Methods for Solving Large Scale Least Squares Problems

Åke Björck
Department of Mathematics, Linköping University

10th of December, 1999

Lecture I: Direct methods.

We survey direct methods for solving least squares problems $\min_x \|Ax-b\|_2^2$, where the matrix $A \in \mathbf{R}^{m\times n}$, $m \geq n$, is large and sparse. From the point of view of numerical stability the method of choice is based on the QR factorization of P_1AP_2, where P_1 and P_2 are permutation matrices chosen to reduce fill-in in R and work. Sometimes the problem has some special structure that can be taken advantage of. The most common such structure is the block angular structure, which can be used to reduce the problem to a sequence of smaller subproblems.

Storage required for Q often is orders of magnitude larger than for R. For example, for a class of grid problems it can be shown that $\mathrm{nnz}(R) = O(n \log n)$ but Q has $O(n\sqrt{n})$ nonzeros. Therefore usually only R is computed in the sparse case, making it more difficult to process additional right hand sides. A new multifrontal code has recently been developed for MATLAB, which stores Q in implicit form in terms of the frontal Householder vectors. For the grid problems mentioned above this representation of Q only requires $O(n \log n)$ storage. We give some more theoretical and experimental results from using this approach.

Sometimes least squares problems occur, which are not sparse but have a Kronecker structure, $\min_x \|(A \otimes B)x - d\|_2$, where the $A \otimes B$ is the Kronecker product of A and B. Problems of Kronecker structure arise in several application areas including signal and image processing, photogrammetry, and multivariate data fitting. Such problems can be solved with great savings in storage and operations. Let $\mathrm{vec}\,(C)$ denote the vector formed by stacking the columns of a matrix C into one long vector. Then the solution to the Kronecker least squares problem can be written $x = \mathrm{vec}\,((B^\dagger)^T D A^\dagger)$, $D = \mathrm{vec}^{\,-1}(d)$.

Lecture II: Iterative methods.

For solving linear least squares problems $\min_x \|Ax - b\|_2$, where A is large and sparse an iterative method can be applied to the factored system of normal equations $A^T(Ax-b)=0$. The explicit formation of the matrix A^TA should be avoided in order to make the method less sensitive to roundoff and also avoid the fill-in which may occur in A^TA. Similar remarks apply to iterative methods for computing minimum norm solutions of consistent underdetermined systems $A^Ty=c$, from the normal equations of the second kind $A^T(Az)=c$.

An important class of methods is the Krylov subspace methods, which in step k seeks an approximation x_k which minimizes a quadratic error functional in the Krylov subspace $x_k \in x^{(0)} + \mathcal{K}_k(A^TA, s^{(0)})$, $s^{(0)} = A^T(b-Ax^{(0)})$. The implementation can either be based on a conjugate gradient algorithm or a Lanczos process. To improve the convergence of iterative methods they can be applied to the (right) preconditioned problem $\min_y \|(AS^{-1})y-b\|_2$, where $Sx=y$. Here S is chosen so that AS^{-1} has a more favorable spectrum than A. Note that if $S=R$, the Cholesky factor of A^TA, then AS^{-1} is orthogonal and CGLS (CG applied to the normal equations) will converge in one iteration. Often S is taken to be an incomplete Cholesky factor of A^TA. i.e., $A^TA=\tilde{S}^T\tilde{S}-E$, where E is a *defect matrix* of small norm.

Iterative methods can be efficient also when the matrix is dense but has a special structure. A rectangular Toeplitz matrix $T \in \mathbf{R}^{m\times n}$ is a matrix with constant entries along each diagonal. Toeplitz least squares problem $\min_x \|Tx-b\|_2$, arise, e.g., in digital signal processing and linear prediction problems. The dimensions of the Toeplitz matrices in such applications are often large and there is a great need for special fast methods.

Iterative solvers are very attractive for Toeplitz least squares problems because the matrix-vector product Tx (and T^Ty) can be computed with two fast Fourier transforms in $O(n\log n)$ operations. For Toeplitz least squares problems fast circulant preconditioners can be constructed, and a preconditioned CGLS method used. Similar ideas can be applied to problems where the least squares matrix T has a general Toeplitz block or block Toeplitz structure. Hence the method can be applied also to two-dimensional, or multidimensional problems.

A Hankel matrix is a Toeplitz matrix in which the rows have been reversed: Hence the methods for solving Toeplitz least squares problems apply also to Hankel least squares problems. Other classes of structured linear systems for which the FFT can be used to compute matrix-vector products include Cauchy systems.