METHODS FOR LARGE SCALE TOTAL LEAST SQUARES PROBLEMS

Å. Björck*, P. Heggernes†, and P. Matstoms*

Abstract

For solving the total least squares problems, $\min_{E,f} \|(E, f)\|_F$ subject to $(A+E)x = b+f$, where A is large and sparse or structured Björck suggested a method based on Rayleigh quotient iteration. This method reduces the problem to the solution of a sequence of symmetric, positive definite linear systems of the form $(A^TA - \bar{\sigma}^2 I)z = g$, where $\bar{\sigma}$ is an approximation to the smallest singular value of $(A,\ b)$. A preconditioned conjugate gradient method, using a sparse, possibly incomplete, Cholesky factor of A^TA can be used for solving these systems. In this paper the method is further developed. The choice of initial approximation and termination criteria are discussed. Numerical results confirm that the method achieves rapid convergence and good accuracy for problems which are not too ill-conditioned.

Key words: Total least squares, Rayleigh quotient iteration, conjugate gradient method, singular values.

1 Introduction.

The estimation of parameters in linear models is a fundamental problem in many scientific and engineering applications. A statistical model that is often realistic is to assume that the parameters x to be determined satisfy a linear relation

$$(A + E)x = b + f, \tag{1.1}$$

where $A \in \mathcal{R}^{m\times n}$, and $b \in \mathcal{R}^m$, are known and $(E,\ f)$ is an error matrix with rows which are independently and identically distributed with zero mean and the same variance. (To satisfy this assumption the data $(A,\ b)$ may need to be premultiplied by appropriate scaling matrices, see Golub and Van Loan [10].) In statistics this model is known as the "errors-in-variables model".

The estimate of the true but unknown parameter vector x in the model (1.1) is obtained from the solution of the total least squares (TLS) problem

$$\min_{E,f} \|(E,\ f)\|_F \quad \text{subject to} \quad (A + E)x = b + f, \tag{1.2}$$

*Department of Mathematics, University of Linköping, S-581 83 Linköping, Sweden. e-mail: akbjo@math.liu.se, pontus.matstoms@vti.se. The work of these authors was supported by the Swedish Research Council for Engineering Sciences, TFR.

†Department of Informatics, University of Bergen, N-5020 Bergen, Norway, email: pinar@ii.uib.no

where $\|\cdot\|_F$ denotes the Frobenius matrix norm. If a minimizing pair $(E,\ f)$ has been found for the problem (1.2) then any x satisfying $(A+E)x = b+f$ is said to solve the TLS problem.

Due to recent advances in data collection techniques LS or TLS problems where A is large and sparse (or structured) frequently arise, e.g., in signal and image processing applications. For the solution of the LS problem both direct methods based on sparse matrix factorizations and iterative methods are well developed, see [2].

An excellent treatment of theoretical and computational aspects of the TLS problem is given in Van Huffel and Vandewalle [25]. Solving the TLS problem requires the computation of the smallest singular value and the corresponding right singular vector of $(A,\ b)$. When A is large and sparse this is a much more difficult problem than that of computing the LS solution. For example, it is usually not feasible to compute the SVD or any other two-sided orthogonal factorization of A since the factors typically are not sparse.

Iterative algorithms for computing the singular subspace of a matrix associated with its smallest singular values, with applications to TLS problems with slowly varying data, have previously been studied by Van Huffel [24]. In [27, 3] a new class of methods based on a Rayleigh quotient iteration was developed for the efficient solution of large scale TLS problems. Related methods for Toeplitz systems were studied by Kamm and Nagy [14]. In this paper the methods in [3] are further developed and numerical results given. Similar algorithms for solving large scale multidimensional TLS problems will be considered in a forthcoming paper [4].

In Section 2 we recall how the solution to the TLS problem can be expressed in terms of the smallest singular value and corresponding right singular vector of the compound matrix $(A,\ b)$. We discuss the conditioning of the LS and TLS problems and illustrate how the TLS problem can rapidly become intractable. Section 3 first reviews a Newton iteration for solving a secular equation. For this method to converge to the TLS solution strict conditions on the initial approximation have to be satisfied. We then derive the Rayleigh quotient method, which ultimately achieves cubic convergence. The choice of initial estimates and termination criteria are discussed. A preconditioned conjugate gradient method is developed in Section 4 for the efficient solution of the resulting sequence of sparse symmetric linear systems. Finally, in Section 5, numerical results are given which confirm the rapid convergence and numerical stability of this class of methods.

2 Preliminaries.

2.1 The TLS problem.

The TLS problem (1.2) is equivalent to finding a perturbation matrix $(E,\ f)$ having minimal Frobenius norm, which lowers the rank of the matrix (A, b).

Hence it can be analyzed in terms of the singular value decomposition

$$(A,\ b) = U\Sigma V^T, \qquad \Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_{n+1}),$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{n+1} \geq 0$, are the singular values of $(A,\ b)$. Note that by the minmax characterization of singular values it follows that the singular values ${\sigma'}_i$ of A interlace those of $(A,\ b)$, i.e.,

$$\sigma_1 \geq {\sigma'}_1 \geq \sigma_2 > \dots \geq \sigma_n \geq {\sigma'}_n \geq \sigma_{n+1}. \tag{2.1}$$

We assume in the following that A has full rank, that is, ${\sigma'}_n > 0$, and that $\sigma_n > \sigma_{n+1}$. Then the minimum is attained for the rank one perturbation

$$(E,\ f) = -(A,\ b)v_{n+1}v_{n+1}^T = -\sigma_{n+1}u_{n+1}v_{n+1}^T,$$

for which $\|(E,\ f)\|_F = \sigma_{n+1}$. A TLS solution is then obtained from the right singular vector

$$v_{n+1} = \begin{pmatrix} z \\ \zeta \end{pmatrix} = -\zeta \begin{pmatrix} x_{TLS} \\ -1 \end{pmatrix}, \tag{2.2}$$

provided that $\zeta \neq 0$. If $\zeta = 0$ the TLS problem is called *nongeneric*, and there is no solution. This case cannot occur if ${\sigma'}_n > \sigma_{n+1}$, and in the following we always assume that this condition holds.

From the characterization (2.2) it follows that $\lambda = \sigma_{n+1}^2$ and $x = x_{TLS}$ satisfy the system of nonlinear equations

$$\begin{pmatrix} A^TA & A^Tb \\ b^TA & b^Tb \end{pmatrix} \begin{pmatrix} x \\ -1 \end{pmatrix} = \lambda \begin{pmatrix} x \\ -1 \end{pmatrix}. \tag{2.3}$$

Putting $\lambda = \sigma_{n+1}^2$ the first block row of this system of equations can be written

$$(A^TA - \sigma_{n+1}^2 I)x = A^Tb, \tag{2.4}$$

which can be viewed as "the normal equations" for the TLS problem. Note that from our assumption that ${\sigma'}_n > \sigma_{n+1}$ it follows that $A^TA - \sigma_{n+1}^2 I$ is positive definite.

2.2 Conditioning of the TLS problem.

For the evaluation of accuracy and stability of the algorithms to be presented we need to know the sensitivity of the TLS problem to perturbations in data. We first recall that if $x_{LS} \neq 0$ the condition number for the LS problem is (see [2, Sec. 1.4])

$$\kappa_{LS}(A, b) = \kappa(A)\left(1 + \frac{\|r_{LS}\|_2}{{\sigma'}_n\|x_{LS}\|_2}\right). \tag{2.5}$$

where $\kappa(A) = {\sigma'}_1/{\sigma'}_n$. Note that the condition number depends on both A and b, and that for large residual problems the second term may dominate.

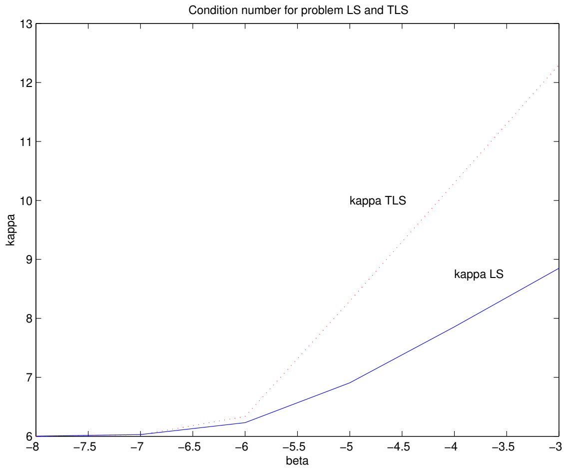

Figure 2.1: Condition numbers κ_{LS} and κ_{TLS} as function of $|\beta| = \|r_{LS}\|_2$.

Equation (2.4) shows that the TLS problem is always worse conditioned than the LS problem. From (2.3), multiplying from the left with $(x^T \ -1)$ we get

$$\|r_{TLS}\|_2^2 = \sigma_{n+1}^2(\|x_{TLS}\|_2^2 + 1), \qquad r_{TLS} = b - Ax_{TLS}.$$

Since $\|r_{LS}\|_2 \leq \|r_{TLS}\|_2$ and $\sigma_{n+1} \leq \sigma'_n$ it follows that

$$\|x_{TLS}\|_2^2 \geq (\|r_{LS}\|_2/\sigma'_n)^2 - 1. \tag{2.6}$$

This inequality is wek, but shows that $\|x_{TLS}\|_2$ will be large when $\|r_{LS}\|_2 \gg \sigma'_n$.

Golub and Van Loan [10] showed that an approximate condition number for the TLS problem is

$$\kappa_{TLS}(A, b) = \frac{\sigma'_1}{\sigma'_n - \sigma_{n+1}} = \kappa(A)\frac{\sigma'_n}{\sigma'_n - \sigma_{n+1}}. \tag{2.7}$$

When $1 - \sigma_{n+1}/\sigma'_n \ll 1$ the TLS condition number can be much greater than $\kappa(A)$. The relation between the two condition numbers (2.5) and (2.7) depend on the relation between the $\|r_{LS}\|_2$ and σ_{n+1}, which is quite intricate. (For a study of this relation in another context see Paige and Strakoš [17].)

As an illustration we consider the following small overdetermined system

$$\begin{pmatrix} \sigma'_1 & 0 \\ 0 & \sigma'_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \beta \end{pmatrix}. \tag{2.8}$$

Trivially, the LS solution is

$$x_{LS} = (c_1/\sigma'_1, c_2/\sigma'_2)^T, \qquad \|r_{LS}\|_2 = |\beta|.$$

If we take in (2.8) $\sigma'_1 = c_1 = 1$, $\sigma'_2 = c_2 = 10^{-6}$, then $x_{LS} = (1,\ 1)^T$ independent of β, and hence does not reflect the illconditioning of A. The TLS solution is of similar size as the LS solution as long as $|\beta| \leq \sigma'_2$. However, when $|\beta| \gg \sigma'_2$ then from (2.6) it follows that $\|x_{TLS}\|_2$ is large.

In Fig. 2.1 the two condition numbers are plotted as a function of $|\beta|$. We note that κ_{LS} increases proportionally to $|\beta|$ because of the second term in (2.5). For $|\beta| > \sigma'_2$ the condition number κ_{TLS} grows proportionally to $|\beta|^2$. It can be verified that $\|x_{TLS}\|_2$ also grows proportionally to $|\beta|^2$.

3 Newton and Rayleigh Quotient methods.

3.1 A Newton method.

Equation (2.3) constitutes a system of $(n+1)$ nonlinear equations in x and λ. One way to proceed (see [14]) is to eliminate x to obtain the rational secular equation for $\lambda = \sigma_{n+1}^2$:

$$g(\lambda) = -b^T(b - Ax(\lambda)) + \lambda = 0, \tag{3.1}$$

where $x(\lambda) = (A^TA - \lambda I)^{-1}A^Tb$. Newton's method applied to (3.1) leads to the iteration

$$\lambda^{(k+1)} = \lambda^{(k)} + \frac{b^T(b - Ax^{(k)}) - \lambda^{(k)}}{1 + \|x^{(k)}\|_2^2}, \tag{3.2}$$

$$x^{(k)} = (A^TA - \lambda^{(k)}I)^{-1}A^Tb. \tag{3.3}$$

This iteration will converge monotonically at a rate that is asymptotically quadratic. The convergence of this method can be improved by using a rational interpolation similar to that in [6] to solve the secular equation. However, in any case, λ will converge to σ_{n+1}^2 and $x^{(k)}$ to the TLS solution only if the initial approximation satisfies

$$\lambda^{(0)} \in (\sigma_{n+1}^2, {\sigma'}_n^2) \tag{3.4}$$

In general it is hard to verify this assumption. For the special case of a Toeplitz TLS problem Kamm and Nagy [14] use a bisection algorithm based on a fast algorithm for factorizing Toeplitz matrices to find an initial starting value satisfying (3.4).

3.2 The Rayleigh quotient method.

The main drawback of the Newton method above is that unless (3.4) is satisfied it will converge to the wrong singular value. A different Newton method is obtained by applying Newton's method to the full system

$$\begin{pmatrix} f(x,\lambda) \\ g(x,\lambda) \end{pmatrix} = \begin{pmatrix} -A^Tr - \lambda x \\ -b^Tr + \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{3.5}$$

where $r = b - Ax$. As remarked in [20] this is closely related to inverse iteration, which is one of the most widely used methods for refining eigenvalues and eigenvectors. Rayleigh quotient iteration (RQI) is inverse iteration with a shift equal to the Rayleigh quotient. RQI has *cubic convergence* for the symmetric eigenvalue problem, see [18, Sec.4–7], and is superior to the standard Newton method applied to (3.5).

For the eigenvalue problem (2.3) the Rayleigh quotient equals

$$\rho(x) = \frac{(x^T A^T - b^T)(Ax - b)}{x^T x + 1} = \frac{r^T r}{x^T x + 1}. \tag{3.6}$$

Let $x^{(k)}$ be the current approximation and ρ_k the corresponding Rayleigh quotient. Then the next approximation $x^{(k+1)}$ in RQI and the scaling factor β_k are obtained from the symmetric linear system

$$\begin{pmatrix} J^{(k)} & A^T b \\ b^T A & \eta_k \end{pmatrix} \begin{pmatrix} x^{(k+1)} \\ -1 \end{pmatrix} = \beta_k \begin{pmatrix} x^{(k)} \\ -1 \end{pmatrix}, \tag{3.7}$$

where

$$J^{(k)} = A^T A - \rho_k I, \quad \eta_k = b^T b - \rho_k.$$

If $J^{(k)}$ is positive definite the solution can be obtained by block Gaussian elimination,

$$\begin{pmatrix} J^{(k)} & A^T b \\ 0 & \tau_k \end{pmatrix} \begin{pmatrix} x^{(k+1)} \\ -1 \end{pmatrix} = \beta_k \begin{pmatrix} x^{(k)} \\ -(z^{(k)})^T x^{(k)} - 1 \end{pmatrix}, \tag{3.8}$$

where

$$J^{(k)} z^{(k)} = A^T b, \qquad \tau_k = b^T (b - A z^{(k)}) - \rho_k. \tag{3.9}$$

It follows that $x^{(k+1)} = z^{(k)} + u^{(k)}$, where

$$J^{(k)} u^{(k)} = \beta_k x^{(k)}, \qquad \beta_k = \tau_k / ((z^{(k)})^T x^{(k)} + 1). \tag{3.10}$$

In [2] a reformulation was made to express the solution in terms of the residual vectors of (3.5)

$$\begin{pmatrix} f^{(k)} \\ g^{(k)} \end{pmatrix} = \begin{pmatrix} -A^T r^{(k)} - \rho_k x^{(k)} \\ -b^T r^{(k)} + \rho_k \end{pmatrix}, \tag{3.11}$$

where $r^{(k)} = b - Ax^{(k)}$. This uses the following formulas to compute τ_k:

$$J^{(k)} w^{(k)} = -f^{(k)}, \qquad z^{(k)} = x^{(k)} + w^{(k)}, \tag{3.12}$$

$$\tau_k = (z^{(k)})^T f^{(k)} - g^{(k)}. \tag{3.13}$$

The RQI iteration is defined by equations (3.10)–(3.13).

3.3 Initial estimate and global convergence.

Parlett and Kahan [19] have shown that for almost all initial vectors the Rayleigh quotient iteration converges to some singular value and vector pair. However, in general we cannot say to *which* singular vector RQI will converge.

If the LS solution is known, a suitable starting approximation for λ may be

$$\rho(x_{LS}) = \frac{\|r_{LS}\|^2}{\|x_{LS}\|^2 + 1} \tag{3.14}$$

Conditions to ensure that RQI will converge to the TLS solution from the starting approximation $(\rho(x_{LS}), x_{LS})$ are in general difficult to verify and often not satisfied in practice. However, in contrast to the simple Newton iteration in Section 3.1, the method may converge to the TLS solution even when $\rho(x_{LS}) \notin (\sigma_{n+1}^2, {\sigma'}_n^2)$.

The Rayleigh quotient $\rho(x_{LS})$ will be a large overestimate of σ_{n+1}^2 when the residual norm $\|r_{LS}\|_2$ is large and $\|x_{LS}\|_2$ does not reflect the illconditioning of A. Note that it is typical for illconditioned least squares problems that the right-hand side is such that $\|x_{LS}\|_2$ is not large! For example, least squares problems arising from ill-posed problems usually satisfy a so called Picard condition, which guarantees that the right-hand side has this property, see [11, Sec. 1.2.3].

Szyld [23] suggested that one or more steps of inverse iteration could be applied initially before switching to RQI, in order to ensure convergence to the smallest eigenvalue. Inverse iteration for σ_{n+1}^2 corresponds to taking $\sigma^2 = 0$ in the RQI algorithm. Starting from $x = x_{LS}$ the first step of inverse iteration simplifies as follows. Using (3.9) and (3.10) with $\rho_k = 0$ and $x^{(k)} = x_{LS}$ we get

$$z^{(k)} = x_{LS}, \qquad \tau_k = \|r_{LS}\|_2^2,$$

and the new approximation becomes

$$x_{INV} = x_{LS} + \beta(A^T A)^{-1} x_{LS}, \qquad \beta = \rho(x_{LS}).$$

Several steps of inverse iteration may be needed to ensure convergence of RQI to the smallest singular value. However, since inverse iteration only converges linearly, taking more than one step will usually just hold up the rapid convergence of RQI. We therefore recommend in general $p = 1$ steps as the default value.

To illustrate the situation consider again the small 3×2 system (2.8) with ${\sigma'}_1 = c_1 = 1$, ${\sigma'}_2 = c_2 = 10^{-6}$. This has the LS solution $x_1 = x_2 = 1$, which does not reflect the illconditioning of A ($\kappa = 10^6$). With $\|r_{LS}\|_2 = \beta$ the initial Rayleigh quotient approximation equals

$$\rho(x_{LS}) = \beta^2/(1+2) = \beta^2/3.$$

By the interlacing property we have that $\sigma_3 \le {\sigma'}_2$. Since $|\beta| \gg {\sigma'}_2$ it is clear that the Rayleigh quotient fails to approximate σ_3^2. This is illustrated in Figure 3.1, where $\rho(x_{LS})^{1/2}$ and σ_3 are plotted as function of $|\beta|$. It is easily verified, however, that after one step of inverse iteration $\rho(x_{INV})$ will be close to ${\sigma'}_2^2$.

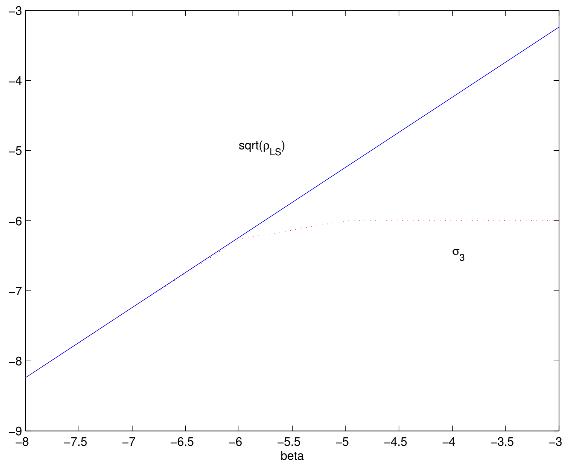

Figure 3.1: Rayleigh quotient approximation and σ_3 for $|\beta| = \|r_{LS}\|_2 = 10^{-k}$.

3.4 Termination criteria for RQI.

The RQI algorithm for the TLS problem is defined by (3.10)–(3.13). When should the RQI iteration be terminated? We suggest two different criteria.

The first is based on the key fact in the proof of global convergence that the normalized residual norm

$$\gamma_k = \left(\frac{\|f_k\|_2^2 + g_k^2}{\|x^{(k)}\|_2^2 + 1} \right)^{1/2}, \qquad \begin{pmatrix} f_k \\ g_k \end{pmatrix} = \begin{pmatrix} -A^T r^{(k)} - \rho_k x^{(k)} \\ -b^T r^{(k)} + \rho_k \end{pmatrix} \tag{3.15}$$

always decreases, $\gamma_{k+1} \le \gamma_k$, for all k. Thus, if an increase in the norm occurs this must be caused by roundoff, and then it makes no sense to continue the iterations. This suggests that we terminate the iterations with x_{k+1} when

$$\gamma_{k+1} > \gamma_k. \tag{3.16}$$

A second criterion is based on the observation that since the condition number for computing σ_{n+1} equals 1, we can expect to obtain σ_{n+1} to full machine precision. Since convergence of RQI is cubic a criterion could be to stop when the change in the approximation to σ_{n+1} is of the order of $\sigma_1 u^{1/p}$, where $p = 3$. (A similar criterion with $p = 2$ is used by Kamm and Nagy [14] for terminating the Newton iteration.) However, as will be evident from the numerical results in Section 5, full accuracy in x_{TLS} in general requires one more iteration after σ_{n+1} has converged. Therefore we recommend to stop when either (3.16) or

$$|\rho(x_{k+1}) - \rho(x_k)| \le Cu, \tag{3.17}$$

is satisfies, where u is the machine unit and C a suitable constant.

We summarize below the RQI algorithm with one step of inverse iteration (cf. [3]):

Algorithm 3.1. **Rayleigh Quotient Iteration.**

$x = x_{LS}$;
$r = b - Ax$;
$\sigma^2 = r^T r/(1 + x^T x)$;
solve $A^T A u = x$;
$x = x + \sigma^2 u$;
for $k = 1, 2, \ldots$
  $r = b - Ax$;
  $\sigma^2 = r^T r/(1 + x^T x)$;
  $f = -A^T r - \sigma^2 x$;
  $g = -b^T r + \sigma^2$;
  solve $(A^T A - \sigma^2 I)w = -f$;
  $z = x + w$;
  $\beta = (z^T f - g)/(z^T x + 1)$;
  solve $(A^T A - \sigma^2 I)u = x$;
  $x = z + \beta u$;
end

3.5 Rounding errors and stability.

If the RQI iteration converges then $f^{(k)}$, $g^{(k)}$, and β_k will tend to zero. Consider the rounding errors which occur in the evaluation of the residuals (3.11). Let $\tilde{u} = 1.06u$, where u is the unit roundoff; see [13, Chap. 3]. Then the computed residual vector satisfies $\bar{r} = r + \delta r$, where

$$\|\delta r\|_2 \leq n\tilde{u}(\|b\|_2 + \|A\|_2\|x\|_2).$$

Obviously convergence will cease when the residuals (3.11) are dominated by roundoff. Assume that we perform one iteration from the exact solution, x_{TLS}, r_{TLS}, and $\lambda = \sigma_{n+1}^2$. Then the first correction to the current approximation is obtained by solving the linear system in (3.13), which now becomes

$$(A^T A - \sigma_{n+1}^2 I)w^{(k)} = -A^T \delta r^{(k)}. \tag{3.18}$$

For the correction this gives the estimate

$$\|w^{(k)}\|_2 = \frac{n\tilde{u}\sigma'_n}{\sigma'^2_n - \sigma_{n+1}^2}(\|b\|_2 + \|A\|_2\|x_{TLS}\|_2). \tag{3.19}$$

This estimate is consistent with the condition estimate for the TLS problem.

We note that the equations (3.18) are of similar form to those that appear in the corrected semi-normal equations for the LS problem; see [1], [2, Sec. 6.6.5]. A detailed roundoff error analysis similar to that done for the LS problem would become very complex and is not attempted here. It seems reasonable to conjecture that if if ${\sigma'}_n^2 - \sigma_{n+1}^2 < u^{1/2}$ it will suffice to solve the linear equations for the correction $w^{(k)}$ using the Cholesky factorization of $(A^TA - \sigma_{n+1}^2 I)$. Methods for the solution of the linear systems are considered in more detail in Section 4.

4 Solving the linear systems.

In the RQI method formulated in the previous section the main work consists of solving in each step two linear systems of the form

$$(A^TA - \sigma^2 I)w = f, \qquad \sigma \approx \sigma_{n+1}. \tag{4.1}$$

Here σ is an approximation to σ_{n+1} and varies from step to step. Provided that $\sigma < {\sigma'}_n$, the system (4.1) is symmetric and positive definite.

4.1 Direct linear solvers.

If $\sigma < {\sigma'}_n$ then the system (4.1) can be solved by computing the (sparse) Cholesky factorization of the matrix $A^TA - \sigma^2 I$. Note that A^TA only has to be formed once and the symbolic phase of the factorization does not have to be repeated. However, it is a big disadvantage that a new numerical factorization has to be computed at each step of the RQI algorithm.

For greater accuracy and stability in solving LS problems it is often preferred to use a QR factorization instead of a Cholesky factorization. However, since in the TLS normal equations the term $\sigma^2 I$ is *subtracted* from A^TA, this is not straightforward. The Cholesky factor of the matrix $A^TA - \sigma^2 I$ can be obtained from the QR factorization of the matrix $\begin{pmatrix} A \\ i\sigma I \end{pmatrix}$, where i is the imaginary unit. This is a downdating problem for the QR factorization and can be performed using stabilized hyperbolic rotations, see [2, pp. 143–144], or hyperbolic Householder transformations, see [22]. However, in the sparse case this is not an attractive alternative, since it would require nontrivial modifications of existing software for sparse QR factorization.

4.2 Iterated deregularization.

To solve the TLS normal equations using only a single factorization of A^TA we can adapt an iterated regularization scheme due to Riley and analyzed by Golub [9]. In this scheme, we solve the TLS normal equations by the iteration $x^{(0)} = 0$, and for $k = 0, 1, \ldots$

$$\begin{aligned} r^{(k)} &= b - Ax^{(k)}, \\ A^TA\delta^{(k)} &= A^Tr^{(k)} + \sigma^2 x^{(k)}, \\ x^{(k+1)} &= x^{(k)} + \delta^{(k)}. \end{aligned}$$

If $\lim_{k\to\infty} x^{(k)} = x$ then $(A^TA - \sigma^2 I)x = A^Tb$. This iteration will converge with linear rate equal to $\rho = \sigma^2/{\sigma'}_n^2$ provided that $\rho < 1$. This iteration may be implemented very efficiently if the QR decomposition of A is available. We do not pursue this method further, since it has no advantage over the preconditioned conjugate gradient method developed in [3].

4.3 A preconditioned conjugate gradient algorithm.

Performing the change of variables $y = Sw$, where S is a given nonsingular matrix, and multiplying from the left with S^{-T} the system (4.1) becomes

$$(S^{-T}A^TAS^{-1} - \sigma^2 S^{-T}S^{-1})y = S^{-T}f, \tag{4.2}$$

This system is symmetric positive definite provided that $\sigma < {\sigma'}_n$, and hence the conjugate gradient method can be applied. We can use for S the same preconditioners as have been developed for the LS problem; for a survey see [2, Ch. 7].

In the following we consider a special choice of preconditioner, the *complete* Cholesky factor R of A^TA (or R from a QR decomposition of A). Unless A is huge this is often a feasible choice, since efficient software for sparse Cholesky and sparse QR factorization are readily available [2, Ch. 7]. Using $AR^{-1} = Q_1$, where $Q_1^TQ_1 = I$, the preconditioned system (4.2) simplifies to

$$(I - \sigma^2 R^{-T}R^{-1})y = R^{-T}f, \qquad w = R^{-1}y. \tag{4.3}$$

(Note that although A and A^T have disappeared from this system of equations matrix-vector multiplications with these matrices are used to compute the right-hand side f!) In the inverse iteration step used in the initialization, $\sigma = 0$, and the solution $w = R^{-1}R^{-T}f$ is obtained by two triangular solves.

The standard conjugate gradient method applied to the system (4.2) can be formulated in terms of the original variables w. The resulting algorithm is a slightly simplified version of the algorithm PCGTLS given in [3] and can be written:

ALGORITHM 4.1. **PCGTLS**

Preconditioned gradient method for solving $(A^TA - \sigma^2 I)w = f$, using the Cholesky factor R of A^TA as preconditioner.

Initialize: $w^{(0)} = 0, \quad p^{(0)} = s^{(0)} = R^{-T}f, \quad \eta_0 = \|s^{(0)}\|_2^2$.

For $j = 0, 1, \ldots, l$, while $\delta_j \neq 0$ compute

$$\begin{aligned}
&q^{(j)} = R^{-1}p^{(j)}\\
&\delta_j = \|p^{(j)}\|_2^2 - \sigma^2\|q^{(j)}\|_2^2\\
&\alpha_j = \eta_j/\delta_j\\
&w^{(j+1)} = w^{(j)} + \alpha_j q^{(j)}\\
&q^{(j)} = R^{-T}q^{(j)}
\end{aligned}$$

$$
\begin{aligned}
&s^{(j+1)} = s^{(j)} - \alpha_j (p^{(j)} - \sigma^2 q^{(j)}) \\
&\eta_{j+1} = \|s^{(j+1)}\|_2^2 \\
&\beta_j = \eta_{j+1}/\eta_j \\
&p^{(j+1)} = s^{(j+1)} + \beta_j p^{(j)}
\end{aligned}
$$

Denote the original and the preconditioned matrix by $C = A^T A - \sigma^2 I$ and $\widetilde{C} = I - \sigma^2 R^{-T} R^{-1}$, respectively. Then a simple calculation shows that for $\sigma = \sigma_{n+1}$ the condition number of the transformed system is reduced by a factor of $\kappa(A)$,

$$\kappa(\widetilde{C}) = \left(\frac{(\sigma'_1)^2 - \sigma_{n+1}^2}{(\sigma'_n)^2 - \sigma_{n+1}^2}\right)\left(\frac{(\sigma'_n)^2}{(\sigma'_1)^2}\right) = \frac{\kappa(C)}{\kappa^2(A)}.$$

The spectrum of $\tilde{C}$ will be clustered close to 1. In particular in the limit when $\sigma \to \sigma_{n+1}$, the eigenvalues of $\widetilde{C}$ will lie in the interval

$$\left[1 - \sigma_{n+1}^2/(\sigma'_n)^2, 1\right]. \tag{4.4}$$

(Note the relation to the condition number κ_{TLS}!) Hence, unless $\sigma'_n \approx \sigma_{n+1}$, we can expect this choice of preconditioner to work very well for solving the shifted system (4.1).

The matrix $R^T R - \sigma^2 I$ is positive definite if $\sigma < \sigma'_n$. In this case $\delta_k > 0$ in PCGTLS, and the division in computing α_k can always be carried out. If $\sigma \geq \sigma'_n$ then the system (4.2) is not positive definite and a division by zero can occur. This can be avoided by including a test to ensure that $\delta_k > 0$. If $\delta_k < 0$, or equivalently $\|p^{(k)}\|_2 < \sigma \|q^{(k)}\|_2$, the CG iterations are considered to have failed. The RQI step is then repeated with a new smaller value of σ_{n+1}^2, e.g.,

$$\sigma^2 = \frac{1}{2}\|p^{(k)}\|_2^2 / \|q^{(k)}\|_2^2. \tag{4.5}$$

The accuracy of TLS solutions computed by Rayleigh Quotient Iteration will basically depend on the accuracy residuals and the stability of the method used to solve the linear systems (4.1). We note that the cg method CGLS1 for the LS problem, which is related to PCGTLS, has been shown to have very good numerical stability properties, see [5].

4.4 Termination criteria in PCGTLS.

The RQI iteration, using PCGTLS as an inner iteration for solving the linear systems, is an inexact Newton method for solving a system of nonlinear equations. Such methods have been studied by Dembo, Eisenstat, and Steihaug [7], who consider the problem of how to terminate the iterative solver so that the rate of convergence of the outer Newton method is preserved.

Consider the iteration

$$F'(x_k)s_k = -F(x_k) + r_k, \quad k = 0, 1, \dots,$$

where r_k is the residual error. In [7] it is shown that maintaining a convergence order of $1 + p$ requires that when $k \to \infty$, the residuals satisfy inequalities

$$\|r_k\| \le \eta_k \|F(x_k)\|, \qquad \eta_k = O(\|F(x_k)\|^p), \tag{4.6}$$

where η_k is a forcing sequence.

In practice the above asymptotic result turns out to be of little practical use in our context. Once the asymptotic cubic convergence is realized, the ultimate accuracy possible in double precision already has been achieved. A more practical, ad hoc termination criterion for the PCGTLS iterations will be described together with the numerical results reported below.

REMARK. In the second linear system to be solved in RQI, $(A^T A - \sigma^2 I)u = x$, the right-hand side converges to x_{TLS}. Hence it is tempting to use the value of u obtained from the last RQI to initialize PCGTLS in the next step. However, our experience is that this slows down the convergence compared to initializing u to zero.

5 Numerical results.

5.1 *Accuracy and termination criteria.*

Numerical tests were performed in MATLAB on a SUN SPARC station 10 using double precision with unit roundoff $u = 2.2 \cdot 10^{-16}$. For the initial testing we used contrived test problems $[A, b] = P(m, n, \epsilon)$, similar to those in [5] and generated in the following way.[1] Let

$$\widetilde{A} = Y \begin{pmatrix} D \\ 0 \end{pmatrix} Z^T \in \mathcal{R}^{m \times n},$$

where Y, Z are random orthogonal matrices and $D = \mathrm{diag}(1, 2^{-1}, \dots, 2^{-n+1})$. Further, let

$$x = (1, 1/2, \dots, 1/n), \qquad \widetilde{b} = \widetilde{A}x.$$

This ensures that the norm of the solution does not reflect the illconditioning of A. We then add random perturbations

$$\begin{aligned} A &= \widetilde{A} + E, & b &= \widetilde{b} + r, \\ E &= \epsilon * \mathrm{rand}(m, n), & r &= \epsilon * \mathrm{rand}(m, 1). \end{aligned}$$

Note that since $\sigma' = 2^{-n+1}$ there is a perturbation E to A with $\|E\|_2 = 2^{n-1}$ which makes A rank deficient. Therefore it is not realistic to consider perturbations with $m\epsilon \ge 2^{-n+1}$. To test the termination criteria for the inner iterations

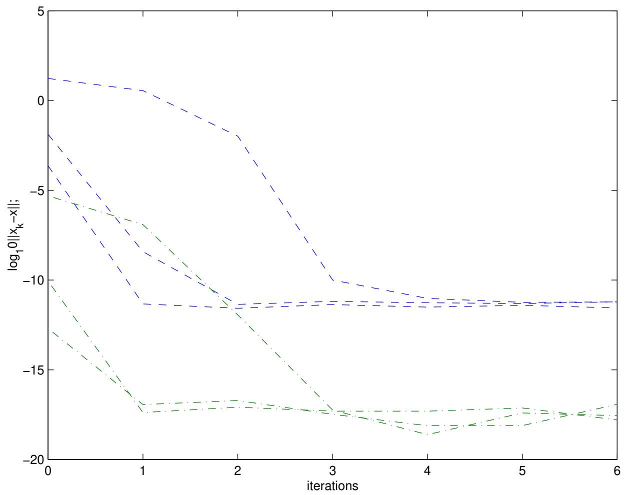

Figure 5.1: Errors $\|x_k - x\|$ and $|\sigma_k - \sigma_{n+1}|$ for problem PS(30,15), with $\epsilon = 10^{-6}$ ($\hat{\sigma}_n = 2^{-15}$). Linear systems solved by PCGTLS with $k, k+1$ and $k+2$ iterations.

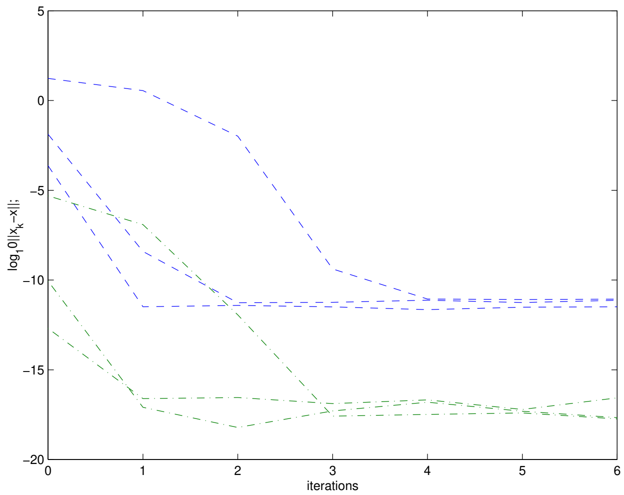

Figure 5.2: Errors $\|x_k - x\|$ and $|\sigma_k - \sigma_{n+1}|$ for problem PS(30,15), $\epsilon = 10^{-8}, 10^{-7}, 10^{-6}$, $\hat{\sigma}_n = 2^{-15}$. Linear systems solved by PCGTLS with $k+1$ iterations.

we used problem $P(30,15)$, $\sigma'_n = 2^{-15} = 3.05 \cdot 10^{-5}$, with error level $\epsilon = 10^{-6}$. The linear systems arising in RQI were solved using PCGTLS with the Cholesky factor of A^TA as preconditioning. The criterion (4.6) shows that the linear systems should be solved more and more accurately as the RQI method converges. The rate of convergence depends on the ratio σ_{n+1}/σ'_n, see (4.4), and is usually very rapid. We have used a very simple strategy where in the kth step of RQI

[1] These test problems are neither large nor sparse!

$k + \nu$ PCGTLS iterations are performed, where $\nu \geq 0$ is a parameter to be chosen.

In Figure 5.1 we show results for $\nu = 0, 1, 2$. The plots for $\nu = 1$ and $\nu = 2$ are almost indistinguishable, whereas $\nu = 0$ gives a slight delay in convergence. Indeed, for this problem taking $k + 1$ iterations in PCGTLS suffices to give the same result as using an exact (direct) solver. Since no refactorizations are performed the object should be to minimize the total number of PCGTLS iterations. Based on these considerations and the test results we recommend taking $\nu = 1$, although $\nu = 0$ should work well for problems where the ratio σ_{n+1}/σ'_n is smaller.

Rarely more than 2–3 RQI iterations will be needed. In Figure 5.2 we show results for problem PS(30,15), and different error levels $\epsilon = 10^{-8}, 10^{-7}, 10^{-6}$. Here $1, 2$, and 3–4 RQI iterations, respectively, were needed to achieve an accuracy of about 10^{-11} in x_{TLS}. Since $\sigma'_n = 2^{-15} = 3.05 \cdot 10^{-5}$, this is equal to the best limiting accuracy that can be expected. Note also that the error in σ_{n+1} converges to machine precision, usually in one less iteration, which supports the use of the criterion (3.17) to terminate RQI.

5.2 Improvement from inverse iteration.

We now show the improvement resulting from including an initial step of inverse iteration. In Figure 5.3 we show results for the problem considered above. For the first two error levels only one RQI iteration now suffices. For the highest error level σ_{n+1} converges in two iterations and x_{TLS} in three.

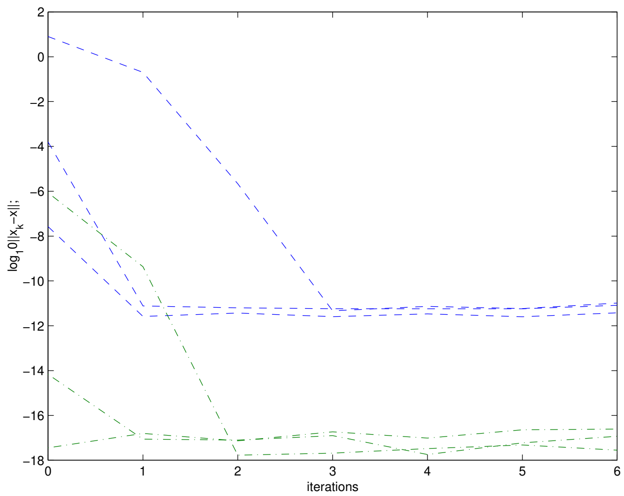

Figure 5.3: Errors $\|x_k - x\|$ and $|\sigma_k - \sigma_{n+1}|$ for problem PS(30,15), $\epsilon = 10^{-6}, 10^{-5}, 10^{-4}$, $\hat{\sigma}_n = 2^{-10}$. One step of inverse iteration. Linear systems solved by PCGTLS with $k+1$ iteration.

We now consider the second test problem in [14], which is defined as

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ . & & & & & & & \\ . & & & & & & & \\ . & & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 2 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ . \\ . \\ . \\ n-1 \end{bmatrix} + e = g + e,$$

where $A \in \mathbf{R}^{n \times n-1}$. Here e is a vector with entries generated randomly from a normal distribution with mean 0.0 and variance 1.0, and scaled so that $||e||_2 = \eta||g||_2$. For $n = 100$ we have $\kappa(A) = 2.62 \cdot 10^3$ and $\eta = 0.01$ the condition numbers in (2.5)–(2.7) are

$$\kappa_{LS} = 3.98 \cdot 10^5, \qquad \kappa_{TLS} = 1.25 \cdot 10^8,$$

respectively. This problem has features similar to those of the small illconditioned example discussed previously in Section 2.2, although here the norm of the solution x_{LS} is large.

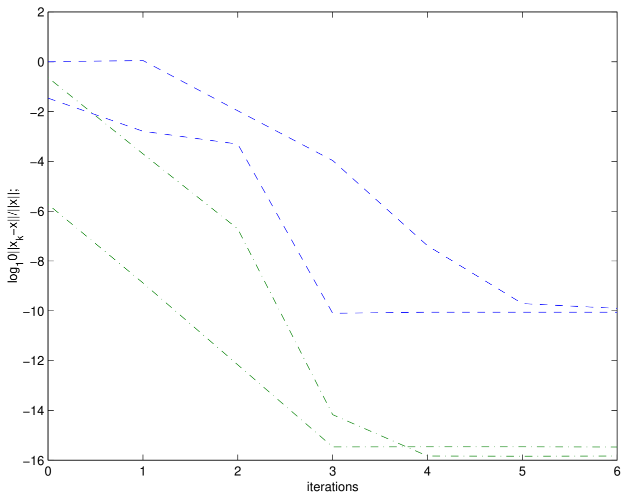

Figure 5.4: Second test problem with $\eta = 0.001$. RQI without/with one step of inverse iteration

Applying the RQI algorithm we obtained the results shown in Figure 5.4. The initial approximation $\rho(x_{LS})$ is here far outside the interval $[\sigma_{n+1}, \sigma'_n)$. Thus the matrix $A^TA - \sigma^2 I$ is initially not positive definite and we cannot guarantee the existence of the Cholesky factor. However, the Algorithm PCGTLS still does not break down, and as shown in Figure 5.4 the limiting accuracy is obtained after five RQI iterations. This surprisingly good performance of RQI can be explained by the fact that even though x_{LS} does not approximate x_{TLS} well, the angle between them is small; the cosine equals 0.98453.

Performing one step of inverse iteration before applying the RQI algorithm gives much improved convergence. The one initial step of inverse iteration here suffices to give an initial approximation in the interval $[\sigma_{n+1}, \sigma'_n)$. This can be compared with 12–23 steps of bisection needed to achieve such a starting approximation, see [14]! Three RQI iterations now give the solution x_{TLS} with an error close to the limiting accuracy, see Fig. 5.4.

We note that in both cases we obtained σ_{n+1} to full machine precision. Also, the relative error norm of in the TLS solution was consistent with the condition number.

5.3 A problem in signal restoration.

The Toeplitz matrix used in this example comes from an application in signal restoration, see [14, Example 3]. Specifically, an $n \times (n - 2\omega)$ convolution matrix $\bar{T}$ is constructed to have entries in the first column given by

$$t_{i,1} = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left[\frac{-(\omega - i + 1)^2}{2\alpha^2}\right] \qquad i = 1, 2, \ldots, 2\omega + 1,$$

and zero otherwise. Entries in the first row given by $t_{1,j} = t_{1,1}$ if $j = 1$, and zero otherwise, where $\alpha = 1.25$ and $\omega = 8$. A Toeplitz matrix T and right-hand side vector g is then constructed as $T = \bar{T} + E$ and $g = \bar{g} + e$, where E is a random Toeplitz matrix with the same structure as T, and e is a random vector. The entries in E and e are generated randomly from a normal distribution with mean 0.0 and variance 1.0, and scaled so that

$$\|e\|_2 = \eta \|\bar{g}\|_2, \qquad \|E\|_2 = \eta \|\bar{T}\|_2$$

In [14] problems with convergence were reported. However, these are due to the choice of right-hand side $\bar{g}_1$, which was taken to be a vector of all ones. For the unperturbed problem ($\gamma = 0$) this vector is orthogonal to the space spanned by the left singular vector corresponding to the smallest singular value. Therefore the magnitude of the component in this direction of the initial vector x_{LS} will be very small, of the order γ. Also, although A is quite well conditioned the least squares residual is large. The TLS problem is therefore close to a nongeneric problem and thus very illconditoned.

Because of the extreme illconditioning for this right-hand side, the behavior of any solution method becomes very sensitive to the particular random perturbation added. We have therefore instead chosen a right-hand side $\bar{g}_2$ given by $\bar{g}(i) = (m - 2i)/m$, $i = 1, \ldots, m$. For this the TLS problem is much better conditioned, see Table 5.1. Convergence is now obtained in just two iterations, see Figure 5.5.

Table 5.1: Condition numbers for test problem 3 for right-hand sides $\bar{g}_i$, $n = 100$.

γ	i	$\kappa(A)$	κ_{LS}	κ_{TLS}
0	1	1.094484e+03	1.968723e+04	> 1.0e+16
	2	1.094484e+03	2.101815e+04	3.069664e+07
0.001	1	1.220696e+03	2.538016e+04	1.692483e+10
	2	1.220696e+03	2.687055e+04	1.202459e+07

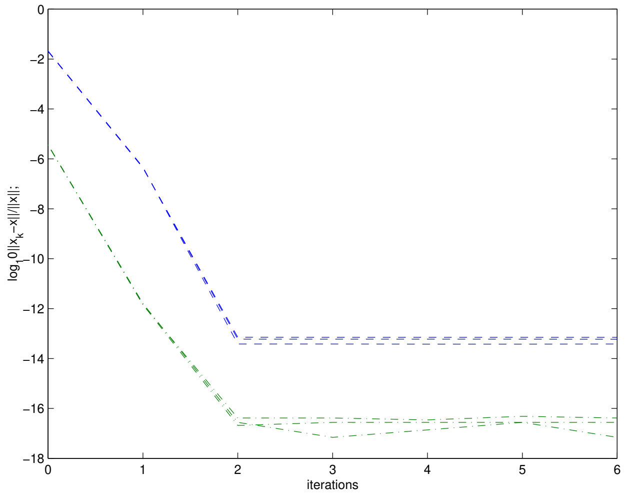

Figure 5.5: Third test problem; RQI with one step of inverse iteration, $n = 100$, $\eta = 0.0001, 0.001, 0.01$.

6 Summary.

We have developed an algorithm for solving large scale TLS problems based on Rayleigh quotient iteration for computing the right singular vector of $(A,\ b)$ defining the solution. The main work in this method consists of solving a sequence of linear systems with matrix $A^TA - \sigma^2 I$, where σ is the current approximation to the smallest singular value of σ_{n+1} of (A, b). For large and sparse TLS problems these linear systems can be solved by a preconditioned conjugate gradient method. An efficient preconditioner is given by a (possibly incomplete) Cholesky factorization of A^TA or QR factorization of A.

Termination criteria for the inner and outer iterations have been given. We conjecture that the described method almost always computes the TLS solution with an accuracy compatible with a backward stable method. Although a detailed error analysis is not given this conjecture is supported by numerical results.

Methods for solving the TLS problem are by necessity more complex than those for the (linear) LS problem. Our algorithm contains several ad hoc choices. On the limited set of test problems we have tried it has only failed for almost singular problems, for which the total least squares model is not relevant and should not be used.

In our method the perturbation E is a rank one matrix which in general is dense. Sometimes it is desired to find a perturbation E that preserves the sparsity structure of A. A Newton method for this more difficult problem has been developed by Rosen, Park, and Glick [21]. However, the complexity of this algorithm limits it to fairly small sized problems. Recently a method, which has the potential to be applied to large sparse problems has been given by Yalamov and Yun Yuan [26]. Their algorithm only converges with linear rate, which may suffice to obtain a low accuracy solution.

REFERENCES

1. Å. Björck, *Stability analysis of the method of semi-normal equations for least squares problems*, Linear Algebra Appl., 88/89 (1987), pp. 31–48.
2. Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
3. Å. Björck, *Newton and Rayleigh quotient methods for total least squares problems*, in Recent Advances in Total Least Squares Techniques and Errors–in–Variables Modeling: Proceedings of the Second International Workshop on Total Least Squares and Errors–in–Variables Modeling, S. Van Huffel, ed., SIAM, Philadelphia, 1997, pp. 149–160.
4. Å. Björck, *Solving large scale multidimensional total least squares problems.* In preparation.
5. Å. Björck, T. Elfving, and Z. Strakos, *Stability of conjugate gradient-type methods for linear least squares problems*, SIAM J. Matrix Anal. Appl., 19:3 (1998), pp. 720–736.
6. J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.
7. R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.
8. J. J. Dongarra, Improving the accuracy of computed singular values. *SIAM J. Sci. Stat. Comput.*, 4 (1983), pp. 712–719.
9. G. H. Golub, *Numerical methods for solving least squares problems*, Numer. Math., 7 (1965), pp. 206–216.
10. G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883—893.
11. P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, 1998.
12. N. J. Higham, *A survey of condition number estimation for triangular matrices*, SIAM Review, 29 (1987), pp. 575–596.
13. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
14. J. Kamm and J. G. Nagy, *A total least squares method for Toeplitz systems of equations*, BIT, 38 (1998), pp. 560–582.

15. W. Mackens and H. Voss, *The minimum eigenvalue of a symmetric positive-definite Toeplitz matrix and rational Hermitian interpolation*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 521–534.

16. P. Matstoms, *Sparse QR factorization in MATLAB*, ACM Trans. Math. Software, 20 (1994), pp. 136–159.

17. C. C. Paige and Z. Strakoš, *Residuals and singular values in linear least squares problems.* Preprint.

18. B. N. Parlett, *The Symmetric Eigenvalue Problem*, 2nd ed., SIAM, Philadelphia, PA, 1998.

19. B. N. Parlett and W. Kahan, *On the convergence of a practical QR algorithm*, in Information Processing 68, Proc. IFIP Congress, Edinburgh, 1968, North-Holland, Amsterdam, 1969, pp.114–118.

20. G. Peters and J. H. Wilkinson, *Inverse iteration, ill-conditioned equations and Newton's method,* SIAM Review, 3 (1979), pp. 339–360.

21. J. B. Rosen, H. Park, and J. Glick, *Total least norm formulation and solution for structured problems,* SIAM J. Matrix Anal. Appl., 17 (1996), pp. 110–126.

22. M. Stewart and G. W. Stewart, *On hyperbolic triangularization: stability and pivoting*, SIAM J. Matrix Anal. Appl., 19:4 (1998), pp. 847–860.

23. D. B. Szyld, *Criteria for combining inverse and Rayleigh quotient iteration,* SIAM J. Numer. Anal., 25 (1988), pp. 1369–1375.

24. S. Van Huffel, *Iterative algorithms for computing the singular subspace of a matrix associated with its smallest singular values,* Linear Algebra Appl., 154/156 (1991), pp. 675–709.

25. S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, *Frontiers in Applied Mathematics* 9, SIAM, Philadelphia, PA, 1991.

26. P. Y. Yalamov and J. Y. Yuan, *A successive least squares method for structured total least squares.* Preprint.

27. T. Yang, *Iterative Methods for Least Squares and Total Least Squares Problems*, Lic. thesis LiU-TEK-LIC-1996-25, Department of Mathematics, University of Linköping, Sweden, 1996.