Error Analysis of Elimination Methods for Equality Constrained Quadratic Programming Problems

E. Galligani and L. Zanni

Department of Mathematics, University of Modena, Via Campi 213/b, I-41100 Modena, Italy.

Abstract. A backward error analysis for the orthogonal factorization method for equality constrained quadratic programming problems has been developed. Furthermore, this method has been experimentally compared with direct elimination method on a class of test problems.

1 Statement of the problem

We consider the equality constrained quadratic programming (QPE) problem

\[
\min_{Cx = d} \frac{1}{2} x^T A x + b^T x + \nu
\]

(1)

where \( C \) is an \( m \times n \) matrix of full rank \((m \leq n)\), \( A \) is an \( n \times n \) symmetric matrix which is positive definite on the subspace \( \mathcal{N}(C) = \{ x \mid C x = 0 \} \), \( d \in \mathbb{R}^m \), \( b \in \mathbb{R}^n \) and \( \nu \in \mathbb{R} \).

It is well known that under these assumptions the problem has a unique solution.

This problem is closely related to the linear equality constrained least squares (LSE) problem

\[
\min_{Cx = d} \frac{1}{2} \|E x - f\|^2
\]

(2)

where \( E \) is a \( k \times n \) matrix \((n \leq k)\), \( f \in \mathbb{R}^k \) and \( \| \cdot \| \) denotes the Euclidean norm. In fact problem (2) is equivalent to (1) if we put \( A = E^T E \), \( b = -E^T f \), \( \nu = \frac{1}{2} f^T f \).

When \( E \) has full rank, \( A \) is positive definite everywhere; when \( \mathcal{N}(E) \cap \mathcal{N}(C) \equiv \{0\} \), \( A \) is positive semidefinite everywhere and positive definite on \( \mathcal{N}(C) \).

A natural way to solve these problems is to use elimination methods [2, 5]. These methods can be interpreted as having the following three stages: i) derive a lower-dimensional unconstrained problem using constraints to eliminate variables; ii) solve the derived problem; iii) transform the solution of the derived problem to obtain the solution of the original constrained problem.
If $Y$ and $Z$ are $n \times m$ and $n \times (n - m)$ matrices respectively such that $[Y \ Z]$ is nonsingular and in addition $CY = I$ and $CZ = 0$ then the solutions $x^*$ of (1) and $y^*$ of (2) have the following explicit expressions:

$$x^* = Yd + Zu^*$$

where $u^*$ is the solution of the positive definite linear system

$$(Z^T AZ)u = -Z^T (b + AYd)$$

and

$$y^* = Yd + Zv^*$$

where $v^*$ is the solution of the unconstrained least squares problem

$$\min \frac{1}{2} ||EZv - (f - EYd)||^2.$$ 

Depending on the choice of $Y$ and $Z$ different elimination methods can be obtained. If $Y$ and $Z$ are obtained by way of QR factorization of the matrix $C^T$

$$Y = Q^T_1 R^{-T} \quad Z = Q^T_2$$

where

$$C^T = Q^T \begin{bmatrix} R \\ 0 \end{bmatrix} \quad Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}_m \quad Q_{n\times n} \text{ orthogonal matrix} \quad R_{m\times m} \text{ nonsingular upper triangular matrix},$$

with $Q$ $n \times n$ orthogonal matrix and $R$ $m \times m$ nonsingular upper triangular matrix, then we have the orthogonal factorization method (OFM) for QPE problems [2] and the null space method for LSE problems [5].

If we assume that the first $m$ columns of $C$ are linearly independent and derive $Y$ and $Z$ from the QR factorization of $C$

$$Y = \begin{bmatrix} R_1^{-1}H^T \\ 0 \end{bmatrix} \quad Z = \begin{bmatrix} -R_1^{-1}R_2 \\ I \end{bmatrix}$$

where

$$C = H \begin{bmatrix} R_1 & R_2 \end{bmatrix}$$

with $H$ $m \times m$ orthogonal matrix, $R_1$ $m \times m$ nonsingular upper triangular matrix and $R_2$ $m \times (n - m)$ matrix, then we have the direct elimination method (DEM) for QPE and LSE problems.

In [3] an error analysis of null space method for LSE problems has been developed. In this work we examine the numerical stability of orthogonal factorization method for QPE problems using backward error analysis approach. Furthermore, the behaviour of the orthogonal factorization method and the direct elimination method is also evaluated experimentally on randomly generated test problems of small and medium scale.
2 On the Numerical Stability of the OFM

In details the orthogonal factorization method for problem (1) consists in the following steps.

1. \( QC^T = \begin{bmatrix} R \\ 0 \end{bmatrix}_{n-m} \).
2. \( R^Ty^* = d. \)
3. \( \bar{A} = QA. \)
4. \( \hat{A} = AQ^T = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix}_{n-m}, \quad \hat{A}_2 = \begin{bmatrix} \hat{A}_{21} \\ \hat{A}_{22} \end{bmatrix}_{n-m}. \)
5. \( \bar{b} = Qb, \quad \bar{b} = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \end{bmatrix}_{n-m}. \)
6. \( \hat{A}_{22}u^* = \bar{b}_2. \)
7. \( x^* = Q^T \begin{bmatrix} y^* \\ u^* \end{bmatrix}_{n-m}. \)

In step 7, the matrix \( \hat{A}_{22} \) is symmetric positive definite and the system can be solved by means of Cholesky factorization or by QR decomposition.

When a computer arithmetic of relative precision \( \epsilon \) is used the effects of rounding errors on each step of the method are well known and by Wilkinson’s backward error analysis it is possible to prove that the computed solution \( x' \) is the exact solution of a perturbed problem.

Suppose to use Householder procedure to perform the orthogonal decomposition required by the method, let \( K(E) \) be the spectral condition number of a matrix \( E \), \( \| E \|_F \) the Frobenius norm of \( E \), \( c'(n, m) = (6n - 3m + 40)m \), \( c(n, j) = c'(n, j) + j \) and \( \phi_E(n, m) = c(n, m)\sqrt{mK(E)} \).

**Theorem 1.** Let \( x' \) be the solution of (1) computed using orthogonal factorization method. If the system in step 7 is solved by QR factorization and

\[
\phi_{C}(n, m)\epsilon < 1
\]

then \( x' \) is the exact solution of the perturbed problem

\[
\min_{(C+\Delta C)x=d+\Delta d} \frac{1}{2} x^T(A + \Delta A)x + (b + \Delta b)^Tx + \nu
\]

where

\[
\|\Delta C\|_F \leq c(n, m)\|C\|_F\epsilon + O(\epsilon^2) \quad (4)
\]

\[
\|\Delta A\|_F \leq (2c(n, m) + c(n - m, n - m))\|A\|_F\epsilon + O(\epsilon^2) \quad (5)
\]
\[ \| \Delta d \| \leq c'(n, m) \| C \|_F \| x' \| \text{eps} + O(\text{eps}^2) \]  \\
\[ \| \Delta b \| \leq (1 + c'(n, m) + c'(n - m, n - m)) \| b \| + \\
+ 2c'(n, m) \| A \|_F \| x' \| + \\
\quad + (1 + c'(n - m, n - m)) K(C) \| A \|_F \| x_m \| \text{eps} + O(\text{eps}^2) \]  

with \( x_m \) minimum length solution of \( Cx = d \).

For the proof see [4]. Moreover, in [4] an analogous result has been proved when the system in step 7 is solved by Cholesky factorization.

In order to obtain bounds for inherent and algorithmic errors we recall the following perturbation formula [4]. Consider the perturbed problem

\[ \min_{(C + \Delta C)x = d \pm \Delta d} \frac{1}{2} x^T (A + \Delta A)x + (b + \Delta b)^T x + \nu \] 

and the matrices \( B = \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix} \) and \( F = \begin{bmatrix} \Delta A & \Delta C^T \\ \Delta C & 0 \end{bmatrix} \).

**Theorem 2.** Let \( A \) be positive definite on \( \mathcal{N}(C) \), \( (A + \Delta A) \) be positive definite on \( \mathcal{N}(C + \Delta C) \), \( \text{rank}(C) = \text{rank}(C + \Delta C) = m \) and \( \|B^{-1}F\| < 1 \). If \( x^* \) and \( \tilde{x} \) denote the solutions of problems (1) and (8) respectively and \( \lambda^* \) denotes the Lagrange multiplier vector associated to \( x^* \) then

\[ \frac{\| x^* - \tilde{x} \|}{\| x^* \|} \leq K_C(A) (\gamma \epsilon_b + \epsilon_A + \omega \epsilon_C) + K_A(C) (\epsilon_C + \epsilon_d) + \\
\quad + \tau_1(\epsilon_A, \epsilon_C, \epsilon_b, \epsilon_d), \]  

where

\[ \epsilon_A = \frac{\| \Delta A \|}{\| A \|}, \quad \epsilon_C = \frac{\| \Delta C \|}{\| C \|}, \quad \epsilon_b = \frac{\| \Delta b \|}{\| b \|}, \quad \epsilon_d = \frac{\| \Delta d \|}{\| d \|}, \]  

\[ K_C(A) = \| H \| \| A \|, \quad K_A(C) = \| S \| \| C \|, \quad \gamma = \frac{\| b \|}{\| A \| \| x^* \|}, \quad \omega = \frac{\| C \| \| \lambda^* \|}{\| A \| \| x^* \|}, \]  

and

\[ H = Z(T^T AZ)^{-1} Z^T, \quad S = Y - Z(T^T AZ)^{-1} Z^T AY, \]

with \( Y \) and \( Z \) \( n \times m \) and \( n \times (n - m) \) matrices respectively such that \([Y \ Z]\) is nonsingular, \( CY = I, \quad CZ = 0 \), and \( \tau_1(\epsilon_A, \epsilon_C, \epsilon_b, \epsilon_d) \) consists of second and higher order terms.

In [4] the mixed condition numbers of the perturbation formula \( K_C(A) \) and \( K_A(C) \) are compared with the condition numbers of the matrices \( A \) and \( C \) and it is proved that if we set \( \rho = \frac{\| Ax^* + b \|}{\| A \| \| x^* \|} \) then

\[ |\rho - \gamma| \leq 1, \quad \rho \leq \omega \leq K(C)\rho. \]  

(10)
If we assume that the data perturbations derive from approximations of real numbers on a computer of relative precision $\epsilon$, then

$$
\epsilon_A \leq \sqrt{n} \epsilon, \quad \epsilon_C \leq \sqrt{m} \epsilon, \quad \epsilon_b \leq \epsilon, \quad \epsilon_d \leq \epsilon,
$$

and from (9) we have the following bound for inherent error

$$
\frac{\|x^* - \tilde{x}\|}{\|x^*\|} \leq \left( K_C(A)(\gamma + \sqrt{n} + \omega\sqrt{m}) + K^A(C)(\sqrt{m} + 1) \right)\epsilon + O(\epsilon^2)
$$

(11)

where $O(\epsilon^2)$ consists of second and higher order terms in $\epsilon$.

A bound for algorithmic error is obtained when in (9) the relative perturbation on data $\epsilon_C, \epsilon_A, \epsilon_d, \epsilon_b$ are bounded by (4), (5), (6), (7). We observe that bounds of $\epsilon_C, \epsilon_A, \epsilon_d$ do not depend on condition numbers of $A$ and $C$; only the bound of $\epsilon_b$ is linearly dependent on $K(C)$. Nevertheless, this dependence can not be considered a potential instability of the algorithm because the coefficient of $\epsilon_b$ in (9) is $K_C(A)\gamma$ and in the bound for inherent error (11) it appears the coefficient $K_C(A)\omega$ that is bounded by $K_C(A)K(C)\rho$ with $|\rho - \gamma| \leq 1$ (see formulae (10)). Thus, we conclude that the bounds of inherent and algorithmic errors can be considered of the same order.

### 3 Computational Experiments

In order to analyse the numerical stability of elimination methods for QPE problems an extensive experimentation has been carried out on HP 9000–730 computer. The algorithms have been implemented in Fortran using LINPACK library [1].

Test problems are constructed using additional precision ($\epsilon \approx 2.22 \times 10^{-16}$) and are solved using relative precision ($\epsilon \approx 1.19 \times 10^{-7}$). In these test problems we prespecify the spectral condition numbers $K(A)$ and $K(C)$, the solution $x^*$ ($\|x^*\| = 1$) of problem (1) and the gradient norm $r = \|Ax^* + b\|$. The matrices $A$ and $C$ are generated in terms of their singular value decompositions using the method given in [6] to generate random orthogonal matrices. The eigenvalues of $A$ and the singular values of $C$ are chosen in geometric progression and in such a way that $\|A\|_2 = 1$ and $\|C\|_2 = 1$.

In the implementation of OFM the orthogonal decomposition of $C^T$ is performed without column pivoting and the linear system in step 7. is solved using Cholesky factorization; for DEM the orthogonal decomposition of $C$ is performed using column pivoting.

Some results of computational experiments are reported in tables 1 and 2 where $E_i$ denotes the bound of inherent error (11) (with $Y$ and $Z$ defined by (3)), $E_{OFM}$ and $E_{DEM}$ denote the relative errors of the computed solution. Results are average values of 50 executions that differ only in the initialization of random-numbers generator used to construct the test problems. In order to have a detailed explanation of further experiments concerning OFM, refer to [4].
Table 1

<table>
<thead>
<tr>
<th>K(A)</th>
<th>E_i</th>
<th>E_{OFM}</th>
<th>E_{DEM}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0(+1)</td>
<td>1.1(-4)</td>
<td>1.9(-6)</td>
<td>2.2(-6)</td>
</tr>
<tr>
<td>1.0(+2)</td>
<td>2.6(-4)</td>
<td>3.0(-6)</td>
<td>3.8(-6)</td>
</tr>
<tr>
<td>1.0(+3)</td>
<td>7.2(-4)</td>
<td>7.2(-6)</td>
<td>1.0(-5)</td>
</tr>
<tr>
<td>1.0(+4)</td>
<td>2.3(-3)</td>
<td>2.3(-5)</td>
<td>3.1(-5)</td>
</tr>
<tr>
<td>1.0(+5)</td>
<td>8.2(-3)</td>
<td>9.5(-5)</td>
<td>1.1(-4)</td>
</tr>
</tbody>
</table>

n = 50, m = 25; K(C) = 100; r = 1;

<table>
<thead>
<tr>
<th>K(A)</th>
<th>E_i</th>
<th>E_{OFM}</th>
<th>E_{DEM}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0(+1)</td>
<td>1.7(-5)</td>
<td>1.1(-6)</td>
<td>1.9(-6)</td>
</tr>
<tr>
<td>1.0(+2)</td>
<td>1.5(-4)</td>
<td>7.1(-6)</td>
<td>8.7(-6)</td>
</tr>
<tr>
<td>1.0(+3)</td>
<td>1.4(-3)</td>
<td>5.4(-5)</td>
<td>5.1(-5)</td>
</tr>
<tr>
<td>1.0(+4)</td>
<td>1.2(-2)</td>
<td>4.1(-4)</td>
<td>3.8(-4)</td>
</tr>
<tr>
<td>1.0(+5)</td>
<td>1.1(-1)</td>
<td>3.3(-3)</td>
<td>2.9(-3)</td>
</tr>
</tbody>
</table>

n = 300, m = 150; K(C) = 100; r = 1;

<table>
<thead>
<tr>
<th>K(A)</th>
<th>E_i</th>
<th>E_{OFM}</th>
<th>E_{DEM}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0(+1)</td>
<td>1.6(-5)</td>
<td>1.1(-6)</td>
<td>2.7(-6)</td>
</tr>
<tr>
<td>1.0(+2)</td>
<td>8.6(-5)</td>
<td>3.8(-6)</td>
<td>6.6(-6)</td>
</tr>
<tr>
<td>1.0(+3)</td>
<td>4.1(-4)</td>
<td>1.5(-5)</td>
<td>2.0(-5)</td>
</tr>
<tr>
<td>1.0(+4)</td>
<td>1.8(-3)</td>
<td>5.7(-5)</td>
<td>6.8(-5)</td>
</tr>
<tr>
<td>1.0(+5)</td>
<td>7.1(-3)</td>
<td>1.5(-4)</td>
<td>2.4(-4)</td>
</tr>
</tbody>
</table>

n = 100, m = 50; K(A) = 1000; r = 1;

<table>
<thead>
<tr>
<th>K(C)</th>
<th>E_i</th>
<th>E_{OFM}</th>
<th>E_{DEM}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0(+1)</td>
<td>1.1(-5)</td>
<td>7.1(-7)</td>
<td>9.2(-7)</td>
</tr>
<tr>
<td>1.0(+2)</td>
<td>2.5(-5)</td>
<td>1.1(-6)</td>
<td>1.2(-6)</td>
</tr>
<tr>
<td>1.0(+3)</td>
<td>5.0(-5)</td>
<td>1.7(-6)</td>
<td>1.8(-6)</td>
</tr>
<tr>
<td>1.0(+4)</td>
<td>8.6(-5)</td>
<td>2.7(-6)</td>
<td>2.5(-6)</td>
</tr>
<tr>
<td>1.0(+5)</td>
<td>1.4(-4)</td>
<td>3.7(-6)</td>
<td>4.0(-6)</td>
</tr>
</tbody>
</table>

Here we conclude that the two methods give almost identical results and they have good numerical stability since, for all the considered values of K(A) and K(C), the relative errors of the computed solution are lower than the bound of inherent error. Moreover, the computational results for OFM are seen to be in keeping with the theoretical conclusions presented in section 2.

References