BIT

SOME FEATURES OF GAUSSIAN ELIMINATION WITH ROOK PIVOTING *

XIAO-WEN CHANG †

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

Abstract.

Rook pivoting is a relatively new pivoting strategy used in Gaussian elimination (GE). It can be as computationally cheap as partial pivoting and as stable as complete pivoting. This paper shows some new attractive features of rook pivoting. We first derive error bounds for the LU factors computed by GE and show rook pivoting usually gives a highly accurate U factor. Then we show accuracy of the computed solution of a linear system by rook pivoting is essentially independent of row scaling of the coefficient matrix. Thus if the matrix is ill-conditioned due to bad row scaling a highly accurate solution can usually be obtained. Finally for a typical inversion method involving the LU factorization we show rook pivoting usually makes both left and right residuals for the computed inverse of a matrix small.

AMS subject classification: 15A23, 65F05, 65G50

Key words: Gaussian elimination, LU factorization, pivoting, error analysis, linear systems, matrix inversion.

1 Introduction.

Gaussian elimination (GE) is one of the most fundamental and effective algorithms in matrix computations. Given a real $n \times n$ matrix A whose n leading principal submatrices are all nonsingular, GE computes the LU factorization

$$A = LU, \tag{1.1}$$

where L is a unit lower triangular matrix and U is an upper triangular matrix. L and U are referred to as the LU factors. Simple examples show that GE is not numerically stable. In order to repair this shortcoming of the algorithm, two well-known pivoting strategies, partial pivoting and complete pivoting, are usually incorporated into the computation.

Recently Neal and Poole [13] presented the so-called rook pivoting strategy, which will be introduced in Section 2. This pivoting strategy appears to be intermediate between partial pivoting and complete pivoting in terms of efficiency and stability. The main purpose of this paper is to shed light on the effect of

*Received April 2000. Revised November 2000. Communicated by Iain S. Duff.

†This research was supported by NSERC of Canada Grant RGPIN217191-99 and FCAR of Quebec Grant NC66487.

rook pivoting on the accuracy of the LU factors, the accuracy of the computed solutions of linear systems and the numerical stability of a typical inversion method.

The LU factorization is widely used in matrix computations. Sometimes the accuracy of the LU factors may affect the accuracy of the final solution. For example, in [6], the accuracy of the LU factorization is crucial in computing the singular value decomposition with high relative accuracy. In Section 4 we will first estimate the accuracy of the LU factors and then discuss how the three pivoting strategies affect the accuracy. We will show that the U factor usually has high accuracy if GE with rook pivoting (GERP) or GE with complete pivoting (GECP) is used.

GE is one of the most popular algorithms for solving a linear system. It has been observed that on average rook pivoting is more accurate than partial pivoting, see Neal and Poole [13, 15]. But the numerical examples given in [15] do not show any significant difference between the accuracy of the two pivoting strategies. In Section 5, we will give some examples to show there can be a significant difference, and give some explanations. We will discuss the componentwise stability of the GERP, and show the accuracy of the computed solution is essentially independent of the row scaling of the coefficient matrix.

One of the important applications of the LU factorization is computing the inverse of a matrix. For four typical inversion methods which all involve GE with partial pivoting (GEPP), Du Croz and Higham [7] showed that only one of the left and right residuals is guaranteed to be usually small. For one of the inversion methods which is used by LINPACK, LAPACK and MATLAB, we will show in Section 6 that *both* left and right residuals will usually be small if rook pivoting or complete pivoting is used instead of partial pivoting.

Before proceeding, let us introduce the notation to be used through out the paper. If $A = (a_{ij})$, then $|A| \equiv (|a_{ij}|)$. A matrix norm $\|\cdot\|$ on $\mathcal{R}^{m\times n}$ is monotone, if $|A| \le |B|$ implies $\|A\| \le \|B\|$. For example, $\|\cdot\|_1$, $\|\cdot\|_\infty$ and $\|\cdot\|_F$ are monotone norms. Obviously for a monotone matrix norm, $\|\,|A|\,\| = \|A\|$. For a nonsingular matrix A, following [12], we denote $\kappa(A) \equiv \|A^{-1}\|\cdot\|A\|$, $\mathrm{cond}(A) \equiv \|\,|A^{-1}|\cdot|A|\,\|$. Although many authors define $\mathrm{cond}(A)$ to be what we have called $\kappa(A)$, we will find it useful to use both $\kappa(A)$ and $\mathrm{cond}(A)$.

2 Rook pivoting.

A traditional quantity used to describe the backward stability of GE is the growth factor ρ. For partial pivoting or complete pivoting, since the elements of L are bounded by 1, the growth factor can be defined by (see for example [5, p.49])

$$\rho = \max_{ij} |u_{ij}| / \max_{ij} |a_{ij}|. \tag{2.1}$$

For the classic definition of the growth factor, see for example [10, p.116] and [12, p.177]. For partial pivoting it is not difficult to show that $\rho \le 2^{n-1}$ and the bound is reachable, see [5, p.49]. Even though ρ usually behaves like n or less, Foster [8] has found an example which plausibly could arise in practice and for which ρ can grow exponentially. For complete pivoting, in [21] it is shown that $\rho \leq 2\sqrt{n} n^{\ln(n)/4}$. No one has been able to find an example where the growth factor for complete pivoting is bigger than, for example, $2n$. So complete pivoting has better numerical stability than partial pivoting. The main disadvantage with complete pivoting is that it requires approximately $n^3/3$ comparisons. So it is not as efficient as partial pivoting, which needs about $n^2/2$ comparisons.

Recently the rook pivoting strategy was introduced in [13]. This pivoting strategy appears to be intermediate between complete pivoting and partial pivoting in terms of efficiency and stability. Its idea is as follows: in each step of the forward elimination of GE, from the remaining matrix choose a pivot element which is the largest both in the row and column it lies in. MATLAB code for the selection of the pivot element at step k in rook pivoting is then:

```
row = k;
col = k;
[colmax, rowindex] = max(abs(A(k:n,k)));
rowmax = 0.0;
while rowmax < colmax
   row = rowindex + k - 1;
   [rowmax, colindex] = max(abs(A(row,k:n)));
   if colmax < rowmax
      col = colindex + k - 1;
      [colmax, rowindex] = max(abs(A(k:n,col)));
   else
      break
   end
end
```

For rook pivoting, Foster [9] showed that the growth factor satisfies $\rho \leq 1.5 n^{3\ln(n)/4}$. His experiments also showed that the average growth factor of rook pivoting is comparable with that of complete pivoting. Under the assumption that in step k of the GE reduction, the elements of $A(k:n, k:n)$ are independent identically distributed random variables from any continuous probability distribution, Foster [9] showed that the expected number of comparisons in step k of rook pivoting is less than or equal to $e(n-k)$, where e is the natural logarithm base. So if the assumption is true, the expected number of comparisons in a complete factorization by rook pivoting would be less than or equal to $en(n-1)/2$. This result is a generalization of a result in Poole and Neal [15] which requires the more restrictive assumption that the elements of $A(k:n, k:n)$ come from a uniform distribution. This theory is empirically supported by [9] and [15]. Foster's numerical experiments in a serial computer environment showed that rook pivoting is close to partial pivoting in efficiency.

3 Some properties of the LU factors.

Let $\tilde{L}$ and $\tilde{U}$ be the computed LU factors of A by GE with some pivoting strategy. If rook pivoting, partial pivoting, or complete pivoting is used in GE, then we have

$$\tilde{l}_{ii} = 1, \quad |\tilde{l}_{ij}| \le 1 \ \text{ for } i > j. \tag{3.1}$$

If rook pivoting or complete pivoting is used, then we have

$$|\tilde{u}_{ii}| \ge |\tilde{u}_{ij}| \ \text{ for } i < j. \tag{3.2}$$

When (3.1) and (3.2) hold, it is easy to show (see [12, p.155])

$$(|\tilde{L}^{-1}|\cdot|\tilde{L}|)_{ij} \le 2^{i-j}, \quad (|\tilde{L}|\cdot|\tilde{L}^{-1}|)_{ij} \le 2^{i-j} \ \text{ for } i \ge j,$$
$$(|\tilde{U}^{-1}|\cdot|\tilde{U}|)_{ij} \le 2^{j-i} \ \text{ for } i \le j.$$

Then it follows that

$$\mathrm{cond}_\infty(\tilde{L}) \le 2^n - 1, \quad \mathrm{cond}_\infty(\tilde{L}^{-1}) \le 2^n - 1, \tag{3.3}$$
$$\mathrm{cond}_\infty(\tilde{U}) \le 2^n - 1. \tag{3.4}$$

We can find an example such that these bounds can be reached. But we believe usually these bounds are very pessimistic. In our numerical experiments, we found that for an $n \times n$ matrix A generated by MATLAB built-in function `randn`, these condition numbers were always smaller than n^2; see Figure 3.1. All our computations were performed in MATLAB 5.2 on a Pentium-II running LINUX. Our simple justification is as follows. In [20, pp.167–170], Trefethen and Bau explain why the elements of $\tilde{L}^{-1}$ are usually not large for the partial pivoting case. It can be seen that similar arguments apply to the L factor when rook pivoting or complete pivoting is used. So with (3.1) we conclude $\mathrm{cond}_\infty(\tilde{L})$ and $\mathrm{cond}_\infty(\tilde{L}^{-1})$ are usually not large for any of these pivoting methods. For the U factor, a little more elaboration is needed. Let $\tilde{U} = D\bar{U}$ where $D = \mathrm{diag}(\tilde{u}_{ii})$ and $|\bar{u}_{ij}| \le 1$. Then for rook pivoting or complete pivoting, $\bar{U}^T$ can be regarded as the computed L factor of A^T. Thus the element of $\bar{U}^{-T}$ are usually not large by applying the Trefethen and Bau argument. But $\mathrm{cond}_\infty(\tilde{U}) = \mathrm{cond}_\infty(\bar{U})$. This justifies our belief that $\mathrm{cond}_\infty(\tilde{U})$ is also usually not large.

Our later analysis will be based on (3.3) and (3.4).

4 Accuracy of the computed LU factors.

If GE applied to a nonsingular matrix A runs to completion then the computed LU factors $\tilde{L}$ and $\tilde{U}$ satisfy

$$A + \Delta A = \tilde{L}\tilde{U}, \qquad |\Delta A| \le \epsilon |\tilde{L}|\cdot|\tilde{U}|, \tag{4.1}$$

where $\epsilon \equiv nu/(1-nu)$ with u being the unit roundoff, see for example [12, Thm 9.3]. The assumption that GE runs to completion guarantees $\tilde{u}_{ii} \ne 0$ for $i = 1, \ldots, n-1$. We assume u is small enough such that $\tilde{u}_{nn} \ne 0$. The study of accuracy of the computed LU factors can be approached via a perturbation analysis of the LU factorization with a special perturbation. For the perturbation analysis of the LU factorization with a general perturbation, see [2, 3, 17, 18, 19].

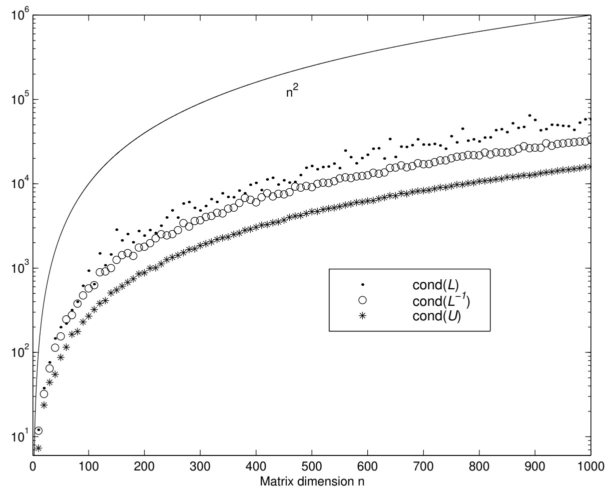

Figure 3.1: Condition numbers of the LU factors of random matrices.

For any $n \times n$ matrix $X = (x_{ij})$, we define the *strictly lower triangular* matrix and *upper triangular* matrix

$$(4.2)\quad \mathrm{slt}(X) \equiv (s_{ij}), \quad s_{ij} \equiv \begin{cases} x_{ij} & \text{if } i > j, \\ 0 & \text{otherwise,} \end{cases} \qquad \mathrm{ut}(X) \equiv X - \mathrm{slt}(X).$$

4.1 First-order error bounds on the computed LU factors.

Let strictly lower triangular $\Delta L \equiv \tilde{L} - L$, and upper triangular $\Delta U \equiv \tilde{U} - U$. Then by (1.1) and (4.1)

$$A = (\tilde{L} - \Delta L)(\tilde{U} - \Delta U) = A + \Delta A - \tilde{L}\Delta U - \Delta L\tilde{U} + \Delta L\Delta U,$$

which, on dropping the second order term, gives a linear matrix equation for the first-order approximations $\widehat{\Delta L}$ (strictly lower triangular) to ΔL and $\widehat{\Delta U}$ to ΔU:

$$\tilde{L}\widehat{\Delta U} + \widehat{\Delta L}\tilde{U} = \Delta A.$$

From this we have

$$\tilde{L}^{-1}\widehat{\Delta L} + \widehat{\Delta U}\tilde{U}^{-1} = \tilde{L}^{-1}\Delta A\tilde{U}^{-1},$$

which with (4.2) gives

$$\tilde{L}^{-1}\widehat{\Delta L} = \mathrm{slt}(\tilde{L}^{-1}\Delta A\tilde{U}^{-1}), \qquad \widehat{\Delta U}\tilde{U}^{-1} = \mathrm{ut}(\tilde{L}^{-1}\Delta A\tilde{U}^{-1}).$$

Thus we obtain

$$(4.3)\qquad \widehat{\Delta L} = \tilde{L}\,\mathrm{slt}(\tilde{L}^{-1}\Delta A\tilde{U}^{-1}),$$

$$(4.4)\qquad \widehat{\Delta U} = \mathrm{ut}(\tilde{L}^{-1}\Delta A\tilde{U}^{-1})\tilde{U}.$$

Let $\tilde U_{n-1}$ denote the leading $(n-1)\times(n-1)$ block of $\tilde U$. If we write $\tilde U = \begin{bmatrix} \tilde U_{n-1} & \tilde u \\ 0 & \tilde u_{nn} \end{bmatrix}$, then from (4.3)

$$\widehat{\Delta L} = \tilde L \,\mathrm{slt}\left(\tilde L^{-1}\Delta A \begin{bmatrix} \tilde U_{n-1}^{-1} & -\tilde U_{n-1}^{-1}\tilde u/\tilde u_{nn} \\ 0 & 1/\tilde u_{nn} \end{bmatrix}\right) = \tilde L \,\mathrm{slt}\left(\tilde L^{-1}\Delta A \begin{bmatrix} \tilde U_{n-1}^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right).$$

Since $|\Delta A| \le \epsilon|\tilde L|\cdot|\tilde U|$ in (4.1), we have the following componentwise bound:

$$|\widehat{\Delta L}| \le |\tilde L|\,\mathrm{slt}\left(|\tilde L^{-1}|\cdot|\tilde L| \begin{bmatrix} |\tilde U_{n-1}|\cdot|\tilde U_{n-1}^{-1}| & 0 \\ 0 & 0 \end{bmatrix}\right)\epsilon.$$

Taking a consistent monotone matrix norm $\|\cdot\|$ on both sides, we obtain

$$\|\widehat{\Delta L}\| \le \|\,|\tilde L|\cdot|\tilde L^{-1}|\cdot|\tilde L|\,\|\cdot\|\,|\tilde U_{n-1}|\cdot|\tilde U_{n-1}^{-1}|\,\|\,\epsilon,$$

or

$$\frac{\|\widehat{\Delta L}\|}{\|\tilde L\|} \le \frac{\|\,|\tilde L|\cdot|\tilde L^{-1}|\cdot|\tilde L|\,\|}{\|\tilde L\|}\mathrm{cond}(\tilde U_{n-1}^{-1})\epsilon. \tag{4.5}$$

The right hand side can be thought of as a measure of the error in the computed L factor, and

$$\chi_L(A) \equiv \frac{\|\,|\tilde L|\cdot|\tilde L^{-1}|\cdot|\tilde L|\,\|}{\|\tilde L\|}\mathrm{cond}(\tilde U_{n-1}^{-1}) \tag{4.6}$$

is the multiplicative factor (like a sensitivity) contributing to the error and will be referred to as the "error indicator" for the computed L factor.

Similarly we can derive the first-order bound for the error in the computed U factor. With $|\Delta A| \le \epsilon|\tilde L|\cdot|\tilde U|$, we obtain from (4.4) the following componentwise bound

$$|\widehat{\Delta U}| \le \mathrm{ut}(|\tilde L^{-1}|\cdot|\tilde L|\cdot|\tilde U|\cdot|\tilde U^{-1}|)|\tilde U|\epsilon,$$

which gives

$$\|\widehat{\Delta U}\| \le \mathrm{cond}(\tilde L)\|\,|\tilde U|\cdot|\tilde U^{-1}|\cdot|\tilde U|\,\|\epsilon,$$

or

$$\frac{\|\widehat{\Delta U}\|}{\|\tilde U\|} \le \mathrm{cond}(\tilde L)\frac{\|\,|\tilde U|\cdot|\tilde U^{-1}|\cdot|\tilde U|\,\|}{\|\tilde U\|}\epsilon. \tag{4.7}$$

Here the multiplicative factor on the right hand side,

$$\chi_U(A) \equiv \mathrm{cond}(\tilde L)\frac{\|\,|\tilde U|\cdot|\tilde U^{-1}|\cdot|\tilde U|\,\|}{\|\tilde U\|}, \tag{4.8}$$

will be referred to as the error indicator for the computed U factor.

The error indicators $\chi_L(A)$ in (4.6) and $\chi_U(A)$ in (4.8) are understandable, but how to estimate $\|\,|\tilde L|\cdot|\tilde L^{-1}|\cdot|\tilde L|\,\|$ and $\|\,|\tilde U|\cdot|\tilde U^{-1}|\cdot|\tilde U|\,\|$ is still a problem we have to solve.

Since

$$\| \, |\tilde{L}| \cdot |\tilde{L}^{-1}| \cdot |\tilde{L}| \, \| \leq \|\tilde{L}\| \mathrm{cond}(\tilde{L}), \ \mathrm{cond}(\tilde{L}^{-1}) \|\tilde{L}\|, \tag{4.9}$$
$$\| \, |\tilde{U}| \cdot |\tilde{U}^{-1}| \cdot |\tilde{U}| \, \| \leq \|\tilde{U}\| \mathrm{cond}(\tilde{U}), \ \mathrm{cond}(\tilde{U}^{-1}) \|\tilde{U}\|, \tag{4.10}$$

we obtain the following simpler bounds:

$$\chi_L(A) \leq \min\{\mathrm{cond}(\tilde{L}), \mathrm{cond}(\tilde{L}^{-1})\} \mathrm{cond}(\tilde{U}_{n-1}^{-1}), \tag{4.11}$$
$$\chi_U(A) \leq \mathrm{cond}(\tilde{L}) \min\{\mathrm{cond}(\tilde{U}), \mathrm{cond}(\tilde{U}^{-1})\}. \tag{4.12}$$

These last two bounds are easy to estimate for typical consistent monotone norms, like the 1, ∞ and F-norms. They indicate that $\chi_L(A)$ is insensitive to row or column scaling on L and column scaling on U, and that $\chi_U(A)$ is insensitive to the row scaling on L and row or column scaling on U. But both of the bounds can be arbitrarily larger than the corresponding error indicators $\chi_L(A)$ and $\chi_U(A)$ due to the inequalities (4.9) and (4.10). For example, assume

$$\tilde{L} = \begin{pmatrix} 1 & 0 \\ \omega & 1 \end{pmatrix},$$

where ω is a very large number. Then for the 1, ∞ and F-norms, $\| \, |\tilde{L}| \cdot |\tilde{L}^{-1}| \cdot |\tilde{L}| \, \| / \|\tilde{L}\| = O(1)$, but $\mathrm{cond}(\tilde{L}) = O(\omega)$ and $\mathrm{cond}(\tilde{L}^{-1}) = O(\omega)$. Therefore we need to give other upper bounds which are easy to estimate, and which approximate $\| \, |\tilde{L}| \cdot |\tilde{L}^{-1}| \cdot |\tilde{L}| \, \|$ and $\| \, |\tilde{U}| \cdot |\tilde{U}^{-1}| \cdot |\tilde{U}| \, \|$ very well.

In the following we consider the ∞-norm. Let

$$D_{Lc} \equiv \mathrm{diag}(\|\tilde{L}(:,j)\|_1, \ j = 1, \ldots, n) \quad \text{and} \quad D_{Lr} \equiv \mathrm{diag}(\|\tilde{L}(i,:)\|_1, \ j = 1, \ldots, n).$$

Then with $e = (1, 1, \ldots, 1)^T \in \mathcal{R}^n$,

$$e^T |\tilde{L} D_{Lc}^{-1}| = e^T, \qquad |D_{Lr}^{-1} \tilde{L}| e = e.$$

We have

$$\begin{aligned}
\| \, |\tilde{L}| \cdot |\tilde{L}^{-1}| \cdot |\tilde{L}| \, \|_\infty &= \| \, |\tilde{L} D_{Lc}^{-1}| \cdot |D_{Lc} \tilde{L}^{-1} D_{Lr}| \cdot |D_{Lr}^{-1} \tilde{L}| e \|_\infty \\
&\leq \|\tilde{L} D_{Lc}^{-1}\|_\infty \|D_{Lc} \tilde{L}^{-1} D_{Lr}\|_\infty \\
&\leq n^2 \, \|\tilde{L} D_{Lc}^{-1}\|_1 \|D_{Lc} \tilde{L}^{-1} D_{Lr}\|_1 \\
&= n^2 \, \|e^T |\tilde{L} D_{Lc}^{-1}| \cdot |D_{Lc} \tilde{L}^{-1} D_{Lr}| \, \|_1 \\
&\leq n^3 \, \| \, |\tilde{L} D_{Lc}^{-1}| \cdot |D_{Lc} \tilde{L}^{-1} D_{Lr}| \, \|_\infty \\
&= n^3 \, \| \, |\tilde{L}| \cdot |\tilde{L}^{-1} D_{Lr}| \cdot |D_{Lr}^{-1} \tilde{L}| e \|_\infty \\
&= n^3 \| \, |\tilde{L}| \cdot |\tilde{L}^{-1}| \cdot |\tilde{L}| \, \|_\infty.
\end{aligned}$$

Therefore $\|\tilde{L} D_{Lc}^{-1}\|_\infty \|D_{Lc} \tilde{L}^{-1} D_{Lr}\|_\infty$ is a good approximation of $\| \, |\tilde{L}| \cdot |\tilde{L}^{-1}| \cdot |\tilde{L}| \, \|_\infty$. By standard matrix ∞-norm estimators, the former can be estimated in $O(n^2)$ flops.

Similarly with $D_{Uc} \equiv \mathrm{diag}(\|\tilde{U}(:,j)\|_1)$ and $D_{Ur} \equiv \mathrm{diag}(\|\tilde{U}(i,:)\|_1)$, we can show

$$\|\,|\tilde{U}|\cdot|\tilde{U}^{-1}|\cdot|\tilde{U}|\,\|_\infty \le \|\tilde{U}D_{Uc}^{-1}\|_\infty \|D_{Uc}\tilde{U}^{-1}D_{Ur}\|_\infty \le n^3\|\,|\tilde{U}|\cdot|\tilde{U}^{-1}|\cdot|\tilde{U}|\,\|_\infty,$$

where $\|\tilde{U}D_{Uc}^{-1}\|_\infty \|D_{Uc}\tilde{U}^{-1}D_{Ur}\|_\infty$ is a good approximation of $\|\,|\tilde{U}|\cdot|\tilde{U}^{-1}|\cdot|\tilde{U}|\,\|_\infty$ and can be estimated in $O(n^2)$ flops.

4.2 Effects of pivoting on the accuracy of the LU factors.

When one of the three pivoting strategies (partial pivoting, complete pivoting and rook pivoting) is used in GE, the computed LU factors satisfy

$$P(A+\Delta A)Q = \tilde{L}\tilde{U}, \qquad |P\Delta AQ| \le \epsilon|\tilde{L}||\tilde{U}|, \tag{4.13}$$

where $\epsilon = nu/(1-nu)$, P and Q are permutation matrices (for partial pivoting $Q = I$). Let the LU factorization of PAQ be $PAQ = LU$. Then the error bounds (4.5) and (4.7) still hold, except that A should be replaced by PAQ.

What effects do these three pivoting strategies have on the accuracy of the LU factors? That is the question we try to answer in this section.

For the L factor, if one of the three pivoting strategies is used in GE, then the two inequalities in (3.3) will hold. Therefore if we use the ∞-norm in (4.11) we have

$$\chi_L(PAQ) \le (2^n - 1)\mathrm{cond}_\infty(\tilde{U}_{n-1}^{-1}).$$

Here $\mathrm{cond}_\infty(\tilde{U}_{n-1}^{-1})$ can still be arbitrarily large, and we cannot say if the pivoting strategies make $\chi_L(PAQ)$ larger or smaller than $\chi_L(A)$. However if A is well-conditioned, i.e., $\kappa_\infty(A)$ is small, then from $\tilde{U} = \tilde{L}^{-1}P(A+\Delta A)Q$ we expect that $\mathrm{cond}_\infty(\tilde{U}^{-1})$ is small since $\kappa_\infty(\tilde{L})$ will usually be small. So the error in $\tilde{L}$ is usually small for this case. Even if A is ill-conditioned, but if it is due to bad column scaling, then $\mathrm{cond}_\infty(A^{-1})$ will be small, and again we expect that $\mathrm{cond}_\infty(\tilde{U}_{n-1}^{-1})$ will be small and $\tilde{L}$ will usually have high accuracy.

If rook pivoting or complete pivoting is used in GE, then (3.4) will hold. Therefore from (4.12) with the ∞-norm we obtain

$$\chi_U(PAQ) \le (2^n - 1)^2.$$

Note the bound is only a function of n. Also we believe usually $\chi_U(PAQ) \ll (2^n-1)^2$ (see the justification given in Section 3). This indicates that we can usually obtain highly accurate $\tilde{U}$ by using rook pivoting or complete pivoting. If we use partial pivoting, we cannot obtain the above result.

5 Solving $Ax = b$ by GE with rook pivoting.

The usual method for computing the solution of the linear system $Ax = b$ is to compute the LU factorization of A by partial pivoting, then solve two triangular systems. Here we consider using the rook pivoting strategy. We will give some examples to show rook pivoting is usually more accurate than partial pivoting and investigate componentwise stability of the former. Poole and Neal [15] also give a lot of numerical examples to show on average rook pivoting produces more accurate solutions than partial pivoting.

The steps of the method with one of the pivoting strategies are as follows:

Step 1: Compute $PAQ = LU$ by GE with some pivoting strategy.
Step 2: Solve the lower triangular system $Ly = Pb$.
Step 3: Solve the upper triangular system $Uz = y$.
Step 4: Compute $x = Qz$.

It follows from [12, Theorems 9.3–9.5] with slight modifications that the computed LU factors $\tilde{L}$ and $\tilde{U}$ and the computed solution $\tilde{x}$ satisfy

$$(5.1)\quad (A+\widehat{\Delta A})\tilde{x} = b, \quad |P\widehat{\Delta A}Q| \le 2\epsilon|\tilde{L}|\cdot|\tilde{U}|, \quad \|\widehat{\Delta A}\|_\infty \le 2n^2\rho\epsilon\,\|A\|_\infty,$$

where $\epsilon = nu/(1-nu)$ and ρ is the growth factor defined by (2.1). Then by the normwise perturbation theory for linear systems (see for example [12, Theorem 7.2]), we have

$$\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} \le 2n^2\rho\,\epsilon\kappa_\infty(A) + O(\epsilon^2).$$

But for partial pivoting, in practice we usually have

$$\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} \approx u\kappa_\infty(A).$$

When the LU factorization is computed by GERP, the computed factors L and U satisfy (3.1) and (3.2). According to [12, Theorem 8.7], both $Ly = Pb$ and $Uz = y$ usually have highly accurate solutions. But for partial pivoting, the computed solution of $Uz = y$ may have poor accuracy. So even if GEPP and GERP have similar growth factors, the latter is usually expected to give a solution to a linear system at least as accurate as the former. See also [14] for the importance of the stability in solving $Uz = y$. We found in our numerical experiments that the relative normwise error in a solution was always bounded by $O(u)\mathrm{cond}_\infty(A)$. Notice $\mathrm{cond}_\infty(A) \le \kappa_\infty(A)$ and the former can be much smaller than the latter if A has bad row scaling.

In order to test the backward componentwise stability of GERP, we need the following quantity, a measure of the componentwise backward error for solving $Ax = b$:

$$\omega(\tilde{x}) = \min\{\delta : (A+\Delta A)\tilde{x} = b+\Delta b,\ |\Delta A| \le \delta|A|,\ |\Delta b| \le \delta|b|\},$$

which, by the Oettli and Prager Theorem (see [12, Theorem 7.3]), is given by

$$\omega(\tilde{x}) = \max_i |(b-A\tilde{x})_i|/(|A|\cdot|\tilde{x}|+|b|)_i.$$

By componentwise perturbation analysis (see [12, p. 134]), we have

$$\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} \le \frac{2\omega(\tilde{x})\mathrm{cond}_\infty(A,x)}{1-\omega(\tilde{x})\mathrm{cond}_\infty(A)} \le \frac{2\omega(\tilde{x})\mathrm{cond}_\infty(A)}{1-\omega(\tilde{x})\mathrm{cond}_\infty(A)},$$

where

$$\mathrm{cond}_\infty(A,x) := \frac{\| |A^{-1}|\cdot|A|\cdot|\tilde{x}| \|_\infty}{\|\tilde{x}\|_\infty} \leq \mathrm{cond}_\infty(A).$$

We found in our numerical experiments that for rook pivoting usually $\omega(\tilde{x}) = O(u)$, i.e., GERP usually has componentwise backward stability. Therefore the error $\|x-\tilde{x}\|_\infty/\|x\|_\infty$ is bounded by $O(u)\mathrm{cond}_\infty(A)$. But for an example of Kahan, we found that $\omega(\tilde{x})$ could be much larger than u. However, the error in the solution was still bounded by $O(u)\mathrm{cond}_\infty(A)$. In the following we first give some examples to illustrate our findings, then give an analysis to justify the findings.

Table 5.1: Errors in computed solutions and backward error bounds.

			Partial pivoting		Rook pivoting	
i,j	$\kappa_\infty(A)$	$\mathrm{cond}_\infty(A)$	$\omega(\tilde{x})$	error	$\omega(\tilde{x})$	error
$1,1$	5.4e+10	3.7e+01	7.2e−17	4.4e−16	8.0e−17	4.4e−16
$1,5$	9.5e+10	6.8e+01	5.6e−07	5.1e−06	6.4e−17	1.1e−15
$1,10$	1.9e+11	1.5e+02	6.4e−08	3.9e−07	9.3e−17	4.4e−16
$5,1$	2.9e+10	2.4e+01	1.9e−16	1.3e−15	9.3e−17	4.4e−16
$5,5$	6.6e+11	5.3e+02	1.6e−07	1.9e−05	6.0e−17	2.7e−15
$5,10$	1.3e+11	1.2e+02	2.8e−07	3.8e−06	6.4e−17	2.2e−16
$10,1$	4.3e+10	3.3e+01	9.3e−17	6.7e−16	1.3e−16	4.4e−16
$10,5$	1.0e+11	7.4e+01	6.9e−08	1.2e−06	1.1e−16	1.2e−15
$10,10$	6.5e+10	5.3e+01	3.0e−07	1.5e−06	9.3e−17	8.9e−16

1. The matrices are generated by a 10×10 random matrix produced by the MATLAB built-in function `randn`, with the (i,j) element changed to 10^{10} for each pair $i,j=1,5,10$. The exact solution x has all unit elements and b is defined by $b=Ax$. The results are displayed in Table 5.1, where error $=\|x-\tilde{x}\|_\infty/\|x\|_\infty$. For each case and for both partial pivoting and rook pivoting, the growth factor $\rho\approx 1$. The results by complete pivoting are similar to those by rook pivoting.
2. Each matrix has the form of $A=DB$, where B is equal to an identity matrix plus very small random entries, around 10^{-7}, i.e.,

```
B=eye(n)+1.e-7*randn(n,n),
```

and D is a diagonal matrix with entries scaled geometrically from 1 up to 10^{14}, i.e, $D=\mathrm{diag}(10^{14(i-1)/(n-1)})$. The A matrices have $\kappa_\infty(A)\approx 10^{14}$ and $\mathrm{cond}_\infty(A)\approx 1.0$. The results for $n=10,20,\ldots,100$ are reported in Table 5.2. The growth factor ρ for each case is close to 1. This example is given in [5, Example 2.5] and it was used to show that sometimes GECP is more accurate than GEPP, but there is no any explanation there.

3. The example of Kahan (see [12, p.136]):

$$A=\begin{bmatrix} 2 & -1 & 1\\ -1 & \delta & \delta\\ 1 & \delta & \delta\end{bmatrix},\qquad x=\begin{bmatrix}\delta\\ -1\\ 1\end{bmatrix},\qquad b=\begin{bmatrix}2(1+\delta)\\ -\delta\\ \delta\end{bmatrix},$$

Table 5.2: Errors in computed solutions and backward error bounds.

	Partial pivoting		Rook pivoting	
n	$\omega(\tilde{x})$	error	$\omega(\tilde{x})$	error
10	1.5e−09	2.9e−09	1.1e−16	2.2e−16
20	6.3e−09	1.3e−08	1.9e−16	4.4e−16
30	1.1e−08	2.2e−08	2.4e−16	4.4e−16
40	6.3e−09	1.3e−08	3.4e−16	5.6e−16
50	2.9e−08	5.8e−08	2.8e−16	4.4e−16
60	1.6e−08	3.3e−08	3.2e−16	6.7e−16
70	6.9e−09	1.4e−08	5.3e−16	1.1e−15
80	9.2e−09	1.8e−08	4.5e−16	8.9e−16
90	1.3e−08	2.7e−08	5.7e−16	1.1e−15
100	1.1e−08	2.2e−08	7.1e−16	1.4e−15

where we take $\delta = 10^{-8}$, then $\kappa_\infty(A) \approx 2.0 \times 10^8$, $\mathrm{cond}_\infty(A) \approx 5.1 \times 10^7$, $\mathrm{cond}_\infty(A, x) \approx 2.5$. For this example, the permutation matrix $Q = I$ in GERP, i.e., GERP and GEPP give the same results. For GERP or GEPP, our computation shows $\omega(\tilde{x}) \approx 1.9 \times 10^{-9}$ and $\|x - \tilde{x}\|_\infty / \|x\|_\infty \approx 1.3 \times 10^{-9}$. Even though this example shows that GERP is not componentwise backward stable, we observe the normwise error in the computed solution is still bounded by $O(u)\mathrm{cond}_\infty(A)$.

Now we give a forward error analysis to explain this phenomenon. From (4.13) we obtain

$$|\tilde{L}| = |P(A + \Delta A)Q\tilde{U}^{-1}| \leq |PAQ| \cdot |\tilde{U}^{-1}| + \epsilon |\tilde{L}| \cdot |\tilde{U}| \cdot |\tilde{U}^{-1}|.$$

Then for the $\widehat{\Delta A}$ in (5.1) we have

$$|\widehat{\Delta A}| \leq 2\epsilon P^T \cdot |\tilde{L}| \cdot |\tilde{U}| \cdot Q^T \leq 2\epsilon \, |A| \cdot Q \cdot |\tilde{U}^{-1}| \cdot |\tilde{U}| \cdot Q^T + O(\epsilon^2). \tag{5.2}$$

From $(A + \widehat{\Delta A})\tilde{x} = b$ and $Ax = b$ it follows that $A(\tilde{x} - x) = -\widehat{\Delta A}\tilde{x}$. This with (5.2) yields the componentwise bound

$$\begin{aligned} |\tilde{x} - x| &\leq |A^{-1}| \cdot |\widehat{\Delta A}| \cdot |\tilde{x}| \\ &= |A^{-1}| \cdot |\widehat{\Delta A}| \cdot (|x| + O(u)) \\ &\leq 2\epsilon \, |A^{-1}| \cdot |A| \cdot Q \cdot |\tilde{U}^{-1}| \cdot |\tilde{U}| \cdot Q^T \cdot |x| + O(\epsilon^2). \end{aligned}$$

Thus taking the ∞-norm gives

$$\frac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty} \leq 2\epsilon \, \mathrm{cond}_\infty(A) \, \mathrm{cond}_\infty(\tilde{U}) + O(\epsilon^2). \tag{5.3}$$

As we know, if we use rook pivoting (or complete pivoting) we have $\mathrm{cond}_\infty(\tilde{U}) \leq 2^n - 1$ (see (3.4)), and usually $\mathrm{cond}_\infty(\tilde{U}) \ll 2^n - 1$. Notice $\epsilon = nu/(1 - nu)$, so the relative normwise error $\|x - \tilde{x}\|_\infty / \|x\|_\infty$ is usually bounded by $O(u)\mathrm{cond}_\infty(A)$.

This is exactly what we have observed. Certainly the error bound can occasionally be much larger than $u\mathrm{cond}_\infty(A)$ if $\mathrm{cond}_\infty(\tilde{U})$ is much larger than 1. Since $\mathrm{cond}_\infty(A)$ is independent of row scaling on A, if A is ill-conditioned due to bad row scaling, we can still get a highly accurate solution if we use GERP (or GECP) and we do not need to scale the rows of A before we apply GE. For partial pivoting, $\mathrm{cond}(\tilde{U})$ can be arbitrarily large. Thus if A has bad row scaling we should not expect to get a highly accurate solution.

Partial pivoting chooses a pivot element in a column, so it is also called column pivoting. Sometimes one uses row pivoting — columns are interchanged so that each pivot element is the largest in its row. Obviously rook pivoting is a combination of column pivoting and row pivoting. In [16] Skeel studied the effect of scaling on stability and accuracy of GE with row pivoting. We found from [16, Theorem 5.6] that a similar conclusion to our findings above could be drawn. But our analysis is much simpler and the forward error bound (5.3) holds for any pivoting strategy.

From Tables 5.1 and 5.2, we see for those examples rook pivoting has better componentwise backward stability than partial pivoting. For the test matrices from the collections of Higham [11], we found that for both partial pivoting and rook pivoting $\omega(\tilde{x}) \approx 10^{-16}$ or $\omega(\tilde{x}) = 0$. i.e., both partial pivoting and rook pivoting are componentwise stable for those matrices. Certainly more numerical experiments and a complete analysis are needed to study the componentwise backward stability of the two pivoting strategies.

Recently Ashcraft et al. [1] extended the rook pivoting strategy to the symmetric case. For the stability analysis and error analysis of this extended pivoting strategy see Cheng [4].

6 Computing the inverse by GE with rook pivoting.

In some applications like statistics, a matrix inverse needs to be computed. There are many different ways to compute matrix inverses, and most of the methods involve the LU factorization. Du Croz and Higham [7] give stability analyses of four typical methods which use the LU factorization with partial pivoting. They showed that only one of the left and right residuals is guaranteed to be usually small; which one depends on whether the method is derived by solving $AX = I$ or $XA = I$, and there is little to choose between the four methods in terms of the error bounds. In this paper we consider one of the four methods which is used by LINPACK's `xGEDI`, LAPACK's `xGETRI`, and MATLAB's `INV` function, but with rook pivoting, and we show that both left and right residuals are usually small. The method is derived by solving $XA = I$ and is as follows:

Method M.
step 1: Compute the LU factorization of A by some pivoting strategy: $PAQ = LU$.
step 2: Compute U^{-1} and then solve for Y the equation $YL = U^{-1}$.
step 3: Compute $X = QYP$.

There are two methods for computing U^{-1}, which can be derived by solving $UX = I$ and solving $XU = I$, respectively.

Method 1.
for $j = n : -1 : 1$
  $x_{jj} = u_{jj}^{-1}$
  $X(1 : j-1, j) = -x_{jj}U(1 : j-1, j)$
  Solve $U(1 : j-1, 1 : j-1)X(1 : j-1, j) = X(1 : j-1, j)$
  by back substitution
end

Method 2.
for $j = 1 : n$
  $x_{jj} = u_{jj}^{-1}$
  $X(1 : j-1, j) = X(1 : j-1, 1 : j-1)U(1 : j-1, j)$
  $X(1 : j-1, j) = -x_{jj}X(1 : j-1, j)$
end

As in [7], we assume the matrix equation $YL = U^{-1}$ is solved by back substitution.

Let $\tilde{L}$ and $\tilde{U}$ be the computed LU factors. Then they satisfy (cf. (4.13))

$$P(A + \varDelta A)Q = \tilde{L}\tilde{U}, \qquad |P\varDelta AQ| \leq c_n u|\tilde{L}|\cdot|\tilde{U}| + O(u^2), \tag{6.1}$$

where c_n is a constant of order n, and for simplicity later we will use it to denote any constant of order n. Let X_U be the computed inverse of $\tilde{U}$. Then it can be shown (cf. [7]) that for Method 1, X_U satisfies the residual bound

$$|\tilde{U}X_U - I| \leq c_n u|\tilde{U}|\cdot|X_U| + O(u^2), \tag{6.2}$$

and for Method 2 X_U satisfies the residual bound

$$|X_U\tilde{U} - I| \leq c_n u|X_U|\cdot|\tilde{U}| + O(u^2). \tag{6.3}$$

We have assumed the matrix equation $YL = U^{-1}$ in step 2 of Method M is solved by back substitution. So the computed solution $\tilde{Y}$ satisfies

$$\tilde{Y}\tilde{L} = X_U + \Delta(\tilde{Y}, \tilde{L}), \tag{6.4}$$

where

$$|\Delta(\tilde{Y}, \tilde{L})| \leq c_n u|\tilde{Y}|\cdot|\tilde{L}| + O(u^2). \tag{6.5}$$

Let $\tilde{X}$ denote the computed inverse of X. From step 3 in Method M, we have

$$\tilde{X} = Q\tilde{Y}P. \tag{6.6}$$

If in Method M, U^{-1} is computed by Method 1, then we can show that the relative *right residual* for $\tilde{X}$ will usually be small when any one of the three pivoting strategies is used. In fact, from (6.1), (6.6) and (6.4) we have

$$P(A+\varDelta A)\tilde{X}P^T = \tilde{L}\tilde{U}Q^T\tilde{X}P^T = \tilde{L}\tilde{U}\tilde{Y} = \tilde{L}\tilde{U}\tilde{Y}\tilde{L}\tilde{L}^{-1} = \tilde{L}\tilde{U}(X_U+\Delta(\tilde{Y}, \tilde{L}))\tilde{L}^{-1}.$$

Thus

$$P(A\tilde{X}-I)P^T = \tilde{L}(\tilde{U}X_U - I)\tilde{L}^{-1} + \tilde{L}\tilde{U}\Delta(\tilde{Y},\tilde{L})\tilde{L}^{-1} - P\Delta A\tilde{X}P^T. \tag{6.7}$$

Then using (6.2), (6.5), (6.1), (6.6) and (6.4), we obtain

$$\begin{aligned}
|P(A\tilde{X}-I)P^T| \;&\leq\; c_n u|\tilde{L}|\cdot|\tilde{U}|\cdot|X_U|\cdot|\tilde{L}^{-1}| + c_n u|\tilde{L}|\cdot|\tilde{U}|\cdot|\tilde{Y}|\cdot|\tilde{L}|\cdot|L^{-1}| \\
&\quad + c_n u|\tilde{L}|\cdot|\tilde{U}|\cdot|\tilde{Y}| + O(u^2). \\
&\leq\; c_n u|\tilde{L}|\cdot|\tilde{U}|(|\tilde{Y}|\cdot|\tilde{L}| + c_n u|\tilde{Y}|\cdot|\tilde{L}|)|\tilde{L}^{-1}| \\
&\quad + c_n u|\tilde{L}|\cdot|\tilde{U}|\cdot|\tilde{Y}|\cdot|\tilde{L}|\cdot|\tilde{L}^{-1}| + c_n u|\tilde{L}|\cdot|\tilde{U}|\cdot|\tilde{Y}| + O(u^2). \\
&\leq\; 3c_n u|\tilde{L}|\cdot|\tilde{U}|\cdot|\tilde{Y}|\cdot|\tilde{L}|\cdot|\tilde{L}^{-1}| + O(u^2),
\end{aligned}$$

Therefore

$$\|A\tilde{X}-I\|_\infty \leq 3c_n u\,\mathrm{cond}_\infty(\tilde{L}^{-1})\|\,|\tilde{L}|\cdot|\tilde{U}|\,\|_\infty\|\tilde{X}\|_\infty + O(u^2).$$

Notice if any one of the three pivoting strategies is used, $\mathrm{cond}_\infty(\tilde{L}^{-1}) \leq n2^n$ (see (3.3)) and usually $\mathrm{cond}_\infty(\tilde{L}^{-1}) \ll n2^n$. Also we know usually $\|\,|\tilde{L}||\tilde{U}|\,\|_\infty \approx \|A\|_\infty$ for any one of the three pivoting strategies. So the right relative residual $\frac{\|A\tilde{X}-I\|_\infty}{\|A\|_\infty\|\tilde{X}\|_\infty}$ is usually small. The numerical experiments given in [7] confirmed this conclusion for partial pivoting.

If in Method M, U^{-1} is computed by Method 2, then following the proof in [7], the computed inverse $\tilde{X}$ of A satisfies the left residual bound

$$\|\tilde{X}A-I\|_\infty \leq c_n' u\|\tilde{X}\|_\infty\|\,|\tilde{L}|\cdot|\tilde{U}|\,\|_\infty + O(u^2).$$

Since usually $\|\,|\tilde{L}||\tilde{U}|\,\|_\infty \approx \|A\|_\infty$, the relative *left residual* $\frac{\|\tilde{X}A-I\|_\infty}{\|\tilde{X}\|_\infty\|A\|_\infty}$ is small.

The numerical examples given in [7] suggest if U^{-1} is computed by Method 1, the relative *left residual* may be large and if it is computed by Method 2, the relative *right residual* may be large. Remember in [7], only partial pivoting is considered. Our numerical experiments showed if rook pivoting is used, these two relative residuals are still small. In the following we give an analysis to explain this.

Suppose U^{-1} is computed by Method 1. We now show that the left residual is usually small. Since the computed inverse X_U of U satisfies (6.2), we have

$$|X_U\tilde{U}-I| = |\tilde{U}^{-1}(\tilde{U}X_U - I)\tilde{U}| \leq c_n u|\tilde{U}^{-1}|\cdot|\tilde{U}|\cdot|X_U|\cdot|\tilde{U}| + O(u^2). \tag{6.8}$$

But for rook pivoting (or complete pivoting), $\mathrm{cond}_\infty(\tilde{U}) \leq 2^n - 1$ and usually $\mathrm{cond}_\infty(\tilde{U}) \ll 2^n - 1$. So we would expect that the left residual for X_U is usually small. That is the key to showing that the left residual for $\tilde{X}$ is usually small. From (6.6), (6.1) and (6.4) we have

$$Q^T\tilde{X}(A+\Delta A)Q = \tilde{Y}\tilde{L}\tilde{U} = X_U\tilde{U} + \Delta(\tilde{Y},\tilde{L})\tilde{U}.$$

Thus

$$Q^T(\tilde{X}A-I)Q = X_U\tilde{U} - I + \Delta(\tilde{Y},\tilde{L})\tilde{U} - Q^T\tilde{X}\Delta AQ.$$

Therefore from (6.8), (6.4), (6.6) and (6.1) we have

$$\begin{aligned}
&|Q^T(\tilde{X}A - I)Q| \\
\le\; & c_n u|\tilde{U}^{-1}|\cdot|\tilde{U}|\cdot|X_U|\cdot|\tilde{U}| + c_n u|\tilde{Y}|\cdot|\tilde{L}|\cdot|\tilde{U}| + c_n u|\tilde{Y}|\cdot|\tilde{L}|\cdot|\tilde{U}| + O(u^2) \\
\le\; & c_n u|\tilde{U}^{-1}|\cdot|\tilde{U}|(|\tilde{Y}|\cdot|\tilde{L}| + c_n u|\tilde{Y}|\cdot|\tilde{L}|)|\tilde{U}| + 2c_n u|\tilde{Y}|\cdot|\tilde{L}|\cdot|\tilde{U}| + O(u^2) \\
\le\; & 3c_n u|\tilde{U}^{-1}|\cdot|\tilde{U}|\cdot|\tilde{Y}|\cdot|\tilde{L}|\cdot|\tilde{U}| + O(u^2).
\end{aligned}$$

This gives

$$\|\tilde{X}A - I\|_\infty \le 3c_n u\, \mathrm{cond}_\infty(\tilde{U})\|\tilde{X}\|_\infty \|\, |\tilde{L}|\cdot|\tilde{U}|\, \|_\infty + O(u^2).$$

Since $\mathrm{cond}_\infty(\tilde{U}) \le 2^n - 1$ and usually $\mathrm{cond}_\infty(\tilde{U}) \ll 2^n - 1$, we would expect that the relative left residual for $\tilde{X}$ is usually small.

Suppose U^{-1} is computed by Method 2, then (6.3) holds. We now show that the right residual is usually small when rook pivoting (or complete pivoting) is used. From (6.7) we have

$$P(A\tilde{X} - I)P^T = \tilde{L}\tilde{U}(X_U\tilde{U} - I)\tilde{U}^{-1}\tilde{L}^{-1} + \tilde{L}\tilde{U}\Delta(\tilde{Y}, \tilde{L})L^{-1} - P\varDelta A\tilde{X}P^T.$$

Then using (6.3), (6.1) and (6.4), we obtain

$$\begin{aligned}
(6.9)\quad |P(A\tilde{X} - I)P^T| \;\le\;& c_n u|\tilde{L}\tilde{U}|\cdot|X_U|\cdot|\tilde{U}|\cdot|(\tilde{L}\tilde{U})^{-1}| \\
&+ c_n u|\tilde{L}\tilde{U}|\cdot|\tilde{Y}|\cdot|\tilde{L}|\cdot|\tilde{L}^{-1}| + c_n u|\tilde{L}|\cdot|\tilde{U}|\cdot|\tilde{Y}| + O(u^2).
\end{aligned}$$

Now we have to give bounds on $|X_U|$ and $|(\tilde{L}\tilde{U})^{-1}|$. Since

$$X_U = (X_U\tilde{U} - I)\tilde{U}^{-1} + \tilde{U}^{-1},$$

we have with (6.3)

$$(6.10)\qquad |X_U| \le |\tilde{U}^{-1}| + O(u).$$

From (6.4) we have

$$\tilde{Y}\tilde{L}\tilde{U} = I + (X_U\tilde{U} - I) + \varDelta(\tilde{Y}, \tilde{L})\tilde{U},$$

which with (6.3) and (6.5) gives

$$(6.11)\qquad |(\tilde{L}\tilde{U})^{-1}| \le |\tilde{Y}| + O(u).$$

Therefore from (6.9), (6.10) and (6.11) it follows that

$$|P(A\tilde{X} - I)P^T| \le c_n u|\tilde{L}\tilde{U}|(|\tilde{U}^{-1}|\cdot|\tilde{U}|\cdot|\tilde{Y}| + |\tilde{Y}|\cdot|\tilde{L}|\cdot|\tilde{L}^{-1}|) + c_n u|\tilde{L}|\cdot|\tilde{U}|\cdot|\tilde{Y}| + O(u^2).$$

This gives

$$\|A\tilde{X} - I\|_\infty \le c_n u\, (\mathrm{cond}_\infty(\tilde{U}) + \mathrm{cond}_\infty(\tilde{L}^{-1}))\|\, |\tilde{L}||\tilde{U}|\, \|_\infty \|X\|_\infty + O(u^2).$$

But $\mathrm{cond}_\infty(\tilde{U}) \le 2^n - 1$ and $\mathrm{cond}_\infty(\tilde{L}^{-1}) \le 2^n - 1$, and usually we have $\mathrm{cond}_\infty(\tilde{U}) \ll 2^n - 1$ and $\mathrm{cond}_\infty(\tilde{L}^{-1}) \ll 2^n - 1$, so we would expect that the relative right residual for $\tilde{X}$ is usually small.

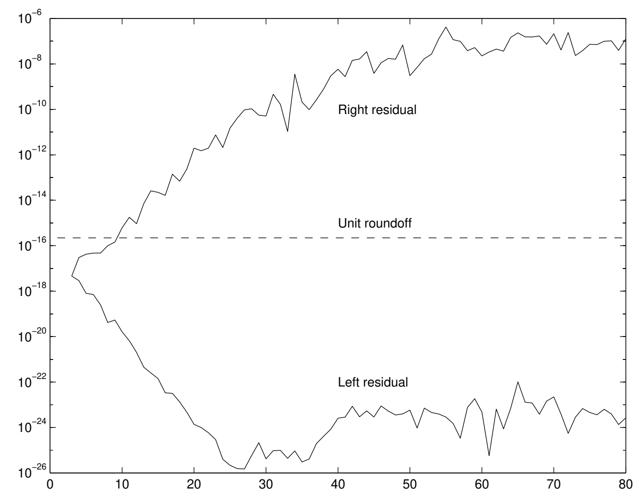

Figure 6.1: Residuals for inverses computed by MATLAB's INV function.

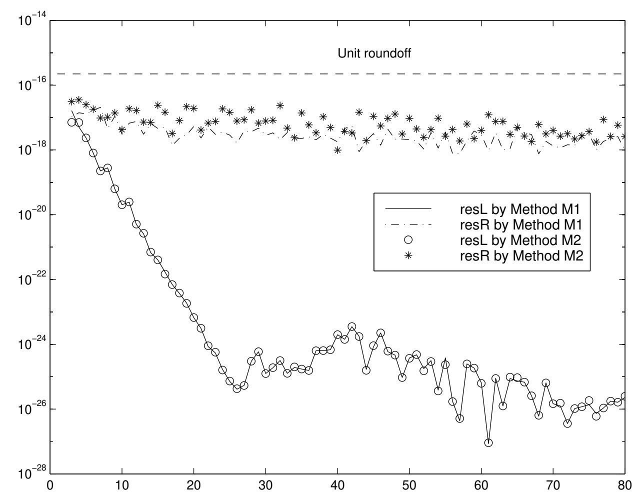

Figure 6.2: Residuals for inverses computed by Methods M1 and M2 with GERP.

We now give some numerical examples illustrating the relative left residual resL and the relative right residual resR, where

$$\mathrm{resL} = \frac{\|\tilde{X}A - I\|_\infty}{\|\tilde{X}\|_\infty \|A\|_\infty}, \qquad \mathrm{resR} = \frac{\|A\tilde{X} - I\|_\infty}{\|A\|_\infty \|\tilde{X}\|_\infty}.$$

We use Method M1 to denote Method M when Method 1 is used to compute U^{-1}, and Method M2 to denote Method M when Method 2 is used to compute U^{-1}.

Table 6.1: Residuals for inverses computed by Methods M1 and M2.

Partial pivoting				Rook pivoting			
Method M1		Method M2		Method M1		Method M2	
resL	resR	resL	resR	resL	resR	resL	resR
2.8e−05	6.7e−18	8.7e−18	6.4e−06	3.9e−18	6.7e−18	3.9e−18	9.5e−18
1.2e−05	5.6e−23	2.3e−17	3.2e−16	1.2e−17	4.4e−23	1.2e−17	8.5e−19
9.5e−07	1.8e−20	3.1e−17	9.5e−09	7.3e−18	4.8e−21	6.9e−18	4.4e−20
6.0e−16	6.9e−18	1.5e−17	2.1e−10	4.6e−18	1.9e−18	4.6e−18	3.8e−18
2.2e−17	8.9e−20	2.3e−17	1.7e−12	5.4e−18	1.0e−19	5.4e−18	5.9e−18
2.1e−14	9.5e−18	6.9e−18	3.5e−12	6.0e−18	5.3e−18	6.0e−18	7.0e−18
2.2e−10	2.5e−28	5.7e−18	1.6e−20	5.2e−18	4.8e−28	5.2e−18	2.4e−28
4.3e−17	9.2e−18	5.4e−18	1.9e−08	6.0e−18	7.5e−18	6.0e−18	5.8e−18
6.4e−05	7.9e−36	3.1e−17	1.9e−32	1.3e−17	8.8e−36	1.3e−17	5.6e−36
2.0e−04	6.1e−22	2.0e−17	1.7e−07	6.4e−18	1.0e−21	6.4e−18	2.4e−20

1. A_n is the upper triangular QR factor of the $n \times n$ Vandermonde matrix based on equispaced points on $[0, 1]$, $n = 1:80$. A_n can be generated by MATLAB command `triu(qr(vand(n))`, where `vand` is a routine from the Test Matrix Toolbox by Higham [11]. The residuals for inverses computed by MATLAB's `INV` function are reported in Figure 6.1, where the horizontal axis represents the matrix dimension n, and the vertical axis represents the residuals for inverses. `INV` adopted Method M2, where partial pivoting is used in the computation of the LU factorization. The residuals for inverses computed by Method M1 and Method M2 are reported in Figure 6.2, where rook pivoting is used in the computation of the LU factorization. The example was given by Higham in [12], and Figure 6.1 appears on both the front cover and p. 264 of [12].

2. $A = LU$, where L is the lower triangular factor from GEPP on a random 10×10 matrix generated by MATLAB's function `randn`, and U is generated as the twentieth power of an upper triangular part of a random 10×10 matrix produced by `randn`. A similar example is given in [7]. The results for 10 different runs are reported in Table 6.1.

These two examples confirm our theoretical finding that rook pivoting can usually make both the left and right residuals small no matter Method 1 or Method 2 is used to compute U^{-1} in Method M.

7 Conclusion.

We have shown that GE with rook pivoting or complete pivoting is superior to GE with partial pivoting in three aspects. If we use rook pivoting (or complete pivoting), usually the U factor of the LU factorization will have high accuracy, the computed solution of a linear system will have high accuracy if the matrix is ill-conditioned due to bad row scaling, and both the left and right residuals for the computed inverse will be small.

Foster's experiments [9] suggested that for a dense problem on a serial computer rook pivoting is usually almost as efficient as partial pivoting, even though in some extreme cases it is more close to complete pivoting in terms of efficiency. His experiments also suggested the stability of rook pivoting is comparable with that of complete pivoting. Our results give other justifications that rook pivoting is a good alternative to partial pivoting and complete pivoting.

Acknowledgments.

The author is grateful to Nick Higham, Chris Paige and two referees for their valuable comments and suggestions on this work. The author is also indebted to Leslie Foster for giving a copy of reference [9] and for a helpful discussion with him, and to one of the referees for pointing out the new reference [15] during the revision process of this paper.

REFERENCES

1. C. Ashcraft, R. G. Grimes, and J. G. Lewis, *Accurate symmetric indefinite linear equation solvers*, SIAM J. Matrix Anal. Appl. 20 (1998), pp. 513–561.
2. A. Barrlund, *Perturbation bounds for the LDL^H and the LU factorizations*, BIT, 31 (1991), pp. 358–363.
3. X.-W. Chang and C. C. Paige, *On the sensitivity of the LU factorization*, BIT, 38 (1998), pp. 486–501.
4. S. H. Cheng, *Symmetric Indefinite Matrices: Linear System Solvers and Modified Inertia Problems*, Ph.D. Thesis, Department of Mathematics, University of Manchester, 1998.
5. J. Demmel, *Applied Numerical Linear Algebra*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997.
6. J. Demmel, M. Gu, S. Eisenstat, I. Slappničar, K. Veselić, and Z. Dramač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl. 299 (1999), pp. 21–80.
7. J. J. Du Croz and N. J. Higham, *Stability of methods for matrix inversion*, IMA J. Numer. Anal., 12 (1992), pp. 1–19.
8. L. V. Foster, *Gaussian elimination with partial pivoting can fail in practice*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1354–1362.
9. L. V. Foster, *The growth factor and efficiency of Gaussian elimination with rook pivoting*, J. Comput. Applied Math., 86 (1997), pp. 177–194.
10. G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
11. N. J. Higham, *The Test Matrix Toolbox for* MATLAB*, version 3.0*, Numerical Analysis Report No. 265, University of Manchester, Manchester, 1995.
12. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.
13. L. Neal and G. Poole, *A geometric analysis of Gaussian elimination, II*, Linear Algebra Appl., 173 (1992), pp. 239–264.

14. G. Poole and L. Neal, *A geometric analysis of Gaussian elimination, I*, Linear Algebra Appl., 149 (1991), pp. 249–272.
15. G. Poole and L. Neal, *The rook's pivoting strategy*, *J. Comp. Appl. Math.*, 123 (2000), pp. 353–369.
16. R. D. Skeel, *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.
17. G. W. Stewart, *On the perturbation of LU, Cholesky, and QR factorizations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1141–1146.
18. G. W. Stewart, *On the perturbation of LU and Cholesky factors*, IMA J. Numer. Anal., 17 (1997), pp. 1–6.
19. J.-G. Sun, *Componentwise perturbation bounds for some matrix decompositions*, BIT, 32 (1992), pp. 702–714.
20. L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
21. J. H. Wilkinson, *Error analysis of direct methods for matrix inversion*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 162–195.