Validated Solution of Large Linear Systems[1]

Siegfried M. Rump

Dedicated to U. Kulisch on the occasion of his 60th birthday

Abstract

Some new methods will be presented for computing verified inclusions of the solution of large linear systems. The matrix of the linear system is typically of band or sparse structure. There are no prerequisites to the matrix such as being M-matrix, symmetric, positive definite or diagonally dominant. For general band matrices of lower, upper bandwidth p, q of dimension n the computing time is less than $n \cdot (pq + p^2 + q^2)$. Examples with up to 1.000.000 unknowns will be presented.

Zusammenfassung

Es werden neuartige Methoden vorgestellt zur Berechnung sicherer Schranken der Lösung großer linearer Gleichungssysteme. Die Matrix des Gleichungssystems hat typischerweise Bandstruktur oder ist spärlich besetzt. Es werden keinerlei Voraussetzungen an die Matrix gestellt wie etwa M-Matrix, symmetrisch, positiv definit oder diagonal dominant. Für Bandmatrizen von oberer bzw. unterer Bandbreite p bzw. q der Dimension n ist die Rechenzeit kleiner als $n \cdot (pq + p^2 + q^2)$. Es werden Beispiele bis Dimension 1.000.000 diskutiert.

0 Notation

Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}^n$ vectors and $\mathbb{R}^{n \times n}$ matrices over those. The letter n is only used for the dimension of vectors and matrices, others then n-vectors and $n \times n$-matrices do not occur in this paper.

$\mathbb{P}T$ denotes the power set over T, $\mathbb{I}T$ the interval extension for $T \in \{\mathbb{R}, \mathbb{R}^n, \mathbb{R}^{n \times n}\}$. Usually hyperrectangulars are used but others are not excluded. It should be stressed that interval operations producing validated bounds are rigorously and very efficiently implementable on digital computers, see [25], [1], [5], [28] for details.

[1] published in R. Albrecht et al. (eds.): Validation numerics: theory and applications, vol. 9 of Computing Supplementum , pp. 191–212, Springer 1993

Intervals are written in brackets $a \pm b$ denotes the interval $[a-b, a+b]$, for some interval $[X]$ is $|[X]| := \max\{\,|x| \,|\, x \in [X]\,\}, \mathrm{mid}([X])$ denotes the midpoint, $\mathrm{rad}([X])$ the radius of an interval $[X]$. Those terms apply to vectors and matrices componentwise.

The interior of a set is denoted by int, ρ denotes the spectral radius of a matrix and $\rho([A]) := \max\{\,\rho(A) \mid A \in [A]\,\}$ for $[A] \in \mathbb{IIR}^{n\times n}$. An interval linear system is sometimes written in short notation $[A] \cdot x = [b]$, solving it means to compute bounds for

$$\sum([A],[b]) := \{\, x \in \mathbb{R}^n \mid \exists\, A \in [A], b \in [b] \quad \text{with} \quad Ax = b\,\}.$$

$\sigma_1, \ldots, \sigma_n$ denote the singular values of a matrix in nonincreasing order such that $\sigma_1 = \|A\|_2$.

If not stated otherwise all operations are real or floating-point operations. We use operations $\triangle$ with upwardly directed rounding, $* \in \{+, -, \cdot, /\}$ having the property

$$a \mathbin{\triangle} b \geq a * b$$

where the latter operation $*$ is the real operation. In case a, b are vectors or matrices the $\leq$-sign applies componentwise.

1 Introduction

Few papers are known dealing with the problem of finding validated inclusions for the solution of sparse linear systems without calculating an approximate inverse of the system matrix. All of the papers known to the author not using an approximate inverse require special properties of the system matrix which is essentially being an M-matrix. The approximate inverse of a sparse matrix is in general full thus limiting the size of the tractable problems significantly. This is because of limitations in memory and because for banded systems the computing time depends quadratically on n. Our goal is to go large sizes, that is 100.000 unknowns and beyond and to keep the computing time for banded systems linearly dependant on n.

There are very interesting papers for condition estimation of sparse matrices (cf. [4], [13]). However, these are yielding estimations rather than verified bounds.

In this paper we describe our method for banded linear systems. The numerical examples are for banded systems, too. For sparse systems the techniques for reducing bandwidth, symbolic pivoting and others (cf. [13]) can be applied. The resulting linear system of reduced bandwidth can be treated by our methods.

There is one yet unpublished method without using an approximate inverse and without prerequisites on the matrix by Jansson [20]. Despite that there are essentially two different approaches known in the literature. The first is the direct extension of some numerical decomposition algorithm by means of replacing every real operation by the corresponding

interval operation. It has been shown that, for example, the interval version of Gaussian elimination is executable in this way for diagonally dominant matrices or M-matrices. In the general case intervals tend to grow in diameter rapidly due to data dependencies such that soon a pivot column only consists of intervals containing zero and the algorithm stops prematurely. This effect depends mainly on the dimension, not on the condition number. As a rule of thumb for general matrices with floating-point input data, for example for random matrices, the range of application of this approach is limited to dimension 50 when calculating in double precision which is roughly 17 decimals. The dimension is even more limited for interval input data.

The other approach uses fixed point methods. We shortly describe this ansatz because it gives insight in the problems we have to deal with.

Let a linear system $Ax = b$ with matrix $A \in \mathbb{R}^{n\times n}$ and right hand side $b \in \mathbb{R}^n$ be given together with some $\widetilde{x} \in \mathbb{R}^n, R \in \mathbb{R}^{n\times n}$. $\widetilde{x}$ is considered to be an approximate solution to the linear system, R an approximate inverse of A. Krawczyk [21], [22] defines for $X \in \mathbb{IR}^n$ the following operator

$$K(X) := \widetilde{x} - R \cdot (A\widetilde{x} - b) + (I - RA) \cdot (X - \widetilde{x}). \tag{1}$$

He shows that

$$\|I - RA\| < 1 \quad \text{and} \quad K(X) \subseteq X \quad \text{implies} \quad \exists\, \widehat{x} \in X : A\widehat{x} = b$$

(see also [26], [27]). In [29] it has been shown that the assumption $\|I - RA\| < 1$ can be replaced by $K(X) \subseteq int(X)$. Algorithms were designed to compute validated inclusions for the solution of general nonlinear systems [30]. There are a number of specializations to specific problems such as polynomial zeros [7], algebraic eigenproblems [29], evaluation of arithmetic expressions [8] and others taking advantage of the special situation. A basic theorem for linear systems is as follows.

Theorem 1.1. Let $\mathcal{A} \in \mathbb{IPIR}^{n\times n}, \mathcal{B} \in \mathbb{IPIR}^n$ be given and let $\widetilde{x} \in \mathbb{R}^n, R \in \mathbb{R}^{n\times n}, \emptyset \neq X \in \mathbb{IPIR}^n, X$ being compact. Define

$$\mathcal{Z} := R \cdot (\mathcal{B} - \mathcal{A}\widetilde{x}) \quad \text{and} \quad \mathcal{C} := I - R \cdot \mathcal{A}, \tag{2}$$

$$L(X) := \mathcal{Z} + \mathcal{C} \cdot X, \tag{3}$$

all operations being power set operations. If

$$L(X) \subseteq \text{int}(X) \tag{4}$$

then R and every $A \in \mathbb{IR}^{n\times n}, A \in \mathcal{A}$ is nonsingular and for every $b \in \mathcal{B}$ the unique solution $\widehat{x} := A^{-1}b$ satisfies

$$\widehat{x} \in \widetilde{x} + L(X). \tag{5}$$

The proof consists of three basic steps. First take fixed but arbitrary $A \in \mathcal{A}, b \in \mathcal{B}$ thus reducing the problem to a point problem. Second, show that $C := I - RA \in \mathcal{C}$ is convergent ($\rho(C) < 1$) and therefore A and R are nonsingular. Moreover, the iteration $x^{k+1} := R(b - A\widetilde{x}) + C \cdot x^k$ has a unique fixed point $\widehat{x} \in X$. Third show that this fixed point is the (unique) solution of $Ax = b$.

Thus theorem 1 already verifies the solvability of the linear system and gives a sufficient criterion for some $X \in \mathbb{IPIR}^n$ for including the solution. To devise an algorithm for *finding* a validated inclusion $[X]$ we have to solve two problems. First the operations have to become executable on the computer and second we need a constructive way to obtain a suitable $[X]$. The first problem is solved by using interval operations rather than power set operations. On the computer floating-point bounds for the intervals are used. Then systems with $[A] \in \mathbb{IIR}^{n\times n}, [b] \in \mathbb{IIR}^n$ can be attacked. This includes for example point matrices the entries of which not being exactly representable on the computer by replacing those by the smallest enclosing machine interval (see [1], [27], [5], [28]).

For the second problem we use an iteration with a so-called ϵ-inflation (see [29], [31]). In this technique for a starting interval $[X] := [Z] := R \cdot ([b] - [A] \cdot \widetilde{x})$ the iterated interval is made "fatter" in every step. This is used in combination with an Einzelschrittverfahren. It can be shown [31] that a validated inclusion will be found

- for a point system $Ax = b$ and *power set operations*
  iff $\rho(I - R \cdot A) < 1$
- for an interval linear system $[A]x = [b]$ and *interval operations*
  iff $\rho(|I - R \cdot [A]|) < 1$.

All of the fixed point methods known in the literature basically use theorem 1, especially (1.2) - (1.4), in one or the other way. Thus in our discussions for sparse linear systems we may concantrate on how to satisfy those conditions.

For simplicity let a point linear system $Ax = b, A \in \mathbb{IR}^{n\times n}, b \in \mathbb{IR}^n$ be given. We do not impose restrictions on A or b. For large banded or sparse linear systems the original approach cannot be used because it needs an approximate inverse R of A which is in general full. We

may omit this by using some decomposition of A. For $A = LU$ and $R = U^{-1}L^{-1}$ we obtain for $x \in \mathbb{R}^n$

$$R \cdot (b - A\widetilde{x}) + (I - RA)x = U^{-1}L^{-1} \cdot (b - A\widetilde{x} + (LU - A) \cdot x). \tag{6}$$

L and U preserve a banded structure of A. In a practical application we would think of replacing U^{-1} and L^{-1} by an efficient algorithm for solving triangular systems. From a mathematical point of view L and U are arbitrary. If for some $L, U \in \mathbb{R}^{n\times n}$ and $[X] \in \mathbb{IR}^n$ we can show that

$$M([X]) := U^{-1}L^{-1}(b - A\widetilde{x} + (LU - A) \cdot [X]) \subseteq \mathrm{int}([X]) \tag{7}$$

then theorem 1 implies that A is nonsingular and the unique solution $\widehat{x} = A^{-1} \cdot b$ satisfies $\widehat{x} \in \widetilde{x} + M([X])$. $LU - A$ can be estimated during the decomposition of A, most simple and without additional cost for example using Crout's variant. Thus we have reduced our problem to computing a validated inclusion of the solution of a linear system with triangular point matrix and interval right hand side.

In (1.7) $b - A\widetilde{x}$ is of order $\epsilon \cdot \|A\| \cdot \|\widetilde{x}\|$ if $\widetilde{x}$ is a reasonable approximate solution, for example the one computed by floating-point Gaussian elimination. Also, numerical error analysis tells us that $LU - A$ will be of the order $\epsilon \cdot \|A\|$. $[X]$ shall contain the error of the approximate solution $\widetilde{x}$ which means that $(LU - A) \cdot [X]$ will be an interval vector of small magnitude. Thus we would not loose too much accuracy going to intervals being symmetric to the origin. This saves us half of the storage per interval vector. Clearly, for $0 < x \in \mathbb{R}^n$

$$U^{-1}L^{-1} \cdot (|b - A\widetilde{x}| + |LU - A| \cdot x) < x \tag{8}$$

implies A being nonsingular and $A^{-1}b \in \widetilde{x} \pm x$.

This reduces our problem to solving a triangular system with right hand side $[b]$ symmetric to the origin and we may further simplify it to $[b] := [-1, 1]$. In other words find

$$\text{validated bounds for } S := \{\, L^{-1} \cdot b \mid -1 \le b \le 1 \,\},\ L \in \mathbb{R}^{n\times n} \text{ lower triangular.} \tag{9}$$

All of the papers [11], [12], [23] using the fixed point approach solve (1.9) using interval backward substitution:

$$\text{for } i = 1 : n \text{ do } [x]_i = ([-1, +1] - \sum_{j=1}^{i-1} L_{ij} \cdot [x]_j)/L_{ii} \tag{10}$$

all operations in (1.10) being interval operations. Thus the intervals $[x]_j$ are symmetric to the origin and (1.10) can be written using absolute values

$$\text{for } i = 1 : n \text{ do } x_i = (1 + \sum_{j=1}^{i-1} |L_{ij}| \cdot x_j)/|L_{ii}| \tag{11}$$

yielding a true inclusion $S \subseteq [-x, +x]$. The overestimation can be estimated observing $x = \langle L \rangle^{-1} \cdot e$ where $e \in \mathbb{IR}^n, e_i = 1$ for $1 \leq i \leq n$ and $\langle L \rangle$ is Ostrowski's comparison matrix (see [28]):

$$\langle L \rangle_{ij} := \begin{cases} |L_{ii}| & \text{for } i = j \\ -|L_{ij}| & \text{otherwise.} \end{cases}$$

For our special right hand side the maximal overestimation is the ratio

$$\|\langle L \rangle^{-1}\|_\infty / \|L^{-1}\|_\infty. \tag{12}$$

If we could estimate $\|L^{-1}\|_\infty$ then our problem (1.9) would be solved. In practical applications the ratio (1.12) is exponentially increasing with n unless L has special properties. Such properties are A and therefore L and U being M-matrices in which case $L = \langle L \rangle, U = \langle U \rangle$. This is the reason why M-matrices can be solved using interval Gaussian elimination without overestimation. To further illustrate the effect consider the following example due to Neumaier:

$$L = \begin{pmatrix} 1 & & & & & & \\ 1 & 1 & & & & & \\ 1 & 1 & 1 & & & & \\ & 1 & 1 & 1 & & & \\ & & & \ddots & & & \\ & & & & 1 & 1 & 1 \end{pmatrix}, \ [b]_i = [-1, +1]. \tag{13}$$

Using interval backward substitution we obtain with $E := [-1, 1]$

$$[x]_1 = E$$

$$[x]_2 = E - [x]_1 = 2 \cdot E$$

$$[x]_3 = E - [x]_1 - [x]_2 = 4 \cdot E$$

$$[x]_4 = E - [x]_2 - [x]_3 = 7 \cdot E$$

with exponentially growing diameter of $[x]_i$. This can also be seen from $\langle L\rangle^{-1}$ which we show for $n = 7$:

$$\langle L\rangle^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 0 & 0 & 0 & 0 \\ 3 & 2 & 1 & 1 & 0 & 0 & 0 \\ 5 & 3 & 2 & 1 & 1 & 0 & 0 \\ 8 & 5 & 3 & 2 & 1 & 1 & 0 \\ 13 & 8 & 5 & 3 & 2 & 1 & 1 \end{pmatrix}.$$

Thus $[x]$ computed by (1.10) is a huge overestimation of the true solution set which computes to

$$(L^{-1} \cdot [b])_i = \pm|L^{-1}| \cdot E = (i - [i/3]) \cdot E \subseteq n \cdot E.$$

This can be seen from

$$L^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 \end{pmatrix}.$$

Unfortunately, this behaviour is typical for practical examples with matrices without special properties.

Methods based on the first approach (replacing floating-point operations by their corresponding interval operations in some numerical decomposition algorithm) are by their nature essentially restricted to diagonally dominant or inverse positive matrices (see for example [1], [28]. See also [33] for an interval version of Bunemann's algorithm for Poisson equation.

As we have just seen the fixed point approach as described in the literature is restricted to a similar class of matrices. This approach is used in [3], [11], [12], [23].

Using a coded version [3] of this algorithm the effect can be demostrated. We used algorithm DSSSB with IWK = 5 which means the maximum possible amount of work is invested. We used $A = 0.1 \cdot LL^T$ with L from (1.13) and right hand side $(1, 0, \cdots, 0)^T$. The factor 0.1 is used to make the factors of A not exactly representable on the computer. Then using double precision floating-point format which is approximately 17 decimal digits the algorithm fails for $n \geq 41$. For $n = 41$ we have $cond(A) = 2.3\text{e}3$. Taking the matrix (4.20) from [16] with $a = 1$ and the same right hand side $(1, 0, \cdots, 0)^T$ the algorithm fails for $n \geq 48$. For $n = 48$ we have $cond(A) = 42$.

The amount of overestimation (1.12) is displayed in the following table.

n	10	20	30	40	50
$\|\langle L\rangle^{-1}\|_\infty / \|L^{-1}\|_\infty$	20.4	1265	1.0e5	9.9e6	9.7e8

Table 1.1 Overestimation of interval Gaussian elimination for L from (1.13)

The figures demonstrate the exponential behaviour of the overestimation.

2 The method

In order to bound (1.9) we may look for the singular values of L. Let U_r be the unit disk of radius r. Then $\|L^{-1} \cdot u\|_2, u \in U_r$ is bounded by $\sigma_n(L)^{-1} \cdot \|u\|_2 = \sigma_n(L)^{-1} \cdot r$. Thus a validated lower bound on the smallest singular value of a triangular matrix would solve the problem. This, in turn, would also yield a validated condition estimator. The problem of finding fast and reliable (although not validated) condition estimators has been attacked by many authors ([9], [10], [15], [17], [18], [2], [6]).

Given an approximation $\widetilde{\lambda}$ of $\sigma_n(L), \widetilde{\lambda}^2$ is an approximate eigenvalue of LL^T. If for some $\kappa \in \mathbb{R}$ being slightly less than one we could prove that $LL^T - \kappa\widetilde{\lambda}^2 \cdot I$ is positive definite then $\kappa^{1/2} \cdot \widetilde{\lambda}$ proved to be a lower bound of $\sigma_n(L)$.

L is a Cholesky factor of LL^T. The change of the Cholesky factor L into G with $GG^T = LL^T - \widetilde{\lambda}^2 I$ is given by the following formulas:

$$\begin{aligned} \sum_{\nu=1}^{i} G_{i\nu}^2 &= \sum_{\nu=1}^{i} L_{i\nu}^2 - \widetilde{\lambda}^2 \quad \text{for } i = j \\ \sum_{\nu=1}^{j} G_{i\nu} G_{j\nu} &= \sum_{\nu=1}^{j} L_{i\nu} L_{j\nu} \quad \text{for } i > j. \end{aligned} \tag{14}$$

We need, however, a validation for the fact that $LL^T - \widetilde{\lambda}^2 I$ is positive semidefinite. When performing an exact Cholesky factorization of $LL^T - \widetilde{\lambda}^2 I$ this is true if the algorithm is executable, i.e. if the diagonal elements stay nonnegative. Using floating-point operations we have to estimate the rounding errors during the computation. Rather than estimating them a priori by replacing the floating-point operations by the corresponding interval operations we estimate them a posteriori by estimating the difference of GG^T and $LL^T - \widetilde{\lambda}^2 I$ for the computed Cholesky factor G and by using perturbation theory.

For the diagonal elements this means

$$\text{computing} \quad G_{ii} := \Big(\sum_{\nu=1}^{i} L_{i\nu}^2 - \sum_{\nu=1}^{i} G_{i\nu}^2 - \widetilde{\lambda}^2\Big)^{1/2} \quad \text{approximatively and}$$

$$\text{estimating} \quad |(LL^T - \widetilde{\lambda}^2 I - GG^T)_{ii}| = |\sum_{\nu=1}^{i} L_{i\nu}^2 - \sum_{\nu=1}^{i} G_{i\nu}^2 - \widetilde{\lambda}^2| \quad \text{rigorously.}$$

For off-diagonal elements this means

$$\text{computing} \quad G_{ij} := (\sum_{\nu=1}^{j} L_{i\nu}L_{j\nu} - \sum_{\nu=1}^{j-1} G_{i\nu}G_{j\nu})/G_{jj} \quad \text{approximatively and}$$

$$\text{estimating} \quad |(LL^T - \widetilde{\lambda}^2 I - GG^T)_{ij}| = |\sum_{\nu=1}^{j} L_{i\nu}L_{j\nu} - \sum_{\nu=1}^{j} G_{i\nu}G_{j\nu}| \quad \text{rigorously.}$$

The computation and the estimation can essentially be done in one step. First the common part of both sums, resp. is evaluated *with* error estimation, then the midpoint is used for the floating-point component G_{ii}, G_{ij} of G, resp. and the interval part for the error estimation.

If only the four basic interval operations, that is IEEE 754 [19] arithmetic, is available that is the best we can do. If a precise scalar product [24], [25] is available then we can do better. For the diagonal elements we compute the exact value

$$dot := \sum_{\nu=1}^{i} L_{i\nu}^2 - \sum_{\nu=1}^{i-1} G_{i\nu}^2 - \widetilde{\lambda}^2$$

and for S being the value of *dot* rounded to nearest we get $G_{ii} := fl(\sqrt{S})$, that is G_{ii} is the floating-point square root of S. Then we use the accumulating feature of the scalar product and compute the exact value of $dot - G_{ii}^2$. This value rounded to the smallest enclosing interval provides a very sharp bound for the error $(LL^T - \widetilde{\lambda}^2 I - GG^T)_{ii}$. For the off-diagonal elements we proceed in a similar way. To avoid to formulate the algorithm twice we simply state in the diagonal case

Compute $S, \Delta S$ such that

$$\sum_{\nu=1}^{i} L_{i\nu}^2 - \sum_{\nu=1}^{i-1} G_{i\nu}^2 - \widetilde{\lambda}^2 \in S \pm \Delta S.$$

For basic interval operations this means S being the midpoint, ΔS the radius of the left hand side computed in naive interval arithmetic. With the precise scalar product we proceed as described before. The off-diagonal elements are treated similarly.

Having an estimation on $E := LL^T - \widetilde{\lambda}^2 I - GG^T$ and assuming the diagonal of G being nonnegative implies that $LL^T - \widetilde{\lambda}^2 I - E$ is positive semidefinite. Hence perturbation theory tells us that the eigenvalues of $LL^T - \widetilde{\lambda}^2 I$ are not smaller than $-\rho(E)$ (cf. [14], Corollary 8.1.3) and those of LL^T not smaller than $\widetilde{\lambda}^2 - \rho(E)$. Now $\rho(E)$ can be estimated conveniently by $\|E\|_\infty$ which is done in the following algorithm. There the ith row sum is stored in e_i. When computing the ij-th component of G the error $(LL^T - \widetilde{\lambda}^2 I - GG^T)_{ij}$ contributes to e_i and e_j due to symmetry. To obtain an upper bound on $\|E\|_\infty$ upward directed rounding is used in the computation of the e_i and e_{max}.

We give the algorithm for full matrix L. It can be altered for band matrices in a straightforward manner. Pivoting is omitted because $LL^T - \widetilde{\lambda}^2 I$ is (hopefully) positive definite.

Given nonsingular lower triangular $L \in \mathbb{R}^{n\times n}$ and $\widetilde{\lambda} \in \mathbb{R}$ do
$e_{\max} := 0$
for $i = 1 : n$ do $e_i := 0$;
for $i = 1 : n$ do
  for $j = 1 : i-1$ do
    Compute $S, \Delta S$ such that
      $\sum_{\nu=1}^{j} L_{i\nu}L_{j\nu} - \sum_{\nu=1}^{j-1} G_{i\nu}G_{j\nu} \in S \pm \Delta S;$
    $G_{ij} := fl(S/G_{jj});$
    Compute ΔT such that
      $|S - G_{ij}G_{jj}| \le \Delta T;$
    $d := \Delta S \boxplus \Delta T; e_i := e_i \boxplus d; e_j := e_j \boxplus d;$
  Compute $S, \Delta S$ such that
    $\sum_{\nu=1}^{i} L_{i\nu}^2 - \sum_{\nu=1}^{i-1} G_{i\nu}^2 - \widetilde{\lambda}^2 \in S \pm \Delta S;$
  $G_{ii} := fl(\sqrt{S});$
  Compute ΔT such that
    $|S - G_{ii}^2| \le \Delta T;$
  $e_i := e_i \boxplus \Delta S \boxplus \Delta T;$
$e_{\max} := \max_i e;$

Algorithm 2.1 Cholesky factorization of $LL^T - \widetilde{\lambda}^2 I$ with lower bound for $\sigma_n(L)$

In precise computation, ΔS and ΔT as well as T would be zero according to (2.1). The main effort in the algorithm goes into the two inner products for computing S together with a validated bound. If L is a lower triangular band matrix of bandwidth p then the vector e needs only to be of length $p+1$ storing the values cyclically. Also, G needs only $(p+1)*(p+1)$ elements of storage.

It should be stressed that G is computed in floating-point arithmetic without presumptions on its accuracy. If the algorithm finishes successfully, i.e. the radicands are nonnegative, then the G_{ii} are nonnegative and therefore GG^T is positive semidefinite with

$$\|(LL^T - \widetilde{\lambda}^2 I) - GG^T\|_\infty \le e_{\max}. \tag{15}$$

The eigenvalues of LL^T are the squared singular values of L and are bounded from below by $\widetilde{\lambda}^2 - e_{\max}$. This establishes the following theorem.

Theorem 2.1. If algorithm 2.1 finishes sucessfully (all square roots real) then $LL^T - (\widetilde{\lambda}^2 - e_{\max})I$ is positive semidefinite. If $\widetilde{\lambda}^2 \ge e_{\max}$ then

$$\sigma_n(L) \geq (\widetilde{\lambda}^2 - e_{\max})^{1/2}.$$

The computing time for L with lower bandwidth p ($L_{ij} = 0$ for $i > j + p$) is less than $n \cdot p^2 + O(np)$ multiplications and additions plus $n(p+1)$ divisions and n square roots.

Proof. The first part has been proved above, the computing time is a straightforward operation count.

■

In our applications we are particularly interested in sparse linear systems. This fact should be taken into account when implementing algorithm 2.1. For example, in case of a band matrix L the scalar products become very short compared to n.

Theorem 2.1 can be applied as follows. Consider some decomposition of A, for example $\widetilde{L}\widetilde{U} \approx A$ with $\widetilde{A} := \widetilde{L}\widetilde{U}$. Then traditional norm estimates can be used to compute validated bounds for the solution together with theorem 2.1.

Theorem 2.2. Let $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$ be given as well as nonsingular $\widetilde{A} \in \mathbb{R}^{n\times n}$ and $\widetilde{x} \in \mathbb{R}^n$. Define $\Delta A := \widetilde{A} - A$ and suppose $\sigma_n(\widetilde{A}) > n^{1/2} \cdot \|\Delta A\|_\infty$.

Then A is not singular and for $\widehat{x} := A^{-1}b$ holds

$$\|\widehat{x} - \widetilde{x}\|_\infty \leq \frac{n^{1/2} \cdot \|b - A\widetilde{x}\|_\infty}{\sigma_n(\widetilde{A}) - n^{1/2} \cdot \|\Delta A\|_\infty}. \tag{16}$$

Proof. Since $\|\widetilde{A}^{-1} \cdot \Delta A\|_2 \leq \sigma_n(\widetilde{A})^{-1} \cdot \|\Delta A\|_2 \leq n^{1/2} \cdot \sigma_n(\widetilde{A})^{-1} \cdot \|\Delta A\|_\infty < 1$ the matrix $I - \widetilde{A}^{-1} \cdot \Delta A = \widetilde{A}^{-1} \cdot A$ and hence A is invertible. Now

$$(I - \widetilde{A}^{-1} \cdot \Delta A)(\widehat{x} - \widetilde{x}) = \widetilde{A}^{-1} \cdot A \cdot (\widehat{x} - \widetilde{x}) = \widetilde{A}^{-1} \cdot (b - A\widetilde{x}).$$

Using $\|(I - F)^{-1}\| \leq (1 - \|F\|)^{-1}$ for convergent $F \in \mathbb{R}^{n\times n}$ this implies

$$\|\widehat{x} - \widetilde{x}\|_\infty \leq \frac{\|\widetilde{A}^{-1} \cdot (b - A\widetilde{x})\|_\infty}{1 - \|\widetilde{A}^{-1} \cdot \Delta A\|_\infty} \tag{17}$$

and with $\|B\|_\infty \leq n^{1/2} \cdot \|B\|_2$ for $B \in \mathbb{R}^{n\times n}$

$$\|\widehat{x} - \widetilde{x}\|_\infty \leq \frac{n^{1/2} \cdot \sigma_n(\widetilde{A})^{-1} \cdot \|b - A\widetilde{x}\|_\infty}{1 - n^{1/2} \cdot \sigma_n(\widetilde{A})^{-1} \cdot \|\Delta A\|_\infty}$$

proving the theorem. ■

In a practical application $\widetilde{A}$ is some floating-point decomposition of A, for example $\widetilde{A} = \widetilde{L}\widetilde{U}$. Then the application of theorem 2.2 runs as follows. The nonsingularity of $\widetilde{A}$ is obvious. Compute an approximate solution $\widetilde{x}$ of $\widetilde{A}x = b$ and a lower bound for $\sigma_n(\widetilde{A})$ by

$$\sigma_n(\widetilde{A}) = \sigma_n(\widetilde{L}\widetilde{U}) \geq \sigma_n(\widetilde{L}) \cdot \sigma_n(\widetilde{U})$$

and theorem 2.1. Then check for $\sigma_n(\widetilde{A}) > n^{1/2} \cdot \|\Delta A\|_\infty$ to satisfy the conditions of theorem 2.1.

In the estimation (2.3) one may try to avoid or diminish the factor $n^{1/2}$. If $B \in \mathbb{R}^{n\times n}$ is sparse with at most $\tau(B)$ elements per row, then it is not difficult to prove

$$\|B\|_\infty \leq \tau(B)^{1/2} \cdot \|B\|_2.$$

One may hope that given B with $\tau(B^{-1})$ small then also the factor $n^{1/2}$ may be decreased. Unfortunately, this is not true. Consider

$$B^{-1} = \begin{pmatrix} 1 & & & & & & \\ & -1 & & & & & \\ & 1 & -1 & & & & \\ & & 1 & \ddots & & & \\ & & & \ddots & -1 & & \\ & & & & 1 & \varepsilon & \\ & & & & & & \end{pmatrix}.$$

Then for small ε

$$\|B\|_2 \approx \varepsilon^{-1} \cdot n^{1/2} \ , \ \|B\|_\infty \approx \varepsilon^{-1} \cdot n \text{ with } \|B\|_\infty / \|B\|_2 \lesssim n^{1/2}.$$

To apply theorem 2.1 in order to obtain a lower bound on the smallest singular value of a triangular matrix L we need an approximation $\widetilde{\lambda} \approx \sigma_n(L)$. There are two ways to obtain such an approximation. First, we could use our favourite condition estimator. This is fast and, according to our experimental results, works fine. The second method would be to apply inverse power iteration to LL^T using forward and backward substitution. The heuristic is that an L occuring in practice has a well separated smallest singular value. Due to our practical results in most cases 2 or 3 iterations sufficed to generate 3 correct decimal digits. This is far more than necessary.

In the applications we have in mind we can do better using a 2-norm estimate and taking advantage of the structure of A. Let $A \in \mathbb{R}^{n\times n}$ be a band matrix of lower, upper bandwidth p, q, that is

$$A_{ij} = 0 \quad \text{for} \quad i > j + p \quad \text{and for} \quad j > i + q.$$

Let $\alpha := \max |A_{ij}|$. Then both $\|A\|_1$ and $\|A\|_\infty$ are bounded by $(p + q + 1) \cdot \alpha$ and using $\|A\|_2^2 \leq \|A\|_1 \cdot \|A\|_\infty$ yields

$$\|A\|_2 \leq (p + q + 1) \cdot \alpha. \tag{18}$$

Hence $\|A\|_2 \leq (\|\Delta A\|_1 \cdot \|\Delta A\|_\infty)^{1/2}$ will in general be smaller than $n^{1/2} \cdot \|\Delta A\|_\infty$. Applying this to (2.4) we obtain the following result.

Theorem 2.3. Let $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$ be given as well as nonsingular $\widetilde{A} \in \mathbb{R}^{n\times n}$ and $\widetilde{x} \in \mathbb{R}^n$. Define $\Delta A := \widetilde{A} - A$ and suppose $\sigma_n(\widetilde{A}) > (\|\Delta A\|_1 \cdot \|\Delta A\|_\infty)^{1/2}$.

Then A is not singular and for $\widehat{x} := A^{-1}b$ holds

$$\|\widehat{x} - \widetilde{x}\|_\infty \leq \|\widehat{x} - \widetilde{x}\|_2 \leq \frac{\|b - A\widetilde{x}\|_2}{\sigma_n(\widetilde{A}) - (\|\Delta A\|_1 \cdot \|\Delta A\|_\infty)^{1/2}}. \tag{19}$$

Theorem 2.3 also follows by a fixed point argument. Using $\widetilde{A} = LU$ and a disk of radius r instead of $[x]$ in (2.7) gives according to theorem 1.1

$$\begin{array}{l} \widetilde{A}^{-1} \cdot (b - A\widetilde{x} + (\widetilde{A} - A) \cdot U_r) \subseteq \mathrm{int}(U_r) \;\Rightarrow \\ A \text{ is not singular and } A^{-1}b \in \widetilde{x} + U_r. \end{array} \tag{20}$$

The inclusion in (2.7) is satisfied if

$$\sigma_n(\widetilde{A})^{-1} \cdot (\|b - A\widetilde{x}\|_2 + r \cdot \|\Delta A\|_2) < r.$$

This yields a bound on r and with a continuouity argument (2.6).

The heuristic is that the elements of ΔA are roughly of the same size, namely $\epsilon\|A\|$. In the application of (2.6) we have to check $\sigma_n(\widetilde{A}) > (\|\Delta A\|_1 \cdot \|\Delta A\|_\infty)^{1/2}$ to verify $\rho(\widetilde{A}^{-1} \cdot \Delta A) < 1$ which is according to (2.5) more likely to happen than $\sigma_n(\widetilde{A}) > n^{1/2} \cdot \|\Delta A\|_\infty$. Moreover, computing $\widetilde{x}$ by Gaussion elimination we know that the residual $\|b - A\widetilde{x}\|$ will be of the order $\epsilon \cdot \|A\| \cdot \|\widetilde{x}\|$ (cf. [14]).

In the following we add some computational hints for specific cases being relevant in practice.

I) *A is M-matrix.* Apply [32].

If A is symmetric positive definite we can use algorithm 2.1 to calculate a lower bound for $\sigma_n(A)$ directly. when replacing

$$\begin{array}{lllllll} \sum_{\nu=1}^{j} L_{i\nu}L_{j\nu} & \text{by} & A_{ij} & \text{in} & \text{row} & 7 & \text{and} \\ \sum_{\nu=1}^{i} L_{i\nu}^2 & \text{by} & A_{ii} & \text{in} & \text{row} & 13 & , \end{array}$$

then obviously

$$\sigma_n(A) \geq (\widetilde{\lambda}^2 - e_{max})^{1/2}.$$

Replacing $\widetilde{A}$ by A in theorem 2.3 then yields

$$\|\widehat{x} - \widetilde{x}\|_\infty \leq \|\widehat{x} - \widetilde{x}\|_2 \leq (\widetilde{\lambda}^2 - e_{max})^{-1/2} \cdot \|b - A\widetilde{x}\|_2. \tag{21}$$

II) *A is symmetric positive definite.* Compute a floating-point Cholesky decomposition $A \approx \widetilde{G}\widetilde{G}^T$ and an approximation $\widetilde{\sigma}$ of the smallest singular value of A. Apply algorithm 2.1 altered as described above with $\widetilde{\lambda} = 0.9 \cdot \widetilde{\sigma}$ to compute a lower bound on $\sigma_n(A)$ and apply (2.8).

In case A is not symmetric positive definite one may use the following method. Having some approximate decomposition $A \approx \widetilde{F} \cdot \widetilde{G}$ compute an approximation $\widetilde{\lambda}$ to the smallest singular value of A by inverse power method applied to $\widetilde{F}\widetilde{G} \cdot (\widetilde{F}\widetilde{G})^T$. If $\widetilde{F}, \widetilde{G}$ are triangular this is inexpensive. Then apply theorem 2.1 with some obvious modifications to $A^TA - \widetilde{\lambda}^2 I$ to bound $\sigma_n(A^TA)$.

This approach is working only for moderate condition numbers because the condition number of A^TA is that of A squared. For working precision ϵ this limits the scope of application to $\mathrm{cond}(A) < \epsilon^{-1/2}$ rather than $\mathrm{cond}(A) < \epsilon^{-1}$.

In contrast we estimate the smallest singular value of the factors of the decomposition separately. We have to take provision that the condition numbers of the factors are of the same order, namely $\mathrm{cond}(A)^{1/2}$. In this case the square of the condition number of the factors is still of the order of $\mathrm{cond}(A)$ and no additional restrictions are imposed on A.

In the case A is symmetric we can do a little bit better than using LDL^T. Instead, let D_1, D_2 be diagonal such that $D_1D_2 = D, |D_1| = |D_2| = |D|^{1/2}$. Then $LDL^T = (LD_1) \cdot (LD_2)^T$ and the usual LDL^T decomposition can be modified in an obvious way to compute $L_1 := L \cdot D_1$ and $L_2 := L \cdot D_2$ directly instead of L and D. Furthermore $D_1 = Q \cdot D_2$ with Q being a diagonal matrix with $+1$ or -1 in the diagonal thus being orthogonal. Therefore LD_1 and LD_2 have the same singular values and lower bound for $\sigma_n(LD_1)$ suffices for our purposes.

Despite saving computing time the heuristic is that $\sigma_n(LD_1)^2$ provides a better lower estimate for $\sigma_n(LDL^T)$ than $\sigma_n(L)^2 \cdot \sigma_n(D)$. Practical examples support this heuristic to a certain point. The same heuristic applies to general nonsymmetric matrices.

III) *A is symmetric.* Compute an approximate $\widetilde{L}_1 \cdot \widetilde{L}_2^T$ decomposition as described above and an approximation σ of the smallest singular value of $\widetilde{L}_1$. Apply algorithm 2.1 with $\widetilde{\lambda} = 0.9 \cdot \widetilde{\sigma}$ to compute a lower bound σ on $\sigma_n(\widetilde{L}_1)$ and apply theorem 2.2 or 2.3 with $\widetilde{A} := \widetilde{L}_1 \cdot \widetilde{L}_2^T$ and $\sigma_n(\widetilde{A}) \geq \sigma^2$.

In the general case we may apply an LU-decomposition. However, L tends to be fairly well-conditioned whereas the condition of A moves into U. Thus we may run into difficulties

trying to estimate $\sigma_n(U^TU)$. On the other hand the LDM^T-decomposition can be altered in an obvious way to distribute $D = D_1 \cdot D_2, |D_1| = |D_2| = |D|^{1/2}$ both in L and M as we did in the $L_1L_2^T$-decomposition in the symmetric case. This yields an LM-decomposition, L and M no longer being unit lower triangular. The heuristic is that then L and M are more or less equally conditioned, the condition number not being much bigger than the square root of the condition number of A.

IV) *A is general nonsymmetric.* Compute an approximate $\widetilde{L} \cdot \widetilde{M}^T$-decomposition of A and approximations $\widetilde{\sigma}_1, \widetilde{\sigma}_2$ for the smallest singular value of $\widetilde{L}, \widetilde{M}$, respectively. Apply algorithm 2.1 with $\widetilde{\lambda}_1 = 0.9 \cdot \widetilde{\sigma}_1, \widetilde{\lambda}_2 = 0.9 \cdot \widetilde{\sigma}_2$ to compute a lower bound σ_1, σ_2 on $\sigma_n(\widetilde{L}), \sigma_n(\widetilde{M})$ and apply theorem 2.2 or 2.3 with $\widetilde{A} = \widetilde{L} \cdot \widetilde{M}^T$ and $\sigma_n(\widetilde{A}) \geq \sigma_1 \cdot \sigma_2$.

It should be pointed out that the heuristic for cases III) and IV) works for many examples but also has its drawbacks. In the moment we do not know a general strategy for choosing a decomposition $A \approx \widetilde{F}\widetilde{G}$ which maximizes $\sigma_n(\widetilde{F}) \cdot \sigma_n(\widetilde{G})$. In case of symmetric positive definite A the method of choice is of course the Cholesky decomposition $A = GG^T$ with $\sigma_n(A) = \sigma_n(G)^2$.

Let $L \in \mathbb{R}^{n \times n}$ be of lower triangular of bandwidth p. Then approximations of the smallest singular value of L are either computed by

- inverse power iteration for LL^T at the cost of $2np$ ops per iteration or
- using some condition estimator at the cost of $c \cdot np$ ops, c small.

As has been pointed out before this is small against np^2. Thus the total computing time for either of the algorithms for a linear system $Ax = b$ with A of lower, upper bandwidth p, q, respectively, $p \ll n, q \ll n$ is

I)	*A is an M-matrix*:	$n \cdot pq$	ops
II)	*A is symmetric positive definite*:	$n \cdot p^2$	ops
III)	*A is symmetric indefinite*:	$\frac{3}{2}n \cdot p^2$	ops
IV)	*A is general matrix*:	$n \cdot (pq + p^2 + q^2)$	ops.

Finally we want to mention how to use our methods in an interval setting, that is to solve $[A]x = [b], [A] \in \mathbb{IIR}^{n \times n}, [b] \in \mathbb{IIR}^n$. Theorem 2.3 extends as follows.

Theorem 2.4. Let $[A] \in \mathbb{IIR}^{n \times n}, [b] \in \mathbb{IIR}^n$ be given as well as nonsingular $\widetilde{A} \in \mathbb{R}^{n \times n}$ and $\widetilde{x} \in \mathbb{R}^n$. Define $\Delta A := |[A] - \widetilde{A}|$ and suppose $\sigma_n(\widetilde{A}) > (\|\Delta A\|_1 \cdot \|\Delta A\|_\infty)^{1/2}$.

Then every $A \in [A]$ is nonsingular and for every $\widehat{x} := A^{-1}b, A \in [A], b \in [b]$ holds

$$\|\widehat{x} - \widetilde{x}\|_\infty \leq \|\widehat{x} - \widetilde{x}\|_2 \leq \frac{\| \, |[b] - [A] \cdot \widetilde{x}| \, \|_2}{\sigma_n(\widetilde{A}) - (\|\Delta A\|_1 \cdot \|\Delta A\|_\infty)^{1/2}} \, . \tag{22}$$

The **proof** follows by applying theorem 2.3 to each $A \in [A], b \in [b]$. ■

We shortly describe an algorithm for solving a general interval linear system. We use the property $A \in [A] \Rightarrow \|A\|_p \leq \|mid([A])\|_p + \|rad([A])\|_p$ for $p \in \{1, \infty\}$.

Let $[A] \in \mathbb{IIR}^{n\times n}, [b] \in \mathbb{IIR}^n$ be given.

1) For $mA := \text{mid}([A])$ compute an approximate decomposition $\widetilde{L} \cdot \widetilde{M}^T := \widetilde{A} \approx mA$ (see IV)) in floating-point arithmetic together with estimates ζ_1, ζ_∞ on $\|\widetilde{A} - mA\|_1, \|\widetilde{A} - mA\|_\infty$, resp.

2) Solve $\widetilde{L} \cdot \widetilde{M}^T \cdot \widetilde{x} = mb, mb := \text{mid}([b])$ by floating-point backward and forward substitution to obtain $\widetilde{x}$

3) Compute approximations for the smallest singular value $\widetilde{\sigma}_1, \widetilde{\sigma}_2$ of $\widetilde{L}, \widetilde{M}$ by floating-point inverse power method applied to $\widetilde{L}\widetilde{L}^T, \widetilde{M}\widetilde{M}^T$, resp. or by some condition estimator

4) Apply algorithm 2.1 to compute lower bounds σ_1, σ_2 on $\sigma_n(\widetilde{L}), \sigma_n(\widetilde{M})$ using $\lambda := 0.9 \cdot \widetilde{\lambda}_i$. If algorithm 2.1 does not finish successfully try smaller values for λ_i.

5) Calculate $z = \sup(|[b] - [A] \cdot \widetilde{x}|)$ and upper bounds $\eta_1 \geq \|\text{rad}([A])\|_1, \eta_\infty \geq \|\text{rad}([A])\|_\infty$ using interval arithmetic (for η_1, η_∞ upwardly directed roundig suffices).

6) If $\mu := \sigma_1 \cdot \sigma_2 - ((\zeta_1 + \eta_1)(\zeta_\infty + \eta_\infty))^{1/2} > 0$ then every $A \in [A]$ is nonsingular and
$$\|\widehat{x} - \widetilde{x}\|_\infty \leq \|\widehat{x} - \widetilde{x}\|_2 \leq \mu^{-1} \cdot \|z\|_2$$
for every $\widehat{x} = A^{-1}b$ with $A \in [A], b \in [b]$.

Algorithm 2.2 Inclusion of the solution of a general interval linear system

If very high accuracy of the inclusion is desived $\widetilde{x}$ may be stored in $\widetilde{x}_1$ and $\widetilde{x}_2$ with $\widetilde{x} = \widetilde{x}_1 + \widetilde{x}_2$ (staggered correction, see [29], [34]). In this case $b - A\widetilde{x}_1 - A\widetilde{x}_2$ should be calculated in double the working precision. Using this method frequently very high or least significant bit accuracy is achieved. A simpler way is to perform a residual iteration

$$x^{k+1} := x^k + \widetilde{M}^{-T}\widetilde{L}^{-1}(b - Ax^k) \tag{23}$$

as usual. Only in the final step the addition is not executed but $\widetilde{x}_1 := x^k$ and $\widetilde{x}_2 := \widetilde{M}^{-T}\widetilde{L}^{-1}(b - Ax^k)$ are stored in separate vectors. This saves computing time and produces similar results to storing $\widetilde{x}$ in two parts $\widetilde{x}_1$, $\widetilde{x}_2$ from the beginning.

3 Computational results

In the following we give numerical results for three different types of our algorithm:

(1) The *symmetric positive definite* case using a Cholesky-decomposition and proceeding as described in (II).

(2) The *symmetric case* using a modified LDL^T-decomposition *without* pivoting as described in (III).

(3) The *general case* using an LU-decomposition *with* pivoting from LAPACK.

In the following tables we display

n	dimension of the matrix
$cond(A)$	approximation of the $\|\cdot\|_\infty$-condition number of A
$iter$	number of inverse power iterations to obtain an approximation for $\sigma_n(A)$
$\sigma_{min}(A)$	lower bound for the smallest singular value of A
$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$	upper bound for the relative error of the approximate solution $\widetilde{x}$.

The condition number is estimated using the vector obtained by the inverse power iteration. Working accuracy is IEEE 754 double precision (approximately 17 decimals). As described in (2.10) we split $\widetilde{x}$ into $\widetilde{x}_1$, $\widetilde{x}_2$ and compute $b - A\widetilde{x}_1 - A\widetilde{x}_2$ in quadruple precision.

In all of the following examples the

- right hand side b is computed such that the solution $\widehat{x}$ of $Ax = b$ is $\widehat{x}_i := (-1)^{i+1} \cdot 1/i$.

This introduces different magnitudes in the solution together with some roughness.

The first example, only displayed for reference purposes, is a discretisation of a Poisson equation

$$n := \begin{pmatrix} 4 & -1 & \\ -1 & 4 & \ddots \\ & \ddots & \ddots \end{pmatrix}; \qquad A := \begin{pmatrix} M & -I & \\ -I & M & \ddots \\ & \ddots & \ddots \end{pmatrix} \tag{24}$$

with I being the identity matrix. We used three different bandwidthes p.

n	p	cond	iter	$\sigma_{min}(A)$	$\|\widehat{x}-\widetilde{x}\|_\infty/\|\widetilde{x}\|_\infty$
200	5	2.90E+01	4	5.24E-01	1.15E-22
2000	5	2.98E+01	3	5.18E-01	1.18E-22
20000	5	2.98E+01	3	5.18E-01	1.18E-22
200	10	7.73E+01	4	3.21E-01	4.49E-22
2000	10	9.78E+01	5	2.86E-01	5.87E-22
20000	10	9.86E+01	3	2.85E-01	5.91E-22
200	20	7.73E+01	4	3.21E-01	8.81E-22
2000	20	3.41E+02	4	1.53E-01	3.94E-21
20000	20	3.57E+02	3	1.50E-01	4.12E-21

Table 3.1 Matrices (3.1) for different bandwidthes.

The second example is (4.16) from Gregory/Karney [16] with bandwidth 2.

$$A := \begin{pmatrix} 5 & -4 & 1 & & & & \\ -4 & 6 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 6 & -4 \\ & & & & 1 & -4 & 5 \end{pmatrix}$$

Example (4.16) from [16]

Here the increasing condition number limits the dimension to the same amount as for a pure floating-point algorithm.

n	cond	iter	σ_{min}	$\|\widehat{x}-\widetilde{x}\|_\infty/\|\widetilde{x}\|_\infty$
100	1.71E+07	2	9.67E-04	2.82E-18
200	2.68E+08	2	2.44E-04	6.26E-17
500	1.03E+10	2	3.93E-05	1.87E-15
1000	1.65E+11	2	9.85E-06	3.95E-14
2000	2.63E+12	2	2.46E-06	7.01E-13
5000	1.03E+14	2	3.95E-07	2.53E-11
10000	1.64E+15	2	9.87E-08	5.38E-10
20000	2.63E+16	2	2.47E-08	1.83E-08
50000	1.03E+18	2	4.05E-09	failed

Table 3.2. Matrix (4.16) from [16]

Another example with high condition numbers are Hilbert matrices, $A_{ij} := 1/(i+j-1)$. The following table shows the results.

n	cond	iter	$\sigma_{min}(A)$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
5	6.94E+05	3	1.81E-03	1.10E-22
6	2.26E+07	3	3.29E-04	4.44E-21
7	7.42E+08	3	5.91E-05	1.76E-19
8	2.45E+10	3	1.05E-05	1.88E-14
9	8.08E+11	3	1.87E-06	72.45E-16
10	2.68E+13	3	3.31E-07	1.86E-11
11	8.84E+14	3	5.83E-08	8.41E-10
12	2.60E+16	3	1.03E-08	2.38E-11
13	2.72E+17	2	1.21E-09	failed

Table 3.3. Hilbert matrices

Using Neumaier's example (1.13) we can show the behaviour for larger dimensions. We used $A = 10^{-1} \cdot LL^T$ producing a matrix of bandwidth 2. The factor 10^{-1} is introduced to make the factors of A not exactly representable. Otherwise a decomposition algorithm would rapidly produce the *exact* Cholesky factors. Here we observe decreasing precision of $\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$ with increasing condition number.

n	cond	iter	$\sigma_{min}(A)$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
100	1.26E+04	3	2.68E-02	3.49E-21
200	4.95E+04	3	1.35E-02	2.71E-20
500	3.06E+05	3	5.43E-03	8.50E-20
1000	1.22E+06	3	2.72E-03	3.40E-19
2000	4.87E+06	3	1.36E-03	1.36E-18
5000	3.04E+07	3	5.44E-04	8.47E-18
10000	1.22E+08	3	2.72E-04	3.39E-17
20000	4.87E+08	3	1.36E-04	1.35E-16
50000	3.04E+09	3	5.44E-05	8.47E-16
100000	1.22E+10	3	2.72E-05	3.39E-15
500000	3.04E+11	3	5.44E-06	8.47E-14
1000000	1.22E+12	3	2.72E-06	3.39E-13

Table 3.4. Neumaier's example with $A = 10^{-1}LL^T$, L from (1.13)

Next we go to the symmetric indefinite case. The first example is taken from [16], (4.20) with $a = 1$, bandwidth 2.

$$A := \begin{pmatrix} -1 & 2 & 1 & & & \\ 2 & 0 & 2 & 1 & & \\ 1 & 2 & 0 & 2 & 1 & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & 1 & 2 & 0 & 2 \\ & & & 1 & 2 & -1 \end{pmatrix}$$

Example (4.20) from [16]

The eigenvalues are $\lambda_k = \left(1 - 2\cos\dfrac{k\pi}{n+1}\right)^2 - 3, \quad 1 \le k \le n$. We also display the computed upper bound on $\|A - \widetilde{L}_1 \cdot \widetilde{L}_2^T\|_2$. It is $\sigma_n(\widetilde{L}_1) = \sigma_n(\widetilde{L}_2)$.

n	cond	$\|A - \widetilde{L}_1 * \widetilde{L}_2^T\|_2$	iter	$\sigma_{min}(\widetilde{L}_1)$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
100	5.37E+01	8.82E-14	3	1.01E-02	8.65E-21
200	9.02E+01	8.82E-14	3	7.66E-03	1.80E-20
500	3.29E+02	8.82E-14	3	2.62E-03	1.28E-19
1000	6.18E+02	2.59E-13	3	1.28E-03	1.02E-18
2000	1.29E+03	6.54E-12	3	3.27E-04	3.43E-18
5000	3.33E+03	6.54E-12	3	1.43E-04	1.82E-17
10000	5.52E+03	6.54E-12	3	9.57E-05	7.57E-17
20000	1.13E+04	6.54E-12	3	5.68E-05	2.62E-16
50000	3.24E+04	7.75E-12	3	2.04E-05	4.13E-15
100000	6.26E+04	7.75E-12	3	1.04E-05	7.62E-14

Table 3.5. Example (4.20) from [16], $a = 1$

The results show that, as before, few inverse power iterations are necessery to obtain an approximation for the smallest singular value of $\widetilde{L}_1$. The iteration is stopped when two successive iterates differ relatively less than 10^{-3}. Note that $\sigma_n(\widetilde{L}_1)$ is fairly small in magnitude. This is due to the fact that the decomposition is performed without pivoting. Nevertheless sharp inclusions of the solution are achieved.

The next two tables show the behaviour for larger bandwidths. We use the abbreviation $M(a, b, c \cdots)$ denoting a symmetric matrix with value a in the diagonal, b in the first subdiagonal, c in the second and so forth.

n	cond	$\|A - \widetilde{L}_1 * \widetilde{L}_2^T\|_2$	iter	$\sigma_{min}(\widetilde{L}_1)$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
100	2.07E+01	6.61E-14	6	2.43E-02	5.06E-20
200	2.82E+01	2.64E-13	6	1.55E-02	1.24E-19
500	1.16E+02	2.64E-13	4	6.78E-03	6.44E-19
1000	1.96E+02	2.64E-13	5	3.66E-03	2.18E-18
2000	4.18E+02	4.08E-12	5	1.12E-03	2.48E-17
5000	1.47E+03	4.46E-12	3	4.49E-04	1.42E-16
10000	1.98E+03	4.46E-12	6	2.93E-04	3.30E-16
20000	4.24E+03	2.15E-11	5	9.86E-05	3.65E-15
50000	1.14E+04	2.15E-11	5	5.06E-05	1.14E-14

Table 3.6. $M(1, -2, 3, 4, -5)$, bandwidth 4

n	cond	$\|A - \widetilde{L}_1 * \widetilde{L}_2^T\|_2$	iter	$\sigma_{min}(\widetilde{L}_1)$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
100	1.08E+01	1.09E-12	4	2.05E-02	1.36E-19
200	4.38E+01	1.83E-12	3	7.91E-03	8.74E-19
500	7.46E+01	1.83E-12	4	5.26E-03	2.06E-18
1000	2.23E+02	6.48E-12	5	2.11E-03	1.24E-17
2000	2.54E+02	6.48E-12	6	1.89E-03	1.56E-17
5000	9.65E+02	6.48E-12	6	6.94E-04	1.28E-16
10000	2.23E+03	1.13E-11	5	2.41E-04	1.01E-15
20000	3.12E+03	1.13E-11	6	1.93E-04	1.54E-15

Table 3.7. $M(1, -2, 3, 4, -5, 5, 4, 3, 2, 1)$, bandwidth 9

Again the comparitatively small values of $\sigma_n(\widetilde{L}_1)$ are due to the lack of pivoting.

Finally we give same examples for the general case. First we show random matrices with upper and lower bandwidth 8 and uniformly distributed entries in the interval [-1,1].

n	cond	$\|A - \widetilde{L}\widetilde{U}\|_2$	$\sigma_{min}(\widetilde{L})$	$\sigma_{min}(\widetilde{U})$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
100	1.2E+03	1.2E-16	7.1E-02	2.5E-02	5.5E-26
200	3.5E+03	1.4E-16	1.0E-01	5.6E-03	2.6E-25
500	9.3E+04	1.4E-16	2.6E-02	1.9E-03	4.1E-23
1000	6.3E+04	2.1E-16	1.3E-02	4.0E-04	2.1E-21
2000	4.4E+05	3.5E-16	1.4E-02	1.0E-04	7.7E-24
5000	4.9E+05	3.2E-16	7.7E-03	1.7E-04	1.3E-23
10000	8.7E+05	2.9E-16	1.8E-02	5.9E-05	6.0E-23
20000	2.2E+05	3.4E-16	2.4E-02	1.0E-04	5.5E-23
50000	1.6E+06	2.8E-16	1.0E-02	2.3E-05	7.6E-22

Table 3.8. Random matrices, upper and lower bandwidth 8

The LU-decomposition is performed using routines DGBTRF and DGBTRS from LAPACK with pivoting. The smallest singular values of $\widetilde{L}$ and $\widetilde{U}$ are not too near. This improves when distributing the diagonal of $\widetilde{U}$ among $\widetilde{L}$ and $\widetilde{U}$. Finally we show an example being unsymmetric in upper and lower bandwidth.

n	cond	$\|A - \widetilde{L}\widetilde{U}\|_2$	$\sigma_{min}(\widetilde{L})$	$\sigma_{min}(\widetilde{U})$	$\|\widehat{x} - \widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty$
100	5.1E+02	1.1E-16	9.1E-02	2.0E-02	1.3E-26
200	1.4E+03	1.8E-16	4.5E-02	1.2E-03	8.1E-24
500	7.8E+06	2.3E-16	4.3E-02	2.7E-06	4.6E-23
1000	3.6E+07	2.2E-16	3.0E-02	3.8E-07	3.8E-22
2000	2.0E+05	5.8E-16	4.0E-02	5.9E-05	3.4E-24
5000	2.3E+06	4.3E-16	1.0E-02	7.7E-06	2.5E-22

Table 3.9. Random matrices, upper/lower bandwidth 8/6

Random matrices with symmetric upper and lower bandwidth are fairly well-conditioned. This changes when the bandwidth becomes unsymmetric. Then for moderate dimension we run into fairly ill-conditioned matrices. Again the numbers become better when distributing of $\widetilde{U}$ among $\widetilde{L}$ and $\widetilde{U}$.

4 Conlusion

The presented algorithm in its different versions for symmetric positive definite, symmetric indefinite and general matrices works for high dimensions. Possible improvements and open questions are the following.

Using a condition estimator instead of inverse power iteration would eventually be cheaper but has not been tested yet. For symmetric indefinite and for general matrices it is not clear how to choose a proper decomposition $A \approx \widetilde{F} \cdot \widetilde{G}$ in order to minimize $\sigma_n(\widetilde{F}) \cdot \sigma_n(\widetilde{G})$. Using an LDL^T or LDM^T decomposition with D equally distributed among the other factors works fine in many cases but also has its drawbacks. The estimations given by the algorithm are ∞-norm estimates on the relative error of an approximate solution $\widetilde{x}$. Componentwise error estimates are not yet available. However, the results obtained up to now look promising.

References

[1] Alefeld, G.; Herzberger, J.: Introduction to Interval Computations, Academic Press (1983)

[2] Anderson, E.: Robust Triangular Solves for Use in Condition Estimation, Cray Research (1991)

[3] High-Accuracy Arithmetic Subroutine Library, Program Description and User's Guide, Release 3, IBM Publications, Document Number SC 33-6164-3 (1986).

[4] Arioli, M.; Demmel, J.W.; Duff, I.S.: Solving Sparse Linear Systems with Backward Error, SIAM J. Matrix Anal. Appl. 10, No. 2, 165–190 (1989)

[5] Bauch, H.; Jahn, K.-U.; Oelschlägel, D.; Süsse, H.; Wiebigke, V.: Intervallmathematik, Theorie und Anwendungen; Mathematisch-naturwissenschaftliche Bibliothek, Bd. 72, B.G. Teubner, Leipzig (1987)

[6] Bischof, Ch. H.; Tang, P.T.P.: Robust Incremental Condition Estimators, Argonne National Lab. (1992)

[7] Böhm, H.: Berechnung von Polynomnullstellen und Auswertung arithmetischer Ausdrucke mit garantierter, maximaler Genauigkeit, Ph.D. dissertation, University of Karlsruhe (1983)

[8] Böhm, H.; Rump, S.M.: Least Significant bit Evaluation for Arithmetic Expressions, Computing 30, 189–199 (1983)

[9] Cline, A.K.; Moler, G.B.; Stewart, G.W.; Wilkinson, J.H.: An Estimate for the Condition Number of a Matrix, SIAM J. Num. Anal. 16, 368–375 (1979)

[10] Cline, A.K.; Conn, A.R.; Van Loan, C.: Generalizing the LINPACK Condition Estimator, in: Numerical Analysis, ed., J. P. Hennart, Lecture Notes in Mathematics, No. 909, Springer-Verlag, New York (1982)

[11] Cordes, D.; Kaucher, E.: Self-Validating Computation for Sparse Matrix Problems, in: E. Kaucher, U. Kulisch, Ch. Ullrich (Eds.): Computerarithmetic: Scientific Computation and Programming Languages. B.G. Teubner Verlag, Stuttgart (1987)

[12] Cordes, D.: Spärlich besetzte Matrizen, in: U. Kulisch (Ed.): Wissenschaftliches Rechnen mit Ergebnisverifikation — Eine Einführung, ausgearbeitet von S. Geörg, R. Hammer und D. Ratz. Akademie Verlag, Berlin, und Vieweg Verlagsgesellschaft, Wiesbaden (1989)

[13] Duff, I.S.; Erisman, A.M.; Reid, J.K.: Direct Methods for Sparse Matrices, Clarendon Press, Oxford (1986).

[14] Golub, G. and v. Loan, C.: Matrix Computations, John Hopkins University Press, second edition (1989)

[15] Grimes, R.G.; Lewis, J.G.: Condition Number Estimation for Sparse Matrices, SIAM J. Sci. and Stat. Comp. 2, 384–388 (1991)

[16] Gregory, R.T.; Karney, D.L.: A Collection of Matrices for Testing Computional Algorithms, John Wiley (1969).

[17] Hager, W.: Condition Estimates, SIAM J. Sci. and Stat. Comp. 5, 311–316 (1984)

[18] Higham, N.J.: Fortran Codes for Estimating the One-norm of a Real or Complex Matrix, with Applications to Condition Estimation, ACM Trans. Math. Soft. 14, 381–396 (1987)

[19] IEEE Standard for Binary Floating-Point Arithmetic, ANSI/IEEE Standard 754 (1985).

[20] Jansson, Ch.: private communication.

[21] Krawczyk, R.: Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken, Computing 4, 187–201 (1969)

[22] Krawczyk, R.: Fehlerabschätzung bei linearer Optimierung, in "Interval Mathematics", edited by K. Nickel, Lecture Notes in Computer Seience 29, Springer (1975).

[23] Krämer, W.: Verified Solution of Eigenvalue Problems with Sparse Matrices. Proceedings of 13th World Congress on Computation and Applied Mathematics, IMACS '91, Dublin, 32–33 (1991)

[24] Kulisch, U.: Grundlagen des numerischen Rechnens, Reihe Informatik 19, BI-Verlag (1976).

[25] Kulisch, U.; Miranker, W.L.: Computer Arithmetic in Theory and Practice, Academic Press (1981)

[26] Moore, R.E.: A Test for Existence of Solutions for Non-Linear Systems, SIAM J. Numer. Anal. 4 (1977)

[27] Moore, R.E.: Methods and Applications of Interval Analysis, SIAM, Philadelphia (1979).

[28] Neumaier, A.: Interval Methods for Systems of Equations, Cambridge University Press (1990)

[29] Rump, S.M.: Kleine Fehlerschranken bei Matrixproblemen, Dissertation, Universitaet Karlsruhe (1980)

[30] Rump, S.M.: Solving Algebraic Problems with High Accuracy, Habilitationsschrift, in: A New Approach to Scientific Computation, Hrsg. U.W. Kulisch und W.L. Miranker, Academic Press, 51–120 (1983)

[31] Rump, S.M.: On the Solution of Interval Linear Systems, COMPUTING 47, p. 337-353 (1992).

[32] Rump, S.M.: Inclusion of the Solution for Large Linear Systems with M-Matrix, Report of the Forschungsschwerpunkt Informations- und Kommunikationstechnik 91.3, Technical University Hamburg-Harburg (1991)

[33] Schwandt, H.: An Interval Arithmetic Approach for the Construction of an almost Globally Convergent Method for the Solution of the Nonlinear Poisson Equation on the Unit Square, SIAM J. Sci. Stat. Comp., 5, No. 2 (1984)

[34] Stetter, H.J.: Sequential Defect Correction in High-Accuracy Floating-Point Arithmetics, in: Numerical Analysis (Proceedings, Dundee 1983), Lecture Notes in Math. 1066, 186–202 (1984)