What Is a Modified Cholesky Factorization?

Nicholas J. Higham[*]

December 22, 2020

Newton methods for minimizing a function $F : \mathbb{R}^n \to \mathbb{R}$ generate a sequence of points x_k, where the step from x_k to x_{k+1} is along a search direction p_k determined from a linear system $G_k p_k = -g_k$, where $g_k = \nabla F(x_k)$ is the gradient and G_k is an approximation to the Hessian matrix $\nabla^2 F(x_k)$. The equation $g_k^T p_k = -p_k^T G_k p_k$ shows that G_k is a descent direction if $p_k^T G_k p_k > 0$, and in order to guarantee that this condition holds for all p_k we need G_k to be positive definite. But even if G_k is the exact Hessian, positive definiteness is not guaranteed far from a minimizer. We can modify the method to ensure definiteness of G_k, as with quasi-Newton methods. Or we can perturb the matrix, if necessary, to make it positive definite. Modified Cholesky factorization perturbs and factorizes the matrix at the same time. It is useful in other situations, too, such as in constructing preconditioners and in bounding the distance to a positive semidefinite matrix.

A *modified Cholesky factorization* of a symmetric matrix A is a factorization $P(A + E)P^T = LDL^T$, where P is a permutation matrix, L is unit lower triangular, and D is diagonal or block diagonal and positive definite. It follows that $A + E$ is a positive definite matrix.

A natural way to compute a modified Cholesky factorization is to modify the Cholesky factorization algorithm. Cholesky factorization fails when it tries to take the square root of a negative quantity or divide by zero. We can avoid both possibilities by increasing nonpositive pivots when they are encountered. This corresponds to making a diagonal perturbation E to A and computing a Cholesky factorization $A + E = R^T R$. However, choosing a suitable E is more difficult than it might seem.

Consider the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1-\epsilon & 2 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}, \quad 0 < \epsilon \ll 1.$$

Since Cholesky factorization generates the same sequence of Schur complements as Gaussian elimination, it suffices to consider Gaussian elimination. The diagonal elements of R are the square roots of the pivots. After one step of elimination the reduced matrix is

$$A^{(2)} = \left[\begin{array}{c|ccc} 1 & 1 & 1 & 1 \\ \hline 0 & -\epsilon & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{array}\right],$$

[*] Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK (nick.higham@manchester.ac.uk).

and the trailing 3×3 matrix (a Schur complement) is clearly indefinite because the $(2,2)$ entry, which is the next pivot, is negative. We can make the (2,2) entry positive by adding 2ϵ to it:

$$A^{(2)} + E = \left[\begin{array}{c|ccc} 1 & 1 & 1 & 1 \\ \hline 0 & \epsilon & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{array}\right] \qquad (E = 2\epsilon e_2 e_2^T).$$

The next stage of the factorization can complete and it yields

$$A^{(3)} = \left[\begin{array}{cc|cc} 1 & 1 & 1 & 1 \\ 0 & \epsilon & 1 & 1 \\ \hline 0 & 0 & -\dfrac{1}{\epsilon} & 1 - \dfrac{1}{\epsilon} \\ 0 & 0 & 1 - \dfrac{1}{\epsilon} & 1 - \dfrac{1}{\epsilon} \end{array}\right],$$

The trailing 2×2 submatrix has elements of order $\epsilon^{-1} \gg 1$. Not only will a perturbation of order ϵ^{-1} be required to the $(3,3)$ element to allow the Cholesky factorization to continue, but the Cholesky factor will have elements of order $\epsilon^{-1/2}$ so numerical stability will likely be lost. Yet the smallest eigenvalue of A is of order 1, so it should have been possible to make only an $O(1)$ perturbation to A in order for the factorization to succeed.

This example shows that if we are to increase a pivot element then we need a more sophisticated strategy that takes account of the size of the resulting elements of the factors and the effect on later stages of the factorization.

A modified Cholesky factorization should satisfy, as far as possible, four objectives.

- If A is "sufficiently positive definite" then E is zero.
- If A is indefinite, $\|E\|$ is not much larger than

  $$\min\{\, \|\mathit{\Delta} A\| : A + \mathit{\Delta} A \text{ is positive semidefinite}\,\}$$

  for some appropriate norm.
- The matrix $A + E$ is reasonably well conditioned.
- The cost of the algorithm is the same as the cost of standard Cholesky factorization, that is, $n^3/3 + O(n^2)$ flops for an $n \times n$ matrix.

Gill and Murray (1974) gave the first modified Cholesky algorithm, which computes $P(A+E)P^T = LDL^T$ with diagonal D and E. It was refined by Gill, Murray, and Wright in 1981. Schnabel and Eskow (1990) gave an algorithm that attempts to produce smaller values of $\|E\|$, partly by exploiting eigenvalue bounds obtained from Gershgorin's theorem. That algorithm was subsequently improved by Schnabel and Eskow (1999).

A different approach was taken by Cheng and Higham (1998), building on an earlier idea by Bunch and Sorensen. This approach computes a block $\mathrm{LDL^T}$ factorization $PAP^T = LDL^T$, were P is a permutation matrix, L is unit lower triangular, and D is block diagonal with diagonal blocks of size 1 or 2. The pivoting strategy is the symmetric rook pivoting strategy of Ashcraft, Grimes, and Lewis (1998), which has the key property of producing a bounded L factor. The cost of pivoting is typically $O(n^2)$ comparisons but can be as large as $O(n^3)$ in the worst case. Cheng and Higham compute the perturbation $\mathit{\Delta} D$ of minimal Frobenius norm such that $D + \mathit{\Delta} D$ has eigenvalues greater than or equal to δ, where $\delta > 0$ is a parameter. The modified Cholesky factorization is then $P(A+E)P^T = L(D + \mathit{\Delta} D)L^T$.

A significant advantage of the block LDL$^{\mathrm{T}}$ approach is that it is modular: any implementation of the factorization can be used and the modification is simply inexpensive postprocessing of the D factor. The other algorithms are not simple modifications of an LDL$^{\mathrm{T}}$ factorization and are not straightforward to implement in an efficient way as a block algorithm. Note that in the block LDL$^{\mathrm{T}}$ approach E is a full matrix and it is not explicitly computed.

Modified Cholesky software is not widely available in libraries. Implementations of the Cheng–Higham algorithm are available in

- the NAG Library routine f01mdf (`real_modified_cholesky`),
- the MATLAB codes in the repository `https://github.com/higham/modified-cholesky`.

Example

We take the 4 × 4 matrix above with $\epsilon = 10^{-2}$:

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0.99 & 2 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}.$$

It has eigenvalues

```
-1.0050e+00  -2.3744e-01   1.0000e+00   4.2325e+00
```

The Gill–Murray–Wright algorithm computes as E the diagonal matrix with diagonal elements

```
0   2.0200e+00   2.0000e+00            0
```

while the Schnabel–Eskow algorithm (1999) computes E with diagonal elements

```
1.0000e+00   1.0050e+00   1.0050e+00   1.0000e+00
```

For the Cheng–Higham algorithm with $\delta = (2u)^{1/2}\|A\|_F = 6.7 \times 10^{-8}$ (where $u \approx 1.11 \times 10^{-16}$ is the unit roundoff), the perturbed matrix $A + E$ is

```
1.0000e+00   1.0000e+00   1.0000e+00            0
1.0000e+00   1.4950e+00   1.4975e+00   9.9749e-01
1.0000e+00   1.4975e+00   1.5000e+00   1.0025e+00
         0   9.9749e-01   1.0025e+00   2.0100e+00
```

The Frobenius norms of the perturbations to A are 2.84, 2.00, and 1.43, respectively, and the 2-norm condition numbers are 33.8, 43.2, and 4.67×10^{8}. The large condition number for the Cheng–Higham algorithm is caused by the value of the parameter δ. With $\delta = 0.1$, the perturbed matrix is

```
1.0000e+00   1.0000e+00   1.0000e+00            0
1.0000e+00   1.5453e+00   1.4475e+00   9.9724e-01
1.0000e+00   1.4475e+00   1.5497e+00   1.0027e+00
         0   9.9724e-01   1.0027e+00   2.1100e+00
```

at Frobenius norm distance 1.57 from A and with 2-norm condition number 327.3. For comparison, the symmetric matrix with all eigenvalues greater than or equal to 0.1 that is closest to A in the Frobenius norm is at a distance 1.15 from A.

In general, there is no clear ordering of the different modified Cholesky methods in terms of their ability to satisfy the four criteria.

References

This is a minimal set of references, which contain further useful references within.

- Sheung Hun Cheng and Nicholas Higham, A Modified Cholesky Algorithm Based on a Symmetric Indefinite Factorization, SIAM J. Matrix Anal. Appl. 19(4), 1097–1110, 1998.
- Haw-Ren Fang and Dianne O'Leary, Modified Cholesky Algorithms: A Catalog with New Approaches, Math. Program. 115, 319–349, 2008
- Philip Gill, Walter Murray, and Margaret Wright, Practical Optimization, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2020. Republication of book first published by Academic Press, 1981.
- Thomas McSweeney, Modified Cholesky Decomposition and Applications, M.Sc. Thesis, The University of Manchester, 2017.
- Robert Schnabel and Elizabeth Eskow, A Revised Modified Cholesky Factorization Algorithm, SIAM J. Optim. 9(4), 1135–1148, 1999.

Related Blog Posts

- What Is a Symmetric Positive Definite Matrix? (2020)
- What Is the Choleksy Factorization? (2020)

This article is part of the "What Is" series, available from `https://nhigham.com/category/what-is` and in PDF form from the GitHub repository `https://github.com/higham/what-is`.