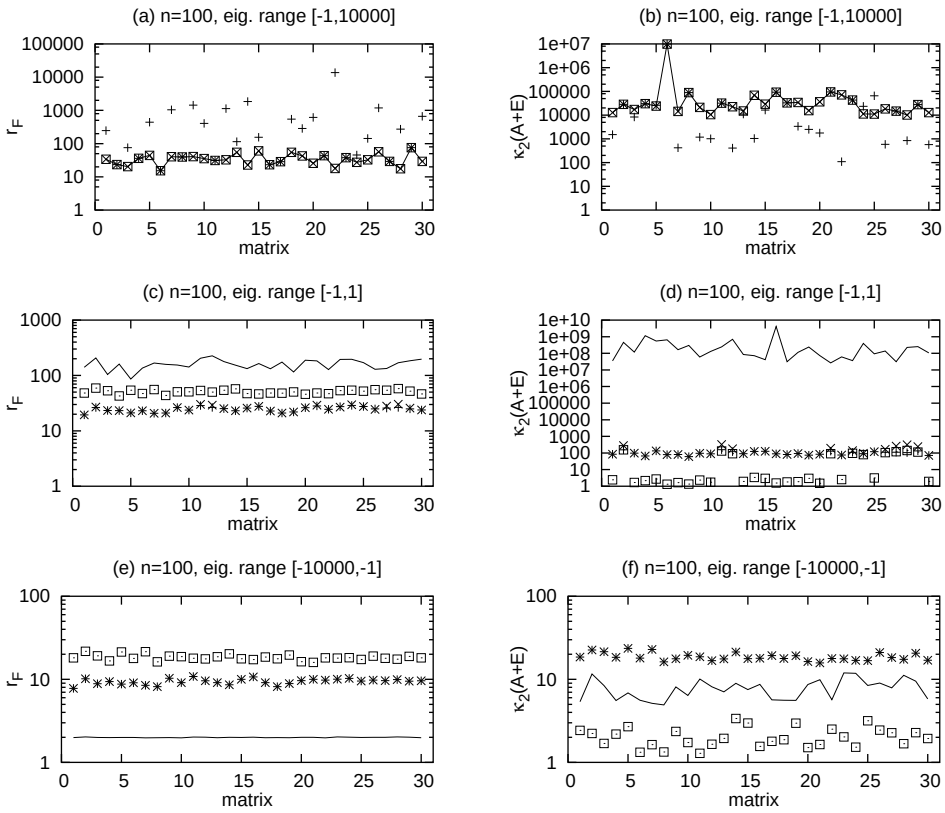

Figure 1: Measures of r_F and $\kappa_2(A+E)$ for the Type-I GMW algorithms for 30 random matrices with $n = 100$. Key: original GMW81 —, with 2-phase strategy $+$, with relaxed 2-phase strategy (GMW-I) $\times$, with relaxed 2-phase and nondecreasing strategy $\square$.

3.2 The GMW-II Algorithm

In this subsection we introduce our GMW-II algorithm, a Type-II variant of the (Type-I) GMW81 algorithm. We apply the nondecreasing strategy and choose δ_k in (1) to be

$$a_k + \delta_k := \max\{\delta, a_k + \delta_{k-1}, \frac{\|c_k\|_\infty^2}{\beta^2}\}$$

for $k = 1, \ldots, n$, where $\beta > 0$ and small tolerance $\delta > 0$ are preset, and $\delta_0 := 0$. The magnitude of the off-diagonal elements in $\bar{L}$ is still bounded by β, where $LDL^T = \bar{L}\bar{L}^T$.

The bound on $\|E\|_2$ for GMW81 is given in (3). For the Type-II GMW algorithm, it is

$$\|E\|_2 \leq (\frac{\xi}{\beta} + (n-1)\beta)^2 + (\eta + (n-1)\beta^2) + \delta =: f(\beta). \tag{23}$$

Equality is attained with $\beta^2 = \frac{\xi}{\sqrt{n^2-n}}$ for $n > 1$. Recall that η and ξ are the maximum magnitudes of the diagonal and off-diagonal elements of A, respectively. The minimum of (23) is

$$\min_\beta f(\beta) = 2\xi(\sqrt{n^2-n} + n - 1) + \eta + \delta \leq 4n\xi + \eta + \delta.$$

The minimum is attained with $\beta^2 = \frac{\xi}{\sqrt{n^2-n}}$ for $n > 1$. Therefore, β is chosen by

$$\beta^2 := \max\{\eta, \frac{\xi}{\sqrt{n^2-n}}, \epsilon_M\}$$

for $n > 1$, where $\beta^2 \geq \eta$ is for satisfying Objective 1 with pivoting. Substituting this into (23), we obtain $\|E\|_2 = O(n^2)$.

The relaxed 2-phase strategy in Algorithm 2 is also incorporated into our GMW-II algorithm. Therefore, the bound on $\|E\|_2$ is

$$\|E\|_2 \leq (\frac{\hat{\xi}}{\beta} + (n-K-1)\beta)^2 + (\hat{\eta} + (n-K-1)\beta^2) + \delta, \tag{24}$$

where K is the number of steps in Phase 1, and $\hat{\eta}$ and $\hat{\xi}$ are the maximum magnitudes of the diagonal and off-diagonal elements of A_{K+1}, respectively. Since $\beta^2 \geq \hat{\eta}$ is not required for satisfying Objective 1, β is determined by

$$\beta^2 := \max\left\{\frac{\hat{\xi}}{\sqrt{(n-K)^2-(n-K)}}, \epsilon_M\right\}$$

for $n - K > 1$. Substituting this into (24), we obtain

$$\|E\|_2 \leq 4(n-K)\hat{\xi} + \hat{\eta} + \delta \leq 4(n-K)(\xi + (1+\mu)\eta) + \eta + \delta = O(n),$$

where we ignore the extreme case $\beta^2 = \epsilon_M$. The last inequality is derived using (18).

The diagonal pivoting strategy can be incorporated into the Type-II GMW algorithms. We pivot on the maximum element for our GMW-II algorithm, as in the GMW-I algorithm. Note that all the *a priori* bounds on $\|E\|_2$ given above for all algorithms in the GMW class are independent of the pivoting strategy applied, if any.

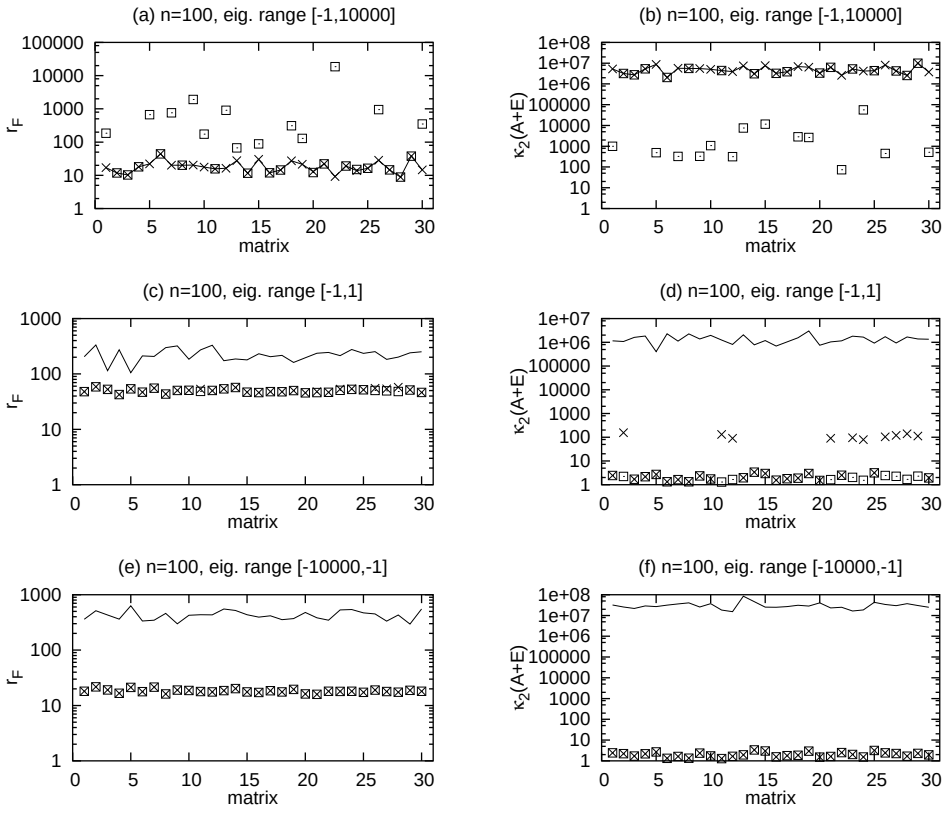

Figure 2: Measures of r_F and $\kappa_2(A+E)$ for the Type-II GMW algorithms for 30 random matrices with $n = 100$, nondecreasing strategy invoked. Key: original Type-II GMW —, with 2-phase strategy $\Box$, with relaxed 2-phase strategy (GMW-II) $\times$.

Recall that GMW81 and GMW-I use $\delta := \epsilon_M$. For the Type-II GMW algorithms, we use $\delta := \sqrt[3]{\epsilon_M^2}\eta$ as in SE99. Our experimental results are shown in Figure 2. Similar to SE90 and the Type-I GMW algorithms, incorporating the 2-phase strategy results in difficulties for the matrices with eigenvalues $[-1, 10000]$, and relaxing is the cure.

For all algorithms in the GMW class, the worst-case condition number is

$$\kappa_2(A+E) = O(n^3(\frac{\xi+\eta}{\delta})^n). \tag{25}$$

The proof uses Theorem 1, Lemma 3, the bounds $[\delta, \eta + (n-1)\beta^2]$ for the diagonal elements in D, and the bound β for the magnitude of the off-diagonal elements in $\bar{L}$, where $P(A+E)P^T = LDL^T = \bar{L}\bar{L}^T$ as denoted in (2).

Whether the 2-phase strategy or the relaxed 2-phase strategy is applied, the bound on $\kappa_2(A+E)$ remains exponential using (7) and (18), respectively. The bounds are not changed when the nondecreasing strategy is applied. All the modified Cholesky algorithms in Sections 2 and 3 are numerically stable, since they can be regarded as the Cholesky factorizations of the symmetric positive definite matrix $A+E$ [5].

3.3 The SE-I Algorithm

Both SE90 and SE99 are Type-II algorithms. In this section we present the Type-I variant corresponding to SE99, denoted by SE-I, by making three changes. First, instead of (14), we determine δ_k by

$$\delta_k := \max\{0, -2a_k, -a_k + \max\{\|a_k\|_1, \bar{\tau}\eta\}\} \leq \max\{2G, G + \bar{\tau}\eta\} \tag{26}$$

for $k = K+1, \ldots, n-2$. Second, instead of (16), the special treatment of the last 2×2 Schur complement in Phase 2 to keep $\|E\|_2$ small is

$$\begin{aligned} \delta_{n-1} = \delta_n &:= \max\{0, -2\lambda_1(A_{n-1}), -\lambda_1(A_{n-1}) + \max\{\frac{\tau(\lambda_2(A_{n-1}) - \lambda_1(A_{n-1}))}{1-\tau}, \bar{\tau}\eta\}\} \\ &\leq \max\{2G, G + \frac{2\tau}{1-\tau}(G+\eta)\}. \end{aligned} \tag{27}$$

Note that $\kappa_2(A_{n-1} + \delta_n I_2) \leq \min\{\kappa_2(A_{n-1}), \frac{1}{\tau}\}$. The derivation is similar to that of (10). Third, if the algorithm switches to Phase 2 at the last step, then δ_n is determined by

$$\begin{aligned} \delta_n &= \max\{0, -2a_n, -a_n + \max\{\frac{-\tau}{1-\tau}a_n, \bar{\tau}\eta\}\} \\ &\leq \max\{2G, G + \frac{\tau}{1-\tau}(G+\eta)\} \end{aligned} \tag{28}$$

instead of (15).

By (26), (27) and (28), we obtain

$$\|E\|_2 \leq \max\{2G, G + \frac{2\tau}{1-\tau}(G+\eta)\}. \tag{29}$$

Comparing (29) with (17), the bound on $\|E\|_2$ for SE-I is less than twice as that for SE99.

Now we investigate the satisfaction of Objective 1 for the GMW and SE algorithms. We begin with a theorem of Ostrowski [12, page 224].

Theorem 1 (Ostrowski) *Suppose we are given a symmetric $M \in C^{n\times n}$ and a nonsingular $S \in C^{n\times n}$. There exists $\theta_k > 0$ such that*

$$\lambda_k(SMS^*) = \theta_k \lambda_k(M),$$

where $\lambda_1(SS^*) \le \theta_k \le \lambda_n(SS^*)$.

Consider the 2-phase strategy presented in Algorithm 1 and the relaxed 2-phase strategy presented in Algorithm 2 with pivoting on the maximum diagonal element. Clearly $E = 0$ if the factorization is done in Phase 1. The derivation of the condition under which the algorithm runs to completion without switching to Phase 2 is by finite induction. We denote the incomplete LDL^T factorization of a symmetric matrix $A \in R^{n\times n}$ after step k by $L_k D_k L_k^T$, where

$$D_k = \begin{array}{c} \\ k \\ n-k \end{array} \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} \bar{D}_k & 0 \\ 0 & S_k \end{bmatrix} \end{array},$$

with $\bar{D}_k$ diagonal and S_k the Schur complement. We claim that the following condition guarantees $E = 0$:

$$\lambda_{\min}(A) \ge \delta \|L_k L_k^T\|_2 \tag{30}$$

for $k = 1, \ldots, n-1$. At the beginning of step k, we assume that the diagonal elements of the Schur complement are all larger than or equal to δ, and investigate whether this condition holds in the next Schur complement[4]. By Theorem 1 and (30),

$$\lambda_{\min}(D_k)\lambda_{\max}(L_k L_k^T) \ge \lambda_{\min}(A) \ge \delta \|L_k L_k^T\|_2 = \delta \lambda_{\max}(L_k L_k^T),$$

and therefore $\lambda_{\min}(D_k) \ge \delta$, so

$$\lambda_{\min}(S_k) \ge \lambda_{\min}(D_k) \ge \delta,$$

which implies $\mathrm{Diag}(S_k) \ge \delta I_{n-k}$. By induction, we stay in Phase 1 during the whole factorization. We conclude that if (30) holds, then $E = 0$.

In the next two lemmas we develop bounds on $\|LL^T\|_2$ and on $\lambda_{\min}(LL^T)$, in order to bound the condition number of $A + E$ for algorithms in Sections 4 and 5.

Lemma 2 *If the positive semidefinite Hermitian matrix $M \in C^{n\times n}$ has a diagonal element equal to 1, (i.e., $m_{kk} = 1$ for some $1 \le k \le n$), then*

$$\lambda_{\min}(M) \le 1 \le \lambda_{\max}(M).$$

[4]For the base case, we have $\lambda_{\min}(A) \ge \delta$ from (30), so $A - \delta I$ is positive definite and therefore $\mathrm{diag}(A) \ge \delta I$.

Proof Let $M = U\Lambda U^*$ denote the spectral decomposition of M, and $a := U^* e_k$. Since $m_{kk} = 1$,

$$1 = e_k^T M e_k = a^* U^* (U\Lambda U^*) U a = a^* \Lambda a.$$

Note that $a^* a = 1$. We conclude that the weighted average of the eigenvalues of M is 1. Therefore, $\lambda_{\min}(M) \leq 1 \leq \lambda_{\max}(M)$. ■

Lemma 3 [5] *For any lower unit triangular matrix $L \in R^{n\times n}$ with $|(L)_{ij}| \leq \gamma$ for $1 \leq j < i \leq n$,*

$$1 \leq \lambda_{\max}(LL^T) \leq n + \frac{1}{2} n(n-1)\gamma^2,$$

and

$$(1+\gamma)^{2-2n} \leq \lambda_{\min}(LL^T) \leq 1.$$

Proof By Lemma 2, $\lambda_{\min}(LL^T) \leq 1 \leq \lambda_{\max}(LL^T)$. Next, an upper bound on $\lambda_{\max}(LL^T)$ is $\lambda_{\max}(LL^T) \leq \text{trace}(LL^T) \leq n + \frac{1}{2} n(n-1)\gamma^2$. Computing the inverse of a lower triangular matrix, we obtain $(L^{-1})_{ii} = 1$ for $i = 1, \ldots, n$ and the bounds $|(L^{-1})_{ij}| \leq \gamma \sum_{k=j+1}^{i} |(L^{-1})_{ik}|$ for $1 \leq j < i \leq n$. The solution to this recursion is

$$|(L^{-1})_{ij}| \leq \gamma(1+\gamma)^{i-j-1}$$

for $1 \leq j < i \leq n$. Therefore,

$$\lambda_{\min}(LL^T)^{-1} = \|(LL^T)^{-1}\|_2 \leq \|L^{-1}\|_2^2 \leq \|L^{-1}\|_1 \|L^{-1}\|_\infty \leq (1+\gamma)^{2n-2}.$$

■

Now we can bound $\|L_k L_k^T\|_2$ in (30). Pivoting on the maximum diagonal element of each Schur complement, the magnitude of the elements in L_k are bounded by 1 for all k. By Lemma 3,

$$\|L_k L_k^T\|_2 \leq \frac{1}{2} n(n+1). \tag{31}$$

Substituting this into (30), we obtain the following result. For algorithms GMW-I, GMW-II, SE90, SE99, and SE-I that use the 2-phase strategy or the relaxed 2-phase strategy, if

$$\lambda_{\min}(A) \geq \frac{1}{2} n(n+1)\delta, \tag{32}$$

then by (30) and (31) we conclude that $E = 0$.

Our experimental results are shown in Figure 3. For the random matrices with eigenvalues in $[-1, 10000]$, SE-I resulted in larger $\|E\|_2$ and $\|E\|_F$ but substantially smaller $\kappa_2(A+E)$ than those of SE99. For the random matrices with eigenvalues in $[-1, 1]$ and $[-10000, -1]$, SE-I had comparable $\|E\|_2$, smaller $\|E\|_F$, but larger $\kappa_2(A+E)$ than SE99.

[5] Cheng and Higham [5] presented this lemma with $\gamma = \frac{7+\sqrt{17}}{4} \approx 2.781$.

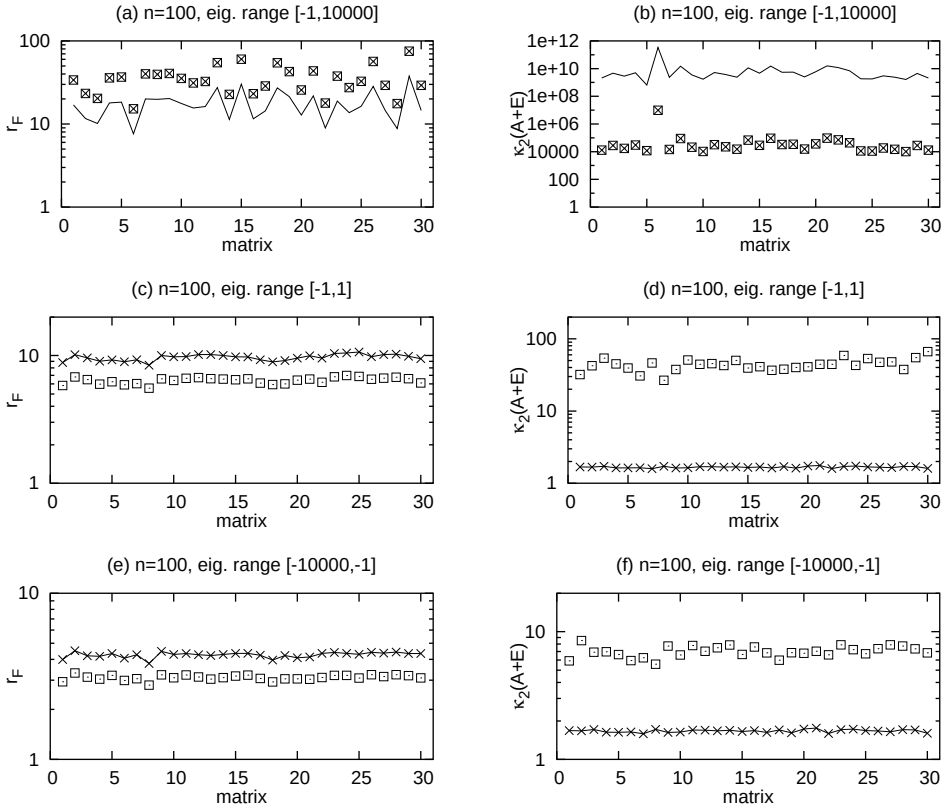

Figure 3: Measures of r_F and $\kappa_2(A+E)$ for the SE algorithms for 30 random matrices with $n = 100$. Key: original SE99 —, Type-I SE99 (SE-I) $\square$, Type-I SE99 with nondecreasing strategy $\times$.

The nondecreasing strategy can be incorporated into the Type-I SE algorithm. The resulting $\|E\|_2$, $\|E\|_F$ and $\kappa_2(A+E)$ were comparable to those of SE-I for the random matrices with eigenvalues in $[-1, 10000]$, and comparable to those of SE99 for the random matrices with eigenvalues in $[-1, 1]$ and $[-10000, -1]$. Incorporating the non-relaxed 2-phase strategy into the Type-I SE algorithms is possible, but it would result in difficulties similar to those of SE90.

For all the algorithms in the SE class, the worst-case condition number is

$$\kappa_2(A+E) = O(\frac{(\xi+\eta)n^3 4^n}{\delta}). \tag{33}$$

The sketch of the proof is similar to that for the GMW algorithms. In practice, the condition number is bounded by about $1/\tau$ and $1/\bar{\tau}$ respectively for SE90 and SE99 [15], and is comparable to $\kappa_2(A)$ for SE-I.

4 Modified LBL^T Algorithms

Any symmetric matrix $A \in R^{n\times n}$ has an LBL^T factorization, where B is block diagonal with block order 1 or 2 [2, 3, 4]. A modified LBL^T algorithm first computes the LBL^T factorization, and then perturbs $\hat{B} = B + \Delta B$ to be positive definite, so that $P(A+E)P^T = L\hat{B}L^T$ is positive definite as well, where P is the permutation matrix for pivoting.

Moré and Sorensen suggested a modified LBL^T algorithm [13] which we call MS79. Each 1×1 block in B, denoted by d, is modified to be $\hat{d} := \max\{\delta, |d|\}$, with $\delta > 0$ the preset small tolerance. For each 2×2 block D, its spectral decomposition $D = U \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} U^T$ is modified to be $\hat{D} := U \begin{bmatrix} \hat{\lambda}_1 & \\ & \hat{\lambda}_2 \end{bmatrix} U^T$, where $\hat{\lambda}_i := \max\{\delta, |\lambda_i|\}$ for $i = 1, 2$.

Cheng and Higham proposed another modified LBL^T algorithm [5] which we call CH98. Each 1×1 block d is modified to be $\hat{d} := \max\{\delta, d\}$, with $\delta > 0$ the preset small tolerance. Each 2×2 block D, with its spectral decomposition denoted by $D = U \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} U^T$, is modified to be $\hat{D} := U \begin{bmatrix} \hat{\lambda}_1 & \\ & \hat{\lambda}_2 \end{bmatrix} U^T$, where $\hat{\lambda}_i = \max\{\delta, \lambda_i\}$ for $i = 1, 2$.

The key distinction is that MS79 is a Type-I algorithm, whereas CH98 is of Type II. The MS79 algorithm was developed before the fast Bunch-Parlett and bounded Bunch-Kaufman pivoting strategies (rook pivoting) for the LBL^T factorization [2], but rook pivoting is also applicable to MS79. For MS79, we set $\delta := \epsilon_M$. Cheng and Higham [5] suggested $\delta := \sqrt{u}\|A\|_\infty$ for CH98, where $u = \epsilon_M/2$ is the unit roundoff.

MS79 predated the four objectives. Cheng and Higham investigated the objectives for CH98 [5], and our analysis for MS79 is similar.

For both MS79 and CH98, if $\lambda_{\min}(B) \geq \delta$, then $E = 0$. By Theorem 1, if A

is positive definite, $\lambda_{\min}(B) \geq \frac{\lambda_{\min}(A)}{\lambda_{\max}(LL^T)}$. Therefore, $E = 0$ is guaranteed when

$$\lambda_{\min}(A) \geq \delta \|LL^T\|_2. \tag{34}$$

Consider $\|E\|_2$ for MS79. By Theorem 1,

$$\begin{aligned} \|E\|_2 &= \lambda_{\max}(E) = \lambda_{\max}(L\Delta BL^T) \leq \lambda_{\max}(LL^T)\lambda_{\max}(\Delta B) \\ &= \lambda_{\max}(LL^T)\max\{\delta - \lambda_{\min}(B), -2\lambda_{\min}(B), 0\}. \end{aligned}$$

By Theorem 1 again, $-\lambda_{\min}(B) \leq -\frac{\lambda_{\min}(A)}{\lambda_{\min}(LL^T)}$ and $-\lambda_{\min}(B) \geq -\frac{\lambda_{\min}(A)}{\lambda_{\max}(LL^T)}$ for $\lambda_{\min}(A) < 0$. Therefore,

$$\|E\|_2 \leq -2\lambda_{\min}(A)\kappa_2(LL^T) \text{ for } \lambda_{\min}(A) \leq -\delta\|LL^T\|_2. \tag{35}$$

Similarly, the bound on $\|E\|_2$ for CH98 is

$$\|E\|_2 \leq \delta\|LL^T\|_2 - \lambda_{\min}(A)\kappa_2(LL^T) \text{ for } \lambda_{\min}(A) \leq 0. \tag{36}$$

Now we assess how well Objective 3 is satisfied for MS79. By Theorem 1,

$$\begin{aligned} \lambda_{\min}(A+E) &\geq \lambda_{\min}(LL^T)\lambda_{\min}(\hat{B}) \\ &= \lambda_{\min}(LL^T)\max\{\delta, \min_{1\leq i\leq n}|\lambda_i(B)|\} &(37)\\ &\geq \lambda_{\min}(LL^T)\max\{\delta, \frac{\min_{1\leq i\leq n}|\lambda_i(A)|}{\lambda_{\max}(LL^T)}\}, &(38) \end{aligned}$$

and

$$\begin{aligned} \lambda_{\max}(A+E) &\leq \lambda_{\max}(LL^T)\lambda_{\max}(\hat{B}) \\ &= \lambda_{\max}(LL^T)\max\{\delta, -\lambda_{\min}(B), \lambda_{\max}(B)\} &(39)\\ &\leq \lambda_{\max}(LL^T)\max\{\delta, \frac{-\lambda_{\min}(A)}{\lambda_{\min}(LL^T)}, \frac{\lambda_{\max}(A)}{\lambda_{\min}(LL^T)}\}. &(40) \end{aligned}$$

By (37) and (39),

$$\kappa_2(A+E) \leq \kappa_2(LL^T)\kappa_2(B).$$

By (38) and (40),

$$\kappa_2(A+E) \leq \kappa_2(LL^T)^2\kappa_2(A). \tag{41}$$

The bound on $\kappa_2(A+E)$ for CH98 [5] is

$$\kappa_2(A+E) \leq \kappa_2(LL^T)\max\{1, \frac{\lambda_{\max}(A)}{\lambda_{\min}(LL^T)\delta}\}. \tag{42}$$

There are four pivoting algorithms for the LBL^T factorization: Bunch-Parlett (complete pivoting) [4], Bunch-Kaufman (partial pivoting) [3], fast Bunch-Parlett and bounded Bunch-Kaufman (rook pivoting) [2], denoted by BP, BK, FBP and BBK, respectively. All these algorithms have a preset argument

$0 < \alpha < 1$. The BK algorithm takes $O(n^2)$ time for pivoting, but the elements in L are unbounded. It is discouraged for the modified LBL^T algorithms because Objectives 1–3 may not be satisfied. For example, the following LBL^T factorization [11] is produced with the BK pivoting strategy for $\epsilon \neq 0$,

$$A = \begin{bmatrix} 0 & \epsilon & 0 \\ \epsilon & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 1/\epsilon & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & \epsilon & \\ \epsilon & 0 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1/\epsilon \\ & 1 & 0 \\ & & 1 \end{bmatrix} = LBL^T.$$

Applying MS79 or CH98 and assuming $0 < \epsilon \leq \delta$, we obtain

$$E = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 1/\epsilon & 0 & 1 \end{bmatrix} \begin{bmatrix} \delta & -\epsilon & \\ -\epsilon & \delta & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1/\epsilon \\ & 1 & 0 \\ & & 1 \end{bmatrix} = \begin{bmatrix} \delta & -\epsilon & \delta/\epsilon \\ -\epsilon & \delta & -1 \\ \delta/\epsilon & -1 & 1+\delta \end{bmatrix}.$$

When $\epsilon \to 0^+$, $\|E\| \to \infty$, so Objective 2 is not satisfied.

From (34)–(42), it is clear that $\lambda_{\min}(LL^T)$, $\lambda_{\max}(LL^T)$ and $\kappa_2(LL^T)$ play an important role for the satisfaction of Objectives 1–3 for both MS79 and CH98.

The BP, BBK and FBP algorithms all produce a bound on the elements in L in terms of the pivoting argument α, suggested to be $\alpha = \frac{1+\sqrt{17}}{8} \approx 0.640$ to minimize the bound on the element growth of the Schur complements [2, 3, 4]. The corresponding element bound of the unit lower triangular matrix L is $\gamma = \frac{7+\sqrt{17}}{4} \approx 2.781$. Alternatively, we could choose $\alpha = 0.5$ to minimize the element bound of L [6, Table 3.1], which is $\gamma = 2$, leading to sharper bounds on $\lambda_{\min}(LL^T)$, $\lambda_{\max}(LL^T)$ and $\kappa_2(LL^T)$. The bounds in Table 3 are obtained using Lemma 3.

Table 3: Bounds for the LBL^T factorization with the BP, BBK or FBP pivoting algorithm.

α	γ	$\lambda_{\min}(LL^T)$	$\lambda_{\max}(LL^T)$	$\kappa_2(LL^T)$
$\frac{1+\sqrt{17}}{8}$	$\frac{7+\sqrt{17}}{4}$	$\geq 3.781^{2-2n}$	$\leq 4n^2 - 3n$	$\leq (4n^2 - 3n)3.781^{2n-2}$
0.5	2	$\geq 3^{2-2n}$	$\leq 2n^2 - n$	$\leq (2n^2 - n)3^{2n-2}$

Although $\alpha = \frac{1}{2}$ results in smaller bounds, $\alpha = \frac{1+\sqrt{17}}{8} \approx 0.640$ is a better choice in practice, as shown in Figure 4.

The BP pivoting strategy takes $\frac{1}{6}n^3 + O(n^2)$ comparisons and does not meet Objective 4. The number of comparisons for the BBK and FBP pivoting strategies are between those of the BK and BP algorithms (i.e., between $O(n^2)$ and $O(n^3)$). There are matrices that require traversing the whole matrix of each Schur complement with either the BBK or the FBP pivoting strategy [2]. Hence they take $\Theta(n^3)$ comparisons for pivoting in worst cases and fail to meet Objective 4.

Here and throughout the remainder of this paper, we assume the pivoting strategy applied to the MS79 and CH98 is BBK, unless otherwise noted.

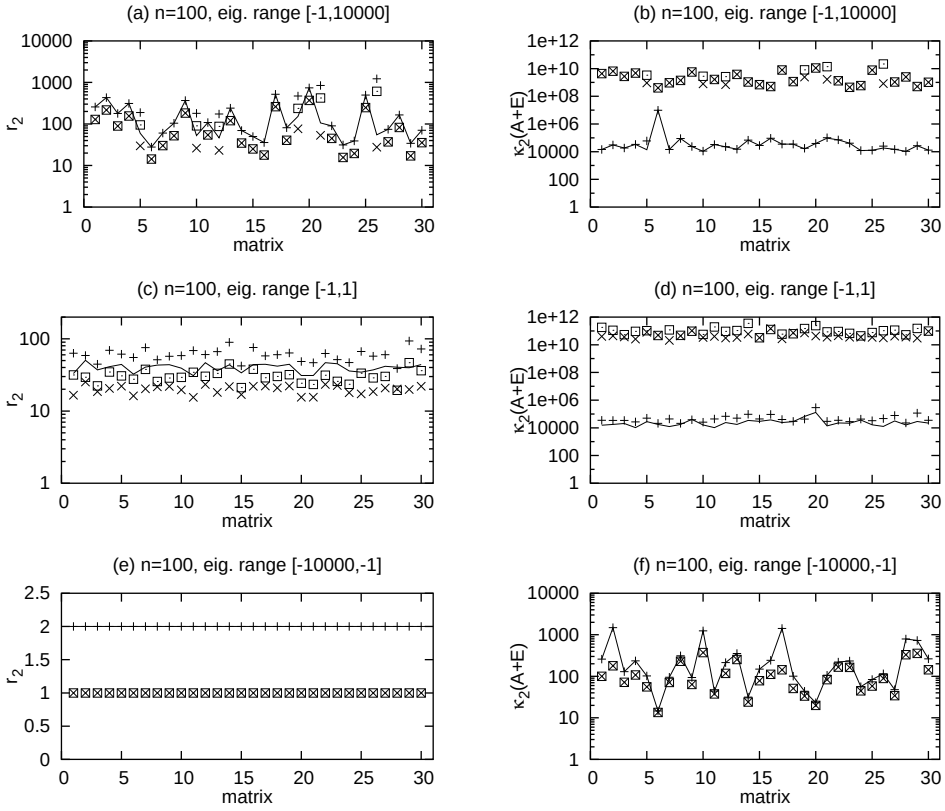

Figure 4: Measures of r_2 and $\kappa_2(A+E)$ for MS79 and CH98 for 30 random matrices with $n = 100$. Key: MS79 $\alpha = 0.640$ —, MS79 $\alpha = 0.5$ $+$, CH98 $\alpha = 0.640$ $\times$, CH98 $\alpha = 0.5$ $\square$.

Three remarks are in order. First, both MS79 and CH98 satisfy Objectives 1–3. Second, the bound on $\|E\|_2$ for MS79 is about twice that for CH98, whereas $A + E$ is generally better conditioned for MS79 than for CH98. Third, both algorithms fail to satisfy Objective 4 in the worst case.

5 A New Approach via Modified LTL^T Factorization

Aasen [1], Parlett and Reid [14] introduced the LTL^T factorization and its application to solving symmetric linear systems. We denote the factorization by $PAP^T = LTL^T$, where T is symmetric tridiagonal, and L is unit lower triangular with the magnitude of its elements bounded by 1 and the off-diagonal elements in the first column all zero.

The work in computing Aasen's LTL^T factorization is about the same as that of the Cholesky factorization, whereas the formulation by Parlett and Reid doubles the cost. In both formulations the storage is the same as that required by the Cholesky factorization, and the numerical stability of its use in solving linear systems is empirically comparable to that of the LBL^T factorization [2].

Our new approach arises from the fact that A is positive definite if and only if T is positive definite. A modified LTL^T algorithm makes $\hat{T} = T + \Delta T$ symmetric positive definite, and the resulting factorization $\hat{A} = P(A + E)P^T = L\hat{T}L^T$ is also symmetric positive definite.

We can apply the modified LDL^T algorithms in Sections 2 and 3 and the modified LBL^T algorithms in Section 4 to the matrix T. The resulting modified LTL^T factorization roughly satisfies Objective 1, assuming the modified Cholesky algorithm applied to T satisfies Objective 1. Our method was inspired by the merits of triadic structure (no more than two off-diagonal elements in every column of a matrix) discussed in [6].

The triadic structure of a symmetric matrix is preserved in the LDL^T or LBL^T factorizations [6, Theorem 2.5]. This implies that the modified LDL^T or LBL^T algorithms in Sections 2–4 applied to a symmetric triadic matrix are very efficient. Recall that both MS79 and CH98 have difficulties in satisfying Objective 4. The potential excessive cost can be reduced to be $O(n^2)$ by instead applying MS79 or CH98 to the symmetric tridiagonal matrix T of the LTL^T factorization. We call the resulting algorithms LTL^T-MS79 and LTL^T-CH98, respectively. For LTL^T-MS79, we use $\delta := \epsilon_M$. For LTL^T-CH98, we use $\delta := \sqrt[3]{\epsilon_M^2}\eta$, as used in SE99.

Table 4 compares the costs of these LBL^T pivoting strategies for symmetric and symmetric tridiagonal matrices. We use the BBK pivoting strategy for both MS79 and CH98, because it is the cheapest pivoting strategy that guarantees a bounded L. Even so, Objective 4 is not satisfied in worst cases. We use the BP pivoting strategy for both LTL^T-CH98 and LTL^T-MS79. Then Objective 4 is satisfied [6], even though BP is the most expensive pivoting strategy.

Given a symmetric matrix $A \in R^{n\times n}$, we denote its LTL^T factorization

Table 4: Comparison costs of various pivoting strategies for the LBL^T factorization.

Symmetric Matrix	*General*		*Tridiagonal*	
case	*worst*	*best*	*worst*	*best*
BP	$O(n^3)$		$O(n^2)$	
FBP	$O(n^3)$	$O(n^2)$	$O(n^2)$	$O(n)$
BBK	$O(n^3)$	$O(n^2)$	$O(n^2)$	$O(n)$

by $PAP^T = LTL^T$, and the LBL^T factorization of T by $\widetilde{P}T\widetilde{P}^T = \widetilde{L}\widetilde{B}\widetilde{L}^T$. The resulting sandwiched factorization is $PAP^T = L\widetilde{P}^T\widetilde{L}\widetilde{B}\widetilde{L}^T\widetilde{P}L^T$. Adding a perturbation $\Delta\widetilde{B}$ to $\widetilde{B}$ to make it positive definite, the modified factorization of T is $\widetilde{P}(T+\Delta T)\widetilde{P}^T = \widetilde{L}(\widetilde{B}+\Delta\widetilde{B})\widetilde{L}^T$. The modified LTL^T factorization is

$$P(A+E)P^T = L\widetilde{P}^T\widetilde{L}(\widetilde{B}+\Delta\widetilde{B})\widetilde{L}^T\widetilde{P}L^T. \tag{43}$$

The matrix L is unit lower triangular with the magnitude of all elements bounded by 1 and all the off-diagonal elements in the first column zero. By Lemma 3, the LTL^T factorization satisfies

$$\begin{array}{rcl} \lambda_{\max}(LL^T) & \leq & \frac{1}{2}n(n-1) \\ \lambda_{\min}(LL^T) & \geq & 2^{4-2n} \end{array} \tag{44}$$

for $n > 1$. Lemma 4 gives the bounds on $\lambda_{\max}(\widetilde{L}\widetilde{L}^T)$ and $\lambda_{\min}(\widetilde{L}\widetilde{L}^T)$, where $\widetilde{L}$ is triadic and unit lower triangular.

Lemma 4 *Let* $F_\gamma(k) = \sum_{i=1}^{\lceil k/2\rceil}\binom{k-i}{i-1}\gamma^{k-i}$ *and* $\Phi_\gamma = \frac{1+\sqrt{1+4/\gamma}}{2}\gamma$ *for* $k \in N$ *and* $\gamma > 0$. *For any triadic and unit lower triangular* $\widetilde{L} \in R^{n\times n}$ *with the magnitude of the off-diagonal elements bounded by* γ,

1. $\frac{1}{1+(1/\gamma)}\Phi_\gamma^{k-1} \leq F_\gamma(k) \leq \Phi_\gamma^{k-1}$ *for* $k \in N$.
2. $(\widetilde{L}^{-1})_{ij} \leq F_\gamma(i-j+1)$ *for* $1 \leq j \leq i \leq n$.
3. $\lambda_{\max}(\widetilde{L}\widetilde{L}^T) \leq n + (2n-3)\gamma^2$ *for* $n > 1$.
4. $\lambda_{\min}(\widetilde{L}\widetilde{L}^T) \geq (\frac{\Phi_\gamma - 1}{\Phi_\gamma^n - 1})^2$.

Proof Part one of this lemma is [6, Lemma 5.4]. Part two is from the proof of [6, Lemma 5.6]. For part three,

$$\lambda_{\max}(\widetilde{L}\widetilde{L}^T) \leq \mathrm{trace}(\widetilde{L}\widetilde{L}^T) = \|\widetilde{L}\|_F^2 \leq n + (2n-3)\gamma^2,$$

for $n > 1$. Finally, by parts one and two,

$$\begin{array}{rcl} \lambda_{\min}(\widetilde{L}\widetilde{L}^T)^{-1} & = & \|(\widetilde{L}\widetilde{L}^T)^{-1}\|_2 = \|\widetilde{L}^{-1}\|_2^2 \leq \|\widetilde{L}^{-1}\|_1\|\widetilde{L}^{-1}\|_\infty \\ & \leq & (\sum_{k=1}^{n} F_\gamma(k))^2 \leq (\sum_{k=1}^{n}\Phi_\gamma^{k-1})^2 = (\frac{\Phi_\gamma^n - 1}{\Phi_\gamma - 1})^2. \end{array}$$

■

Now we can assess the satisfaction of Objectives 1–3 for our LTL^T-MS79 and LTL^T-CH98 algorithms. To ensure a bounded L of the LBL^T factorization, we can use BP, FBP or BBK, but not BK. By (34), $\lambda_{\min}(T) \geq \delta\|\widetilde{L}\widetilde{L}^T\|_2$ implies $E = 0$. By Theorem 1, if A is positive definite, $\lambda_{\min}(A) \geq \lambda_{\min}(T)\lambda_{\min}(LL^T)$. We conclude that $E = 0$ if

$$\lambda_{\min}(A) \geq \delta\|\widetilde{L}\widetilde{L}^T\|_2\lambda_{\min}(LL^T). \tag{45}$$

For LTL^T-MS79, by Theorem 1 and (35),

$$\begin{aligned} \|E\|_2 &= \lambda_{\max}(E) = \lambda_{\max}(L\Delta TL^T) \\ &\leq \lambda_{\max}(LL^T)\lambda_{\max}(\Delta T) = \|LL^T\|_2\|\Delta T\|_2 \\ &\leq -2\lambda_{\min}(A)\kappa_2(LL^T)\kappa_2(\widetilde{L}\widetilde{L}^T) \end{aligned} \tag{46}$$

for $\lambda_{\min}(A) \leq -\delta\|LL^T\|_2\|\widetilde{L}\widetilde{L}^T\|_2$. For LTL^T-CH98, by Theorem 1 and (36),

$$\|E\|_2 \leq \delta\|LL^T\|_2\|\widetilde{L}\widetilde{L}^T\|_2 - \lambda_{\min}(A)\kappa_2(LL^T)\kappa_2(\widetilde{L}\widetilde{L}^T) \tag{47}$$

for $\lambda_{\min}(A) \leq 0$. For LTL^T-MS79, by Theorem 1 and (41),

$$\begin{aligned} \kappa_2(A+E) &\leq \kappa_2(LL^T)\kappa_2(T+\Delta T) \\ &\leq \kappa_2(LL^T)\kappa_2(\widetilde{L}\widetilde{L}^T)^2\kappa_2(T) \\ &\leq \kappa_2(LL^T)^2\kappa_2(\widetilde{L}\widetilde{L}^T)^2\kappa_2(A). \end{aligned} \tag{48}$$

For LTL^T-CH98, by Theorem 1 and (42),

$$\begin{aligned} \kappa_2(A+E) &\leq \kappa_2(LL^T)\kappa_2(\widetilde{L}\widetilde{L}^T)\max\{1, \frac{\lambda_{\max}(T)}{\lambda_{\min}(\widetilde{L}\widetilde{L}^T)\delta}\} \\ &\leq \kappa_2(LL^T)\kappa_2(\widetilde{L}\widetilde{L}^T)\max\{1, \frac{\lambda_{\max}(A)}{\lambda_{\min}(LL^T)\lambda_{\min}(\widetilde{L}\widetilde{L}^T)\delta}\}. \end{aligned} \tag{49}$$

Note that the pivoting argument used in $\widetilde{P}T\widetilde{P}^T = \widetilde{L}\widetilde{B}\widetilde{L}^T$ is $\alpha = \frac{\sqrt{5}-1}{2} \approx 0.618$ for symmetric triadic matrices [6, Theorem 4.1]. The corresponding element bound of L is $\gamma = \frac{\sqrt{5}+3}{2} \approx 2.618$. One may choose $\alpha = 0.5$ to obtain the minimum element bound of L [6, Table 3.1], which is $\gamma = 2$, but it could result in an excessive $\|E\|_2$ for random matrices with eigenvalues $[-1, 10000]$ as shown in Figure 5.

The bounds on $\|LL^T\|_2$ and $\lambda_{\min}(LL^T)$ are given in (44). The bounds on $\|\widetilde{L}\widetilde{L}^T\|_2$ and $\lambda_{\min}(\widetilde{L}\widetilde{L}^T)$ are in Lemma 4 with $\gamma = \frac{\sqrt{5}+3}{2} \approx 2.618$. We conclude that

$$\begin{aligned} \|LL^T\|_2\|\widetilde{L}\widetilde{L}^T\|_2 &\leq 7.5n^3 - 17.5n^2 + 10.5n \\ \lambda_{\min}(LL^T)\lambda_{\min}(\widetilde{L}\widetilde{L}^T) &\geq \tfrac{91}{4^n(3.4^n-1)^2} \end{aligned} \tag{50}$$

for $n > 1$.

Comparing (50) with Table 3 with $\alpha = \frac{1+\sqrt{17}}{8}$, $\|LL^T\|_2$ and $\lambda_{\min}(LL^T)$ for MS79 and CH98 have sharper bounds than $\|LL^T\|_2\|\widetilde{L}\widetilde{L}^T\|_2$ and $\lambda_{\min}(LL^T)\lambda_{\min}(\widetilde{L}\widetilde{L}^T)$

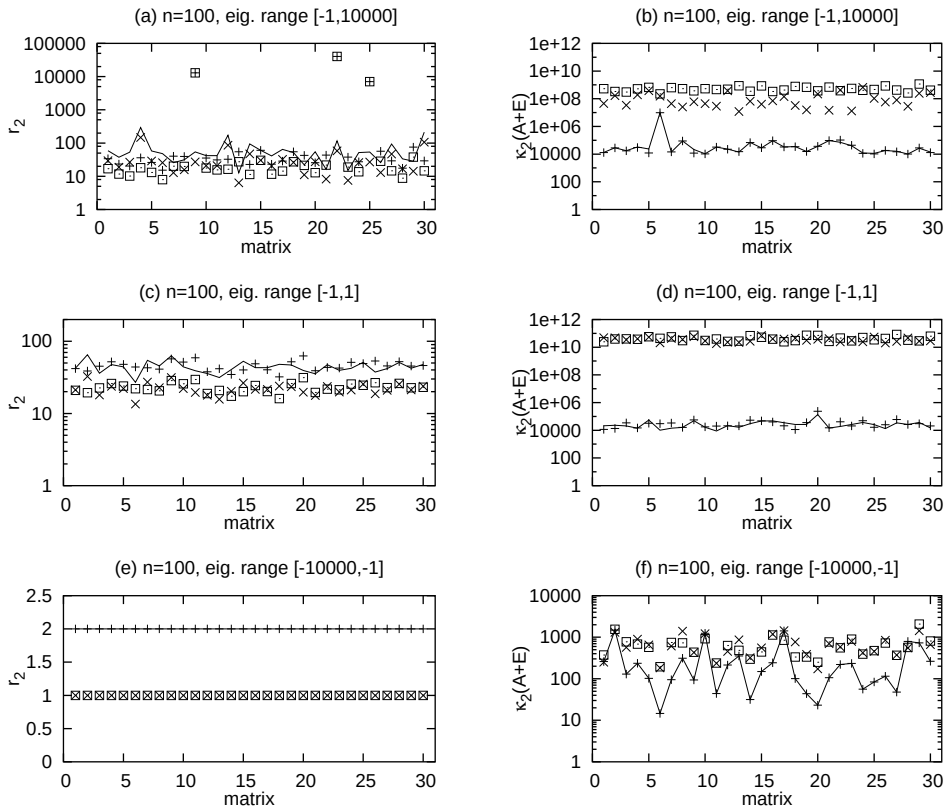

Figure 5: Measures of r_2 and $\kappa_2(A+E)$ for LTL^T-MS79 and LTL^T-CH98 for 30 random matrices with $n = 100$. Key: LTL^T-MS79 $\alpha = 0.618$ —, LTL^T-MS79 $\alpha = 0.5$ $+$, LTL^T-CH98 $\alpha = 0.618$ $\times$, LTL^T-CH98 $\alpha = 0.5$ $\square$.

for LTL^T-MS79 and LTL^T-CH98, respectively. Comparing (35) and (36) with (46) and (47), MS79 and CH98 have sharper bounds on $\|E\|_2$ than LTL^T-MS79 and LTL^T-CH98, respectively. Comparing (41) and (42) with (48) and (49), MS79 and CH98 have sharper bounds on $\kappa_2(A + E)$ than LTL^T-MS79 and LTL^T-CH98, respectively. In our experiments, however, our LTL^T-MS79 and LTL^T-CH98 algorithms usually performed as well as (and sometimes better than) MS79 and CH98, respectively.

In our experiments on the random matrices with eigenvalues in $[-1, 1]$ and $[-10000, -1]$, $\|E\|_2$ produced by LTL^T-MS79 and LTL^T-CH98 were comparable to those from MS79 and CH98, respectively. For the random matrices with eigenvalues in $[-1, 10000]$, our LTL^T-MS79 and LTL^T-CH98 algorithms slightly outperformed MS79 and CH98 by keeping $\|E\|_2$ smaller on average, respectively. Figures 6 and 7 show the result of MS79 and LTL^T-MS79 and that

of CH98 and LTL^T-CH98, respectively.

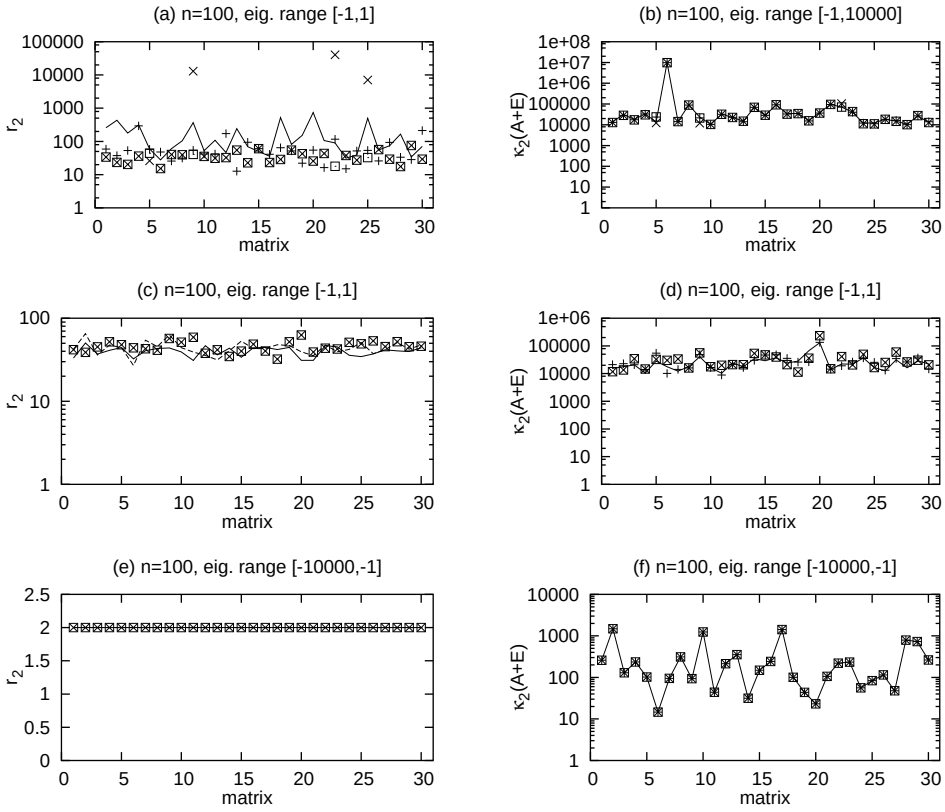

Figure 6: Measures of r_2 and $\kappa_2(A+E)$ for MS79, LTL^T-MS79, 2-phase LTL^T-MS79, and relaxed 2-phase LTL^T-MS79 for 30 random matrices with $n = 100$. Key: MS79 —, LTL^T-MS79 $+$, 2-phase LTL^T-MS79 $\times$, relaxed 2-phase LTL^T-MS79 $\square$.

The 2-phase strategy can also be incorporated into LTL^T-MS79 and LTL^T-CH98. However, this results in the potential problem of large $\|E\|$, similar to that of SE90. The problem was roughly resolved by relaxing in our experiments, as shown in Figures 6 and 7. Unfortunately, the problem was not extinguished with the relaxed 2-phase strategy. See the discussion in Section 6.2. Therefore, we do not advise incorporating the 2-phase or the relaxed 2-phase strategy into LTL^T-MS79 or LTL^T-CH98.

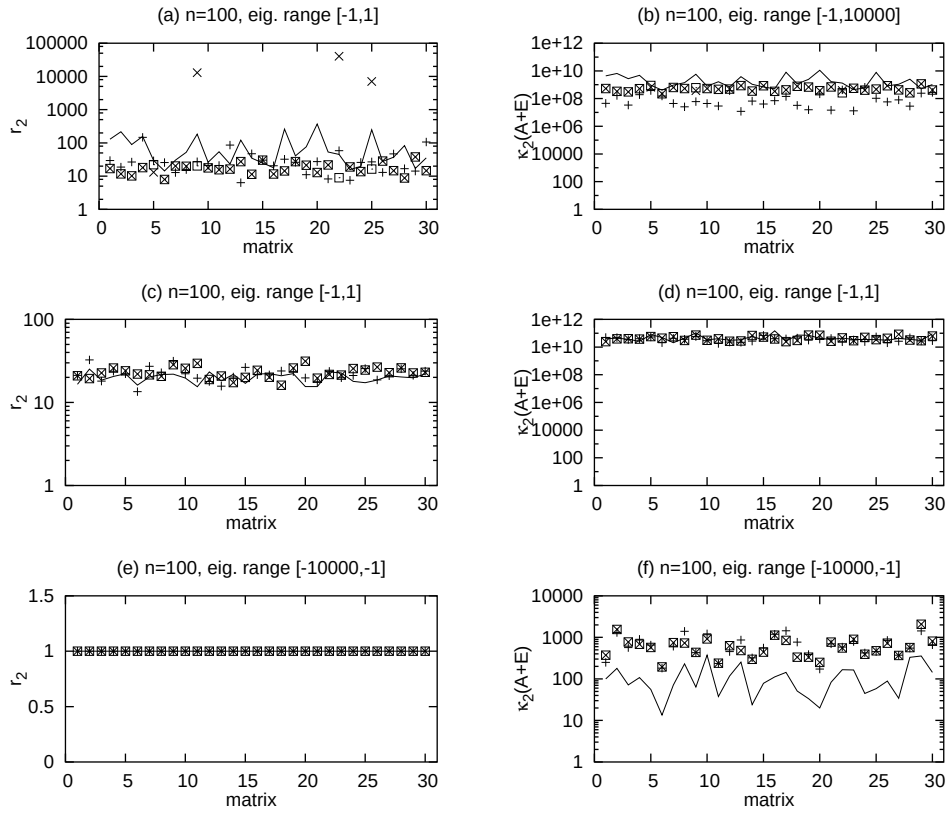

Figure 7: Measures of r_2 and $\kappa_2(A+E)$ for CH98, LTL^T-CH98, 2-phase LTL^T-CH98, and relaxed 2-phase LTL^T-CH98 for 30 random matrices with $n = 100$. Key: CH98 —, LTL^T-CH98 $+$, 2-phase LTL^T-CH98 $\times$, relaxed 2-phase LTL^T-CH98 $\square$.

6 Additional Numerical Experiments

Our previous experiments provided good values for the parameters in our methods. Now we present more extensive comparisons among the methods.

We ran three tests in our experiments. The first test concerns random matrices similar to those in [5, 15, 16]. The second test was on the first matrix in [15] for which SE90 had difficulties. The third test was on the 33 matrices used in [16]. Our experiments were on a laptop with a Intel Celeron 2.8GHz CPU using IEEE standard arithmetic with machine epsilon $\epsilon_M = 2^{-52} \approx 2.22 \times 10^{-16}$.

6.1 Random Matrices

To investigate the behaviors of the factorization algorithms, we experimented on the random matrices with eigenvalues in $[-1, 10000]$, $[-1, 1]$, and $[-10000, -1]$ for dimensions $n = 25, 50, 100$. The random matrices were generated as described in Section 3. We compare the performance of the four Type-I algorithms, GMW-I, SE-I, MS79 and LTL^T-MS79, and the four Type-II algorithms, GMW-II, SE99, CH98 and LTL^T-CH98.

Figures 8–10 show the results of the Type-I algorithms, whereas Figures 11–13 show results of the Type-II algorithms. We measure $\|E\|_2$ by $r_2 = \frac{\|E\|_2}{|\lambda_{\min}(A)|}$ as defined in (20).

Consider the Type-I algorithms. MS79 and LTL^T-MS79 generally produced comparable $\|E\|_2$ and condition numbers, but for matrices with eigenvalues in $[-1, 10000]$, LTL^T-MS79 achieved a smaller $\|E\|_2$ than MS79 in several cases. For matrices with eigenvalues in $[-1, 1]$, SE-I outperformed the other Type-I algorithms by not only producing smaller $\|E\|_2$ but also smaller $\kappa_2(A+E)$. For matrices with eigenvalues in $[-10000, -1]$, the GMW-II produced larger $\|E\|_2$ than the others.

Now compare the Type-II algorithms. In experiments on matrices with eigenvalues in $[-1, 10000]$, GMW-II and SE99 produced $\|E\|_2$ smaller than the others on average. The LTL^T-CH98 algorithm outperformed CH98 by usually achieving a smaller $\|E\|_2$. For the random matrices with eigenvalues in $[-1, 1]$, SE99 remains the best. For the random matrices with eigenvalues in $[-10000, -1]$, CH98 and LTL^T-CH98 achieved the minimal $\|E\|_2$.

6.2 The Benchmark Matrix

Schnabel and Eskow [15] identified a matrix that gives SE90 difficulties:

$$A = \begin{bmatrix} 1890.3 & -1705.6 & -315.8 & 3000.3 \\ -1705.6 & 1538.3 & 284.9 & -2706.6 \\ -315.8 & 284.9 & 52.5 & -501.2 \\ 3000.3 & -2706.6 & -501.2 & 4760.8 \end{bmatrix}. \tag{51}$$

It became one of the benchmark matrices for the modified Cholesky algorithms [5, 16]. This matrix has eigenvalues $\{-0.378, -0.343, -0.248, 8.24 \times 10^3\}$.

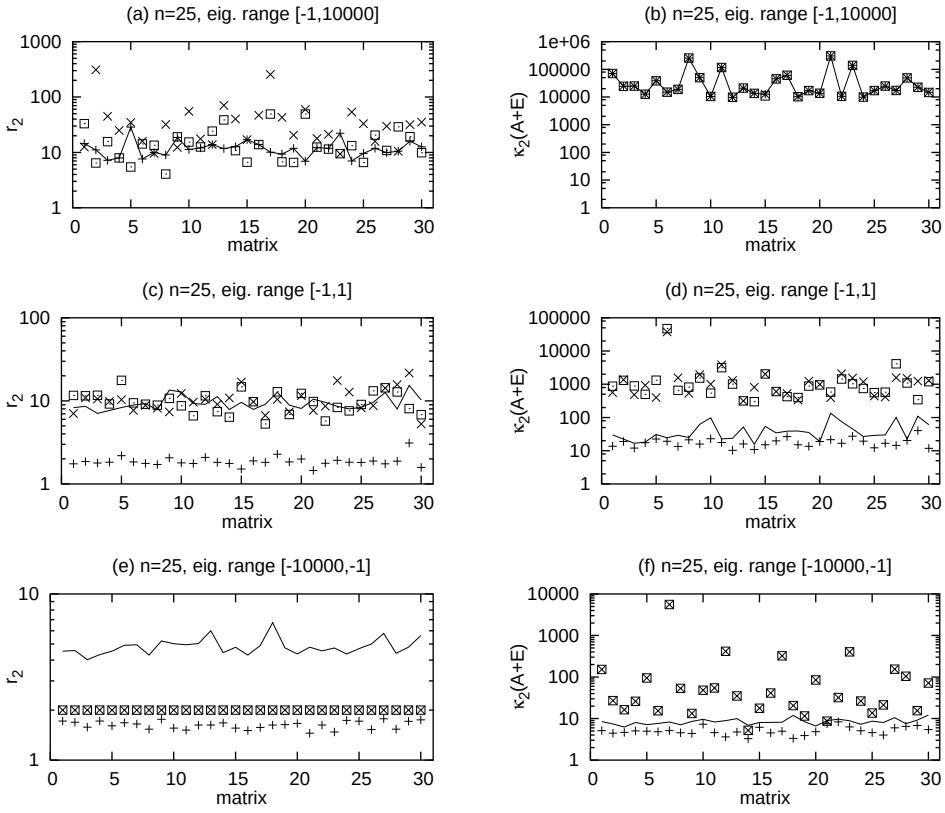

Figure 8: Measures of r_2 and $\kappa_2(A+E)$ for GMW-I, SE-I, MS79, and LTL^T-MS79 for 30 random matrices with $n = 25$. Key: GMW-I —, SE-I $+$, CH98 $\times$, LTL^T-CH98 $\square$.

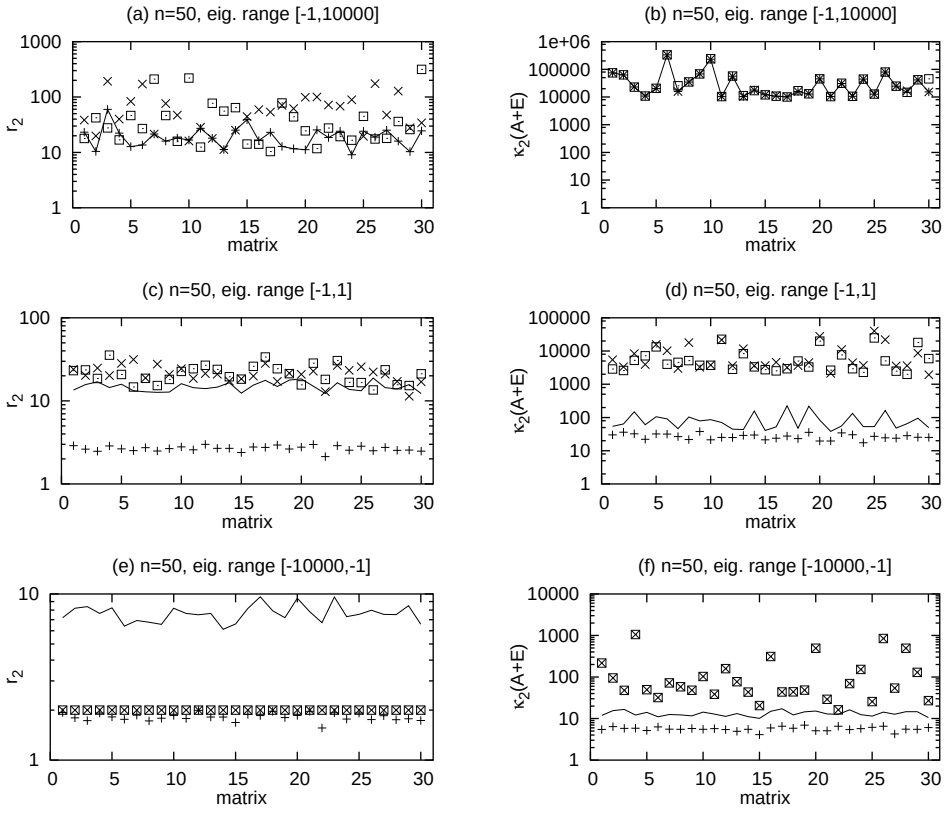

Figure 9: Measures of r_2 and $\kappa_2(A+E)$ for GMW-I, SE-I, MS79, and LTL^T-MS79 for 30 random matrices with $n = 50$. Key: GMW-I —, SE-I $+$, CH98 $\times$, LTL^T-CH98 $\square$.

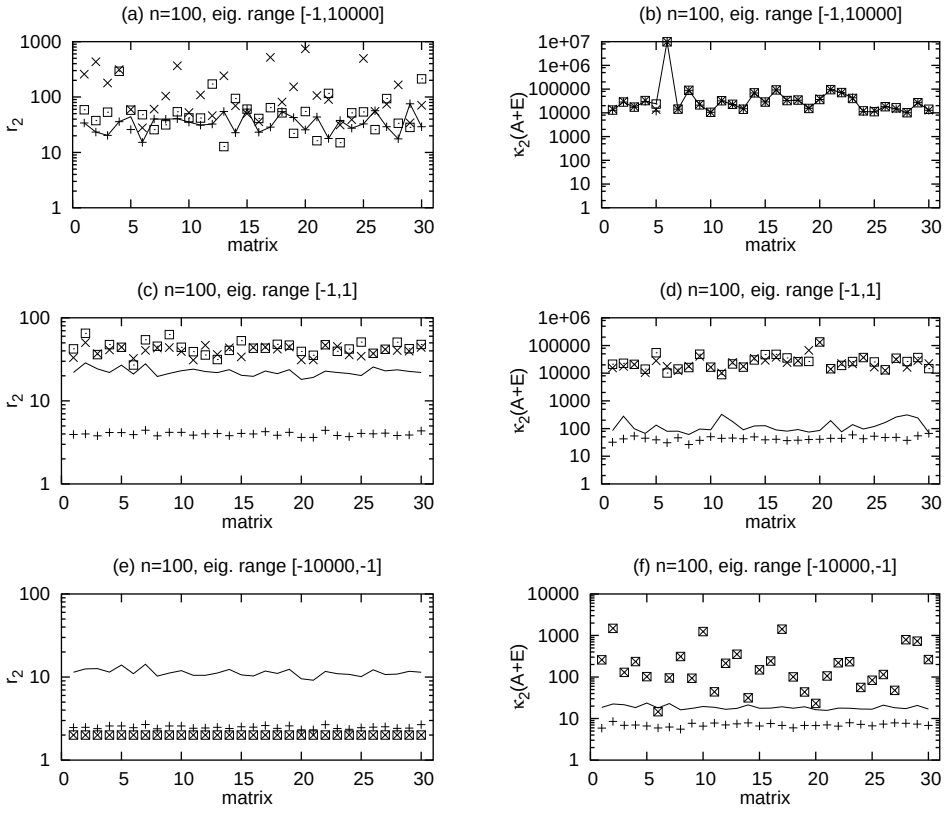

Figure 10: Measures of r_2 and $\kappa_2(A+E)$ for GMW-I, SE-I, MS79, and LTL^T-MS79 for 30 random matrices with $n = 100$. Key: GMW-I —, SE-I $+$, CH98 $\times$, LTL^T-CH98 $\square$.

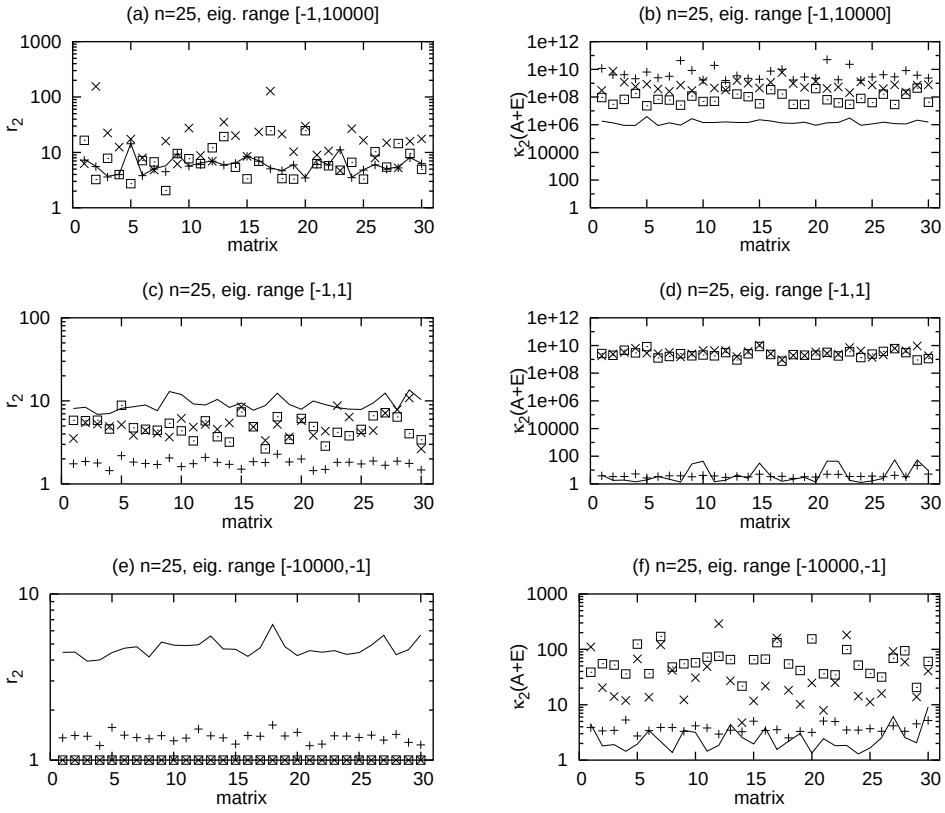

Figure 11: Measures of r_2 and $\kappa_2(A+E)$ for GMW-II, SE99, CH98, and LTL^T-CH98 for 30 random matrices with $n = 25$. Key: GMW-II —, SE99 $+$, CH98 $\times$, LTL^T-CH98 $\square$.

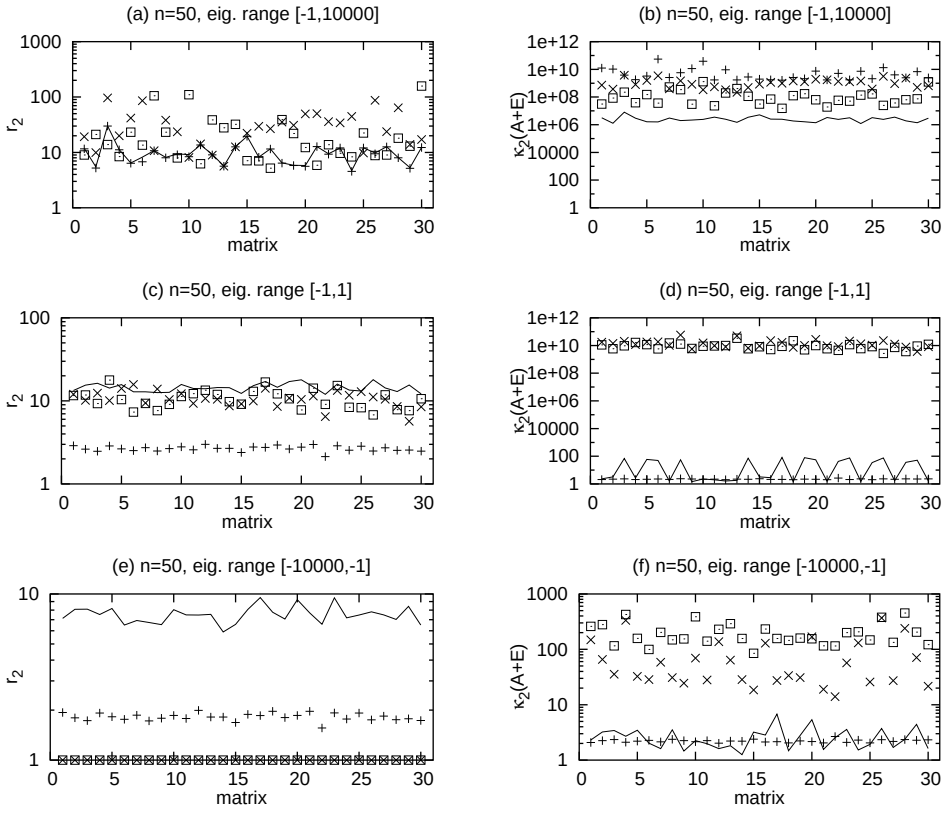

Figure 12: Measures of r_2 and $\kappa_2(A+E)$ for GMW-II, SE99, CH98, and LTL^T-CH98 for 30 random matrices with $n = 50$. Key: GMW-II —, SE99 $+$, CH98 $\times$, LTL^T-CH98 $\square$.

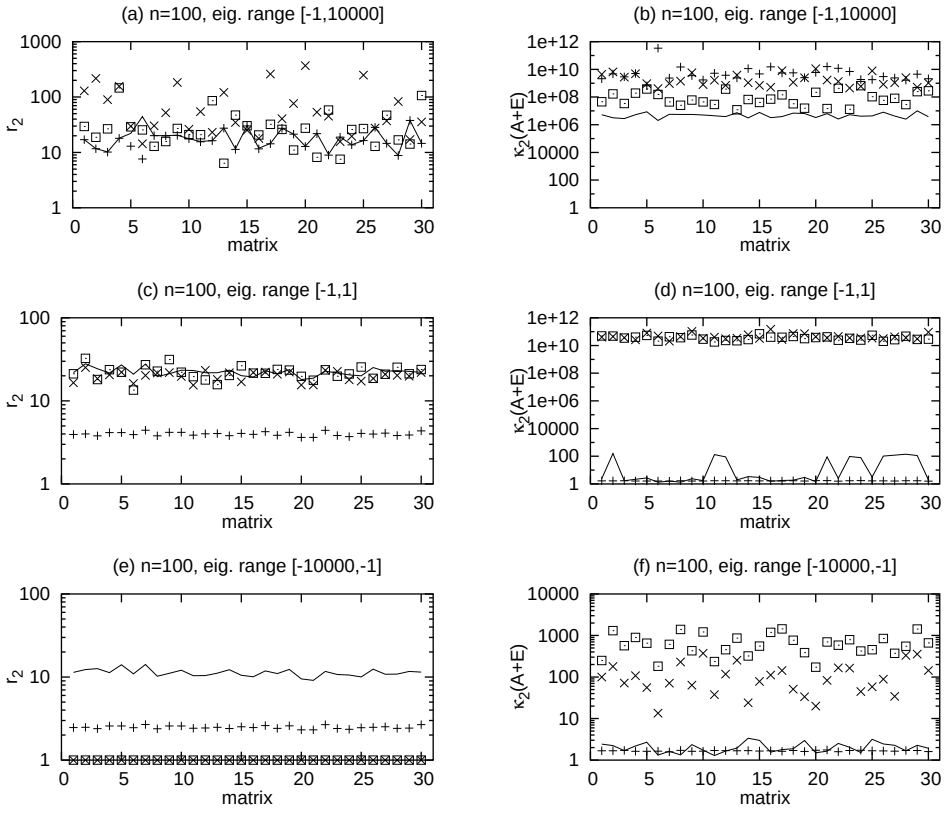

Figure 13: Measures of r_2 and $\kappa_2(A+E)$ for GMW-II, SE99, CH98, and LTL^T-CH98 for 30 random matrices with $n = 100$. Key: GMW-II —, SE99 $+$, CH98 $\times$, LTL^T-CH98 $\square$.

Table 5: Measures of $\|E\|$ and $\kappa_2(A+E)$ for the benchmark matrix (51).

Algorithm	r_2	r_F	$\kappa_2(A+E)$
GMW81	2.733	2.674	4.50×10^4
GMW-I	3.014	2.739	4.51×10^4
GMW-II	2.564	2.489	1.64×10^5
SE90	2.78×10^3	3.70×10^3	8.858
SE99	1.759	1.779	1.04×10^{10}
SE-I	3.346	3.289	3.61×10^4
MS79	3.317	2.689	3.33×10^4
CH98	1.659	1.345	9.88×10^7
LTL^T**-MS79**	3.317	2.689	3.33×10^4
LTL^T**-CH98**	1.658	1.344	6.74×10^{10}
LTL^T**-MS79**, 2-phase	3.317	2.689	3.33×10^4
LTL^T**-CH98**, 2-phase	1.658	1.344	8.59×10^{10}
LTL^T**-MS79**, relaxed 2-phase	2.15×10^4	2.03×10^4	3.68×10^4
LTL^T**-CH98**, relaxed 2-phase	2.15×10^4	2.03×10^4	7.47×10^{10}

The measures of $\|E\|_2$ and $\|E\|_F$ in terms of r_2 and r_F, and the condition numbers $\kappa_2(A+E)$ are listed in Table 5 for various modified Cholesky algorithms, where the new methods are in boldface. This illustrates the instability of incorporating the relaxed 2-phase strategy into LTL^T-CH98 and LTL^T-MS79, where the relaxation factor was $\mu = 0.1$. In this case the instability can be resolved by dropping the relaxation factor down to $\mu = 10^{-4}$. However, the instability was not extinguished for the matrices A15_1, A15_2, and A15_3 in Section 6.3, after trying several different relaxation factors.

6.3 The 33 Matrices

33 matrices, generated by Gay, Overton, and Wright from optimization problems where GMW81 outperformed SE90, were used by Schnabel and Eskow [16] to evaluate modified Cholesky algorithms.

Table 6 summarizes $r_2 = \frac{\|E\|_2}{|\lambda_{\min}(A)|}$ and $\zeta = \lfloor \log_{10}(\kappa_2(A+E)) \rfloor$ for the existing algorithms in the literature, whereas Tables 7 gives the result of the new algorithms. Matrix B13_1 is positive definite but extremely ill-conditioned, so that we measure E by $\|E\|_2$ instead of r_2. We see that SE90 did not perform well on several matrices, and the r_2 for CH98 is somewhat large on a few matrices (e.g., A6_7). The other methods produced a reasonable E in all cases. For these 33 matrices, Type-I algorithms generally resulted in better conditioning of $A+E$, whereas Type-II algorithms generally produced smaller $\|E\|$, except for SE90 and CH98.

Incorporating the special treatments from SE99 for the last 1×1 and 2×2 Schur complements (see (15) and (16)) into GMW-II can often produce slightly smaller

$\|E\|_2$ for matrices close to positive definite. Similarly, the special treatments for SE-I in (27) and (28) can help GMW-I reduce $\|E\|_2$. The detailed discussion is omitted for simplicity.

7 Concluding Remarks

The modified Cholesky algorithms in this paper are categorized in Table 8, where the new methods are in boldface. Our conclusions are listed below.

1. The rationale for the algorithms in the GMW class is to bound the off-diagonal elements in $\bar{L}$. The rationale for the algorithms in the SE class is to control the Gerschgorin circles in the Schur complements.

2. The nondecreasing strategy can be incorporated into virtually all algorithms which confine the modification to the diagonal. The rationale is that it does not increase $\|E\|_2$ at each stage, and it may keep the subsequent modifications smaller. It is especially favored by the Type-II algorithms, since it can also empirically improve the conditioning of $A + E$.

3. The 2-phase and relaxed 2-phase strategies are incorporated into SE90 and SE99 respectively for satisfying Objective 1, whereas they are not required for GMW81.

4. GMW81 and its Type-II variant have $\|E\| = O(n^2)$. The 2-phase strategy can drop the bound to be $\|E\| = O(n)$. However, it may result in excessive $\|E\|_2$ for matrices close to being positive definite. The problem can be solved by relaxing. The situation is similar to that of SE90 and SE99. The relaxed 2-phase strategy usually improves the modified LDL^T algorithms.

5. For algorithms in the GMW class and in the SE class, the theoretical bounds on $\|E\|_2$ and $\kappa_2(A + E)$ do not rely on pivoting. In practice, pivoting reduces $\|E\|_2$.

6. Our GMW-II algorithm outperforms GMW81 and GMW-I by generally keeping $\|E\|_2$ smaller for the random matrices with eigenvalues in $[-1, 10000]$, whereas GMW81 outperforms our GMW-I and GMW-II for the random matrices with eigenvalues in $[-10000, -1]$.

7. In our experiments, SE99 and GMW-II are the best modified LDL^T algorithms with respect to $\|E\|$ for matrices with eigenvalues in $[-1, 10000]$, whereas SE-I generally produces $\|E\|$ smaller than those for SE90 and SE99 for matrices with eigenvalues in $[-10000, -1]$ and $[-1, 1]$.

8. In experiments on distance matrix completion problems we noted that increasing the relaxation factor μ from 0.75 to 1.0 can significantly improve the performance of GMW-II not only for random problems with 65% or more unspecified entries but also for protein problems [7]. With these changes, however, Objective 2 was not satisfied as well for the 33 matrices in Section 6.3.

Table 6: $r_2 = \frac{\|E\|_2}{|\lambda_{\min}(A)|}$ and $\zeta = \lfloor \log_{10}(\kappa_2(A+E)) \rfloor$ of the existing methods.

Method	GMW81		SE90		SE99		MS79		CH98	
	r_2	ζ	r_2	ζ	r_2	ζ	r_2	ζ	r_2	ζ
A6_1	1.365	5	3.6e+2	1	1.079	9	2.188	5	1.094	8
A6_2	4.844	3	1.175	5	1.180	7	2.304	3	1.152	7
A6_3	4.847	4	1.200	5	1.208	6	2.328	3	1.164	7
A6_4	2.501	5	1.275	5	1.270	8	2.541	4	1.271	8
A6_5	2.347	5	6.503	3	1.448	9	4.512	5	2.257	8
A6_6	1.693	8	2.947	5	1.201	10	2.757	8	1.384	8
A6_7	1.953	12	4.6e+4	5	1.334	10	2.033	12	3.6e+2	8
A6_8	1.953	8	6.611	5	1.138	10	2.033	8	1.030	8
A6_9	1.958	8	47.221	5	1.125	10	2.031	9	1.131	8
A6_10	5.887	8	5.4e+6	0	1.076	11	6.636	8	3.675	8
A6_11	2.334	8	7.3e+6	0	1.648	7	6.049	8	3.570	8
A6_12	4.847	4	1.200	5	1.208	6	2.328	3	1.164	7
A6_13	2.180	2	1.322	5	1.322	6	3.115	2	1.558	8
A6_14	4.847	4	1.200	5	1.208	6	2.328	3	1.164	7
A6_15	5.188	1	1.090	5	1.090	5	2.146	1	1.073	7
A6_16	2.180	2	1.322	5	1.322	6	3.115	2	1.558	8
A6_17	1.527	2	1.246	5	1.246	6	2.752	2	1.376	8
A13_1	2.253	10	8.9e+3	5	1.183	10	3.847	9	57.944	8
A13_2	2.599	8	1.5e+4	5	1.317	10	2.805	8	4.716	8
A15_1	2.421	9	2.5e+7	5	1.895	11	4.165	10	5.954	8
A15_2	2.375	9	3.9e+5	3	1.449	10	2.834	10	9.948	8
A15_3	1.957	6	2.183	5	1.503	10	3.991	7	2.021	8
B6_1	4.901	3	52.418	0	1.773	8	3.024	2	1.512	8
B6_2	4.495	2	45.866	0	2.315	7	4.200	3	2.100	8
B7_1	1.666	2	3.450	2	1.067	2	2.263	2	1.131	8
B7_2	1.932	2	11.005	0	1.309	7	3.320	2	1.660	7
B7_3	1.967	2	6.998	0	1.227	6	2.669	2	1.334	7
B7_4	1.929	2	5.325	1	1.189	6	2.619	2	1.310	7
B8_1	4.164	12	8.7e+2	5	1.279	10	4.164	12	9.705	8
B13_1 (abs.)	0	9	27.15	5	0	9	0	9	0.215	7
B13_2	1.762	7	7.846	5	1.291	10	3.887	7	1.949	8
B26_1	9.833	1	2.234	3	2.364	7	28.293	2	14.146	8
B55_1	3.504	1	1.714	5	1.714	6	95.603	3	47.802	9

Table 7: $r_2 = \frac{\|E\|_2}{|\lambda_{\min}(A)|}$ and $\zeta = \lfloor \log_{10}(\kappa_2(A+E)) \rfloor$ of the new methods.

Method	GMW-I		GMW-II		SE-I		LTL^T–MS79		LTL^T–CH98	
	r_2	ζ	r_2	ζ	r_2	ζ	r_2	ζ	r_2	ζ
A6_1	1.989	4	2.111	5	2.181	4	2.153	5	1.077	11
A6_2	4.265	3	3.881	2	2.360	3	2.306	3	1.153	10
A6_3	5.160	2	2.528	11	2.416	2	2.323	3	1.162	10
A6_4	2.574	3	1.290	10	2.541	3	2.756	4	1.378	10
A6_5	3.120	4	1.647	10	2.895	4	3.111	4	1.556	10
A6_6	2.363	7	1.418	5	2.403	6	2.754	8	1.377	10
A6_7	2.189	11	1.331	10	2.109	10	2.032	11	1.352	10
A6_8	2.189	7	1.051	10	2.277	7	2.032	8	1.016	10
A6_9	2.179	8	1.064	10	2.249	8	2.031	9	1.016	10
A6_10	4.883	7	4.399	6	2.031	7	2.438	8	1.219	11
A6_11	2.550	7	2.311	7	1.737	7	3.604	8	1.802	11
A6_12	5.160	2	2.528	11	2.416	2	2.323	3	1.162	10
A6_13	3.289	1	2.971	1	2.643	1	2.911	2	1.455	11
A6_14	5.160	2	2.528	11	2.416	2	2.323	3	1.162	10
A6_15	5.338	1	2.666	11	2.181	1	3.195	1	1.597	10
A6_16	3.289	1	2.971	1	2.643	1	2.911	2	1.455	11
A6_17	2.713	1	2.461	1	2.492	1	2.519	2	1.259	11
A13_1	2.288	10	1.198	10	2.257	9	2.258	9	1.184	10
A13_2	2.767	8	1.406	10	2.627	8	2.642	8	1.324	10
A15_1	5.718	9	5.372	8	3.815	8	4.886	10	2.444	11
A15_2	2.925	8	2.728	8	2.887	8	2.834	10	1.432	10
A15_3	3.953	6	3.789	6	3.006	6	2.689	7	1.344	11
B6_1	2.817	2	2.512	2	3.545	2	2.224	2	1.112	11
B6_2	3.367	2	3.061	2	4.630	2	2.398	2	1.199	11
B7_1	2.062	2	1.663	2	2.005	2	2.019	2	1.010	11
B7_2	2.721	2	1.449	11	2.618	1	7.217	2	3.609	11
B7_3	2.610	2	1.377	11	2.453	1	6.795	2	3.397	11
B7_4	2.538	2	1.337	11	2.378	1	2.683	2	1.342	10
B8_1	4.164	12	2.087	11	2.548	10	4.022	12	2.017	11
B13_1	0	9	0	9	0	9	0	9	0	9
B13_2	5.273	7	4.859	5	2.581	6	2.405	7	1.203	10
B26_1	6.639	1	3.721	2	5.827	1	17.386	2	8.693	11
B55_1	3.504	1	1.752	10	3.428	1	11.289	1	5.645	10

Table 8: Categories of various modified Cholesky algorithms.

Category	*Type I*	*Type II*
LDL^T	GMW81, **GMW-I**, **SE-I**	**GMW-II**, SE90, SE99
LBL^T	MS79	CH98
LTL^T	LTL^T**-MS79**	LTL^T**-CH98**

For the modified LBL^T factorizations and our new approach via the LTL^T factorization, the concluding remarks are as follows.

1. In worst cases, MS79 and CH98 take $\Theta(n^3)$ time more than the standard Cholesky factorization and therefore do not satisfy Objective 4, whereas our LTL^T-MS79 and LTL^T-CH98 algorithms guarantee $O(n^2)$ modification expense.

2. In experiments on random matrices with eigenvalues $[-1, 10000]$, LTL^T-MS79 and LTL^T-CH98 usually produce an $\|E\|_2$ smaller than MS79 and CH98, respectively. Our new approach outperforms the modified LBL^T algorithms in the literature, not only by guaranteeing the $O(n^2)$ modification cost, but also by usually producing a smaller $\|E\|_2$ for matrices close to being positive definite.

3. It is possible to incorporate the 2-phase strategy or the relaxed 2-phase strategy into LTL^T-MS79 and LTL^T-CH98, but the resulting algorithms may produce unreasonably large $\|E\|$, as shown in Figure 7 and discussed in Section 6.2, respectively.

4. The modification arguments δ listed in Table 1 aimed at the satisfaction the four objectives. In practice, especially for Type-II algorithms, they could be too small and affect the conditioning, from which difficulty may arise. In our experiments on random distance matrix completion problems [7], difficulty was apparent for CH98 and LTL^T-CH98. To amend the problem, we increased the modification tolerance parameter δ to $\tau\eta$ (used by SE90) for both CH98 and LTL^T-CH98.

The best algorithm, modification tolerance δ, and relaxation factor μ for the relaxed 2-phase strategy depend on the optimization problem. Experiments may be required to tune δ and μ for each application.

Acknowledgement

The authors thank Betty Eskow for very helpful discussions and for making her code from [16] available to us.

References

[1] J. O. Aasen. On the reduction of a symmetric matrix to tridiagonal form. *BIT*, 11(3):233–242, 1971.

[2] C. Ashcraft, R. G. Grimes, and J. G. Lewis. Accurate symmetric indefinite linear equation solvers. *SIAM J. Matrix Anal. Appl.*, 20(2):513–561, 1998.

[3] J. R. Bunch and L. Kaufman. Some stable methods for calculating inertia and solving symmetric linear systems. *Math. Comp.*, 31:163–179, 1977.

[4] J. R. Bunch and B. N. Parlett. Direct methods for solving symmetric indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 8(4):639–655, 1971.

[5] S. H. Cheng and N. J. Higham. A modified Cholesky algorithm based on a symmetric indefinite factorization. *SIAM J. Matrix Anal. Appl.*, 19(4):1097–1110, 1998.

[6] H.-r. Fang and D. P. O'Leary. Stable factorizations of symmetric tridiagonal and triadic matrices. *SIAM J. Matrix Anal. Appl., to Appear*, 2006.

[7] H.-r. Fang and D. P. O'Leary. Dimensional relaxation for distance matrix completion problems. Technical report, Computer Science Department, Univ. of Maryland, College Park, MD, in preparation.

[8] A. Forsgren, P. E. Gill, and W. Murray. Computing modified Newton directions using a partial Cholesky factorization. *SIAM J. Sci. Comput.*, 16(1):139–150, 1995.

[9] P. E. Gill and W. Murray. Newton-type methods for unconstrained and linearly constrained optimization. *Math. Programming*, 28:311–350, 1974.

[10] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, 1981.

[11] N. J. Higham. Stability of the diagonal pivoting method with partial pivoting. *SIAM J. Matrix Anal. Appl.*, 18(1):52–65, 1997.

[12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[13] J. J. Moré and D. C. Sorensen. On the use of directions of negative curvature in a modified Newton method. *Math. Programming*, 16:1–20, 1979.

[14] B. N. Parlett and J. K. Reid. On the solution of a system of linear equations whose matrix is symmetric but not definite. *BIT*, 10(3):386–397, 1970.

[15] R. B. Schnabel and E. Eskow. A new modified Cholesky factorization. *SIAM J. Sci. Stat. Comput.*, 11:1136–1158, 1990.

[16] R. B. Schnabel and E. Eskow. A revised modified Cholesky factorization algorithm. *SIAM J. Optim.*, 9(4):1135–1148, 1999.

[17] G. W. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimation. *SIAM J. Numer. Anal.*, 17:403–409, 1980.