Sensitivity Analyses for Factorizations of Sparse or Structured Matrices

Xiao-Wen Chang[*]

Department of Computer Science
University of British Columbia
Vancouver, British Columbia
Canada V6T 1Z4. Email: chang@cs.ubc.ca

and

Christopher C. Paige[†]

School of Computer Science
McGill University
Montreal, Quebec
Canada H3A 2A7. Email: paige@cs.mcgill.ca

Submitted by Dianne P. O'Leary

ABSTRACT

For a unique factorization of a matrix B, the effect of sparsity or other structure on measuring the sensitivity of the factors of B to some change G in B is considered. In particular, norm-based analyses of the QR and Cholesky factorizations are examined. If B is structured but G is not, it is shown that the expressions for the condition numbers are identical to those when B is not structured, but because of the structure the condition numbers may be easier to estimate. If G is structured, whether B is or not, then the expressions for the condition numbers can change, and it is shown how to derive the new expressions. Cases where B and G have the same sparsity structure occur often: here, for the QR factorization an example shows the value of the new expression can be arbitrarily smaller, but for the Cholesky factorization of a tridiagonal matrix and perturbation the value of the new expression cannot be significantly different from the value of the old one. Thus taking account of sparsity can show the condition is much better than would be suggested by ignoring it, but only for some classes of problems, and perhaps only for some types of factorization. The generalization of these ideas to other factorizations is discussed.

[*] Supported by Killam and NSERC postdoctoral fellowships

[†] Supported by NSERC of Canada Grant OGP0009236

655 Avenue of the Americas, New York, NY 10010 0024-3795/92/$5.00

1. INTRODUCTION

For any unique factorization of a matrix B, for example the QR factorization of full column rank B when R is chosen to have positive diagonal elements, we will be interested in how sensitive the factors are to changes in B. Recent work by Chang [1], see also [2, 3, 4, 5, 6, 7], gave an approach to finding and analyzing exact expressions for the condition numbers of such factorizations.

This approach is ideal for taking account of sparsity and other structure in the original matrix B and the change to B represented by G. This paper will attempt to clarify the basic ideas and produce initial meaningful results in this area.

Before proceeding it is important to be clear about the terminology we use. Here a 'condition number' of some factor of B (with respect to the factorization) will always come from an inequality for which equality can be attained for any given matrix B having a unique factorization, see for example (6) and its following sentence. So at no time will we use the term 'condition number' loosely. Throughout the text the term 'structure' will refer to any known structure in a matrix, including any form of sparsity known *a priori*. If the *sparsity* of a matrix has some very regular structure, for example band form, we will either use the standard name, or refer to it as structured sparsity. Thus structure is the most general term, sparsity more specific, and structured sparsity more specific still. We will show how to handle element structure (by which we mean that some equality relationships involving elements hold, as for example in Toeplitz matrices), general sparsity, and structured sparsity in finding, and sometimes analyzing, condition numbers.

The simplest approach to the sensitivity analysis of a unique factorization of B appears to be to consider the factorization of $B(t) \equiv B+tG$, and to take the derivative with respect to t of some matrix equation at $t=0$ in order to relate the derivatives of the factors to the derivative $\dot{B} = G$, see for example the paragraph containing (4). We will use this approach.

There are then two main objects whose structures are important in this analysis, B, and $\dot{B} = G$. Keep in mind these have different possible effects:

$$\begin{array}{rcl} \text{Structure in } B & \rightarrow & \text{structure in the factors of } B, \\ \text{Structure in } \dot{B} = G & \rightarrow & \text{structure in the } \textit{derivatives} \text{ of the factors.} \end{array}$$

The case of structured B but unstructured G is straightforward: $B+tG$ has no element or sparsity structure for $t > 0$, so its factors, and their derivatives even at $t = 0$, have no more structure than those in the unstructured case, so we suspect in general the expressions for the condition numbers will be identical to those for unstructured B. We will see for the QR and Cholesky factorizations that the expressions for the condition numbers of the upper triangular factor R do not change, but the values of these expressions may be easier to estimate compared with the unstructured case. The same observations apply to other factorizations in [1]—[7].

When G has structure we will see the expressions for the condition numbers will usually change. This again applies to other factorizations in [1]—[7]. We will show how to take account of element structure or general sparsity in deriving the new condition numbers.

One of the most common cases of structured G is where B and G have related structure. This can arise when we consider meaningful physical changes, for example a Toeplitz change in Toeplitz B. It can also arise when G corresponds to the equivalent backward rounding error term resulting from a numerically stable finite precision computation for a sparse B, see for example [8]. Cases like this where B and G have related *sparsity* lead to considerable changes in the expression for the condition number, and by using simple examples, we show how to take account of such cases to derive the new condition numbers. Two questions then arise: Can these new condition numbers have significantly different values from the condition numbers for unstructured perturbations? Is it worthwhile going to the extra effort of taking account of sparsity?

To reach meaningful conclusions here, among all the available sparsity patterns, we examine very simple problems having as much sparsity and structure as possible while remaining nontrivial. For if we obtain no significant advantage in a very sparse and structured case, we cannot expect advantages in more complex cases (that is, cases closer to the general unstructured case). In the QR factorization $B = QR$ where B exhibits an important practical sparsity pattern leading to upper bidiagonal R, and G has the same sparsity pattern as B, we show the condition number for R which takes account of this structure can be arbitrarily smaller than that which does not. For the Cholesky factorization $A = R^T R$ of tridiagonal A (again leading to upper bidiagonal R) with a tridiagonal perturbation M, we prove the improvement in value of the new condition number for R can never be great. This suggests for less sparse A and M, such as band A and M, and perhaps even for generally sparse A with M having the same envelope, for the Cholesky factorization the value of the condition number for R will not be improved much by taking account of the sparsity in M.

In Section 2 we will give a short motivation for examining the sensitivity of factorizations, introduce a practical structured sparse problem of the type we will use later as an example, and present some notation. In Section 3 we examine the QR factorization of full column rank B, treating a practical form of sparse B with a sparse perturbation in Section 3.1. Section 3.2 uses the QR factorization to illustrate how in general we can handle some other perturbation structures. In Section 4 we examine the Cholesky factorization of symmetric positive definite A, treating structured A in Section 4.1, structured A with a structured perturbation in Section 4.2, and commenting on our findings in Section 4.3. We give some overall thoughts in Section 5. The Appendix contains a somewhat long proof of (20) and (21) which are required in Section 4.2.

2. A PRACTICAL EXAMPLE OF STRUCTURE

Sensitivity analysis of factorizations is important for at least two reasons. It is important when the factors have some meaning in their own right, and also where the analysis is useful as part of a larger analysis, for example in explaining the high accuracy of some computations. We give a simple example of the former that will also show why we might want to examine the sensitivity of factorizations of sparse and structured matrices. Consider the estimation problem in which we know y and full column rank B so that

$$y = Bx + v, \qquad \mathcal{E}(v) = 0, \qquad \mathcal{E}(vv^T) = \sigma^2 I,$$

where v is an unknown noise vector and $\mathcal{E}(\cdot)$ denotes the expected value. If we obtain the QR factorization of B

$$B = Q_1 R, \qquad Q_1^T Q_1, \qquad R \text{ upper triangular},$$

then solving $R\hat{x} = Q_1^T y$ gives the best linear unbiased estimate (BLUE) $\hat{x}$ of x, and

$$R\mathcal{E}\{(\hat{x} - x)(\hat{x} - x)^T\}R^T = \sigma^2 I,$$

so $\sigma^{-1}R$ is the factor of $[\mathcal{E}\{(\hat{x} - x)(\hat{x} - x)^T\}]^{-1}$, which has sometimes been called the "information matrix". This is important in its own right, and we are interested in how changes in B affect R.

There is a large class of problems of this form that have strong structure, see for example [9, 11]. Suppose we have a discrete Kalman filtering problem (because of the form of the noise vectors u_k and v_k, this is a restricted formulation designed to keep the illustration simple)

$$\begin{aligned} y_k &= C_k x_k + u_k, & \mathcal{E}(u_k) = 0, \qquad & \mathcal{E}(u_k u_k^T) = \sigma^2 I, \\ x_{k+1} &= A_k x_k + v_k, & \mathcal{E}(v_k) = 0, \qquad & \mathcal{E}(v_k v_k^T) = \sigma^2 I, \end{aligned} \tag{1}$$

for $k = 1, 2, \ldots$, with uncorrelated noise vectors. The y_k are known, and we want to estimate the x_k. This becomes a linear least squares problem

$y = Bx + v$ with $y^T \equiv (y_1^T, 0^T, y_2^T, 0^T, \ldots)$, $x^T \equiv (x_1^T, x_2^T, \ldots)$, and $v^T \equiv (u_1^T, v_1^T, u_2^T, v_2^T, \ldots)$, which we can solve via the QR factorization $B = Q_1 R$, where B, and the resulting factor R of the 'information matrix' have the block structure:

$$B \equiv \begin{pmatrix} C_1 & & & & \\ A_1 & -I & & & \\ & C_2 & & & \\ & A_2 & -I & & \\ & & C_3 & & \\ & & A_3 & -I & \\ & & & \cdot & \\ & & & \cdot & \cdot \end{pmatrix} \rightarrow R = \begin{pmatrix} R_{11} & R_{12} & & & \\ & R_{22} & R_{23} & & \\ & & R_{33} & R_{34} & \\ & & & \cdot & \cdot \end{pmatrix}. \tag{2}$$

Note how the column structure of R comes from the column structure of B. For us, a crucial point in such problems is that perturbations only occur in the nonzero blocks of B, and so can only alter the nonzero blocks of R.

We finish this section with an indication of the notation we will use. For any matrix $C \equiv (c_{ij}) \equiv [c_1, \ldots, c_n] \in \mathbf{R}^{n \times n}$, denote by $c_j^{(i)}$ the vector of the first i elements of c_j. With these, we define ("u" denotes "upper")

$$\mathrm{vec}(C) \equiv \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ c_n \end{bmatrix}, \qquad \mathrm{uvec}(C) \equiv \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ \cdot \\ c_n^{(n)} \end{bmatrix}.$$

The second one is the vector formed by stacking the columns of the upper triangular part of C into one long vector. The norms we will use are $\|X\|_2$ the largest singular value $\sigma_{\max}(X)$, and $\|X\|_F \equiv \sqrt{\mathrm{trace}(X^T X)}$.

3. THE QR FACTORIZATION

Let $B \in \mathbf{R}^{m \times n}$ have full column rank. Then B has a unique QR factorization $B = Q_1 R$, where $Q_1 \in \mathbf{R}^{m \times n}$ has orthonormal columns and $R \in \mathbf{R}^{n \times n}$ is upper triangular with positive diagonal entries. Suppose G is a real $m \times n$ matrix such that $\|Q_1^T G\|_2 < \sigma_{\min}(B)$. Then $B + tG$ still has full column rank for $|t| \leq 1$ and has a unique QR factorization

$$B(t) \equiv B + tG = Q_1(t) R(t), \qquad Q_1^T(t) Q_1(t) = I. \tag{3}$$

Here $B(0) = B$, $Q_1(0) = Q_1$ and $R(0) = R$. If we differentiate $R(t)^T R(t) = B(t)^T B(t)$ with respect to t and set $t = 0$, and use $B = Q_1 R$, we obtain with obvious notation the $n \times n$ *symmetric* matrix equation

$$R^T \dot{R} + \dot{R}^T R = R^T Q_1^T G + G^T Q_1 R. \tag{4}$$

It was shown in [7], and it is easy to see that elements $(1,1)$; $(1,2)$, $(2,2)$; $\ldots$; $(1,n)$, $(2,n)$, $\ldots$, (n,n) give, in that order, the $n(n+1)/2$ equations

$$Z_R \cdot \text{vec}(\dot{R}) = Z_R \cdot \text{vec}(Q_1^T G), \qquad Z_R \equiv \left[\begin{array}{cccc} r_1^T & & & \\ \hline r_2^T & r_1^T & & \\ & r_2^T & & \\ \hline \cdot & \cdot & \cdot & \\ \hline r_n^T & & & r_1^T \\ & r_n^T & & r_2^T \\ & & \cdot & \cdot \\ & & & r_n^T \end{array}\right],$$

with r_j being the jth column of R. But $\dot{R}$ must be upper triangular, so removing the strictly lower triangular elements of $\dot{R}$ and the corresponding elements of Z_R on the left gives

$$\begin{array}{rcl} W_R \cdot \text{uvec}(\dot{R}) & = & Z_R \cdot \text{vec}(Q_1^T G), \\ W_R & \equiv & \left[\begin{array}{cccccccc} r_{11} & & & & & & & \\ r_{12} & r_{11} & & & & & & \\ & r_{12} & r_{22} & & & & & \\ \cdot & \cdot & \cdot & \cdot & & & & \\ r_{1n} & & & & r_{11} & & & \\ & r_{1n} & r_{2n} & & r_{12} & r_{22} & & \\ & & & \cdot & \cdot & \cdot & \cdot & \\ & & & & r_{1n} & r_{2n} & \cdot & r_{nn} \end{array}\right]. \end{array} \tag{5}$$

This has a unique solution, and since $\|\text{uvec}(\dot{R})\|_2 = \|\dot{R}\|_F$ *etc.*

$$\frac{\|\dot{R}\|_F}{\|R\|_2} \leq \|W_R^{-1} Z_R\|_2 \frac{\|Q_1^T G\|_F}{\|B\|_2}. \tag{6}$$

For general G, $Q_1^T G$ and therefore $\text{vec}(Q_1^T G)$ may be chosen arbitrarily in $\text{uvec}(\dot{R}) = W_R^{-1} Z_R \text{vec}(Q_1^T G)$. So for any B, the upper bound is attainable, and $\|W_R^{-1} Z_R\|_2$ is the condition number (using this choice of norms) for the R factor with respect to $\|Q_1^T G\|_F$, which measures that part of the perturbation lying in $\mathcal{R}(B)$, the range of B. For general B, as far as we know, it is expensive to estimate $\|W_R^{-1} Z_R\|_2$ directly. So the following upper bound on $\|W_R^{-1} Z_R\|_2$ was obtained in [7]:

$$\|W_R^{-1} Z_R\|_2 \leq \inf_{D>0} \sqrt{1 + \zeta_D^2}\, \kappa_2(D^{-1} R),$$

where $D = \mathrm{diag}(\delta_i)$ and $\zeta_D \equiv \max_{1\leq i<j\leq n} \delta_j/\delta_i$. Notice $\sqrt{2}\kappa_2(B)$, the best of the earlier known upper bounds on the condition number of R in the QR factorization (see [10] and [12]), is an upper bound on this (corresponding to $D = I$ in the above). In practice we do not seek the infimum, but choose D to equilibrate the rows of R as far as possible while keeping $\zeta_D \leq 1$, then use standard condition estimators to estimate $\kappa_2(D^{-1}R)$. This can be done cheaply. For such a choice of D, experiments in [7] suggest that $\sqrt{1+\zeta_D^2}\kappa_2(D^{-1}R)$ is a good approximation to $\|W_R^{-1}Z_R\|_2$.

When G has some structure, we usually cannot choose a G such that the upper bound in (6) is attained. Unfortunately the above approach (from [7]) does not generalize successfully to such structured problems. So now we replace Q_1R in (4) by B, and obtain

$$R^T\dot{R} + \dot{R}^TR = B^TG + G^TB, \tag{7}$$

and then

$$W \cdot \mathrm{uvec}(\dot{R}) = Z \cdot \mathrm{vec}(G), \tag{8}$$

where with the same W_R as in (5)

$$W = W_R, \qquad Z = Z_B = \left[\begin{array}{cccc} b_1^T & & & \\ \hline b_2^T & b_1^T & & \\ & b_2^T & & \\ \hline \cdot & \cdot & \cdot & \\ \hline b_n^T & & & b_1^T \\ & b_n^T & & b_2^T \\ & & \cdot & \cdot \\ & & & b_n^T \end{array}\right].$$

Again this has a unique solution and

$$\frac{\|\dot{R}\|_F}{\|R\|_2} \leq \|W^{-1}Z\|_2 \frac{\|G\|_F}{\|B\|_2}.$$

Once again for *any* structure or sparsity in B, we can choose *unstructured* G such that the upper bound is attained. Thus $\|W^{-1}Z\|_2$ can be regarded as the condition number for the R factor (but now with respect to the full perturbation $\|G\|_F$) for unstructured perturbations in B, no matter what sparsity or structure B has.

When structure in B leads to structure in R (for example band R), then both the condition numbers $\|W_R^{-1}Z_R\|_2$ and $\|W^{-1}Z\|_2$ may be estimated more cheaply than when R has no sparsity.

The applications of this approach to [1]—[7] also lead to well-determined equations similar to the form of (8). Whenever we meet this form it is clear that whatever structure B has, if G is unstructured so the vector on the right hand side can be chosen arbitrarily, then the above remarks on the condition number will also hold.

But in general if the perturbation G is sparse or otherwise structured, then we cannot usually choose G to achieve the upper bound, and the condition number in a case of unstructured G becomes an upper bound on the condition number for the case of structured G. However the new approach (8) to the QR factorization *does* generalize to the case of structured G. We first illustrate this with an interesting and practical example where B and G have the same sparsity. The ideas are simple, and should be easy to apply to any similar analysis dealing with sparse or otherwise structured perturbations, whether B is structured or not.

3.1. The QR factorization for B and G with the same structured sparsity Suppose $(2n-2)\times n$ B has the structure (illustrated here for $n=4$)

$$B = \begin{pmatrix} b_{11} & & & \\ b_{21} & b_{22} & & \\ & b_{32} & & \\ & b_{42} & b_{43} & \\ & & b_{53} & \\ & & b_{63} & b_{64} \end{pmatrix}, \tag{9}$$

and that G has the same structure. Such structures arise naturally in Kalman filtering problems, as can be seen by taking the vectors x_k, y_k, u_k and v_k to be scalars in (1), see (2). Because of the structure, $R(t)$ in (3) and so $\dot{R}(t)$ will be upper bidiagonal, and (7) will be tridiagonal, so we need only include the $(1,1)$ element, and for $j=2,\ldots n$ elements $(j-1,j)$ and (j,j), in deriving the new version of (8). That is in (8) we can drop all but row 1, and for $j=2,\ldots n$ rows $j(j+1)/2-1$ and $j(j+1)/2$. We also drop each column of W corresponding to elements of $\mathrm{uvec}(\dot{R})$ which are necessarily zero (so we drop the same columns as rows above), and drop each column of Z corresponding to elements of $\mathrm{vec}(G)$ which are necessarily zero.

Thus we obtain the reduced system, ("ub" denotes "upper bidiagonal"),

$$W_S \cdot \mathrm{ubvec}(\dot{R}) \;=\; W_S \begin{pmatrix} \dot{r}_{11} \\ \dot{r}_{12} \\ \dot{r}_{22} \\ \dot{r}_{23} \\ \dot{r}_{33} \\ \dot{r}_{34} \\ \dot{r}_{44} \end{pmatrix} \;=\; Z_S \begin{pmatrix} g_{11} \\ g_{21} \\ g_{22} \\ g_{32} \\ g_{42} \\ g_{43} \\ g_{53} \\ g_{63} \\ g_{64} \end{pmatrix}, \tag{10}$$

(for the $n = 4$ case of course) where

$$W_S = \begin{pmatrix} r_{11} & & & & & & \\ r_{12} & r_{11} & & & & & \\ & r_{12} & r_{22} & & & & \\ & & r_{23} & r_{22} & & & \\ & & & r_{23} & r_{33} & & \\ & & & & r_{34} & r_{33} & \\ & & & & & r_{34} & r_{44} \end{pmatrix},$$

$$Z_S = \begin{pmatrix} b_{11} & b_{21} & & & & & & & \\ & b_{22} & b_{21} & & & & & & \\ & & b_{22} & b_{32} & b_{42} & & & & \\ & & & & b_{43} & b_{42} & & & \\ & & & & & b_{43} & b_{53} & b_{63} & \\ & & & & & & & b_{64} & b_{63} \\ & & & & & & & & b_{64} \end{pmatrix}. \qquad (11)$$

It follows that

$$\frac{\|\dot{R}\|_F}{\|R\|_2} \leq \|W_S^{-1} Z_S\|_2 \frac{\|G\|_F}{\|B\|_2},$$

and it is clear that the allowably nonzero elements of G may be chosen to achieve the upper bound, so this is a condition number for this *structured* problem. Here we can estimate $\|W_S^{-1} Z_S\|_2$ in $O(n)$ flops.

It can be shown via (A.3) in the Appendix that $W_S^{-1} Z_S$ is a submatrix of a row and column permutation of $W^{-1}Z$ (the proof is not trivial, but noting W_R and W_E in (A.3) are just $W \equiv W_R$ and W_S here helps), so $\|W_S^{-1} Z_S\|_2 \leq \|W^{-1}Z\|_2$. This suggests the new condition number $\|W_S^{-1} Z_S\|_2$ is an improvement on $\|W^{-1}Z\|_2$. We can give simple examples to show this improvement can be significant, for example

$$B = \begin{pmatrix} 10^{-8} & & & \\ 1 & 1 & & \\ & 10^{-4} & & \\ & 10^{-4} & 10^{-4} & \\ & & 10^{-4} & \\ & & 1 & 1 \end{pmatrix},$$

$$\|W_S^{-1} Z_S\|_2 \approx 1.0000, \qquad \|W^{-1}Z\|_2 \approx 1.4142 \times 10^4,$$

$$\|W_R^{-1} Z_R\|_2 \approx 1.4142 \times 10^4, \qquad \sqrt{2}\kappa_2(B) \approx 2.8284 \times 10^4.$$

3.2. Other forms of structure

We briefly indicate how to handle other structure in the perturbation G.

Suppose in Section 3.1 we are interested in how changes in just one element of B, for example $b_{21} \to b_{21} + t\gamma$, affect $\dot{R}$. From (10) we see $\mathrm{ubvec}(\dot{R}) = W_S^{-1} Z_S e_2 \gamma$, and we not only have the easy to compute condition number $\|W_S^{-1} Z_S e_2\|_2$, but we also have the rates of change for individual elements. We can handle this case for unstructured B in (5) similarly. Such results may not be new, but it is nice to see how easily they fit into the approach here, *and* take account of any structure in B too.

It is obvious how this extends to handling possible changes in any number of selected elements of B. In the structured case (10) we just eliminate those columns of Z_S which correspond to zero elements in the g vector, that is corresponding to unchanging (nonzero) elements of B, giving Z_S^0, and the new condition number is $\|W_S^{-1} Z_S^0\|_2$.

This approach also allows us to handle element structure in G easily. Here is a simple illustrative example. In what is called the constant coefficient case in (1), $A_1 = A_2 = \ldots$, and $C_1 = C_2 = \ldots$, and in our $n = 4$ example (9) we would have $b_{22} = b_{43} = b_{64} = -1$ with no error, $b_{11} = b_{32} = b_{53} = c$ say, and $b_{21} = b_{42} = b_{63} = a$ say. If we are then only considering changes in the coefficients a and c, a meaningful G would then have $g_{22} = g_{43} = g_{64} = 0$, $g_{11} = g_{32} = g_{53}$, and $g_{21} = g_{42} = g_{63}$. The condition number is then $\|W_S^{-1} Z_S [e_1 + e_4 + e_7, e_2 + e_5 + e_8]\|_2$, which is easy to compute, even in the case of general n.

4. THE CHOLESKY FACTORIZATION

Let $A \in \mathbf{R}^{n\times n}$ be a symmetric positive definite matrix. Then A has a unique Cholesky factorization $A = R^T R$, where R is an upper triangular matrix with positive diagonal entries. If $A = B^T B$ in Sections 2 or 3, the R here is identical to the R there.

Suppose M is symmetric and $\|M\|_2 < \sigma_{\min}(A)$, the minimum singular value (here eigenvalue) of A. Then $A + tM$ is still symmetric positive definite for $|t| \le 1$ and has a unique Cholesky factorization

$$A(t) \equiv A + tM = R(t)^T R(t). \tag{12}$$

Here $A(0) = A$ and $R(0) = R$. Write $\dot{A} \equiv \{\frac{d}{dt} A(t)\}_{t=0} = M$ and $\dot{R} \equiv \{\frac{d}{dt} R(t)\}_{t=0} = \dot{R}(0)$. Differentiating (12) with respect to t at $t = 0$ gives

$$\dot{A} = R^T \dot{R} + \dot{R}^T R = M. \tag{13}$$

It was shown in [6] (and it is straightforward to see via the argument following (4)) that the upper triangle of (13) can be written as a linear equation whose solution is the vector of upper triangular elements of $\dot{R}$

$$W_R \cdot \mathrm{uvec}(\dot{R}) = D \cdot \mathrm{uvec}(M), \tag{14}$$

where W_R has the same form as that in (5), and

$$D = \mathrm{diag}(\frac{1}{2}, \underbrace{1, \frac{1}{2}}_{2}, \ldots, \underbrace{1, \ldots, 1, \frac{1}{2}}_{n}) \in \mathbf{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}.$$

For a norm-based analysis, we can multiply by

$$\tilde{D} = \mathrm{diag}(2, \underbrace{\sqrt{2}, 2}_{2}, \ldots, \underbrace{\sqrt{2}, \ldots, \sqrt{2}, 2}_{n}) \in \mathbf{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}$$

on both sides of (14), and define $\widehat{W}_R \equiv \tilde{D}W_R$ and $\widehat{D} \equiv \tilde{D}D$, to obtain

$$\widehat{W}_R \cdot \mathrm{uvec}(\dot{R}) = \widehat{D} \cdot \mathrm{uvec}(M), \tag{15}$$

where $\|\mathrm{uvec}(\dot{R})\|_2 = \|\dot{R}\|_F$ and $\|\widehat{D} \cdot \mathrm{uvec}(M)\|_2 = \|M\|_F$. Then since $\mathrm{uvec}(\dot{R}) = \widehat{W}_R^{-1}[\widehat{D} \cdot \mathrm{uvec}(M)]$ and $\|A\|_2 = \|R^TR\|_2 = \|R\|_2^2$, we obtain

$$\frac{\|\dot{R}\|_F}{\|R\|_2} \leq \kappa_C(A)\frac{\|M\|_F}{\|A\|_2}, \tag{16}$$

where

$$\kappa_C(A) \equiv \|\widehat{W}_R^{-1}\|_2\|R\|_2.$$

Since it is clear that for *any* symmetric positive definite A, symmetric $M \neq 0$ can be chosen to give equality in (16), $\kappa_C(A)$ is the condition number (for the choice of norms in (16)) for the Cholesky factorization. In [6] the following bounds on $\kappa_C(A)$ were derived:

$$\frac{1}{2}\kappa_2^{1/2}(A) \leq \kappa_C(A) \leq \frac{1}{\sqrt{2}}\kappa_2(A), \tag{17}$$

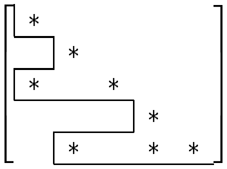
$$\kappa_C(A) \leq \inf_{D>0} \kappa_2(D^{-1}R)\kappa_2(R). \tag{18}$$

In proving the lower bound in (17), $\|R^{-1}\|_2 \leq \|W_R^{-1}\|_2$ was used, since R^T is at the bottom right-hand corner of lower triangular W_R. The expression $\kappa_2(A)/\sqrt{2}$ in (17) was derived in [12, 10], and was then the best of the known estimates for the condition of the problem. It can be seen from (18) that if the ill conditioning of R is mostly due to the bad scaling of the rows, then correct choice of D can give $\kappa_2(D^{-1}R)$ very near one, and $\kappa_C(A)$ will be close to the lower bound $\frac{1}{2}\kappa_2^{1/2}(A)$ since $\kappa_2(R) = \kappa_2^{1/2}(A)$.

4.1. Cholesky factorization with structured A but unstructured M

Note from the previous paragraph that for *unstructured* symmetric perturbations M, $\kappa_C(A) \equiv \|\widehat{W}_R^{-1}\|_2\|R\|_2$ is still the condition number even if

A is structured. For general A, it is unreasonably expensive to estimate $\kappa_C(A)$ directly, so we estimate the upper bound in (18) instead. This can be done cheaply, and usually gives a reasonable approximation to $\kappa_C(A)$. But if A is structured so R is too, then the estimation of $\kappa_C(A)$ might not be difficult, and approximation techniques may not be necessary. We can see this from examining the case of sparse A. It is known that R^T has the same lower envelope as A, that is, if in the following diagram the lower triangle of A has nonzero elements denoted by $*$, then R^T can only have nonzeros in the indicated region (envelope)



.

If A has small envelope, for example band, then R will have many zeros, $\widehat{W}_R$ will have many more zero elements, and $\kappa_C(A) \equiv \|\widehat{W}_R^{-1}\|_2\|R\|_2$ will be easier to estimate.

4.2. *Cholesky factorization with structured M*

If the perturbation M has a fixed structure, then uvec(M) will not be fully general in (14), and $\kappa_C(A)$ may no longer represent the condition number for the factor R. A common example is where A is sparse and M has the same envelope as A. This is true for instance if M is the equivalent backward rounding error term resulting from finite precision computation of the Cholesky factor, for in this case the computed factor $\tilde{R}$ satisfies [8, Thm. 10.3, p. 206]

$$A + M = \tilde{R}^T\tilde{R}, \qquad |M| \leq \epsilon|\tilde{R}^T||\tilde{R}|, \qquad \epsilon = (n+1)u/[1-(n+1)u],$$

where u is the unit roundoff. Since $\tilde{R}^T$ has the same lower envelope as A, M here has the same envelope as A.

We will show here how to derive new condition numbers in such cases. In fact, now some elements of uvec$(\dot{R})$ and uvec(M) in (15) are necessarily zero, so we can drop the corresponding columns in the related matrices in (15). Also we drop the equations that come from elements outside the (upper) envelope in (13), to give ("E" and "env" denote "envelope")

$$\widehat{W}_E \cdot \text{uenv.vec}(\dot{R}) = \widehat{D}_E \cdot \text{uenv.vec}(M),$$

where we give an example to illustrate this shortly. Again, diagonal scaling has been chosen to ensure $\|\widehat{D}_E \cdot \text{uenv.vec}(M)\|_2 = \|M\|_F$, so that

$$\frac{\|\dot{R}\|_F}{\|R\|_2} \leq \kappa_{CE}(A)\frac{\|M\|_F}{\|A\|_2}.$$

Obviously the allowably nonzero elements of M may be chosen to achieve the upper bound, so (for this choice of norms)

$$\kappa_{CE}(A) \equiv \|\widehat{W}_E^{-1}\|_2 \|R\|_2$$

is the condition number for this *structured perturbation* problem.

When A is tridiagonal, $r_{ij} = 0$ for $j \geq i+2$, i.e., W_R in (14) becomes

$$W_R = \left[\begin{array}{c|cc|ccc|cccccc}
r_{11} & & & & & & & & & & & \\
\hline
r_{12} & r_{11} & & & & & & & & & & \\
 & r_{12} & r_{22} & & & & & & & & & \\
\hline
 & & & r_{11} & & & & & & & & \\
 & & r_{23} & r_{12} & r_{22} & & & & & & & \\
 & & & & r_{23} & r_{33} & & & & & & \\
\hline
 & & & & & & \ddots & & & & & \\
 & & & & & & & r_{11} & & & & \\
 & & & & & & & r_{12} & r_{22} & & & \\
 & & & & & & & & r_{23} & r_{33} & & \\
 & & & & & & \cdot & & & \ddots & & \\
 & & & & & & & & & & \cdot & r_{nn}
\end{array}\right].$$

If the perturbation M is also tridiagonal, then in (13), $\mathrm{uvec}(\dot{R})$ and $\mathrm{uvec}(M)$, we see for $i = 1, \ldots, n-2$ that $\dot{r}_{ij} = m_{ij} = 0$ for $j \geq i+2$. Therefore in (14), for $i = 3, \ldots, n$ and $j = 1, \ldots, i-2$, we can discard rows $\frac{1}{2}(i-1)i + j$ of $\mathrm{uvec}(\dot{R})$ and $\mathrm{uvec}(M)$, also rows and columns $\frac{1}{2}(i-1)i + j$ of W_R and D. The perturbation equation (14) then becomes ("ub" denotes "upper bidiagonal")

$$W_E \cdot \mathrm{ubvec}(\dot{R}) = D_E \cdot \mathrm{ubvec}(M),$$

where $W_E \in \mathbf{R}^{(2n-1)\times(2n-1)}$ has the form

$$W_E = \begin{bmatrix}
r_{11} & & & & & & & \\
r_{12} & r_{11} & & & & & & \\
 & r_{12} & r_{22} & & & & & \\
 & & r_{23} & r_{22} & & & & \\
 & & & r_{23} & \cdot & & & \\
 & & & & \cdot & \cdot & & \\
 & & & & & r_{n-1,n} & r_{n-1,n-1} & \\
 & & & & & & r_{n-1,n} & r_{nn}
\end{bmatrix}, \qquad (19)$$

$$D_E = \mathrm{diag}(1/2, 1, 1/2, 1, 1/2, \ldots, 1, 1/2) \in \mathbf{R}^{(2n-1)\times(2n-1)},$$

and

$$\mathrm{ubvec}(C) = [c_{11}, c_{12}, c_{22}, \ldots, c_{i-1,i}, c_{ii}, \ldots, c_{n-1,n}, c_{nn}]^T \in \mathbf{R}^{2n-1}$$

for any $C = (c_{ij}) \in \mathbf{R}^{n\times n}$. Because of this structure in W_E, the condition number $\kappa_{CE}(A) = \|\widehat{W}_E^{-1}\|_2\|R\|_2 = \|(\tilde{D}_E W_E)^{-1}\|_2\|R\|_2$ with $\tilde{D}_E = \mathrm{diag}(2, \sqrt{2}, 2, \sqrt{2}, \ldots, \sqrt{2}, 2)$ can be estimated in $O(n)$ flops, and no further approximation techniques are required.

We want to know if structure can significantly improve our measure of sensitivity. That is, can we have $\kappa_{CE}(A) \ll \kappa_C(A)$, meaning $\|\widehat{W}_E^{-1}\|_2 \ll \|\widehat{W}_R^{-1}\|_2$? We examine the tridiagonal case closely here, for if we obtain no significant improvement for this very sparse and structured case, we cannot reasonably expect significant improvement for less sparse or structured cases. Since W_E (or $\widehat{W}_E$) is obtained by deleting the columns and rows of W_R (or $\widehat{W}_R$) which have the same indices, we certainly have $\|W_E^{-1}\|_2 \le \|W_R^{-1}\|_2$ (or $\|\widehat{W}_E^{-1}\|_2 \le \|\widehat{W}_R^{-1}\|_2$), which suggests an improvement. Also we knew $\|R^{-1}\|_2 \le \|W_R^{-1}\|_2$, and used this in proving the lower bound in (17). But now we cannot say $\|R^{-1}\|_2 \le \|W_E^{-1}\|_2$, so we might have an even *lower* bound on $\kappa_{CE}(A)$ than in (17). Unfortunately, neither improvement is significant. Let $\|X\|_M \equiv \max_{ij}|x_{ij}|$, then we can show (see Appendix)

$$\|W_R^{-1}\|_M = \|W_E^{-1}\|_M, \tag{20}$$
$$\|R^{-1}\|_M \le \|W_E^{-1}\|_M, \tag{21}$$

so we cannot get a much better condition number or lower bound.

In fact from (20), with $\widehat{W}_R = \tilde{D}W_R$ and $\widehat{W}_E = \tilde{D}_E W_E$, we have

$$\begin{aligned}\|\widehat{W}_R^{-1}\|_2 &\le \frac{1}{\sqrt{2}}\|W_R^{-1}\|_2 \le \frac{n(n+1)}{2\sqrt{2}}\|W_R^{-1}\|_M = \frac{n(n+1)}{2\sqrt{2}}\|W_E^{-1}\|_M \\ &\le \frac{n(n+1)}{2\sqrt{2}}\|W_E^{-1}\|_2 \le \frac{n(n+1)}{\sqrt{2}}\|\widehat{W}_E^{-1}\|_2,\end{aligned}$$

which with $\|\widehat{W}_E^{-1}\|_2 \le \|\widehat{W}_R^{-1}\|_2$ gives

$$\frac{\sqrt{2}}{n(n+1)}\kappa_C(A) \le \kappa_{CE}(A) \le \kappa_C(A),$$

showing $\kappa_{CE}(A)$ cannot be very much smaller than $\kappa_C(A)$. From (20) and (21) we have

$$\begin{aligned}\kappa_{CE}(A) &= \|(\tilde{D}_E W_E)^{-1}\|_2\|R\|_2 \ge \frac{1}{2}\|W_E^{-1}\|_M\|R\|_2 \ge \frac{1}{2}\|R^{-1}\|_M\|R\|_2 \\ &\ge \frac{1}{2n}\|R^{-1}\|_2\|R\|_2 = \frac{1}{2n}\kappa_2^{1/2}(A),\end{aligned}$$

showing that $\kappa_{CE}(A)$ cannot be very much smaller than the lower bound in (17).

4.3. Comments on the sensitivity of the Cholesky factorization

So far our thorough analysis for the Cholesky factorization with sparse A and M has only been for symmetric tridiagonal A and M. Since this is among the most sparse of (irreducible positive definite) matrix forms, and does not lead to $\kappa_{CE}(A)$ being significantly smaller than $\kappa_C(A)$, we suspect the same result will hold for all other band or block structures. We have not examined this further, however the following fact would appear to be useful in studying this, and is of practical use. When A and M are banded with width $2p+1$, lower triangular $\widehat{W}_E$ is *block* banded, with block bandwidth $p+1$. Then $\|\widehat{W}_E^{-1}\|$ can be estimated in $O(np^2)$ flops, and so can $\kappa_{CE}(A)$, and for small enough p no further approximation will be needed.

5. CONCLUSIONS

We saw the approach used by Chang [1] for deriving exact expressions for condition numbers of matrix factorizations can also be used to examine the effects of structure in the matrices on these condition numbers. Broadly we saw that there were two main effects in QR, Cholesky, and related factorizations, see for example [1]—[7].

The first effect occurs whenever the original matrix B has some structure, for then the factors may have more than the usual structure, and this can lead to the condition numbers being easier to estimate than for the unstructured case. This happens because the condition numbers are expressed in terms of the elements of the factors, and possibly of the elements of B. But if the perturbation has no structure, then the *expressions* for the condition numbers do not change. In particular we looked at the Cholesky and QR factorizations, and showed if the matrix is structured but the perturbation is not, then the expressions for the condition numbers (see [1, 6, 7]) are unchanged, but in the sparse case the condition numbers may be far easier to estimate than in the full case.

The second effect occurs if the perturbation has some structure. In this case the expressions for the condition numbers derived for unstructured perturbations may no longer give the condition numbers of the structured problem. We showed how to take account of element or sparsity structure, or both, in deriving the new expressions for the condition numbers. If the present approach is used, the comments here and the techniques we exhibited appear to be generally applicable to the factorizations in [1]—[7] and elsewhere.

Two important cases are where the perturbation has the same sparsity structure as B, which can occur in examining the effects of perturbations in the physical coefficients of some problem, or where the perturbation matrix has the same envelope as B, which can occur if we are examining the effect of rounding errors in the computation of the factorization on the factors. We showed how the new condition numbers could be derived for such problems. In these cases the new expressions for the condition numbers can often be estimated directly and efficiently, often obviating the need for the approximation techniques that appear to be needed in the full case. Then we examined the question: Could the new expressions give greatly improved condition numbers?

For the QR factorization we gave a practical example of structure in both the original and perturbation matrices where the value of the new expression for the condition number was never greater than that of the old expression, and showed with particular numbers that it could be very much less. This is a very encouraging result, showing that factors of certain sparse matrices have even better condition than we previously thought. This pleasing result might for example extend to the accuracy of the information matrix factor R in (2) for more general Kalman filtering problems. We conclude that structure must be taken into account when assessing the condition of the QR factorization.

For the Cholesky factorization we closely examined the case where both the original and perturbation matrices were symmetric tridiagonal, and showed that while the value of the new expression for the condition number was always bounded above by that for the old one, the difference could never be significant. Since this is the most sparse of (irreducible positive definite) matrix forms, we suspect similar results will hold for the Cholesky factorizations of all other band or block structures. Of course structure should still be taken into account to facilitate estimating the condition.

A. Appendix

Proof of (20) *and* (21) in Section 4.2.

First we prove the easier (21), then use it to prove (20). The proofs use the simple fact that for $k \geq i \geq j \geq 1$ the (i,j)th block of the inverse of

$$\begin{pmatrix} M_1 & & & \\ N_2 & M_2 & & \\ & \cdot & \cdot & \\ & & N_k & M_k \end{pmatrix},$$

with M_i nonsingular, is

$$(-1)^{i-j} M_i^{-1} N_i M_{i-1}^{-1} N_{i-1} \cdots M_{j+1}^{-1} N_{j+1} M_j^{-1}. \tag{A.1}$$

Since R is upper bidiagonal, for $i \geq j$ the (i,j)th element of R^{-T} is

$$(-1)^{i-j}\,\frac{\phi_{ij}}{r_{ii}}, \quad \phi_{ij} \equiv \frac{r_{j,j+1}}{r_{jj}} \cdot \frac{r_{j+1,j+2}}{r_{j+1,j+1}} \cdots \frac{r_{i-1,i}}{r_{i-1,i-1}}, \;\; i > j; \;\; \phi_{jj} \equiv 1, \quad (A.2)$$

while from (19) for $n \geq i \geq j \geq 1$ the $(2i-1, 2j-1)$ element of W_E^{-1} is ϕ_{ij}^2/r_{ii}. In particular the $(2i-1, 2i-1)$ element of W_E^{-1} is $1/r_{ii}$, so

$$|[R^{-T}]_{ij}| = \frac{|\phi_{ij}|}{|r_{ii}|} \leq \frac{\max\{\phi_{ij}^2, 1\}}{|r_{ii}|} = \max\{|[W_E^{-1}]_{2i-1,2j-1}|, |[W_E^{-1}]_{2i-1,2i-1}|\},$$

and (21) holds.

To prove (20), let R_i denote the leading principal $i \times i$ submatrix of R. If we permute to the top left the rows and columns of W_R that we previously discarded to get W_E, we can obtain

$$P^T W_R P = \left[\begin{array}{c|c} D_R & 0 \\ \hline F & W_E \end{array}\right] \equiv \left[\begin{array}{ccc|c} R_1^T & & & \\ & \cdot & & 0 \\ & & R_{n-2}^T & \\ \hline F_1 & \cdot & F_{n-2} & W_E \end{array}\right], \quad (A.3)$$

where $F_i \equiv e_{2i+2} r_{i,i+1} e_i^T$ is $(2n-1) \times i$. $P^T W_R P$ then has inverse

$$\left[\begin{array}{c|c} D_R^{-1} & 0 \\ \hline -W_E^{-1} F D_R^{-1} & W_E^{-1} \end{array}\right],$$

so every element of W_E^{-1} is also an element of W_R^{-1}, giving

$$\|W_E^{-1}\|_M \leq \|W_R^{-1}\|_M. \quad (A.4)$$

Since each R_i is a leading principal submatrix of upper triangular R, we also have with (21) that

$$\|D_R^{-1}\|_M \leq \|R^{-1}\|_M \leq \|W_E^{-1}\|_M. \quad (A.5)$$

Finally we examine the ith block of $W_E^{-1} F D_R^{-1}$, which is $(2n-1) \times i$,

$$W_E^{-1} F_i R_i^{-T} = W_E^{-1} e_{2i+2} r_{i,i+1} e_i^T R_i^{-T}.$$

Now W_E is lower bidiagonal, so $W_E^{-1} e_{2i+2}$ has elements $1, 2, \ldots, 2i+1$ zero, element $2i+2$ is $1/r_{i+1,i+1}$, and for $k \geq i+2$

$$(W_E^{-1})_{2k,2i+2} = \frac{r_{i+1,i+2}}{r_{i+1,i+1}} \cdot [\frac{r_{i+2,i+3}}{r_{i+2,i+2}} \cdots \frac{r_{k-1,k}}{r_{k-1,k-1}}]^2 \cdot \frac{r_{k,k+1}}{r_{kk}} \cdot \frac{1}{r_{k,k}},$$

where for $k = i+2$ the squared term is replaced by unity. Combining this last expression with (A.2) shows for $k-2 \geq i \geq j \geq 1$

$$
\begin{aligned}
&(-1)^{i-j}(W_E^{-1}e_{2i+2}r_{i,i+1}e_i^T R_i^{-T})_{2k,j} = \\
&\qquad \frac{r_{j,j+1}}{r_{jj}} \cdots \frac{r_{i+1,i+2}}{r_{i+1,i+1}} \cdot \left[\frac{r_{i+2,i+3}}{r_{i+2,i+2}} \cdots \frac{r_{k-1,k}}{r_{k-1,k-1}}\right]^2 \cdot \frac{r_{k,k+1}}{r_{kk}} \cdot \frac{1}{r_{k,k}}.
\end{aligned}
$$

$$
\begin{aligned}
&\text{But} \qquad -(W_E^{-1})_{2k,2i+3} = \\
&\qquad\qquad \left[\frac{r_{i+2,i+3}}{r_{i+2,i+2}} \cdots \frac{r_{k-1,k}}{r_{k-1,k-1}}\right]^2 \cdot \frac{r_{k,k+1}}{r_{kk}} \cdot \frac{1}{r_{k,k}},
\end{aligned}
$$

$$
\begin{aligned}
&\text{while} \qquad -(W_E^{-1})_{2k,2j-1} = \\
&\qquad \left[\frac{r_{j,j+1}}{r_{jj}} \cdots \frac{r_{i+1,i+2}}{r_{i+1,i+1}}\right]^2 \cdot \left[\frac{r_{i+2,i+3}}{r_{i+2,i+2}} \cdots \frac{r_{k-1,k}}{r_{k-1,k-1}}\right]^2 \cdot \frac{r_{k,k+1}}{r_{kk}} \cdot \frac{1}{r_{k,k}},
\end{aligned}
$$

so for $k-2 \geq i \geq j \geq 1$

$$
|(W_E^{-1}e_{2i+2}r_{i,i+1}e_i^T R_i^{-T})_{2k,j}| \leq \max\{|(W_E^{-1})_{2k,2i+3}|, |(W_E^{-1})_{2k,2j-1}|\}.
$$

For the $k = i+1$ case we can show that for $i \geq j \geq 1$

$$
\begin{aligned}
&|(W_E^{-1}e_{2i+2}r_{i,i+1}e_i^T R_i^{-T})_{2i+2,j}| \\
&\qquad \leq \max\{|(W_E^{-1})_{2i+1,2j-1}|, |(W_E^{-1})_{2i+2,2i+2}|\}.
\end{aligned}
$$

Similarly we can show for $k > i \geq j \geq 1$

$$
\begin{aligned}
&|(W_E^{-1}e_{2i+2}r_{i,i+1}e_i^T R_i^{-T})_{2k+1,j}| \\
&\qquad \leq \max\{|(W_E^{-1})_{2k+1,2i+3}|, |(W_E^{-1})_{2k+1,2j-1}|\},
\end{aligned}
$$

and so $\|W_E^{-1}F_iR_i^{-T}\|_M \leq \|W_E^{-1}\|_M$. But this result holds for all $i = 1, 2, \ldots, n-2$, so $\|W_E^{-1}FD_R^{-1}\|_M \leq \|W_E^{-1}\|_M$, which with (A.5) shows $\|W_R^{-1}\|_M \leq \|W_E^{-1}\|_M$. Combining this with (A.4) proves (20).

ACKNOWLEDGEMENT

The presentation was improved greatly as a result of suggestions by the referees.

REFERENCES

1 X.-W. Chang, *Perturbation Analysis of Some Matrix Factorizations*, PhD thesis, Computer Science, McGill University, Montreal, Canada, February 1997.

2 X.-W. Chang, Perturbation analyses for the Cholesky factorization with backward rounding errors, *Scientific Computing: Proceedings of the Workshop, 10-12 March 1997, Hong Kong*, ed. G. Golub, F. Luc and R. Plemmons, Springer, Singapore, pp. 180-187, 1997.

3 X.-W. Chang and C. C. Paige, A perturbation analysis for R in the QR factorization, School of Computer Science SOCS-95.7, McGill University, Montreal, Canada, 1995. 20 pages.

4 X.-W. Chang and C. C. Paige, Perturbation analyses for the Cholesky downdating problem, *SIAM J. Matrix Anal. Appl.*, to appear in 1998. 15 pages.

5 X.-W. Chang and C. C. Paige, On the sensitivity of the LU factorization, submitted to *BIT*, 1997. 15 pages.

6 X.-W. Chang, C. C. Paige, and G. W. Stewart, New perturbation analyses for the Cholesky factorization, *IMA J. Numer. Anal.*, 16:457–484 (1996).

7 X.-W. Chang, C. C. Paige, and G. W. Stewart, Perturbation analyses for the QR factorization, *SIAM J. Matrix Anal. Appl.*, 18:775–791 (1997).

8 N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

9 C. C. Paige and M. A. Saunders, Least squares estimation of discrete linear dynamic systems using orthogonal transformations, *SIAM J. Numer. Anal.*, 14:180–193 (1977).

10 G. W. Stewart, On the perturbation of LU, Cholesky, and QR factorizations, *SIAM J. Matrix Anal. Appl.*, 14:1141–1145 (1993).

11 G. Strang and K. Borre, *Linear Algebra, Geodesy and GPS*, Wellesley-Cambridge Press, Wellesley, Massachusetts 02181, 1997.

12 J.-G. Sun, Perturbation bounds for the Cholesky and QR factorizations, *BIT*, 31:341–352 (1991).