Acta Numerica (2004), pp. 1–53
DOI: 10.1017/S0962492904000169

The calculation of linear least squares problems

Åke Björck
Department of Mathematics,
Linköping University,
SE-581 83 Linköping, Sweden
E-mail: akbjo@math.liu.se

Dedicated to Michael A. Saunders on the occasion of his sixtieth birthday.

We first survey componentwise and normwise perturbation bounds for the standard least squares (LS) and minimum norm problems. Then some recent estimates of the optimal backward error for an alleged solution to an LS problem are presented. These results are particularly interesting when the algorithm used is not backward stable.

The QR factorization and the singular value decomposition (SVD), developed in the 1960s and early 1970s, remain the basic tools for solving both the LS and the total least squares (TLS) problems. Current algorithms based on Householder or Gram–Schmidt QR factorizations are reviewed. The use of the SVD to determine the numerical rank of a matrix, as well as for computing a sequence of regularized solutions, is then discussed. The solution of the TLS problem in terms of the SVD of the compound matrix $(b \;\; A)$ is described.

Some recent algorithmic developments are motivated by the need for the efficient implementation of the QR factorization on modern computer architectures. This includes blocked algorithms as well as newer recursive implementations. Other developments come from needs in different application areas. For example, in signal processing rank-revealing orthogonal decompositions need to be frequently updated. We review several classes of such decompositions, which can be more efficiently updated than the SVD.

Two algorithms for the orthogonal bidiagonalization of an arbitrary matrix were given by Golub and Kahan in 1965, one using Householder transformations and the other a Lanczos process. If used to transform the matrix $(b \;\; A)$ to upper bidiagonal form, this becomes a powerful tool for solving various LS and TLS problems. This bidiagonal decomposition gives a core regular subproblem for the TLS problem. When implemented by the Lanczos process it forms the kernel in the iterative method LSQR. It is also the basis of the partial least squares (PLS) method, which has become a standard tool in statistics.

We present some generalized QR factorizations which can be used to solve different generalized least squares problems. Many applications lead to LS problems where the solution is subject to constraints. This includes linear equality and inequality constraints. Quadratic constraints are used to regularize solutions to discrete ill-posed LS problems. We survey these classes of problems and discuss their solution.

As in all scientific computing, there is a trend that the size and complexity of the problems being solved is steadily growing. Large problems are often sparse or structured. Algorithms for the efficient solution of banded and block-angular LS problems are given, followed by a brief discussion of the general sparse case. Iterative methods are attractive, in particular when matrix-vector multiplication is cheap.

CONTENTS

1. Introduction

The method of least squares has been the standard procedure for the analysis of data from the beginning of 1800s. A famous example of its use is when Gauss successfully predicted the orbit of the asteroid Ceres in 1801. Two hundred years later, least squares remains a widely used computational principle in science and engineering.

In the simplest case the problem is, given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, to find a vector $x \in \mathbb{R}^n$ such that

$$\min_x \|b - Ax\|_2, \qquad (1.1)$$

where $\|\cdot\|_2$ denotes the Euclidean norm. A least squares solution x is characterized by $r \perp \mathcal{R}(A)$, where $r = b - Ax$ is the residual and $\mathcal{R}(A)$ the range space of A. The residual r is uniquely determined and the solution x is unique if and only if $\operatorname{rank}(A) = n$. If $\operatorname{rank}(A) < n$, we seek the unique least squares solution $x \perp \mathcal{N}(A)$, which is called the pseudo-inverse solution.

Under-determined systems arise from problems where there are more variables than needed to match the observed data. The model problem for this case is to find $y \in \mathbb{R}^m$ such that

$$\min \|y\|_2, \qquad A^T y = c, \tag{1.2}$$

where $c \in \mathbb{R}^n$. Here $y \in \mathbb{R}^m$, the minimum norm solution of the consistent under-determined system $A^T y = c$, is characterized by $y \perp \mathcal{N}(A^T)$. If the system $A^T y = c$ is not consistent we compute the pseudo-inverse solution.

When uncertainties are present also in the matrix A, the total least squares (TLS) model is more appropriate. The TLS problem is

$$\min \| \begin{pmatrix} E & r \end{pmatrix} \|_F, \qquad (A+E)x = b + r, \tag{1.3}$$

where $\| \cdot \|_F$ denotes the Frobenius matrix norm.

Models where the parameters x occur nonlinearly are common, but in this survey we will take the simplistic view that nonlinear problems can be solved by linearization.

From the time of Gauss until the computer age the basic computational tool for solving (1.1) was to form the normal equations $A^T Ax = A^T b$ and solve these by symmetric Gaussian elimination (which Gauss did), or later by the Cholesky factorization (Benoit 1924). This approach has the drawback that forming the matrix $A^T A$ will square the condition number of the original problem. This can lead to difficulties since least squares problems are frequently ill-conditioned.

In the 1950s algorithms based on Gram–Schmidt orthogonalization were widely used, although their numerical properties were not well understood at the time. Björck (1967*b*) analysed the modified Gram–Schmidt algorithm and showed its stability for solving linear least squares problems.

A breakthrough came with the seminal paper by Golub (1965), where it was shown how to compute a QR factorization of A using Householder transformations. A backward stable algorithm for the linear least squares problems was given. Another important development, which took place around the same time, was that of a stable algorithm for computing the singular value decomposition (SVD); see Golub and Kahan (1965) and Golub (1968), and the Algol program for computing the SVD in Golub and Reinsch (1970).

Modern numerical methods for solving least squares problems are surveyed in the two comprehensive monographs by Lawson and Hanson (1995) and Björck (1996). The latter contains a bibliography of 860 references, indicating the considerable research interest in these problems. Hansen (1998) gives an excellent survey of numerical methods for the treatment of numerically rank-deficient linear systems arising, for example, from discrete ill-posed problems. A comprehensive discussion of theory and methods for solving TLS problems is found in Van Huffel and Vandewalle (1991).

Although methods continue to evolve, variations of the QR factorization and SVD remain the basic tools for solving least squares problems. Much of the algorithmic development taking place has been motivated by needs in different application areas, *e.g.*, statistics, signal processing and control theory. For example, in signal processing data is often analysed in real time and estimates need to be updated at each time step. Other applications lead to generalized least squares problems, where the solution is subject to linear or quadratic constraints. A common trend, as in all scientific computing, is that the size and complexity of the problems being solved are steadily growing. There is also an increased need to take advantage of any structure that may exist in the model. Geodetic networks lead to huge sparse structured least squares problems, that have to be treated by sparse factorization methods. Other large-scale problems are better handled by a combination of direct and iterative methods.

The following survey of some areas of recent progress represents a highly subjective selection. Hopefully it will show that many interesting developments still take place in this field.

2. Perturbation analysis and stability

2.1. Perturbation analysis

Consider the least squares problem (1.1) with $\operatorname{rank}(A) = n$ and solution x and residual vector $r = b - Ax$. Let the data A, b be perturbed to $A + \delta A$, $b + \delta b$ where $\operatorname{rank}(A + \delta A) = n$. The perturbed solution by $x + \delta x$ and $r + \delta r$ satisfies the normal equations

$$(A + \delta A)^T (A + \delta A)(x + \delta x) = (A + \delta A)^T (b + \delta b).$$

Subtracting $A^T A x = A^T b$ and solving for δx gives

$$\delta x \approx A^\dagger (\delta b - \delta A\, x) + (A^T A)^{-1} \delta A^T r, \quad A^\dagger = (A^T A)^{-1} A^T,$$

where $r = b - Ax$ is the residual and second-order terms have been neglected. For $r = 0$ this reduces to the well-known first-order perturbation bound for a square nonsingular linear system. For the residual we have $\delta r \approx (\delta b - \delta A x) - A \delta x$ and hence

$$\delta r \approx P_{\mathcal{N}(A^T)} (\delta b - \delta A\, x) + (A^\dagger)^T \delta A^T\, r, \quad P_{\mathcal{N}(A^T)} = I - AA^\dagger.$$

Here $P_{\mathcal{N}(A^T)}$ is the orthogonal projection onto $\mathcal{N}(A^T)$. These equations yield the componentwise estimates (see Björck (1991))

$$|\delta x| \lessapprox |A^\dagger|\,(|\delta b| + |\delta A|\,|x|) + |(A^T A)^{-1}|\,|\delta A|^T\,|r|, \tag{2.1}$$

$$|\delta r| \lessapprox |P_{\mathcal{N}(A^T)}|\,(|\delta b| + |\delta A|\,|x|) + |(A^\dagger)^T|\,|\delta A|^T\,|r|, \tag{2.2}$$

where the inequalities are to be interpreted componentwise. Taking norms in (2.1) and using

$$\|A^\dagger\|_2 = 1/\sigma_n, \quad \|(A^TA)^{-1}\|_2 = 1/\sigma_n^2,$$

where σ_n is the smallest singular value of A, we obtain the approximate upper bound

$$\|\delta x\|_2 \lessapprox \frac{1}{\sigma_n}(\|\delta b\|_2 + \|\delta A\|_2\|x\|_2) + \frac{1}{\sigma_n^2}\|\delta A\|_2\|r\|_2. \tag{2.3}$$

It can be shown that for an arbitrary matrix A and vector b there are perturbations δA and δb such that this upper bound is almost attained. Note that when the residual $r \neq 0$ there is an additional term not present for consistent linear systems. The presence of this term, which will dominate if $\|r\|_2 > \sigma_n\|x\|_2$, was first pointed out by Golub and Wilkinson (1966).

Setting $\delta b = 0$ and assuming $x \neq 0$, we get for the normwise relative perturbation in x

$$\frac{\|\delta x\|_2}{\|x\|_2} \lessapprox \kappa(A)\frac{\|\delta A\|_2}{\|A\|_2}\left(1 + \frac{\|r\|_2}{\sigma_n\|x\|_2}\right), \tag{2.4}$$

where $\kappa(A) = \sigma_1/\sigma_n$ is the condition number of A.

For the minimum norm problem (1.2) with A^T of full row rank, the solution can be expressed in terms of the normal equation as $y = Az$, where $A^TAz = c$. Proceeding as before and neglecting second-order terms in the perturbation we obtain

$$\delta y \approx P_{\mathcal{N}(A^T)}\delta A\, A^\dagger y + (A^\dagger)^T(\delta c - \delta A^T y),$$

giving the componentwise approximate bound

$$|\delta y| \lessapprox |P_{\mathcal{N}(A^T)}|\,|\delta A|\,|A^\dagger|\,|y| + |(A^\dagger)^T|(|\delta c| + |\delta A|^T\,|y|). \tag{2.5}$$

Taking norms we get

$$\|\delta y\|_2 \lessapprox \frac{1}{\sigma_n}(\|\delta c\|_2 + 2\|\delta A\|_2\|y\|_2). \tag{2.6}$$

The statistical model leading to the least squares problem (1.1) is that the vector b of observations is related to the solution x by a linear relation $Ax = b + \epsilon$, where ϵ is a random error vector with zero mean and whose components are uncorrelated and have equal variance. More generally, if the covariance matrix of ϵ equals a symmetric positive definite matrix W, then the best linear unbiased estimate of x is the solution to the least squares problem $\min_x(Ax - b)^TW^{-1}(Ax - b)$, or equivalently

$$\min_x \|W^{-1/2}(b - Ax)\|_2. \tag{2.7}$$

If the errors are uncorrelated then W is a diagonal matrix and we set $D = \text{diag}(d_1, \dots, d_m) = W^{-1/2}$. Then (2.7) is a weighted least squares problem. If some components of the error vector have much smaller variance than the rest, $\kappa(DA) \gg \kappa(A) \geq 1$. The perturbation bound (2.4) then seems to indicate that the problem is ill-conditioned. This is not necessarily so and for such problems it is preferable to use the componentwise bounds (2.1)–(2.2). Special methods for weighted problems are discussed in Björck (1996, Section 4.4).

2.2. Backward error and stability

Consider an algorithm for solving the linear least squares problem (1.1). The algorithm is said to be numerically stable if, for any data A and b, there exist small perturbation matrices and vectors δA and δb, such that the computed solution $\bar{x}$ is the *exact* solution to

$$\min_x \|(A + \delta A)x - (b + \delta b)\|_2, \tag{2.8}$$

where $\|\delta A\| \leq \tau$, $\|\delta b\| \leq \tau$, with τ being a small multiple of the unit round-off u. Any computed solution $\bar{x}$ is called a stable solution if it satisfies (2.8). This does not mean that $\bar{x}$ is close to the exact solution x. If the least squares problem is ill-conditioned then a stable solution can be very different from x. For a stable solution the error $\|x - \bar{x}\|$ can be estimated using the perturbation results given in Section 2.1.

The method by Golub (1965) based on Householder QR factorization is known to be numerically stable with $\delta b = 0$ (Higham 2002, Theorem 20.3). Methods which explicitly form the normal equations are not backward stable. This is because *round-off errors that occur in forming A^TA and A^Tb are not in general equivalent to small perturbations in A and b.* Although the method of normal equations gives results of sufficient accuracy for many applications, its use can result in errors in the computed solution, which are of much larger size than for a stable method.

Many fast methods exist for solving structured least squares problems, *e.g.*, when A is a Toeplitz or Cauchy matrix. These are not in general backward stable (see Gu (1998*b*)), which is one reason why the following results are of interest.

Given an alleged solution $\tilde{x}$, a backward error is a perturbation δA, such that $\tilde{x}$ is the exact solution to the perturbed problem

$$\min_x \|(b + \delta b) - (A + \delta A)x\|_2. \tag{2.9}$$

If we could find the backward error of smallest norm, this could be used to verify numerically the stability properties of an algorithm. There is not much loss in assuming that $\delta b = 0$ in (2.10). Then the optimal backward error in the Frobenius norm is

$$\eta_F(\tilde{x}) = \min\{\|\delta A\|_F \mid \tilde{x} \text{ solves } \min_x \|b - (A + \delta A)x\|_2\}. \tag{2.10}$$

How to find the optimal backward error for the linear least squares problem was an open problem for many years, until it was elegantly answered by Waldén, Karlsson and Sun (1995). They solved the problem by characterizing the set of all backward perturbations and by giving an optimal bound, which minimizes the Frobenius norm $\|\delta A\|_F$; see also Higham (2002, pp. 392–393). Their main result can be stated as follows.

Theorem 1. Let $\tilde{x}$ be an alleged solution and $\tilde{r} = b - A\tilde{x} \neq 0$. The optimal backward error in the Frobenius norm is

$$\eta_F(\tilde{x}) = \begin{cases} \|A^T\tilde{r}\|_2/\|\tilde{r}\|_2, & \text{if } \tilde{x} = 0, \\ \min\{\eta, \sigma_{\min}([A \;\; C])\}, & \text{otherwise}, \end{cases} \tag{2.11}$$

where

$$\eta = \|\tilde{r}\|_2/\|\tilde{x}\|_2, \qquad C = I - (\tilde{r}\tilde{r}^T)/\|\tilde{r}\|_2^2,$$

and $\sigma_{\min}([A \;\; C])$ denotes the smallest (nonzero) singular value of the matrix $[A \;\; C] \in \mathbb{R}^{m\times(n+m)}$.

The task of computing $\eta_F(\tilde{x})$ is thus reduced to that of computing $\sigma_{\min}(\mathcal{A})$. Since this is expensive, approximations that are accurate and less costly have been derived. Karlsson and Waldén (1997) assume that a QR factorization of A is available and give lower and upper bounds for $\eta_F(\tilde{x})$ that only require $O(mn)$ operations. Gu (1998*a*) gives several approximations to $\eta_F(\tilde{x})$ that are optimal up to a factor less than 2. His bounds are formulated in terms of the singular value decomposition of A but his Corollary 2.2 can also be stated as follows.

Let $r_1 = P_{\mathcal{R}(A)}\tilde{r}$ be the orthogonal projection of $\tilde{r}$ onto the range of A. If $\|r_1\|_2 \leq \alpha\|r\|_2$ it holds that

$$\frac{\sqrt{5}-1}{2}\,\tilde{\sigma}_1 \leq \eta_F(\tilde{x}) \leq \sqrt{1+\alpha^2}\,\tilde{\sigma}_1, \tag{2.12}$$

where

$$\tilde{\sigma}_1 = \left\|(A^TA + \eta I)^{-1/2}A^T\tilde{r}\right\|_2/\|\tilde{x}\|_2. \tag{2.13}$$

Since $\alpha \to 0$ for small perturbations $\tilde{\sigma}_1$ is an asymptotic upper bound.

Optimal backward perturbation bounds for under-determined systems are derived in Sun and Sun (1997). The extension of backward error bounds to the case of constrained least squares problems is discussed by Cox and Higham (1999*b*).

3. Orthogonal decompositions

3.1. Algorithms using Householder reflections

The QR factorization of a matrix $A \in \mathbb{R}^{m\times n}$ is

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{3.1}$$

where $R \in \mathbb{R}^{n\times n}$ is upper triangular and $Q \in \mathbb{R}^{m\times m}$ is orthogonal. If A has linearly independent columns, *i.e.*, $\operatorname{rank}(A) = n$, then R is nonsingular. If we partition

$$Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}, \quad Q_1 \in \mathbb{R}^{m\times n}, \quad Q_2 \in \mathbb{R}^{m\times(m-n)},$$

we obtain the compact form $A = Q_1R$ of the QR factorization. In the full rank case Q_1 and R are uniquely determined, provided R is normalized to have positive diagonal elements. Q_1 gives an orthogonal basis for $\mathcal{R}(A)$. Q_2, which is not uniquely determined, gives an orthogonal basis for $\mathcal{N}(A^T)$.

The standard method to compute the QR factorization (3.1) is to premultiply A with a product of Householder reflections $Q^T = P_n \cdots P_2P_1$, where

$$P_j = I - 2v_jv_j^T/\|v_j\|_2^2, \quad j = 1 : n,$$

is constructed to zero out the elements below the main diagonal in the jth column of A. Since a Householder reflection is symmetric and orthogonal,

$$Q = P_1P_2 \cdots P_n. \tag{3.2}$$

There is usually no need to form Q explicitly, since the matrix–vector products Qy and Q^Tz can be efficiently formed using only the Householder vectors $v_1, v_2, \ldots, v_n$. Since v_j only has nonzero elements in positions $j : m$, these can be stored in an $m \times n$ lower trapezoidal matrix. In the dense case this is the most compact representation possible of Q and Q^T.

Given the QR factorization (3.1), the solution x to the linear least squares problem (1.1) and the corresponding residual $r = b - Ax$ is computed:

$$\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = Q^Tb, \qquad x = R^{-1}d_1, \quad r = Q\begin{pmatrix} 0 \\ d_2 \end{pmatrix} = Q_2d_2. \tag{3.3}$$

This algorithm is backward stable (with $\delta b = 0$) both for computing the solution x and the residual $r = b - Ax$; see Higham (2002, Theorem 20.3).

Note that the residual r solves the problem of computing the orthogonal projection of b onto $\mathcal{N}(A^T)$:

$$\min_r \|b - r\|_2 \quad \text{subject to} \quad A^Tr = 0.$$

In some applications we are more interested in the residual r than in the solution x. From the stability (see also the error analysis in Björck (1967*a*)) it follows that the computed residual $\bar{r}$ using (3.3) satisfies a relation

$$(A+E)^T\bar{r} = 0, \quad \|E\|_2 \leq cu\|A\|_2. \tag{3.4}$$

Here and in the following c is a generic constant that grows slowly with n. This implies

$$\|A^T\bar{r}\|_2 \leq cu\|\bar{r}\|_2\|A\|_2, \tag{3.5}$$

that is, the computed residual is accurately orthogonal to $\mathcal{R}(A)$. On the other hand, if $\bar{r} = fl(b - Ax)$, then the best bound we can guarantee is of the form $\|A^T\bar{r}\|_2 \leq cu\|b\|_2\|A\|_2$, even if x is the *exact* least squares solution, When $\|\bar{r}\|_2 \ll \|b\|_2$ this is a much weaker bound than (3.5).

The solution to the minimum norm problem (1.2) can be computed from the QR factorization (3.1) using

$$z = R^{-T}c, \qquad y = Q\begin{pmatrix} z \\ 0 \end{pmatrix} = Q_1 z. \tag{3.6}$$

The fact that this algorithm is backward stable is a relatively new result and the first proof was published in Higham (1995, Theorem 20.3).

An implementation of Householder QR factorization is given in Businger and Golub (1965) (see Wilkinson and Reinsch (1971, Contribution I/8)). A more general implementation, that also solves least squares problems with linear constraints and performs a stable form of iterative refinement of the solution, is given in Björck and Golub (1967).

3.2. Algorithms using modified Gram–Schmidt

In Gram–Schmidt orthogonalization the kth column of Q in the QR factorization is computed as a linear combination of the first k columns in A. This is equivalent to computing the compact QR factorization[1]

$$A = (a_1, a_2, \ldots, a_n) = (q_1, q_2, \ldots, q_n)\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{2n} \end{pmatrix}.$$

Gram–Schmidt QR factorization can also be described as employing a sequence of elementary orthogonal projections to orthogonalize a given sequence of vectors For any nonzero vector $a \in \mathbb{R}^m$ the orthogonal projector P onto the orthogonal complement of a is given by

$$P = I_m - qq^T, \quad q = a/\|a\|_2. \tag{3.7}$$

[1] Trefethen and Bau, III (1997) aptly calls Householder QR orthogonal triangularization and Gram–Schmidt QR triangular orthogonalization.

Two versions of the Gram–Schmidt algorithm exist, usually called the Classical Gram–Schmidt (CGS) and the Modified Gram–Schmidt (MGS) algorithms. Although these only differ in the order in which the operations are performed, MGS has much better numerical stability properties.

Setting $a_j = a_j^{(1)}$, $j = 1 : n$, in MGS at the beginning of step k, $k = 1 : n$, we have computed

$$(q_1, \ldots, q_{k-1}, a_k^{(k)}, \ldots, a_n^{(k)}), \tag{3.8}$$

where $a_k^{(k)}, \ldots, a_n^{(k)}$ are orthogonal to $q_1, \ldots, q_{k-1}$. First the vector q_k is obtained by normalizing $a_k^{(k)}$. The remaining columns are then made orthogonal[2] to q_k, using orthogonal projections

$$a_j^{(k+1)} = (I - q_k q_k^T) a_j^{(k)} = a_j^{(k)} - q_k (q_k^T a_j^{(k)}), \quad j = k+1 : n.$$

Owing to rounding errors the computed $Q_1 = (q_1, q_2, \ldots, q_n)$ will not be orthogonal to working accuracy. For MGS the loss of orthogonality can be bounded in terms of the condition number of A, namely,

$$\|I - Q_1^T Q_1\|_2 \le c_1 u \kappa(A),$$

where u is the unit round-off; see Björck (1967*b*), Björck and Paige (1992).

Because of the loss of orthogonality care is needed in using the MGS factorization. Using a remarkable connection between MGS and Householder QR factorization, Björck and Paige (1992) were able to analyse MGS and rigorously prove the stability of several algorithm based on the MGS factorization. If these algorithms are used with MGS there is *no need for reorthogonalization of the q vectors* for computing least squares solutions, orthogonal projections or solving minimum norm problems. Since few textbooks describe these stable algorithms we present them again here.

Linear least squares solution by MGS

Carry out MGS on $A \in R^{m \times n}$, rank$(A) = n$, to give $Q_1 = (q_1, \ldots, q_n)$ and R, and put $b^{(1)} = b$. Compute the vector $z = (z_1, \ldots, z_n)^T$ by

```
for $k = 1 : n$
    $z_k = q_k^T b^{(k)}$;    $b^{(k+1)} = b^{(k)} - z_k q_k$;
end
$r = b^{(n+1)}$;
solve $Rx = z$;
```

[2] MGS can also be organized so that all previous projections to a_k are applied in the kth step, but this version is not suitable for column pivoting.