Fast Enclosure for Solutions of Generalized Least Squares Problems*

Shinya Miyajima
Faculty of Engineering, Gifu University, 1-1 Yanagido,
Gifu-shi, Gifu 501-1193, Japan
miyajima@gifu-u.ac.jp

Abstract

Fast algorithms for enclosing solutions of generalized least squares problems are proposed. To develop these algorithms, theories for obtaining error bounds on numerical solutions are established. The error bounds obtained by these algorithms are "verified" in the sense that all the possible rounding errors have been taken into account. A technique for obtaining smaller error bounds is introduced. Numerical results show the properties of the proposed algorithms.

Keywords: Generalized least squares problems, Numerical enclosure, Verified error bound
AMS subject classifications: 15A09, 65F20, 65G20

1 Introduction

The generalized least squares problem considered in this paper is to find the n-vector x that minimizes

$$(Ax - b)^T B^{-1}(Ax - b), \qquad A \in \mathbb{R}^{m\times n}, \quad b \in \mathbb{R}^m, \quad B \in \mathbb{R}^{m\times m}, \tag{1}$$

where $m \geq n$, A, b and B are given, A has full column rank, and B is symmetric positive definite. This problem arises in finding the least squares estimate of the vector x when we are given the linear model $b = Ax + w$ with w an unknown noise vector of zero mean and covariance B. In several practical problems in econometrics [5] and engineering [2], A and B will have special block structure. It is known that the vector minimizing (1) is $(A^T B^{-1} A)^{-1} A^T B^{-1} b$.

Since B is symmetric positive definite, there exists a matrix L satisfying $B = LL^T$, which can be obtained by Cholesky decomposition or eigen-decomposition. In several applications, L is more basic and important than B, so that it is assumed in several papers (e.g. [10, 11]) that L is given. Then, the solution can be written as $(L^{-1}A)^+ L^{-1} b$, where $(L^{-1}A)^+$ denotes the Moore-Penrose inverse of $L^{-1}A$. In this paper, we treat both of the cases when B is given and L is given.

*Submitted: May 29, 2014; Revised: January 7, 2015; Accepted: February 16, 2015.

Stable algorithms for solving (1) have been proposed in [10, 11]. These algorithms are based on the idea that (1) is equivalent to the problem of finding x which minimizes v^Tv on the equality constraint $b = Ax + Lv$. In these algorithms, the equivalent problem is solved via orthogonal transformation.

In this paper, we consider numerically enclosing $(A^TB^{-1}A)^{-1}A^TB^{-1}b$, specifically, computing error bounds of $\tilde{x}$ using floating point operations, where $\tilde{x}$ denotes a numerical result for $(A^TB^{-1}A)^{-1}A^TB^{-1}b$. To the author's knowledge, an algorithm for enclosing solutions designed specifically for (1) has not appeared in the literature. Direct approaches for enclosing $(A^TB^{-1}A)^{-1}A^TB^{-1}b$ and $(L^{-1}A)^+L^{-1}b$ use the INTLAB [13] routine `verifylss` such that

```
Res = verifylss(A'*verifylss(B,A), A'*verifylss(B,b));
```

and `Res = verifylss(verifylss(L,A),verifylss(L,b));`, respectively. When we use these approaches, on the other hand, the computing times are not always small (see Sections 6 and 7 for details). The solution $(A^TB^{-1}A)^{-1}A^TB^{-1}b$ can be obtained by solving the augmented linear system

$$\begin{pmatrix} A & -B \\ O_n & A^T \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix},$$

where O_n denotes the $n \times n$ zero matrix, since this system implies $x = (A^TB^{-1}A)^{-1}A^T B^{-1}b$ and $y = B^{-1}(A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b)$. Therefore, inclusion of $(A^TB^{-1}A)^{-1}\ A^TB^{-1}b$ can be obtained by enclosing the solution of this system. When A or B are not well-conditioned, on the other hand, this approach does not always give a tight enclosure (see Sections 6 and 7). When B is the identity matrix, (1) reduces to the "standard" least squares problem. For standard least squares problems, effective and efficient algorithms for enclosing the solutions have been proposed in the literature [7, 12, 13, 14, 15]. On the other hand, these references do not mention how to extend these algorithms to generalized least squares problems.

The purpose of this paper is to propose fast algorithms for enclosing the solution of (1) in both of the cases when B is given and L is given. These algorithms allow the presence of underflow in floating point arithmetic. To develop these algorithms, we establish theories for computing error bounds on $\tilde{x}$. The error bounds obtained by the proposed algorithms are "verified" in the sense that all the possible rounding errors have been taken into account. In the case when B is given, the proposed algorithms do not assume but prove A and B to have full rank and to be positive definite, respectively. In the case when L is given, the algorithms do not assume but prove A and L to have full rank and to be nonsingular, respectively. We refer to and introduce techniques for accelerating the enclosure and obtaining smaller error bounds, respectively.

This paper is organized as follows: In Section 2, notations and theories utilized in this paper are introduced. In Section 3, the theories for enclosing the solution of (1) are established. In Sections 4 and 5, the techniques for accelerating the enclosure and obtaining smaller error bounds are referenced and introduced, respectively. In Sections 6 and 7, numerical results are reported. Section 8 finally summarizes the results in this paper and highlights possible extensions and future work.

2 Preliminaries

In this section, we define some notation and introduce theories utilized hereafter. For $M = (M_{ij}) \in \mathbb{R}^{m\times n}$, M^+ denotes the Moore-Penrose inverse of M, $M^T := (M_{ji})$ and

$|M| := (|M_{ij}|)$. For a nonsingular real matrix S, $S^{-T} := (S^T)^{-1}$. Let I_m and $e^{(i)}$ be the $m \times m$ identity matrix and i-th column of I_m, respectively, and $s^{(n)} := (1, \ldots, 1)^T \in \mathbb{R}^n$. For $v, w \in \mathbb{R}^n$, v_i denotes the i-th component of v and $v \le w$ means that $v_i \le w_i$ follows for all $i = 1, \ldots, n$. Let $\mathbf{u}$ and $\underline{\mathbf{u}}$ be unit roundoff and underflow unit (especially $\mathbf{u} = 2^{-53}$ and $\underline{\mathbf{u}} = 2^{-1074}$ in IEEE 754 double precision), respectively, and $\gamma_m := m\mathbf{u}/(1 - m\mathbf{u})$. The notation $\mathrm{fl}(\cdot)$ denotes a result of floating point operations, where all inside parenthesis are executed by ordinary floating point arithmetic in rounding-to-nearest mode. For $X_c, X_r \in \mathbb{R}^{m \times n}$ with $\min_{i,j}(X_r)_{ij} \ge 0$, $\langle X_c, X_r \rangle$ denotes the interval matrix whose center and radius are X_c and X_r, respectively. For $C \in \mathbb{R}^{m \times n}$, define the condition number $\kappa(C) := \|C\|_2 \|C^+\|_2$. When $m = n$, let $\varrho(C)$ be the spectral radius of C.

We cite Lemmas 1, 3 and 4, and present Lemma 2, which are used in Section 3.

Lemma 1 (e.g. Golub and Van Loan [3]) *Let $S \in \mathbb{R}^{n \times n}$ and $1 \le p \le \infty$. If $\|S\|_p < 1$, $I_n - S$ is nonsingular.*

Lemma 2 is a modification of [16, Theorem 3].

Lemma 2 *Let $F \in \mathbb{R}^{m \times n}$, $S \in \mathbb{R}^{n \times n}$ and $f \in \mathbb{R}^n$. If $\|S\|_\infty < 1$, it holds that*

$$|FS(I_n - S)^{-1} f| \le \frac{\|f\|_\infty}{1 - \|S\|_\infty} |F||S| s^{(n)}.$$

Proof. The inequality $\|S\|_\infty < 1$ and the Neumann series give

$$\begin{aligned}
|FS(I_n - S)^{-1} f| &= |FS(I_n + S + S^2 + \cdots) f| = |F(S + S^2 + S^3 + \cdots) f| \\
&\le |F|(|S| + |S|^2 + |S|^3 + \cdots)|f| \\
&= |F|(|S||f| + |S||S||f| + |S||S|^2|f| + \cdots) \\
&\le |F|(\|f\|_\infty |S| s^{(n)} + \||S||f|\|_\infty |S| s^{(n)} + \||S|^2|f|\|_\infty |S| s^{(n)} + \cdots) \\
&\le (\|f\|_\infty + \|S\|_\infty \|f\|_\infty + \|S\|_\infty^2 \|f\|_\infty + \cdots)|F||S| s^{(n)} \\
&= \|f\|_\infty (1 + \|S\|_\infty + \|S\|_\infty^2 + \cdots)|F||S| s^{(n)} \\
&= \frac{\|f\|_\infty}{1 - \|S\|_\infty} |F||S| s^{(n)}. \qquad \square
\end{aligned}$$

Lemma 3 (e.g. Higham [4]) *If the floating point Cholesky decomposition applied to a symmetric matrix $B \in \mathbb{R}^{m \times m}$ runs to completion, the computed Cholesky factor $\tilde{L}$ satisfies*

$$\begin{aligned}
\tilde{L}\tilde{L}^T &= B + \Delta B, \\
|\Delta B| &\le \gamma_{m+1} |\tilde{L}||\tilde{L}|^T + \frac{\underline{\mathbf{u}}}{1 - m\mathbf{u}} (m s^{(m)} + v_{\tilde{L}}) s^{(m)^T},
\end{aligned} \tag{2}$$

also in the presence of underflow, where $v_{\tilde{L}} := (|\tilde{L}_{11}|, \ldots, |\tilde{L}_{mm}|)^T$.

Remark 1 *The second term in the right hand side of (2) is devised by the author. By adding this term, Lemma 3 holds also in the presence of underflow.*

Lemma 4 (Oishi and Rump [9]) *Let a nonsingular triangular $m \times m$ matrix $\tilde{L}$ be given and $v_{\tilde{L}}$ be as in Lemma 3. Suppose the columns $Me^{(i)}$ of an approximate inverse M are computed by substitution, in any order, of m linear systems $L(Me^{(i)}) = e^{(i)}$. Then, including possible underflow,*

$$|M\tilde{L} - I_m| \le \gamma_m |M||\tilde{L}| + \frac{\underline{\mathbf{u}}}{1 - m\mathbf{u}} (m s^{(m)} + v_{\tilde{L}}) s^{(m)^T}.$$

3 Enclosure Theories

In this section, we establish theories for enclosing the solutions of (1). Let L be a matrix satisfying $B = LL^T$, $\tilde{L}$ and M be approximations of L and $\tilde{L}^{-1}$, respectively, $MA \approx QR$ be an economy size floating point QR factorization of MA, S be an approximate inverse of R, and $X := MAS$. Then X and MBM^T are expected to be not too far from orthogonality and identity, respectively. If $\tilde{x}$, $\tilde{w}$ and $\tilde{y}$ are approximations of $(A^TB^{-1}A)^{-1}A^TB^{-1}b$, $A(A^TB^{-1}A)^{-1}A^TB^{-1}b-b$ and $B^{-1}(A(A^TB^{-1}A)^{-1}A^TB^{-1}b-b)$, respectively, we can expect $A\tilde{x}-b-\tilde{w} \approx 0$, $\tilde{w}-B\tilde{y} \approx 0$ and $A^T\tilde{y} \approx 0$.

Consider first the case when B is given. We formulate and prove Theorems 1 and 2 for enclosing $(A^TB^{-1}A)^{-1}A^TB^{-1}b$. The outline of the derivation of Theorem 1 is to transform $\tilde{x}-(A^TB^{-1}A)^{-1}A^TB^{-1}b$ such that the residuals $A\tilde{x}-b-\tilde{w}$, $\tilde{w}-B\tilde{y}$ and $A^T\tilde{y}$ appear, and to apply Lemma 2.

Theorem 1 *Let $m \geq n$, $A \in \mathbb{R}^{m\times n}$, $M, B \in \mathbb{R}^{m\times m}$, $S \in \mathbb{R}^{n\times n}$, $\tilde{x} \in \mathbb{R}^n$ and $b, \tilde{w}, \tilde{y} \in \mathbb{R}^m$ be given, and B be symmetric. Define*

$$\begin{aligned} \rho_{\tilde{x}} &:= A\tilde{x}-b-\tilde{w}, \quad \rho_{\tilde{w}} := \tilde{w}-B\tilde{y}, \quad \rho_{\tilde{y}} = A^T\tilde{y}, \quad X := MAS, \\ F &:= I_m - MBM^T, \quad v_E := |I_n - X^TX|s^{(n)} + \frac{\|X\|_\infty}{1-\|F\|_\infty}|X|^T|F|s^{(m)}. \end{aligned}$$

If $\|F\|_\infty < 1$, M and B are nonsingular and positive definite, respectively. If $\|v_E\|_\infty < 1$, additionally, S is nonsingular, A has full column rank, and $|\tilde{x}-(A^TB^{-1}A)^{-1}A^TB^{-1}b| \leq d_B^{(1)}$ follows, where

$$\begin{aligned} d_B^{(1)} &:= |S(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}})| + \frac{\|X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}}\|_\infty}{1-\|v_E\|_\infty}|S|v_E \\ &+\frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty}{1-\|F\|_\infty}|S||X|^T|F|s^{(m)} \\ &\qquad +\frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty\||X|^T|F|s^{(m)}\|_\infty}{(1-\|v_E\|_\infty)(1-\|F\|_\infty)}|S|v_E. \end{aligned}$$

Proof. The inequality $\|F\|_\infty < 1$ and Lemma 1 implies $I_m - F$ is nonsingular, which implies the nonsingularity of M. Similarly to [8, Proof of Theorem 4], moreover, $\|F\|_\infty < 1$ shows that B is positive definite. Let $E := I_n - X^T(MBM^T)^{-1}X$. It holds from $\|F\|_\infty < 1$ and Lemma 2 that

$$\begin{aligned} |E|s^{(n)} &= |I_n - X^T(I_m-F)^{-1}X|s^{(n)} = |I_n - X^T(I_m + F(I_m-F)^{-1})X|s^{(n)} \\ &= |I_n - X^TX - X^TF(I_m-F)^{-1}X|s^{(n)} \\ &\leq |I_n - X^TX|s^{(n)} + |X^TF(I_m-F)^{-1}X|s^{(n)} \leq v_E. \end{aligned}$$

This and $\|v_E\|_\infty < 1$ yield $\|E\|_\infty < 1$, so $I_n - E$ is nonsingular. Hence X has full column rank, showing that A has also full column rank and S is nonsingular. We obtain

$$\begin{aligned}
&\tilde{x} - (A^TB^{-1}A)^{-1}A^TB^{-1}b \\
&\quad = (A^TB^{-1}A)^{-1}A^TB^{-1}(A\tilde{x}-b) \\
&\quad = (A^TB^{-1}A)^{-1}A^TB^{-1}(\rho_{\tilde{x}}+\tilde{w}) \\
&\quad = (A^TB^{-1}A)^{-1}A^T(B^{-1}(\rho_{\tilde{x}}+\rho_{\tilde{w}})+\tilde{y}) \\
&\quad = (A^TB^{-1}A)^{-1}(A^TB^{-1}(\rho_{\tilde{x}}+\rho_{\tilde{w}})+\rho_{\tilde{y}}) \\
&\quad = (S^{-T}X^TM^{-T}B^{-1}M^{-1}XS^{-1})^{-1}(A^TM^TM^{-T}B^{-1}M^{-1}M(\rho_{\tilde{x}}+\rho_{\tilde{w}})+\rho_{\tilde{y}}) \\
&\quad = S(X^T(MBM^T)^{-1}X)^{-1}S^T(A^TM^T(MBM^T)^{-1}M(\rho_{\tilde{x}}+\rho_{\tilde{w}})+\rho_{\tilde{y}}) \\
&\quad = S(I_n-E)^{-1}(X^T(I_m-F)^{-1}M(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}}) \\
&\quad = S(I_n+E(I_n-E)^{-1})(X^T(I_m+F(I_m-F)^{-1})M(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}}) \\
&\quad = S(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}}) + SE(I_n-E)^{-1}(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}}) \\
&\qquad +SX^TF(I_m-F)^{-1}M(\rho_{\tilde{x}}+\rho_{\tilde{w}}) + SE(I_n-E)^{-1}X^TF(I_m-F)^{-1}M(\rho_{\tilde{x}}+\rho_{\tilde{w}}).
\end{aligned}$$

It finally follows from this, $\|E\|_\infty \le \|v_E\|_\infty < 1$, $\|F\|_\infty < 1$ and Lemma 2 that

$$\begin{aligned}
&|\tilde{x} - (A^TB^{-1}A)^{-1}A^TB^{-1}b| \\
&\quad \le |S(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}})| + |SE(I_n-E)^{-1}(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}})| \\
&\qquad +|SX^TF(I_m-F)^{-1}M(\rho_{\tilde{x}}+\rho_{\tilde{w}})| \\
&\qquad\qquad +|SE(I_n-E)^{-1}||X^TF(I_m-F)^{-1}M(\rho_{\tilde{x}}+\rho_{\tilde{w}})| \\
&\quad \le |S(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}})| + \frac{\|X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}}\|_\infty}{1-\|E\|_\infty}|S||E|s^{(n)} \\
&\qquad +\frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty}{1-\|F\|_\infty}|S||X|^T|F|s^{(m)} \\
&\qquad\qquad +|SE(I_n-E)^{-1}|\left(\frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty}{1-\|F\|_\infty}|X|^T|F|s^{(m)}\right) \\
&\quad \le |S(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}})| + \frac{\|X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}})+S^T\rho_{\tilde{y}}\|_\infty}{1-\|v_E\|_\infty}|S|v_E \\
&\qquad +\frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty}{1-\|F\|_\infty}|S||X|^T|F|s^{(m)} \\
&\qquad\qquad +\frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty\||X|^T|F|s^{(m)}\|_\infty}{(1-\|E\|_\infty)(1-\|F\|_\infty)}|S||E|s^{(n)} \\
&\quad \le d_B^{(1)}. \qquad \square
\end{aligned}$$

By adding some assumptions, we can avoid executing the matrix multiplication MBM^T in Theorem 1. We establish Theorem 2 to achieve this. The main point of the proof of Theorem 2 is to transform $\tilde{x} - (A^TB^{-1}A)^{-1}A^TB^{-1}b$ such that the residuals $\tilde{L}\tilde{L}^T - B$ and $I_m - M\tilde{L}$ appear, and to estimate these residuals via Lemmas 3 and 4.

Theorem 2 *Let $\tilde{L}$ and $v_{\tilde{L}}$ be as in Lemma 3, M be as in Lemma 4, A, B, S, $\tilde{x}$, b,*

$\tilde{w}$, $\tilde{y}$, $\rho_{\tilde{x}}$, $\rho_{\tilde{w}}$, $\rho_{\tilde{y}}$ and X be as in Theorem 1, and

$$
\begin{aligned}
v_G &:= \gamma_m |M||\tilde{L}|s^{(m)} + \frac{m\underline{\mathbf{u}}}{1-m\mathbf{u}}(ms^{(m)} + v_{\tilde{L}}),\\
v_{G^T} &:= \gamma_m |\tilde{L}|^T|M|^T s^{(m)} + \frac{\underline{\mathbf{u}}(m^2 + v_{\tilde{L}}^T s^{(m)})}{1-m\mathbf{u}} s^{(m)},\\
v_M &:= \gamma_{m+1}|M||\tilde{L}||\tilde{L}|^T|M|^T s^{(m)} + \frac{\underline{\mathbf{u}}({s^{(m)}}^T|M|^T s^{(m)})}{1-m\mathbf{u}}|M|(ms^{(m)} + v_{\tilde{L}}),\\
v_{MG^T} &:= \gamma_{m+1}|M||\tilde{L}||\tilde{L}|^T|M|^T v_{G^T} + \frac{\underline{\mathbf{u}}({s^{(m)}}^T|M|^T v_{G^T})}{1-m\mathbf{u}}|M|(ms^{(m)} + v_{\tilde{L}}),\\
v_H &:= v_M + \frac{\|v_M\|_\infty}{1-\|v_G\|_\infty}v_G + \frac{1}{1-\|v_{G^T}\|_\infty}v_{MG^T}\\
&\qquad + \frac{\|v_{MG^T}\|_\infty}{(1-\|v_G\|_\infty)(1-\|v_{G^T}\|_\infty)}v_G,\\
v_{GH} &:= \gamma_m|M||\tilde{L}|v_H + \frac{\underline{\mathbf{u}}({s^{(m)}}^T v_H)}{1-m\mathbf{u}}(ms^{(m)} + v_{\tilde{L}}),\\
v_{HG^T} &:= v_{MG^T} + \frac{\|v_{MG^T}\|_\infty}{1-\|v_G\|_\infty}v_G + \frac{\|v_{G^T}\|_\infty}{1-\|v_{G^T}\|_\infty}v_{MG^T}\\
&\qquad + \frac{\|v_{MG^T}\|_\infty\|v_{G^T}\|_\infty}{(1-\|v_G\|_\infty)(1-\|v_{G^T}\|_\infty)}v_G,\\
v_{GG^T} &:= \gamma_m|M||\tilde{L}|v_{G^T} + \frac{\underline{\mathbf{u}}({s^{(m)}}^T v_{G^T})}{1-m\mathbf{u}}(ms^{(m)} + v_{\tilde{L}}),\\
v_{GHG^T} &:= \gamma_m|M||\tilde{L}|v_{HG^T} + \frac{\underline{\mathbf{u}}({s^{(m)}}^T v_{HG^T})}{1-m\mathbf{u}}(ms^{(m)} + v_{\tilde{L}}),\\
v_P &:= v_G + v_H + v_{G^T} + v_{GH} + v_{HG^T} + v_{GG^T} + v_{GHG^T},\\
v_Q &:= |I_n - X^TX|s^{(n)} + \frac{\|X\|_\infty}{1-\|v_P\|_\infty}|X|^Tv_P.
\end{aligned}
$$

Suppose $\|v_G\|_\infty < 1$, $\|v_{G^T}\|_\infty < 1$, $\|v_P\|_\infty < 1$ and $\|v_Q\|_\infty < 1$. Then S, $\tilde{L}$ and M are nonsingular, A has full column rank, B is positive definite, and $|\tilde{x} - (A^TB^{-1}A)^{-1}A^TB^{-1}b| \le d_B^{(2)}$ holds, where

$$
\begin{aligned}
d_B^{(2)} &:= |S(X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}})| + \frac{\|X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}}\|_\infty}{1-\|v_Q\|_\infty}|S|v_Q\\
&\quad + \frac{\|M(\rho_{\tilde{x}} + \rho_{\tilde{w}})\|_\infty}{1-\|v_P\|_\infty}|S||X|^Tv_P + \frac{\|M(\rho_{\tilde{x}} + \rho_{\tilde{w}})\|_\infty\||X|^Tv_P\|_\infty}{(1-\|v_Q\|_\infty)(1-\|v_P\|_\infty)}|S|v_Q.
\end{aligned}
$$

Proof. Let $G := I_m - M\tilde{L}$. From Lemma 4, we have $|G|s^{(m)} \le v_G$ and $|G|^Ts^{(m)} \le v_{G^T}$. The first inequality, $\|v_G\|_\infty < 1$ and Lemma 1 give that $I_m - G$ is nonsingular, showing the nonsingularities of $\tilde{L}$ and M. Let $H := I_m - \tilde{L}^{-1}B\tilde{L}^{-T}$ and $P := G +$

$H + G^T - GH - HG^T - GG^T + GHG^T$. We have

$$\begin{aligned} H &= \tilde{L}^{-1}(\tilde{L}\tilde{L}^T - B)\tilde{L}^{-T} = \tilde{L}^{-1}M^{-1}M(\tilde{L}\tilde{L}^T - B)M^T M^{-T}\tilde{L}^{-T} \\ &= (I_m - G)^{-1}M(\tilde{L}\tilde{L}^T - B)M^T(I_m - G^T)^{-1} \\ &= (I_m + G(I_m - G)^{-1})M(\tilde{L}\tilde{L}^T - B)M^T(I_m + G^T(I_m - G^T)^{-1}) \\ &= M(\tilde{L}\tilde{L}^T - B)M^T + G(I_m - G)^{-1}M(\tilde{L}\tilde{L}^T - B)M^T \\ &\quad +M(\tilde{L}\tilde{L}^T - B)M^T G^T(I_m - G^T)^{-1} \\ &\quad +G(I_m - G)^{-1}M(\tilde{L}\tilde{L}^T - B)M^T G^T(I_m - G^T)^{-1}. \end{aligned}$$

This, $|G|s^{(m)} \le v_G$, $|G|^T s^{(m)} \le v_{G^T}$, $\|v_G\|_\infty < 1$, $\|v_{G^T}\|_\infty < 1$, and Lemmas 2, 3 and 4 yield

$$\begin{aligned} |H|s^{(m)} &\le |M||\tilde{L}\tilde{L}^T - B||M|^T s^{(m)} \\ &\quad +|G(I_m - G)^{-1}||M||\tilde{L}\tilde{L}^T - B||M|^T s^{(m)} \\ &\quad +|M||\tilde{L}\tilde{L}^T - B||M|^T|G^T(I_m - G^T)^{-1}|s^{(m)} \\ &\quad +|G(I_m - G)^{-1}||M||\tilde{L}\tilde{L}^T - B||M|^T|G^T(I_m - G^T)^{-1}|s^{(m)} \\ &\le v_M + \frac{\|v_M\|_\infty}{1 - \|G\|_\infty}|G|s^{(m)} \\ &\quad + \frac{1}{1 - \|G^T\|_\infty}|M||\tilde{L}\tilde{L}^T - B||M|^T|G|^T s^{(m)} \\ &\quad +|G(I_m - G)^{-1}|\left(\frac{1}{1 - \|G^T\|_\infty}|M||\tilde{L}\tilde{L}^T - B||M|^T|G|^T s^{(m)}\right) \\ &\le v_M + \frac{\|v_M\|_\infty}{1 - \|v_G\|_\infty}v_G + \frac{1}{1 - \|v_{G^T}\|_\infty}|M||\tilde{L}\tilde{L}^T - B||M|^T v_{G^T} \\ &\quad + \frac{\||M||\tilde{L}\tilde{L}^T - B||M|^T v_{G^T}\|_\infty}{(1 - \|G\|_\infty)(1 - \|v_{G^T}\|_\infty)}|G|s^{(m)} \\ &\le v_H, \\ |G||H|s^{(m)} &\le |G|v_H \le v_{GH}, \\ |H||G|^T s^{(m)} &\le |H|v_{G^T} \\ &\le |M||\tilde{L}\tilde{L}^T - B||M|^T v_{G^T} + |G(I_m - G)^{-1}||M||\tilde{L}\tilde{L}^T - B||M|^T v_{G^T} \\ &\quad +|M||\tilde{L}\tilde{L}^T - B||M|^T|G^T(I_m - G^T)^{-1}|v_{G^T} \\ &\quad +|G(I_m - G)^{-1}||M||\tilde{L}\tilde{L}^T - B||M|^T|G^T(I_m - G^T)^{-1}|v_{G^T} \\ &\le v_{HG^T}, \\ |G||G|^T s^{(m)} &\le |G|v_{G^T} \le v_{GG^T}, \\ |G||H||G|^T s^{(m)} &\le |G|v_{HG^T} \le v_{GHG^T}, \end{aligned}$$

$$\begin{aligned} |P|s^{(m)} &\le |G|s^{(m)} + |H|s^{(m)} + |G|^T s^{(m)} + |G||H|s^{(m)} + |H||G|^T s^{(m)} \\ &\quad +|G||G|^T s^{(m)} + |G||H||G|^T s^{(m)} \\ &\le v_P. \end{aligned}$$

This and $\|v_P\|_\infty < 1$ yield $\|P\|_\infty < 1$, so that $I_m - P$ is nonsingular. The inequalities $|H|s^{(m)} \le v_H$ and $\|v_P\|_\infty < 1$ moreover imply $\|H\|_\infty < 1$, so that B is positive definite. Let $Q := I_n - X^T(I_m - P)^{-1}X$. The inequality $\|v_P\|_\infty < 1$ and Lemma 2 give

$$\begin{aligned}|Q|s^{(m)} &\le |I_n - X^T(I_m + P(I_m - P)^{-1})X|s^{(m)} \\ &\le |I_n - X^TX|s^{(n)} + |X^TP(I_m - P)^{-1}X|s^{(n)} \\ &\le |I_n - X^TX|s^{(n)} + \frac{\|X\|_\infty}{1-\|P\|_\infty}|X|^T|P|s^{(m)} \le v_Q.\end{aligned}$$

This and $\|v_Q\|_\infty < 1$ yield $\|Q\|_\infty < 1$, so that $I_n - Q$ is nonsingular, which shows that X has full column rank, implying that A has also full column rank and S is nonsingular. From the proof of Theorem 1, we obtain

$$\begin{aligned}&\tilde{x} - (A^TB^{-1}A)^{-1}A^TB^{-1}b \\ &= (S^{-T}X^TM^{-T}B^{-1}M^{-1}XS^{-1})^{-1}(A^TM^TM^{-T}B^{-1}M^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + \rho_{\tilde{y}}) \\ &= S(X^TM^{-T}\tilde{L}^{-T}\tilde{L}^TB^{-1}\tilde{L}\tilde{L}^{-1}M^{-1}X)^{-1}S^T \\ &\quad \times(A^TM^TM^{-T}\tilde{L}^{-T}\tilde{L}^TB^{-1}\tilde{L}\tilde{L}^{-1}M^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + \rho_{\tilde{y}}) \\ &= S(X^T(I_m - G^T)^{-1}(I_m - H)^{-1}(I_m - G)^{-1}X)^{-1}S^T \\ &\quad \times(A^TM^T(I_m - G^T)^{-1}(I_m - H)^{-1}(I_m - G)^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + \rho_{\tilde{y}}) \\ &= S(X^T((I_m - G)(I_m - H)(I_m - G^T))^{-1}X)^{-1}S^T \\ &\quad \times(A^TM^T((I_m - G)(I_m - H)(I_m - G^T))^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + \rho_{\tilde{y}}) \\ &= S(X^T(I_m - P)^{-1}X)^{-1}S^T(A^TM^T(I_m - P)^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + \rho_{\tilde{y}}) \\ &= S(I_m - Q)^{-1}(X^T(I_m - P)^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}}) \\ &= S(X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}}) + SQ(I_n - Q)^{-1}(X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}}) \\ &\quad +SX^TP(I_m - P)^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + SQ(I_n - Q)^{-1}X^TP(I_m - P)^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}}).\end{aligned}$$

It finally holds from $\|P\|_\infty \le \|v_P\|_\infty < 1$, $\|Q\|_\infty \le \|v_Q\|_\infty < 1$ and Lemma 2 that

$$\begin{aligned}&|\tilde{x} - (A^TB^{-1}A)^{-1}A^TB^{-1}b| \\ &\le |S(X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}})| + |SQ(I_n - Q)^{-1}(X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}})| \\ &\quad +|SX^TP(I_m - P)^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}})| \\ &\quad +|SQ(I_n - Q)^{-1}||X^TP(I_m - P)^{-1}M(\rho_{\tilde{x}} + \rho_{\tilde{w}})| \\ &\le |S(X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}})| + \frac{\|X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}}\|_\infty}{1-\|Q\|_\infty}|S||Q|s^{(n)} \\ &\quad +\frac{\|M(\rho_{\tilde{x}} + \rho_{\tilde{w}})\|_\infty}{1-\|P\|_\infty}|S||X|^T|P|s^{(m)} \\ &\quad +|SQ(I_n - Q)^{-1}|\left(\frac{\|M(\rho_{\tilde{x}} + \rho_{\tilde{w}})\|_\infty}{1-\|P\|_\infty}|X|^T|P|s^{(m)}\right) \\ &\le |S(X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}})| + \frac{\|X^TM(\rho_{\tilde{x}} + \rho_{\tilde{w}}) + S^T\rho_{\tilde{y}}\|_\infty}{1-\|v_Q\|_\infty}|S|v_Q \\ &\quad +\frac{\|M(\rho_{\tilde{x}} + \rho_{\tilde{w}})\|_\infty}{1-\|v_P\|_\infty}|S||X|^Tv_P + \frac{\|M(\rho_{\tilde{x}} + \rho_{\tilde{w}})\|_\infty\||X|^Tv_P\|_\infty}{(1-\|Q\|_\infty)(1-\|v_P\|_\infty)}|S||Q|s^{(n)} \\ &\le d_B^{(2)}. \qquad \square\end{aligned}$$

Consider next the case when L is given. We present Corollaries 1 and 2 for enclosing $(L^{-1}A)^+L^{-1}b$, which can be obtained analogously to Theorems 1 and 2.

Corollary 1 *Let A, M, S, $\tilde{x}$, b, $\tilde{w}$, $\tilde{y}$, $\rho_{\tilde{x}}$, $\rho_{\tilde{y}}$ and X be as in Theorem 1, $L \in \mathbb{R}^{m\times m}$ be given, $\rho_{\tilde{w}} := \tilde{w} - LL^T\tilde{y}$, $Y := ML$, $V := I_m - YY^T$ and*

$$v_U := |I_n - X^TX|s^{(n)} + \frac{\|X\|_\infty}{1-\|V\|_\infty}|X|^T|V|s^{(m)}.$$

If $\|V\|_\infty < 1$, M and L are nonsingular. If $\|v_U\|_\infty < 1$, additionally, S is nonsingular, A has full column rank, and $|\tilde{x} - (L^{-1}A)^+L^{-1}b| \le d_L^{(1)}$, where

$$\begin{aligned} d_L^{(1)} \quad := \quad & |S(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}}) + S^T\rho_{\tilde{y}})| + \frac{\|X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}}) + S^T\rho_{\tilde{y}}\|_\infty}{1-\|v_U\|_\infty}|S|v_U \\ & + \frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty}{1-\|V\|_\infty}|S||X|^T|V|s^{(m)} + \frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty \||X|^T|V|s^{(m)}\|_\infty}{(1-\|v_U\|_\infty)(1-\|V\|_\infty)}|S|v_U. \end{aligned}$$

Proof. The result holds by setting $B = LL^T$ in the proof of Theorem 1. □

Corollary 2 *Let $L \in \mathbb{R}^{m\times m}$ be given and triangular, M be an approximate inverse of L computed similarly to Lemma 4, A, S, $\tilde{x}$, b, $\tilde{w}$, $\tilde{y}$, $\rho_{\tilde{x}}$, $\rho_{\tilde{y}}$ and X be as in Theorem 1, v_G, v_{G^T} and v_{GG^T} be as in Theorem 2, $\rho_{\tilde{w}}$ be as in Corollary 1, $v_K := v_G + v_{G^T} + v_{GG^T}$ and*

$$v_N := |I_n - X^TX|s^{(n)} + \frac{\|X\|_\infty}{1-\|v_K\|_\infty}|X|^Tv_K.$$

Assume $\|v_G\|_\infty < 1$, $\|v_{G^T}\|_\infty < 1$, $\|v_K\|_\infty < 1$ and $\|v_N\|_\infty < 1$. Then S, L and M are nonsingular, A has full rank, and $|\tilde{x} - (L^{-1}A)^+L^{-1}b| \le d_L^{(2)}$, where

$$\begin{aligned} d_L^{(2)} \quad := \quad & |S(X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}}) + S^T\rho_{\tilde{y}})| + \frac{\|X^TM(\rho_{\tilde{x}}+\rho_{\tilde{w}}) + S^T\rho_{\tilde{y}}\|_\infty}{1-\|v_N\|_\infty}|S|v_N \\ & + \frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty}{1-\|v_K\|_\infty}|S||X|^Tv_K + \frac{\|M(\rho_{\tilde{x}}+\rho_{\tilde{w}})\|_\infty \||X|^Tv_K\|_\infty}{(1-\|v_N\|_\infty)(1-\|v_K\|_\infty)}|S|v_N. \end{aligned}$$

Proof. The result follows by putting $B = LL^T$ and $\tilde{L} = L$ in the proof of Theorem 2. Note that H in the proof becomes zero matrix in this case. □

4 A Technique for Accelerating the Enclosure

In this section, we review a technique for accelerating the enclosure that appears in [6, Section 3]. Let L, $\tilde{L}$ and R be as in Section 3, $\tilde{x}$, $\tilde{w}$, $\tilde{y}$, M, S, X, F and $d_B^{(1)}$ be as in Theorem 1, Y and V be as in Corollary 1.

The proposed algorithm based on Theorem 1 computes $d_B^{(1)}$ taking rounding errors into account. In order to compute $d_B^{(1)}$ considering rounding errors, rigorous upper bounds for $|I_n - X^TX|s^{(n)}$ and $|F|s^{(m)}$ are necessary. The computation of the upper bound on $|I_n - X^TX|s^{(n)}$ can be accelerated by the following technique: Let X_c, $X_r \in \mathbb{R}^{m\times n}$ with $\min_{i,j}(X_r)_{ij} \ge 0$ satisfy $X \in \langle X_c, X_r\rangle$. The matrices X_c and X_r can be obtained by rounding mode-controlled floating point computation. From the center-radius interval arithmetic evaluation (e.g. [1]), we have

$$I_n - X^TX \in \langle I_n - X_c^TX_c, |X_c^T|X_r + X_r^T|X_c| + X_r^TX_r\rangle. \tag{3}$$

Utilizing an a priori error estimation (e.g. [4]), it holds that

$$|I_n - X_c^TX_c - \mathrm{fl}(I_n - X_c^TX_c)| \le \gamma_{m+1}(I_n + |X_c^T||X_c|) + m\underline{\mathbf{u}}s^{(n)}s^{(n)^T}. \tag{4}$$

Note that (4) holds also in the presence of underflow. The relations (3) and (4) yield $I_n - X^T X \in \langle \mathrm{fl}(I_n - X_c^T X_c), X_R \rangle$, where

$$X_R := \gamma_{m+1}(I_n + |X_c^T||X_c|) + m\underline{\mathbf{u}}s^{(n)}s^{(n)^T} + |X_c^T|X_r + X_r^T|X_c| + X_r^T X_r.$$

This shows

$$\begin{aligned} |I_n - X^T X|s^{(n)} \quad \leq \quad & |\mathrm{fl}(I_n - X_c^T X_c)|s^{(n)} + \gamma_{m+1}(s^{(n)} + |X_c^T||X_c|s^{(n)}) \\ & + mn\underline{\mathbf{u}}s^{(n)} + |X_c^T|X_r s^{(n)} + X_r^T|X_c|s^{(n)} + X_r^T X_r s^{(n)}. \end{aligned}$$

From this, we need to execute the matrix multiplication $X_c^T X_c$ only once in rounding-to-nearest mode for calculating the rigorous upper bound of $|I_n - X^T X|s^{(n)}$, if X_c and X_r are given. The computations of the upper bounds for $|F|s^{(m)}$ and $|V|s^{(m)}$ can be accelerated similarly (see Appendix with respect to $|F|s^{(m)}$).

We finally estimate the computational costs of the algorithms based on the theories in Section 3. Note that the costs for obtaining $\tilde{x}$, $\tilde{w}$ and $\tilde{y}$ are excluded in those of the algorithms shown below. By adopting the technique above, the cost of the algorithm based on Theorem 1 is $20m^3/3 + 4m^2n + 10mn^2 - n^3/3 + \mathcal{O}(m^2 + n^2)$, divided into

$m^3/3$ the floating point Cholesky decomposition of B to obtain $\tilde{L}$,

$m^3/3$ the floating point inversion of $\tilde{L}$ to obtain M,

$4m^2n$ the inclusion of MA,

$2n^2(m - n/3)$ the economy size floating point QR factorization of the mid point of the inclusion of MA to obtain R,

$n^3/3$ the floating point inversion of R to obtain S,

$6mn^2$ the inclusion of X reusing the inclusion of MA,

$2mn^2$ the computation of the upper bound of $|I_n - X^T X|s^{(n)}$,

$4m^3$ the inclusion of MB,

$2m^3$ the computation of the upper bound of $|F|s^{(m)}$.

The cost of the algorithm based on Theorem 2 is $2m^3/3 + 4m^2n + 10mn^2 - n^3/3 + \mathcal{O}(m^2 + n^2)$. The cost of the algorithm based on Corollary 1 is $19m^3/3 + 4m^2n + 10mn^2 - n^3/3 + \mathcal{O}(m^2 + n^2)$, divided into

$m^3/3$ the inversion of L,

$4m^2n$ the inclusion of MA,

$2n^2(m - n/3)$ the QR factorization,

$n^3/3$ the inversion of R,

$6mn^2$ the inclusion of X,

$2mn^2$ the computation of the upper bound of $|I_n - X^T X|s^{(n)}$,

$4m^3$ the inclusion of Y,

$2m^3$ the computation of the upper bound of $|V|s^{(m)}$.

The cost of the algorithm based on Corollary 2 is $m^3/3 + 4m^2n + 10mn^2 - n^3/3 + \mathcal{O}(m^2 + n^2)$.

5 A Technique for Obtaining Smaller Error Bounds

For reducing each component of $d_B^{(1)}$, $d_B^{(2)}$, $d_L^{(1)}$ and $d_L^{(2)}$ in Section 3, we need to reduce the absolute values of each component of the residuals $\rho_{\tilde{x}}$, $\rho_{\tilde{w}}$ and $\rho_{\tilde{y}}$. For obtaining the residuals whose components are small in the sense of absolute value, accurate approximations $\tilde{x}$, $\tilde{w}$ and $\tilde{y}$ to $(A^TB^{-1}A)^{-1}A^TB^{-1}b$, $A(A^TB^{-1}A)^{-1}A^TB^{-1}b-b$ and $B^{-1}(A(A^TB^{-1}A)^{-1}A^TB^{-1}b-b)$, respectively, are necessary. Such accurate approximations can be obtained via iterative refinement. In this section, we introduce the iterative refinement in the case when B is given. In the case when L is given, the iterative refinement can be achieved analogously by considering LL^T instead of B.

Let $M \in \mathbb{R}^{m\times m}$ and $S \in \mathbb{R}^{n\times n}$ be given and $X := MAS$. Assume M and S are nonsingular, $\|I_m - MBM^T\|_\infty < 1$ and $\|I_n - X^T(MBM^T)^{-1}X\|_\infty < 1$. The approximations $\tilde{x}$, $\tilde{w}$ and $\tilde{y}$ can be improved into $\tilde{x}-\delta_{\tilde{x}}$, $\tilde{w}-\delta_{\tilde{w}}$ and $\tilde{y}-\delta_{\tilde{y}}$ by the following residual iteration step:

$$\begin{aligned}
\rho_{\tilde{x}} &:= A\tilde{x} - b - \tilde{w}, \quad \rho_{\tilde{w}} := \tilde{w} - B\tilde{y}, \quad \rho_{\tilde{y}} := A^T\tilde{y}, \\
\delta_{\tilde{x}} &:= (A^TB^{-1}A)^{-1}(A^TB^{-1}(\rho_{\tilde{x}}+\rho_{\tilde{w}})+\rho_{\tilde{y}}), \quad \delta_{\tilde{w}} := A\delta_{\tilde{x}} - \rho_{\tilde{x}}, \\
\delta_{\tilde{y}} &:= -B^{-1}(\rho_{\tilde{w}} - \delta_{\tilde{w}}).
\end{aligned} \tag{5}$$

We then have $\tilde{x}-\delta_{\tilde{x}} = (A^TB^{-1}A)^{-1}A^TB^{-1}b$, $\tilde{w}-\delta_{\tilde{w}} = A(A^TB^{-1}A)^{-1}A^TB^{-1}b-b$ and $\tilde{y}-\delta_{\tilde{y}} = B^{-1}(A(A^TB^{-1}A)^{-1}A^TB^{-1}b-b)$, so that the residuals $\rho_{\tilde{x}}$, $\rho_{\tilde{w}}$ and $\rho_{\tilde{y}}$ vanish after one iteration. In theory, $(A^TB^{-1}A)^{-1} = S(X^T(MBM^T)^{-1}X)^{-1}S^T$ holds. Since X and MBM^T are expected to be not too far from orthogonality and identity, respectively, we change (5) by replacing $(A^TB^{-1}A)^{-1}$ by SS^T and introducing approximations of $B^{-1}(\rho_{\tilde{x}}+\rho_{\tilde{w}})$ and $B^{-1}(\rho_{\tilde{w}}-\delta_{\tilde{w}})$. Then the numerical iteration is defined as follows:

$$\begin{aligned}
\rho_{\tilde{x}}^{(i)} &:= A\tilde{x}^{(i)} - b - \tilde{w}^{(i)}, \quad \rho_{\tilde{w}}^{(i)} := \tilde{w}^{(i)} - B\tilde{y}^{(i)}, \quad \rho_{\tilde{y}}^{(i)} := A^T\tilde{y}^{(i)}, \\
\delta_{\tilde{x}}^{(i)} &:= SS^T(A^T\tilde{u}^{(i,k_u)} + \rho_{\tilde{y}}^{(i)}), \quad \tilde{x}^{(i+1)} := \tilde{x}^{(i)} - \delta_{\tilde{x}}^{(i)}, \quad \delta_{\tilde{w}}^{(i)} := A\delta_{\tilde{x}}^{(i)} - \rho_{\tilde{x}}^{(i)}, \\
\tilde{w}^{(i+1)} &:= \tilde{w}^{(i)} - \delta_{\tilde{w}}^{(i)}, \quad \delta_{\tilde{y}}^{(i)} := -\tilde{t}^{(i,k_t)}, \quad \tilde{y}^{(i+1)} := \tilde{y}^{(i)} - \delta_{\tilde{y}}^{(i)}, \quad i = 1,2,\dots,
\end{aligned}$$

where $\tilde{u}^{(i,k_u)}$ and $\tilde{t}^{(i,k_t)}$ are the approximations of $B^{-1}(\rho_{\tilde{x}}+\rho_{\tilde{w}})$ and $B^{-1}(\rho_{\tilde{w}}-\delta_{\tilde{w}})$, and results for the inner iterations

$$\begin{aligned}
\rho_{\tilde{u}}^{(i,k_u)} &:= B\tilde{u}^{(i,k_u)} - (\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)}), \quad \delta_{\tilde{u}}^{(i,k_u)} := M^TM\rho_{\tilde{u}}^{(i,k_u)}, \\
\tilde{u}^{(i,k_u+1)} &:= \tilde{u}^{(i,k_u)} - \delta_{\tilde{u}}^{(i,k_u)}, \quad k_u = 1,2,\dots,
\end{aligned}$$

and

$$\begin{aligned}
\rho_{\tilde{t}}^{(i,k_t)} &:= B\tilde{t}^{(i,k_t)} - (\rho_{\tilde{w}}^{(i)} - \delta_{\tilde{w}}^{(i)}), \quad \delta_{\tilde{t}}^{(i,k_t)} := M^TM\rho_{\tilde{t}}^{(i,k_t)}, \\
\tilde{t}^{(i,k_t+1)} &:= \tilde{t}^{(i,k_t)} - \delta_{\tilde{t}}^{(i,k_t)}, \quad k_t = 1,2,\dots,
\end{aligned}$$

respectively. The initial values $\tilde{x}^{(1)}$, $\tilde{w}^{(1)}$, $\tilde{y}^{(1)}$, $\tilde{u}^{(i,1)}$ and $\tilde{t}^{(i,1)}$ are defined such that $\tilde{x}^{(1)} := \tilde{x}$, $\tilde{w}^{(1)} := \tilde{w}$, $\tilde{y}^{(1)} := \tilde{y}$, $\tilde{u}^{(i,1)} := M^TM(\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)})$ and $\tilde{t}^{(i,1)} := M^TM(\rho_{\tilde{w}}^{(i)} - \delta_{\tilde{w}}^{(i)})$, respectively.

We present Theorem 3, which shows the convergence of the approximations.

Theorem 3 *Let M, S, X, $\tilde{x}^{(i)}$, $\tilde{w}^{(i)}$, $\tilde{y}^{(i)}$, $\rho_{\tilde{x}}^{(i)}$, $\rho_{\tilde{w}}^{(i)}$, $\delta_{\tilde{w}}^{(i)}$, $\tilde{u}^{(i,k_u)}$ and $\tilde{t}^{(i,k_t)}$ be as the above. Assume M and S are nonsingular, $\|I_m - MBM^T\|_\infty < 1$ and $\|I_n - X^T(MBM^T)^{-1}X\|_\infty < 1$. It then holds that*

$$\begin{aligned}
\lim_{k_u\to\infty} \tilde{u}^{(i,k_u)} &= B^{-1}(\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)}), \quad \lim_{k_t\to\infty} \tilde{t}^{(i,k_t)} = B^{-1}(\rho_{\tilde{w}}^{(i)} - \delta_{\tilde{w}}^{(i)}),\\
\lim_{i\to\infty}\left(\lim_{k_u\to\infty} \tilde{x}^{(i)}\right) &= (A^TB^{-1}A)^{-1}A^TB^{-1}b,\\
\lim_{i\to\infty}\left(\lim_{k_u\to\infty} \tilde{w}^{(i)}\right) &= A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b,\\
\lim_{i\to\infty}\left(\lim_{k_u\to\infty}\left(\lim_{k_t\to\infty} \tilde{y}^{(i)}\right)\right) &= B^{-1}(A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b).
\end{aligned}$$

Proof We obtain

$$\begin{aligned}
\tilde{u}^{(i,k_u+1)} &= M^T(I_m - MBM^T)M^{-T}\tilde{u}^{(i,k_u)} + M^TM(\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)}) =: C\tilde{u}^{(i,k_u)} + z_{\tilde{u}}^{(i)},\\
\tilde{t}^{(i,k_t+1)} &= M^T(I_m - MBM^T)M^{-T}\tilde{t}^{(i,k_t)} + M^TM(\rho_{\tilde{w}}^{(i)} - \delta_{\tilde{w}}^{(i)}) =: C\tilde{t}^{(i,k_t)} + z_{\tilde{t}}^{(i)}.
\end{aligned}$$

By the assumption, $\varrho(C) = \varrho(I_m - MBM^T) \le \|I_m - MBM^T\|_\infty < 1$, so that

$$\begin{aligned}
\lim_{k_u\to\infty} \tilde{u}^{(i,k_u)} &= (I_m - C)^{-1}z_{\tilde{u}}^{(i)} = (M^TMB)^{-1}M^TM(\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)}) = B^{-1}(\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)}),\\
\lim_{k_t\to\infty} \tilde{t}^{(i,k_t)} &= (I_m - C)^{-1}z_{\tilde{t}}^{(i)} = (M^TMB)^{-1}M^TM(\rho_{\tilde{w}}^{(i)} - \delta_{\tilde{w}}^{(i)}) = B^{-1}(\rho_{\tilde{w}}^{(i)} - \delta_{\tilde{w}}^{(i)}).
\end{aligned}$$

Let $C_{\tilde{x}} := S(I_n - X^T(MBM^T)^{-1}X)S^{-1}$ and $z_{\tilde{x}} := SS^TA^TB^{-1}b$. From the assumption, we have

$$\varrho(C_{\tilde{x}}) = \varrho(I_n - X^T(MBM^T)^{-1}X) \le \|I_n - X^T(MBM^T)^{-1}X\|_\infty < 1.$$

It holds from this and the convergence of $\tilde{u}^{(i,k_u)}$ that

$$\begin{aligned}
\lim_{i\to\infty}\left(\lim_{k_u\to\infty} \tilde{x}^{(i)}\right) &= \lim_{i\to\infty}\left(\lim_{k_u\to\infty} \tilde{x}^{(i+1)}\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty} (\tilde{x}^{(i)} - SS^T(A^T\tilde{u}^{(i,k_u)} + \rho_{\tilde{y}}^{(i)}))\right)\\
&= \lim_{i\to\infty} (\tilde{x}^{(i)} - SS^T(A^TB^{-1}(\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)}) + \rho_{\tilde{y}}^{(i)}))\\
&= \lim_{i\to\infty} (C_{\tilde{x}}\tilde{x}^{(i)} + z_{\tilde{x}}) = (I_n - C_{\tilde{x}})^{-1}z_{\tilde{x}}\\
&= (SS^TA^TB^{-1}A)^{-1}SS^TA^TB^{-1}b = (A^TB^{-1}A)^{-1}A^TB^{-1}b.
\end{aligned}$$

It follows from the convergence of $\tilde{u}^{(i,k_u)}$ and $\tilde{x}^{(i)}$ that

$$\begin{aligned}
\lim_{i\to\infty}\left(\lim_{k_u\to\infty}\tilde{w}^{(i)}\right) &= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}\tilde{w}^{(i+1)}\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}(\tilde{w}^{(i)} - ASS^T(A^T\tilde{u}^{(i,k_u)} + \rho_{\tilde{y}}^{(i)}) + \rho_{\tilde{x}}^{(i)})\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}(\tilde{w}^{(i)} - ASS^T(A^TB^{-1}(\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)}) + \rho_{\tilde{y}}^{(i)}) + \rho_{\tilde{x}}^{(i)})\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}(-ASS^T(A^TB^{-1}(A\tilde{x}^{(i)} - b)) + A\tilde{x}^{(i)} - b)\right)\\
&= -ASS^T(A^TB^{-1}(A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b))\\
&\quad +A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b\\
&= A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b.
\end{aligned}$$

The convergences of $\tilde{u}^{(i,k_u)}$, $\tilde{t}^{(i,k_t)}$ and $\tilde{x}^{(i)}$ finally yield

$$\begin{aligned}
&\lim_{i\to\infty}\left(\lim_{k_u\to\infty}\left(\lim_{k_t\to\infty}\tilde{y}^{(i)}\right)\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}\left(\lim_{k_t\to\infty}\tilde{y}^{(i+1)}\right)\right) = \lim_{i\to\infty}\left(\lim_{k_u\to\infty}\left(\lim_{k_t\to\infty}(\tilde{y}^{(i)} + \tilde{t}^{(i,k_t)})\right)\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}(\tilde{y}^{(i)} + B^{-1}(\rho_{\tilde{w}}^{(i)} - \delta_{\tilde{w}}^{(i)}))\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}B^{-1}(\tilde{w}^{(i)} - ASS^T(A^T\tilde{u}^{(i,k_u)} + \rho_{\tilde{y}}^{(i)}) + \rho_{\tilde{x}}^{(i)})\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}B^{-1}(\tilde{w}^{(i)} - ASS^T(A^TB^{-1}(\rho_{\tilde{x}}^{(i)} + \rho_{\tilde{w}}^{(i)}) + \rho_{\tilde{y}}^{(i)}) + \rho_{\tilde{x}}^{(i)})\right)\\
&= \lim_{i\to\infty}\left(\lim_{k_u\to\infty}B^{-1}(-ASS^TA^TB^{-1}(A\tilde{x}^{(i)} - b) + A\tilde{x}^{(i)} - b)\right)\\
&= B^{-1}(-ASS^TA^TB^{-1}(A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b)\\
&\quad +A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b)\\
&= B^{-1}(A(A^TB^{-1}A)^{-1}A^TB^{-1}b - b). \qquad \square
\end{aligned}$$

The iteration benefits substantially by using extra-precise evaluations of the residuals. For this purpose, a so-called error-free transformation is available (see [14]). In our application, in particular the amplification of the correction $\delta_{\tilde{x}}^{(i)}$ by SS^T is of the order $(A^TB^{-1}A)^{-1}$, thus it is beneficial to store $\tilde{x}$ in two terms, which can be achieved similarly to [14, Section 5] (see the appendix for details).

6 Numerical Results when B is Given

In this section, we consider the case when B is given and report numerical results to illustrate the properties of the proposed algorithms and performance of our implementation. We used a computer with Intel Xeon 2.66GHz Dual CPU, 4.00GB RAM and MATLAB 7.5 with Intel Math Kernel Library and IEEE 754 double precision.

We denote the compared algorithms as follows [1]:

- `Ci`: The algorithm based on Theorem 1 with the acceleration and iterative refinement in Sections 4 and 5, respectively (see Appendix),
- `Fi`: The algorithm based on Theorem 2 with the acceleration and iterative refinement (see Appendix),
- `I1`: The code `Res = verifylss([A,-B;zeros(n),A'],[b;zeros(n,1)]);` using the INTLAB routine `verifylss`,
- `I2`: The code `Res = verifylss(A'*verifylss(B,A),A'*verifylss(B,b));`.

The computations of the upper bounds for $|I_n - X^TX|s^{(n)}$ and $|F|s^{(m)}$, and $|I_n - X^TX|s^{(n)}$ only were accelerated in `Ci` and `Fi`, respectively. In `Ci` and `Fi`, we stored $\tilde{x}$ in two terms and set k_u and k_t in Section 5 such that $k_u = k_t = 1$. See the appendix for how to obtain M, S, $\tilde{x}$, $\tilde{w}$ and $\tilde{y}$ in Theorem 1, where the INTLAB codes of `Ci` and `Fi` are displayed.

Let $x^c, x^r \in \mathbb{R}^n$ with $\min_i x_i^r \geq 0$ satisfy $(A^TB^{-1}A)^{-1}A^TB^{-1}b \in \langle x^c, x^r\rangle$. In order to assess the quality of the enclosure, we define the relative radii $\xi_i := x_i^r/(|x_i^c| + x_i^r)$, $i = 1, \ldots, n$. We can regard $-\log_{10}\xi_i$ as the number of correct significant decimal digits, since it roughly corresponds to the number of digits to which the upper and the lower bounds coincide, i.e., the number of significant digits we know to be correct for each entry. Maximum relative radius MRR and average relative radius ARR are defined as $\mathrm{MRR} := \max_i \xi_i$ and $\mathrm{ARR} := \left(\prod_{i=1}^n \xi_i\right)^{\frac{1}{n}}$, respectively. Hence $-\log_{10}\mathrm{MRR}$ and $-\log_{10}\mathrm{ARR}$ represent the minimum and arithmetic mean of the correct digits, respectively. In `Ci` and `Fi`, we repeated the iterative refinement until either $\mathrm{MRR} \leq 10^{-11}$ or the number of iterations was 10.

The algorithms `Ci` and `Fi` verified that A had full column rank and B was positive definite for examples in which these algorithms succeeded. In Sections 6 and 7, for each parameter, we treated 100 problems and took the median of obtained radii or computing times, and the notation `NaN` means that `I2` returned `NaN` in all the problems. In some cases, the algorithms failed in parts of the 100 problems. In these cases, the tables below show the median of the radii obtained within the problems that succeeded.

6.1 Example 1

In this example, we observe the magnitudes of radii and computing times of the algorithms for various m and n. Consider (1) where A, b and B are generated by

$$\begin{aligned} &\mathtt{A} = \mathtt{randn(m,n)}; \quad \mathtt{b} = \mathtt{randn(m,1)}; \\ &\mathtt{B} = \mathtt{randn(m)}; \quad \mathtt{B} = \mathtt{m*eye(m)} + (\mathtt{B}+\mathtt{B}')/2;. \end{aligned} \tag{6}$$

The function `randn` generates a matrix whose elements are normally distributed pseudo random numbers. Table 1 displays the obtained radii and computing times of the algorithms for various m and n. Note that the computing times of `Ci` and `Fi` in Tables 1 and 4 include the times necessary for the computation of $\tilde{x}$, $\tilde{w}$ and $\tilde{y}$.

[1] The solution $(A^TB^{-1}A)^{-1}A^TB^{-1}b$ may be able to be enclosed by the code.

```
invB = verinverse(B); Res = verinverse(A'*invB*A)*(A'*(invB*b));
```

using the VERSOFT [12] routine `verinverse`. On the other hand, this approach required prohibitively large computing times in the examples below. Hence we excluded this algorithm from the comparisons in Sections 6 and 7.

Table 1: Obtained radii and computing times (sec) in Section 6.1

		`Ci`	`Fi`	`I1`	`I2`	`Ci`	`Fi`	`I1`	`I2`
m	n	MRR	MRR	MRR	MRR	time	time	time	time
		ARR	ARR	ARR	ARR				
1000	100	4.9e–13	4.9e–13	3.3e–12	3.4e–10	2.16	1.02	2.01	3.89
		1.9e–15	1.9e–15	9.3e–15	8.9e–13				
2000	100	5.7e–13	5.7e–13	1.4e–11	4.3e–10	13.2	4.78	11.5	22.9
		2.5e–15	2.5e–15	3.7e–14	1.1e–12				
3000	100	5.9e–13	5.9e–13	2.2e–11	3.5e–10	39.3	12.3	34.2	68.7
		3.2e–15	3.2e–15	9.3e–14	1.5e–12				
1000	200	5.7e–13	5.7e–13	5.9e–12	1.5e–9	2.43	1.27	2.50	4.81
		1.9e–15	1.9e–15	1.0e–14	2.8e–12				
1000	400	9.4e–13	9.4e–13	7.8e–12	8.6e–9	3.12	1.98	3.77	6.93
		2.1e–15	2.1e–15	1.2e–14	1.3e–11				
1000	600	1.5e–12	1.5e–12	1.9e–11	5.6e–8	4.12	3.00	5.49	9.39
		2.3e–15	2.3e–15	1.7e–14	4.9e–11				

The radii by `Ci` were approximately equal to those by `Fi`. The reason is that $d_B^{(1)}$ and $d_B^{(2)}$ in Section 3 have the same first term, and this term was dominant in the magnitude of these error bounds. The algorithm `Fi` was faster than `Ci`. This result coincides with the fact that the computational cost of the algorithm based on Theorem 2 is smaller than that based on Theorem 1.

6.2 Example 2

In this example, we observe the radii for various $\kappa(A)$. Consider (1) where $m = 60$, $n = 30$, A is generated by `A = gallery('randsvd', [60,30],cndA);`, and b and B are obtained by (6). We used the Higham's test matrix `randsvd` [4]. Then it holds approximately that $\kappa(A) \approx$ `cndA`. Table 2 shows the obtained radii for various `cndA`. When `cndA` $= 10^{14}$ and 10^{15}, the proposed algorithms failed in 9 and all the problems, respectively. The reasons for the failure of `Ci` and `Fi` are that $\|v_E\|_\infty < 1$ and $\|v_Q\|_\infty < 1$ could not be verified, respectively. When `cndA` $= 10^{15}$, `I1` returned `NaN` in 92 problems.

The radii by the proposed algorithms when `cndA` $= 10^9$ and 10^{14} were smaller and approximately equal to those when `cndA` $= 10^4$, respectively. The reason is that the numbers of the iterations when `cndA` $= 10^9$ and 10^{14} were larger than those when `cndA` $= 10^4$.

6.3 Example 3

In this example, we observe the radii for various $\kappa(B)$. Consider (1) where $m = 60$, $n = 30$, A and b are obtained by (6), and B is generated by `B = gallery('randsvd',60, -cndB);`. Then B is expected to be symmetric positive definite with $\kappa(B) \approx$ `cndB`. Table 3 displays the similar quantities to Table 2 for various `cndB`. When `cndB` $= 10^{13}$, `Fi` failed in all the problems. The reason of the failure is similar to that in Section 6.2.

Table 2: Obtained radii in Section 6.2

cndA	Ci	Fi	I1	I2
	MRR	MRR	MRR	MRR
	ARR	ARR	ARR	ARR
1e+4	2.2e–12	2.2e–12	6.1e–14	3.1e–5
	1.8e–13	1.8e–13	5.5e–16	2.9e–7
1e+9	8.6e–14	8.6e–14	7.9e–9	NaN
	1.7e–15	1.7e–15	9.9e–11	NaN
1e+14	2.0e–12	2.0e–12	4.5e–4	NaN
	1.9e–13	1.9e–13	1.2e–5	NaN
1e+15	failed	failed	6.2e–4	NaN
	failed	failed	3.4e–5	NaN

When cndB $= 10^{12}$, I2 returned NaN in 30 problems. The result in this section shows that Ci and I1 are robust for ill-conditioned B.

Table 3: Obtained radii in Section 6.3

cndB	Ci	Fi	I1	I2
	MRR	MRR	MRR	MRR
	ARR	ARR	ARR	ARR
1e+4	8.5e–13	8.5e–13	2.2e–14	8.0e–10
	1.7e–14	1.7e–14	4.3e–16	1.4e–11
1e+8	6.2e–13	6.2e–13	4.4e–9	2.5e–4
	5.2e–14	5.2e–14	9.6e–13	1.6e–6
1e+12	1.1e–13	6.2e–13	5.0e–8	8.6e–1
	1.2e–15	1.5e–14	3.1e–10	3.6e–1
1e+13	1.6e–12	failed	4.2e–8	NaN
	8.6e–14	failed	9.1e–10	NaN

As far as we see Tables 1, 2 and 3, it is recommended to execute Ci and Fi when B is and is not ill-conditioned, respectively.

7 Numerical Results when L is Given

In this section, we report numerical results when L is given. Let MRR and ARR be as in Section 6. We used the same computer as that in Section 6. The compared algorithms are as follows:

Ci: The algorithm based on Corollary 1 with the acceleration and iterative refinement,

Fi: The algorithm based on Corollary 2 with the acceleration and iterative refinement,

- I1: The code `Res=verifylss([A,-intval(L)*L';zeros(n),A'],[b;zeros(n,1)]);`,
- I2: The code `Res=verifylss(verifylss(L,A),verifylss(L,b));`.

The computations of the upper bounds for $|I_n - X^T X|s^{(n)}$ and $|V|s^{(m)}$, and $|I_n - X^T X|s^{(n)}$ only were accelerated in `Ci` and `Fi`, respectively. In `Ci` and `Fi`, we computed M, S, $\tilde{x}$, $\tilde{w}$, $\tilde{y}$, $\rho_{\tilde{x}}$ and $\rho_{\tilde{y}}$, and stored $\tilde{x}$ similarly to Section 6. Using the INTLAB routine `Dot_`, the enclosure of $\rho_{\tilde{w}}$ was calculated such that

```
Ly = Dot_(L',y,-2);  setround(1);  radLLy = abs(L)*rad(Ly);
rho_w = Dot_(1,w,-L,mid(Ly),-2);  rho_w = rho_w + midrad(0,radLLy);.
```

The iterative refinement was repeated until MRR $\leq 10^{-8}$ held or the number of the iteration became 10. The algorithms `Ci` and `Fi` verified that A had full column rank and L was nonsingular for examples where these algorithms succeeded.

7.1 Example 1

In this example, we observe the radii and computing times of the algorithms for various m and n. Consider (1) where A, b and L are generated by

$$\begin{aligned} &\mathtt{A} = \mathtt{randn}(\mathtt{m},\mathtt{n}); \quad \mathtt{b} = \mathtt{randn}(\mathtt{m},1);\\ &\mathtt{B} = \mathtt{randn}(\mathtt{m}); \quad \mathtt{B} = \mathtt{m} * \mathtt{eye}(\mathtt{m}) + (\mathtt{B}+\mathtt{B}')/2; \quad \mathtt{L} = \mathtt{chol}(\mathtt{B})'; . \end{aligned} \tag{7}$$

Table 4 displays the similar quantities to Table 1.

Table 4: Obtained radii and computing times (sec) in Section 7.1

		Ci	Fi	I1	I2	Ci	Fi	I1	I2
m	n	MRR	MRR	MRR	MRR	time	time	time	time
		ARR	ARR	ARR	ARR				
1000	100	4.2e–12	4.2e–12	3.9e–10	5.2e–11	2.31	1.25	3.28	6.12
		2.3e–14	2.3e–14	2.3e–12	3.0e–13				
2000	100	4.9e–12	4.9e–12	1.0e–9	7.4e–11	13.0	5.39	20.2	36.8
		2.8e–14	2.8e–14	5.7e–12	4.1e–13				
3000	100	1.5e–11	1.5e–11	4.5e–9	2.2e–10	37.2	13.2	62.1	111
		3.3e–14	3.3e–14	9.9e–12	5.1e–13				
1000	200	1.2e–11	1.2e–11	9.6e–10	2.2e–10	2.61	1.48	3.95	7.50
		3.2e–14	3.2e–14	2.7e–12	6.2e–13				
1000	400	2.9e–11	2.9e–11	1.9e–9	7.4e–10	3.23	2.17	5.62	10.8
		4.6e–14	4.6e–14	3.7e–12	1.5e–12				
1000	600	2.1e–10	2.1e–10	1.3e–8	7.1e–9	4.39	3.33	7.83	14.6
		5.5e–14	5.5e–14	5.2e–12	2.8e–12				

Comparing with `I1` in Section 6, `I1` in this section gave larger radii and was slower. The reason is guessed that `I1` in this section includes the interval arithmetic evaluation of $-LL^T$.

7.2 Example 2

In this example, we observe the radii for various $\kappa(A)$. Consider (1) where $m = 60$, $n = 30$, A is generated similarly to Section 6.2, and b and L are obtained by (7). Table 5 shows the similar quantities to Table 2 for various `cndA`. When `cndA` $= 10^{14}$ and 10^{15}, the proposed algorithms failed in 8 and all the problems, respectively. The reasons for the failure of `Ci` and `Fi` are that $\|v_U\|_\infty < 1$ and $\|v_N\|_\infty < 1$ could not be verified, respectively. When `cndA` $= 10^{15}$, `I1` returned `NaN` in 96 problems. The algorithms `Ci` and `Fi` gave smaller radii than those by `I1` and `I2`. When `cndA` $= 10^{14}$, on the other hand, the proposed algorithms failed in some problems, although `I1` and `I2` succeeded in all the problems.

Table 5: Obtained radii in Section 7.2

	`Ci`	`Fi`	`I1`	`I2`
`cndA`	MRR	MRR	MRR	MRR
	ARR	ARR	ARR	ARR
1e+4	9.6e–11	9.6e–11	3.2e–9	4.8e–9
	5.9e–13	5.9e–13	2.3e–11	3.5e–11
1e+9	1.3e–12	1.3e–12	4.4e–4	6.4e–4
	6.3e–15	6.3e–15	1.9e–6	2.9e–6
1e+14	1.9e–9	1.9e–9	6.2e–1	6.9e–1
	2.3e–10	2.3e–10	1.4e–1	1.9e–1
1e+15	failed	failed	9.6e–1	`NaN`
	failed	failed	5.5e–1	`NaN`

7.3 Example 3

In this example, we observe the radii for various $\kappa(L)$. Consider (1) where $m = 60$, $n = 30$, A and b are obtained by (7), and L is generated by

```
P = qr(gallery('randsvd',m,cndL));  L = triu(P)';
L = L*sparse(diag(sign(diag(L))));.
```

Then L is the Cholesky factor of LL^T with $\kappa(L) \approx$ `cndL`. Table 6 displays the similar quantities to Table 2 for various `cndL`. When `cndL` = 1e+14, `Ci` and `Fi` failed in 4 and all the problems, respectively. The reason of the former failure is similar to that in Section 7.2. The reason of the latter failure is that $\|v_G\|_\infty < 1$, $\|v_{G^T}\|_\infty < 1$ or $\|v_K\|_\infty < 1$ could not be verified. When `cndL` $= 10^{13}$ and 10^{14}, `I1` returned `NaN` in 5 and 95 problems, respectively. The result in this section shows that `I2` is robust for ill-conditioned L.

From Tables 4, 5 and 6, it is recommended to apply `Fi` when L is not ill-conditioned. When L is ill-conditioned, the recommendation depends on either faster algorithm or smaller radii are required. In the former and latter cases, we recommend executing `Ci` and `I2`, respectively.

Table 6: Obtained radii in Section 7.3

cndL	`Ci`	`Fi`	`I1`	`I2`
	MRR	MRR	MRR	MRR
	ARR	ARR	ARR	ARR
1e+4	5.2e–10	5.2e–10	1.1e–7	5.2e–11
	3.4e–12	3.4e–12	4.0e–10	2.0e–13
1e+8	6.5e–10	6.5e–10	2.2e–4	2.9e–11
	1.7e–11	1.7e–11	3.9e–6	4.8e–13
1e+13	3.0e–9	8.4e–9	7.7e–1	8.1e–11
	4.8e–11	1.4e–10	2.7e–1	9.5e–13
1e+14	6.2e–8	failed	1.0e+0	1.4e–10
	2.4e–10	failed	9.3e–1	1.3e–12

8 Conclusion

In this paper, we proposed algorithms for enclosing the solutions of (1), established Theorems 1 and 2, and Corollaries 1 and 2 for developing these algorithms, reviewed and introduced the techniques for accelerating the enclosure and obtaining smaller error bounds, respectively, and reported numerical results. By modifying these algorithms slightly, enclosing the solutions where A, b, and/or B are complex and/or interval is also possible. As was suggested by one of the referees, the techniques presented here can also be applied to the enclosure of the covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ of $x = (A^TB^{-1}A)^{-1}A^TB^{-1}b$, where $\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$ and $\mu_i = E[x_i]$ is the expected value of x_i, if we can evaluate $E[x_i]$ via interval arithmetic. In several cases, B has special structure, such as diagonal. When B is diagonal, the floating point Cholesky decomposition, inversions of $\tilde{L}$ or L, and their error estimations are not required, so the enclosure can be obtained more efficiently and effectively. Our future work will be to develop a robust algorithm when L is given and ill-conditioned.

Acknowledgments

The author wishes to thank Prof. Siegfried M. Rump in Hamburg University of Technology for fruitful discussions, and acknowledges the referees for the comments which improved the first version of this paper. This research was partially supported by Grant-in-Aid for Scientific Research (C) (23560066, 2011–2015) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Appendix

In what follows we display the INTLAB codes of `Ci` and `Fi` in Section 6 in order to clarify the implementation. The codes of `Ci` and `Fi` in Section 7 are analogous.

```
function [x1,x2,dB1] = Ci(A,b,B)
% Ci encloses the solutions of the generalized least squares problems
% based on Theorem 1 with the iterative refinement.
% A: m * n real matrix (m >= n),
% b: real m-vector,
% B: m * m real symmetric matrix (expected to be positive definite),
% x1 + x2: approximate solution,
% dB1: error bound of x1 + x2.

setround(0);  [m,n] = size(A);  s_n = ones(n,1);  s_m = ones(m,1);
ur = 2^(-53);  uu = 2^(-1074);  % unit roundoff and underflow unit

% computation of approximate solutions
L = chol(B)';  M = L\speye(m);  MA = M*intval(A);  P = qr(mid(MA),0);
R = triu(P(1:n,:));  clear P;  S = R\speye(n);
x1 = R\(R'\A'*(L'\(L\b)));  clear R;  x2 = zeros(n,1);  w = A*x1 - b;
y = L'\(L\w);

% enclosure of X and B*M [MA is an interval]
X = MA*S;  clear MA;  Xc = mid(X);  Xr = rad(X);
BM = intval(B)*M';  BMc = mid(BM);  BMr = rad(BM);  clear BM;

% fl(I_m - M*B*M') and fl(I_n - X_c'*X_c)
fl_IMBMc = eye(m) - M*BMc;  fl_IXX = eye(n) - Xc'*Xc;

% gamma_m+1 [denominators are computed as if rounding to -inf mode]
setround(1);  gam_m1 = (m+1)*ur/(-((m+1)*ur - 1));

% rigorous upper bound of abs(F)*ones(m,1) and norm(F,inf)
Fs = abs(fl_IMBMc)*s_m + gam_m1*(s_m + abs(M)*(abs(BMc)*s_m)) ...
     + m*m*uu*s_m + abs(M)*(BMr*s_m);
clear fl_IMBMc BMc BMr;  normF = max(Fs);

if normF >= 1  % norm(F,inf) < 1 could not be verified
    disp('Error: Ci (normF >= 1)');
    x1 = NaN;  x2 = NaN;  dB1 = NaN;  return;
end

% rigorous upper bounds of abs(I_n - X'*X)*ones(n,1), abs(X),
% norm(X,inf), abs(X)'*abs(F)*ones(m,1), v_E and norm(v_E,inf)
IXXs = abs(fl_IXX)*s_n + gam_m1*(s_n + abs(Xc)'*(abs(Xc)*s_n)) ...
       + m*n*uu*s_n + abs(Xc)'*(Xr*s_n) + Xr'*(abs(Xc)*s_n) ...
       + Xr'*(Xr*s_n);  clear fl_IXX Xc Xr;
absX = mag(X);  normX = max(sum(absX,2));  XFs = absX'*Fs;
clear absX;  vE = IXXs + (normX*XFs)/(-(normF - 1));  normvE = max(vE);

if normvE >= 1  % norm(v_E,inf) < 1 could not be verified
    disp('Error: Ci (normvE >= 1)');
    x1 = NaN;  x2 = NaN;  dB1 = NaN;  return;
end

% rigorous upper bounds of abs(S)*v_E, norm(abs(X)'*abs(F)*ones(m,1),inf)
% and abs(S)*abs(X)'*abs(F)*ones(m,1)
```

```
SvE = abs(S)*vE;  normXFs = max(XFs);  SXFs = abs(S)*XFs;

% accurate computations of rho_x, rho_w and rho_y (enclosure)
rho_x = Dot_(A,x1,A,x2,-1,b,-1,w,-2);  rho_w = Dot_(1,w,-B,y,-2);
rho_y = Dot_(A',y,-2);

% computation of dB1 and the iterative refinement

setround(0);  % u and t are required previously
u = M'*(M*(mid(rho_x) + mid(rho_w)));
delta_x = S*(S'*(A'*u + mid(rho_y)));
delta_w = A*delta_x - mid(rho_x);  t = M'*(M*(mid(rho_w) - delta_w));

for loop = 1:10  % at most 10 iterations
    % enclosure of M*(rho_x + rho_w), X'*M*(rho_x + rho_w) + S'*rho_y
    % and S*(X'*M*(rho_x + rho_w) + S'*rho_y)
    % [rho_x, rho_w and rho_y are intervals]
    Mrxw = M*(rho_x + rho_w);  XMrxwSry = X'*Mrxw + S'*rho_y;
    SXMrxwSry = S*XMrxwSry;

    % rigorous upper bounds of norm(X'*M*(rho_x + rho_w) + S'*rho_y,inf)
    % and norm(M*(rho_x + rho_w),inf)
    normXMrxwSry = max(mag(XMrxwSry));  normMrxw = max(mag(Mrxw));

    % rigorous upper bounds of d_B^(1) and the relative radii
    setround(1);
    dB1 = mag(SXMrxwSry) + (normXMrxwSry*SvE)/(-(normvE - 1)) ...
          + (normMrxw*SXFs)/(-(normF - 1)) ...
          + ((normMrxw*normXFs)*SvE)/(-((-(normvE - 1))*(normF - 1)));
    RR = dB1 ./ (-((-abs(x2) - dB1) - abs(x1)));

    if max(RR) <= 1e-11,  break;  % iteration terminates since MRR <= 1e-11
    else
        setround(0);  % iterative refinement
        rho_u = Dot_(B,u,-1,mid(rho_x),-1,mid(rho_w),2);
        delta_u = M'*(M*rho_u);  u = u - delta_u;
        delta_x = S*(S'*(A'*u + mid(rho_y)));
        delta_w = A*delta_x - mid(rho_x);
        rho_t = Dot_(B,t,-1,mid(rho_w),1,delta_w,2);  delta_t = M'*(M*rho_t);
        t = t - delta_t;  delta_y = -t;  w = w - delta_w;  y = y - delta_y;
        [tmp,e1] = TwoSum(x2,-delta_x);  [x1,e2] = TwoSum(tmp,x1);
        x2 = e1 + e2;  % x is stored by x1 and x2

        % update of rho_x, rho_w and rho_y (enclosure)
        rho_x = Dot_(A,x1,A,x2,-1,b,-1,w,-2);  rho_w = Dot_(1,w,-B,y,-2);
        rho_y = Dot_(A',y,-2);
    end
end

setround(0)
```

```
function [x1,x2,dB2] = Fi(A,b,B)
% Fi encloses the solutions of the generalized least squares problems
% based on Theorem 2 with the iterative refinement.

setround(0);  [m,n] = size(A);  s_n = ones(n,1);  s_m = ones(m,1);
ur = 2^(-53);  uu = 2^(-1074);

L = chol(B)';  M = L\speye(m);  MA = M*intval(A);  P = qr(mid(MA),0);
R = triu(P(1:n,:));  clear P;  S = R\speye(n);  x1 = R\(R'\A'*(L'\(L\b)));
clear R;  x2 = zeros(n,1);  w = A*x1 - b;  y = L'\(L\w);

X = MA*S;  clear MA;  Xc = mid(X);  Xr = rad(X);  fl_IXX = eye(n) - Xc'*Xc;
setround(1);  gam_m1 = (m+1)*ur/(-((m+1)*ur - 1));
IXXs = abs(fl_IXX)*s_n + gam_m1*(s_n + abs(Xc)'*(abs(Xc)*s_n)) ...
       + m*n*uu*s_n ...
       + abs(Xc)'*(Xr*s_n) + Xr'*(abs(Xc)*s_n) + Xr'*(Xr*s_n);
clear fl_IXX Xc Xr;  absX = mag(X);  normX = max(sum(absX,2));
absL = abs(L);  absM = abs(M);
gam_m = m*ur/(-(m*ur - 1));  vL = diag(abs(L));  % gamma_m and v_L

% rigorous upper bounds of v_G, v_G^T, v_M, v_MG^T, v_H, v_GH, v_HG^T,
% v_GG^T, v_GHG^T, v_P, v_Q and their norms
vG = gam_m*(absM*(absL*s_m)) + (m*uu)*(m*s_m + vL)/(-(m*ur - 1));
normvG = max(vG);

if normvG >= 1  % norm(v_G,inf) < 1 could not be verified
    disp('Error: Fi (normvG >= 1)');
    x1 = NaN;  x2 = NaN;  dB2 = NaN;  return;
end

vGT = gam_m*(absL'*(absM'*s_m)) + ((uu*(m*m + vL'*s_m))*s_m)/(-(m*ur - 1));
normvGT = max(vGT);

if normvGT >= 1  % norm(v_G^T,inf) < 1 could not be verified
    disp('Error: Fi (normvGT >= 1)');
    x1 = NaN;  x2 = NaN;  dB2 = NaN;  return;
end

vM = gam_m1*(absM*(absL*(absL'*(absM'*s_m)))) ...
     + (uu*(s_m'*(absM'*s_m)))*(absM*(m*s_m + vL))/(-(m*ur- 1));
vMGT = gam_m1*(absM*(absL*(absL'*(absM'*vGT)))) ...
       + (uu*(s_m'*(absM'*vGT)))*(absM*(m*s_m + vL))/(-(m*ur- 1));
vH = vM + (max(vM)*vG)/(-(normvG - 1)) + vMGT/(-(normvGT - 1)) ...
     + (max(vMGT)*vG)/(-((-(normvG - 1))*(normvGT - 1)));
vGH = gam_m*(absM*(absL*vH)) + (uu*sum(vH))*(m*s_m + vL)/(-(m*ur - 1));
vHGT = vMGT + (max(vMGT)*vG)/(-(normvG - 1)) + (normvGT*vMGT)/(-(normvGT - 1)) ...
       + ((max(vMGT)*normvGT)*vG)/(-((-(normvG - 1))*(normvGT - 1)));
vGGT = gam_m*(absM*(absL*vGT)) + (uu*sum(vGT))*(m*s_m + vL)/(-(m*ur - 1));
vGHGT = gam_m*(absM*(absL*vHGT)) + (uu*sum(vHGT))*(m*s_m + vL)/(-(m*ur - 1));
clear absL absM;
```

```
vP = vG + vH + vGT + vGH + vHGT + vGGT + vGHGT;  normvP = max(vP);

if normvP >= 1  % norm(v_P,inf) < 1 could not be verified
    disp('Error: Fi (normvP >= 1)');
    x1 = NaN;  x2 = NaN;  dB2 = NaN;  return;
end

XvP = absX'*vP;  vQ = IXXs + (normX*(XvP))/(-(normvP - 1));  normvQ = max(vQ);

if normvQ >= 1  % norm(v_Q,inf) < 1 could not be verified
    disp('Error: Fi (normvQ >= 1)');
    x1 = NaN;  x2 = NaN;  dB2 = NaN;  return;
end

% rigorous upper bounds of abs(S)*v_Q, norm(abs(X)'*v_P,inf)
% and abs(S)*abs(X)'*v_P
SvQ = abs(S)*vQ;  normXvP = max(XvP);  SXvP = abs(S)*XvP;

rho_x = Dot_(A,x1,A,x2,-1,b,-1,w,-2);  rho_w = Dot_(1,w,-B,y,-2);
rho_y = Dot_(A',y,-2);

setround(0);
u = M'*(M*(mid(rho_x) + mid(rho_w)));  delta_x = S*(S'*(A'*u + mid(rho_y)));
delta_w = A*delta_x - mid(rho_x);  t = M'*(M*(mid(rho_w) - delta_w));
for loop = 1:10
    Mrxw = M*(rho_x + rho_w);  XMrxwSry = X'*Mrxw + S'*rho_y;
    SXMrxwSry = S*XMrxwSry;  normXMrxwSry = max(mag(XMrxwSry));
    normMrxw = max(mag(Mrxw));

    setround(1);
    dB2 = mag(SXMrxwSry) + (normXMrxwSry*SvQ)/(-(normvQ - 1)) ...
          + (normMrxw*SXvP)/(-(normvP - 1)) ...
          + ((normMrxw*normXvP)*SvQ)/(-((-(normvQ - 1))*(normvP - 1)));
    RR = dB2 ./ (-((-abs(x2) - dB2) - abs(x1)));

    if max(RR) <= 1e-11,  break;
    else
        setround(0);  rho_u = Dot_(B,u,-1,mid(rho_x),-1,mid(rho_w),2);
        delta_u = M'*(M*rho_u);  u = u - delta_u;
        delta_x = S*(S'*(A'*u + mid(rho_y)));  delta_w = A*delta_x - mid(rho_x);
        rho_t = Dot_(B,t,-1,mid(rho_w),1,delta_w,2);  delta_t = M'*(M*rho_t);
        t = t - delta_t;  delta_y = -t;  w = w - delta_w;  y = y - delta_y;
        [tmp,e1] = TwoSum(x2,-delta_x);  [x1,e2] = TwoSum(tmp,x1);
        x2 = e1 + e2;

        rho_x = Dot_(A,x1,A,x2,-1,b,-1,w,-2);  rho_w = Dot_(1,w,-B,y,-2);
        rho_y = Dot_(A',y,-2);
    end
end
setround(0)
```

References

[1] H.R. Arndt. On the interval systems $[x] = [A][x] + [b]$ and the powers of interval matrices in complex interval arithmetics. *Reliab. Comput.*, 13:245–259, 2007.

[2] D.B. Duncan and S.D. Horn. Linear dynamic recursive estimation from the viewpoint of regression analysis. *J. Amer. Statist. Assoc.*, 67:815–821, 1972.

[3] G.H. Golub and C.F. Van Loan. *Matrix Computations, third ed.* The Johns Hopkins University Press, Baltimore and London, 1996.

[4] N.J. Higham. *Accuracy and Stability of Numerical Algorithms, second ed.* SIAM Publications, Philadelphia, 2002.

[5] J. Johnston. *Econometric Methods, second ed.* McGraw-Hill, New York, 1972.

[6] S. Miyajima. Fast enclosure for the minimum norm least squares solution of the matrix equation $AXB = C$, 2013. Submitted for publication.

[7] S. Miyajima. Componentwise enclosure for solutions of least squares problems and underdetermined systems. *Linear Algebra Appl.*, 444:28–41, 2014.

[8] S. Miyajima, T. Ogita, S.M. Rump, and S. Oishi. Fast verification for all eigenpairs in symmetric positive definite generalized eigenvalue problems. *Reliab. Comput.*, 14:24–45, 2010.

[9] S. Oishi and S.M. Rump. Fast verification of solutions of matrix equations. *Numer. Math.*, 90:755–773, 2002.

[10] C.C. Paige. Computer solution and perturbation analysis of generalized linear least squares problems. *Math. Comp.*, 33:171–183, 1979.

[11] C.C. Paige. Fast numerically stable computations for generalized linear least squares problems. *SIAM J. Numer. Anal.*, 16:165–171, 1979.

[12] J. Rohn. VERSOFT: Verification software in MATLAB / INTLAB. `http://www.nsc.ru/interval/index.php?j=Programing/index`.

[13] S.M. Rump. INTLAB - INTerval LABoratory. In *Developments in Reliable Computing (T. Csendes ed.)*, pages 77–104. Kluwer Academic, Dordrecht, 1999.

[14] S.M. Rump. Verified bounds for least squares problems and underdetermined linear systems. *SIAM J. Matrix Anal. Appl.*, 33:130–148, 2012.

[15] S.M. Rump. Improved componentwise verified error bounds for least squares problems and underdetermined linear systems. *Numer. Algorithms*, 66:309–322, 2014.

[16] T. Yamamoto. Error bounds for approximate solutions of systems of equations. *Japan J. Indust. Appl. Math.*, 1:157–171, 1984.