The effect of reordering strategies on rounding errors in large, sparse equation systems

A. Ernst and W.-D. Schuh

Abstract The effect of reordering strategies on the rounding errors is considered for the factorization and solution of sparse symmetric systems. On the one hand, a reduction of rounding errors can be expected, because the number of floating point operations decreases. On the other hand, the clustering of neighboring parameters and therefore the fixing of the sequence of parameter elimination may result in numerical instabilities. These effects are demonstrated for sparse covariance matrices in Wiener filtering. In particular Cholesky factorization and profile reordering in conjunction with envelope storage schemes are examined.

1 Introduction

In this work we investigate the hypothesis that reordering the sequence of unknown parameters of a sparse equation system has no negative effect on the rounding errors. In priniciple the sequence of the elimination of unknowns is subject to an appropriate pivoting strategy to deal with numerical instabilities. Strongly correlated parameters are separated by reordering the sequence of parameter elimination. In contrast to the pivoting strategy the reordering scheme for sparse systems aims at a clustering of neighbored data points. This yields a small profile and only few fill-ins during the solution process (Ernst, 2009). From the numerical point of view reordering counteracts pivoting. As a typical and also most critical application we have a look at Wiener filtering and other prediction processes where large covariance matrices are generated. Compactly supported covariance functions in 2D (Sansò and Schuh, 1987) and 3D (Gaspari and Cohn, 1999; Gaspari *et al.*, 2006; Moreaux, 2008) allow for a sparse representation of the covariance information considering the positive definiteness. Naturally, we exploit the sparse structure of the covariance matrices as much as possible by an efficient reordering algorithm (e.g. reversed Cuthill-McKee (Gibbs *et al.*, 1976) or banker's algorithm (Snay, 1976)) and an appropriate storage schema (Ernst, 2009). As outlined in Schuh (1991) the numerical stability of covariance matrices in prediction procedures is basically influenced by the shape of the covariance function, the variance of the uncorrelated noise and the data distribution. Especially neighboring data points cause numerical problems. To study the numerical behaviour in detail a specific rounding error analysis is necessary. Whereas norm-based perturbation bounds (Stewart, 1973) are focused on the global assessment of algorithmic processes and ignore the sparsity of a system, a stochastic approach (Meissl, 1980) allows for an individual handling.

The paper is organized as follows. Section 2 defines some fundamental terms concerning rounding error analysis. Section 3 presents norm-based rounding error analysis applied to Cholesky's algorithm and introduces the stochastic approach for a precise rounding error analysis and an algorithmic procedure to overcome the recursive variance propagation within Cholesky factorization. Section 4 gives an example. The paper finishes with conclusions.

2 Rounding error analysis

For an efficient solution on a computer each floating point number d is represented by its machine repre-

Andreas Ernst, Wolf-Dieter Schuh
Institut für Geodäsie und Geoinformation der Universität Bonn, D-53115 Bonn, Nussallee 17, Germany

sentation $\overline{d}$. Today the widely-used IEEE standard 754 defines the representation, rounding algorithms, mathematical operations and exception handling for floating point arithmetics (IEEE , 2008). A floating point number d in binary coded 64-bit (double precision) representation consists of

$$\overline{d} = (-1)^s \cdot m \cdot b^q \,, \tag{1}$$

where s denotes a binary digit for the sign of the number, m the mantissa with $\tau = 53$ binary digits, b the basis 2, and q the exponent with 10 binary digits and a given bias. The relative error

$$\left| \frac{\overline{d} - d}{d} \right| \leq \varepsilon_m, \tag{2}$$

defines the unit roundoff or machine epsilon ε_m. This quantity depends on the number of digits of the mantissa τ and the rounding procedure. True rounding (rounding to nearest) yields $\varepsilon_m = 2^{-\tau}$. A mapping error occurs also during each arithmetic operation. The computer evaluates the computed function $\overline{f}(\overline{d})$ instead of the mathematical function $f(d)$.

A rounding error analysis provides information about the perturbance measured by the size of $|f(d) - \overline{f}(\overline{d})|$, the difference between the mathematically rigorous result $f(d)$ and the function $\overline{f}(\overline{d})$ evaluated with machine numbers. Expanding this norm by plus minus $f(\overline{d})$ we get the inequality

$$|f(d) - \overline{f}(\overline{d})| \leq |f(d) - f(\overline{d})| + |f(\overline{d}) - \overline{f}(\overline{d})| \,. \tag{3}$$

The first absolute term on the right-hand side of inequality 3 characterizes the *stability of the problem* closely connected with the condition of the problem, whereas the second term contains information about the *stability of the algorithm*, where beside the condition also the order and number of operations in the algorithm has to be taken into account (Dahmen and Reusken , 2008).

3 Rounding error analysis applied to Cholesky's algorithm

Without restricting the generality we focus our investigation on the Cholesky solution of an n-dimensional equation system $\mathbf{N}\mathbf{x} = \mathbf{y}$, where the positive definite, symmetric matrix $\mathbf{N}$ is factorized by $\mathbf{N} = \mathbf{R}^T\mathbf{R}$ into a unique upper triangular matrix $\mathbf{R}$ with positive diagonal elements. For a given right hand side $\mathbf{y}$ the unknown parameter vector $\mathbf{x}$ is computed by the solution of two triangular systems. In the forward substitution step $\mathbf{R}^T\mathbf{z} = \mathbf{y}$ the auxiliary vector $\mathbf{z}$ is determined and after this the unknown parameter vector $\mathbf{x}$ results from the backward substitution step $\mathbf{R}\mathbf{x} = \mathbf{z}$.

In general the effect of rounding errors in a triangular factorization process can be measured indirectly by an estimation of the coefficients of the disturbed system $(\mathbf{N} + \Delta\mathbf{N})\overline{\mathbf{x}} = \mathbf{y}$, which are given by

$$|\Delta n_{ij}| \leq (c_1 n + 2c_2 n^2 + c_2^2 n^3 \varepsilon_m) \max_{i,j} |n_{ij}| \; g \;\; \varepsilon_m \tag{4}$$

where c_1 and c_2 are constants of the order unity and g denotes the growth factor, which is defined by half of the magnitude of the largest number occurring during the whole computation divided by the largest absolute value in $\mathbf{N}$ (Stewart, 1973, Theorem 5.3, p. 155). Applying the propagation of relative errors in linear equation systems

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{N}\|\|\mathbf{N}^{-1}\| \left(\frac{\|\Delta\mathbf{N}\|}{\|\mathbf{N}\|} + \frac{\|\Delta\mathbf{y}\|}{\|\mathbf{y}\|} \right) \tag{5}$$

(cf. Kreyszig, 1993, p. 998) the disturbances in $\Delta\mathbf{N}$ of eq. (4) can be propagated to the relative disturbances of the solution vector. Introducing the norm $\|\mathbf{N}\|$ by the *infinity norm* $\|\mathbf{N}\| = n \max_{i,j} |n_{ij}|$ and substitute eq. (4) in eq. (5) yields

$$\frac{\|\mathbf{x} - \overline{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|\mathbf{N}\|\|\mathbf{N}^{-1}\|(c_1 + 2c_2 n + c_2^2 n^2 \varepsilon_m) \; g \; \varepsilon_m \,. \tag{6}$$

In contrast to LU factorization strategies, the growth factor of Cholesky decomposition is not affected by the pivoting strategy and is bounded by $g \leq 1$. The influence of rounding errors is dominated by the linear term $2c_2 n$. This term is mainly caused by the accumulation of the scalar products, and can be reduced by a higher precision in the computation of the scalar product (Stewart, 1973, p. 156). However, also sparsity reduces the number of operations and may have a positive influence on the rounding errors.

To allow for an individual analysis Meissl (1980) introduced a stochastic approach to estimate the rounding error for very large networks in particular for the adjustment of the US ground-control network. The rounding error ε is considered a random variable and defined by its expectation $E\{\varepsilon\}$ and variance $\sigma^2\{\varepsilon\}$. Table 1 contains the expectation and variance for the

arithmetic operations used in the Cholesky algorithm. The expectation depends on the rounding algorithm. In the IEEE 754 definitions true rounding is implemented, so in this case no bias occurs. The variances $\sigma^2\{\varepsilon\}$ of the individual operations are given in the last column of table 1. The variance depends on the factor c, which is an operation dependent number, and on the machine epsilon ε_m. The factor $\frac{1}{\sqrt{12}}$ is defined by the variance of a uniformly distributed random variable. The factor c characterizes the maximum number of digits that are lost during the operation and depends for the addition/subtraction on the maximum of the input values as well as on the result of the operation. Within the other operations of multiplication, division and square root the factor c depends only on the magnitude of the result.

Table 1 Stochastical description for rounding errors of arithmetical operations. γ denotes the smallest integer power satisfying the inequality.

Operation	ε	$E\{\varepsilon\}$	$\sigma\{\varepsilon\} = \frac{c}{\sqrt{12}}\,\varepsilon_m$
summation	$\varepsilon^{(a)}$	0	$c = 2^\gamma > \max(\lvert a\rvert, \lvert b\rvert, \lvert a+b\rvert)$
subtraction	$\varepsilon^{(s)}$	0	$c = 2^\gamma > \max(\lvert a\rvert, \lvert b\rvert, \lvert a-b\rvert)$
multiplication	$\varepsilon^{(m)}$	0	$c = 2^\gamma > \lvert a\cdot b\rvert$
division	$\varepsilon^{(d)}$	0	$c = 2^\gamma > \lvert a/b\rvert$
square root	$\varepsilon^{(sq)}$	0	$c = 2^\gamma > \lvert\sqrt{a}\rvert$

In contrast to Meissl's approach where a rough estimation of the number of operations and the magnitude of the quantities is used to propagate the rounding error for the large system, we consider each individual computing step. All functional dependencies during the Cholesky decomposition are taken into account and we perform a rigorous variance propagation for the whole solution process. The rounding errors in each operation are modeled individually by the size of the actual operators,

$$\overline{f}(\overline{a},\overline{b}) = f(a+\varepsilon_a, b+\varepsilon_b) + \varepsilon_{f(\overline{a},\overline{b})}. \tag{7}$$

Here ε_a and ε_b denotes the perturbance of the coefficients and $\varepsilon_{f(\overline{a},\overline{b})}$ the rounding error during the operation. Applying linear perturbation theory we get

$$\overline{f}(\overline{a},\overline{b}) = f(a,b) + c_1\varepsilon_a + c_2\varepsilon_b + \varepsilon_{f(\overline{a},\overline{b})}\,. \tag{8}$$

Collecting the ε-quantities in the variable $\varepsilon_{f(a,b)}$ yields

$$\overline{f}(\overline{a},\overline{b}) = f(a,b) + \varepsilon_{f(a,b)}\,. \tag{9}$$

The Cholesky factorization is a recursive evaluation process. All elements r_{ij}, r_{ii}, z_i and x_i depend on previous evaluated elements and all these elements are correlated. To show the principle approach we pick out a special operation, the computation of

$$r_{ij} = \left(n_{ij} - \sum_{k=1}^{i-1} r_{ki}r_{kj}\right)/r_{ii},\;\; i = 1\ldots j,\;\; j = 1\ldots n \tag{10}$$

(Meissl, 1980, eq. 3.31). For the evaluation the disturbed values $\overline{r}_{ij}$ and $\overline{n}_{ij}$ (ref. eq. (9)) as well as the basic rounding errors $\varepsilon^{(m)}$, $\varepsilon^{(s)}$ and $\varepsilon^{(d)}$ caused by the arithmetic operations have to be taken into account,

$$\overline{r}_{ij} = \frac{\overline{n}_{ij} - \sum_{k=1}^{i-1}\left(\left(\overline{r}_{ki}\overline{r}_{kj} + \varepsilon_k^{(m)}\right) + \varepsilon_k^{(s)}\right)}{\overline{r}_{ii}} + \varepsilon^{(d)}\,. \tag{11}$$

Applying linear pertubation theory the individual basic errors can be summarized by $\varepsilon_{\varepsilon_{r_{ij}}}$

$$\varepsilon_{\varepsilon_{r_{ij}}} = \frac{1}{\overline{r}_{ii}}\sum_{k=1}^{i-1}\left(\varepsilon_k^{(m)} + \varepsilon_k^{(s)}\right) + \varepsilon^{(d)}. \tag{12}$$

It should be mentioned that the order of the computing steps is important because the rounding errors are not commutative as they depend on the size of the result. Taken into account also the disturbances of the input quantities we get

$$r_{ij} + \varepsilon_{r_{ij}} = \frac{\left(n_{ij}+\varepsilon_{n_{ij}}\right) - \sum_{k=1}^{i-1}\left(r_{ki}+\varepsilon_{r_{ki}}\right)\left(r_{kj}+\varepsilon_{r_{kj}}\right)}{r_{ii}+\varepsilon_{r_{ii}}} + \varepsilon_{\varepsilon_{r_{ij}}}. \tag{13}$$

By expanding this equation and disregarding second order terms of ε we get a linearized construction for the evaluated Cholesky-element $\overline{r}_{ij}$,

$$r_{ij} + \varepsilon_{r_{ij}} = r_{ij} + \frac{1}{r_{ii}}\varepsilon_{n_{ij}} - \sum_{k=1}^{i-1}\left(\frac{r_{kj}}{r_{ii}}\varepsilon_{r_{ki}} + \frac{r_{ki}}{r_{ii}}\varepsilon_{r_{kj}}\right) - \frac{1}{2r_{ii}}\varepsilon_{r_{ii}} + \varepsilon_{\varepsilon_{r_{ij}}}\,. \tag{14}$$

We end up with an implicit formulation of $\varepsilon_{r_{ij}}$, which depends on the already computed quantities $\varepsilon_{n_{ij}}$, $\varepsilon_{r_{kj}}$, and $\varepsilon_{\varepsilon_{r_{kj}}}$. The same approach is also applied to the quantities of the forward and backward substitution step, ε_{z_i} and ε_{x_i}.

The structure of the complete implicit equation system is shown in Fig. 1. The system shows the functional dependencies of the rounding errors as they are formulated in eq. (14). The system is ordered column wise by the errors of the derived quantities ($\varepsilon_{r_{ij}}$, ε_{z_i}, ε_{x_i}) followed by the input errors ($\varepsilon_{n_{ij}}$, ε_{y_i}) and the individ-

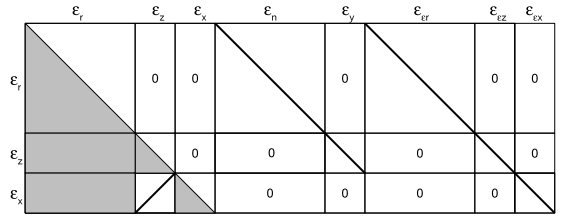

Fig. 1 Structure of the implicit error formulation of the complete solution process with Cholesky's algorithm. The filled parts depict dense matrix structures, the fat lines represent diagonal matrix entries, and the white parts contain just zeros.

ually processed errors of the operations ($\varepsilon_{\varepsilon_{r_{ij}}}, \varepsilon_{\varepsilon_{z_i}}, \varepsilon_{\varepsilon_{x_i}}$) defined by eq. (12). Out of these implicit equations an explicit formulation for the unknown rounding errors is needed. Therefore, the system is factorized by the Gauss-Jordan algorithm, which solves for the dependencies of the Cholesky quantities (see Fig. 2).

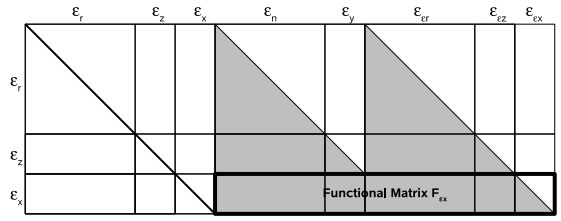

Fig. 2 Structure of the explicit rounding error formulation after reduction by the Gauss-Jordan algorithm. Dependencies shifted to the right blocks of the known rounding errors.

To compute the rounding error covariance matrix of the unknown parameters $\mathbf{x}$ the bordered block in Fig. 2 is needed. This block is the functional matrix $\mathbf{F}_{\varepsilon_\mathbf{x}}$ in the variance propagation $\Sigma\{\varepsilon_x\} = \mathbf{F}_{\varepsilon_x}\Sigma\{\varepsilon_{basic}\}\mathbf{F}_{\varepsilon_x}^T$. $\Sigma\{\varepsilon_{basic}\}$ contains the uncorrelated basic rounding errors that arise during the solution, e.g $\varepsilon_{\varepsilon_{r_{ij}}}$ of eq. (12) and the *a priori* error information of the normal equation system. The result is a full covariance matrix $\Sigma\{\varepsilon_\mathbf{x}\}$ where the variances describe the stochastic rounding errors of the solution of the equation system. The covariances also contain information concerning the correlations between the single rounding errors.

4 Simulations

The algorithm outlined in sect. 3 is tested with a Wiener-Kolmogorov filtering of Bouguer anomalies

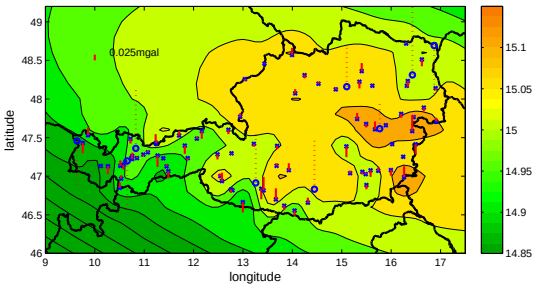

Fig. 3 Wiener-Kolmogorov prediction of Austrian Bouguer anomalies (data set Ruess 1986, intern comunication). Red bars display the residuals. Red dotted bars display identified outliers.

derived at irregular positions. Figure 3 shows the spatial data distribution with the residuals and identified outliers. The measurements are reduced by a polynomial of second order to ensure stationarity. The residual signal $\mathbf{s}$ is predicted by $\mathbf{s} = \Sigma\{\mathbf{s}, \Delta\mathbf{l}\}\Sigma\{\Delta\mathbf{l}\}^{-1}\Delta\mathbf{l}$, where $\Delta\mathbf{l}$ denotes the vector with the trend reduced measurements and $\Sigma\{\Delta\mathbf{l}\}$ the covariances. The matrix is deduced from the analytic covariance function, where the empirical covariances are approximated by a Bessel function combined with a compactly supported function (Sansò and Schuh, 1987; Moreaux, 2008). This leads to a sparse matrix $\Sigma\{\Delta\mathbf{l}\}$. The matrix $\Sigma\{\mathbf{s}, \Delta\mathbf{l}\}$ defines the covariances between the data points and the prediction points.

We focus our attention on the inversion process $\mathbf{w} = \Sigma\{\Delta\mathbf{l}\}^{-1}\Delta\mathbf{l}$. This is equivalent to the solution of the linear equation system $\Sigma\{\Delta\mathbf{l}\}\mathbf{w} = \Delta\mathbf{l}$. The matrix $\Sigma\{\Delta\mathbf{l}\}$ is reordered with three different numbering schemes. The first scheme is the natural form given by the ordering of the data. The second scheme is produced by a reordering with the banker's algorithm (Snay, 1976). The third system is generated by a randomized ordering. The three profiles are shown in figure 4.

To analyze the different rounding errors dependent on the condition of the system we vary the condition of the system by arbitrary choices of the uncorrelated noise in the data points, which is defined by the difference between the empirical and the analytic covariance function at the distance zero. The noise is added on the main diagonal of $\Sigma\{\Delta\mathbf{l}\}$ and stabilizes the system. These systems are tested with the developed algorithm and the covariance matrix of the rounding errors of the so-

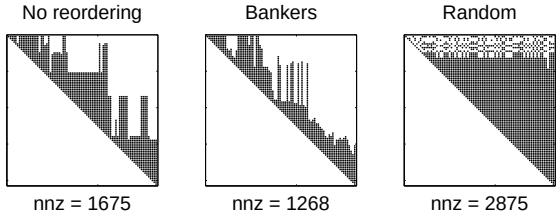

Fig. 4 Profiles of test system after Cholesky factorization with different reordering strategies

lution vector **w** is computed. Results from the simulation are shown in Fig. 5. The maximum rounding errors are plotted for the different numbering schemes and systems. The errors have almost the same size for the same condition number. They do not differ significantly because of the reordering strategy. The reordering with the banker's algorithm influences the rounding errors positively at higher condition numbers. There the randomized ordering produces the highest rounding errors. But in general the size of the rounding errors is essentially influenced by the condition number. A clustering of numerical instabilities cannot be observed. For well conditioned systems the algorithm gives a more optimistic approximation of the rounding errors than the formula by Stewart up to the factor 100. For bad conditioned systems both approximations show similar results, but for instable systems Stewart's approximation is smaller than the stochastic errors by a factor of five but this depends basically on the choice of the constants c_1 and c_2 in eq. (6). For the stochastic approach besides the absolute rounding errors also individual values including the correlations between the rounding errors can be analyzed. It can be observed that the correlations depend very strongly on the size of the rounding errors. The larger they are the higher they are correlated.

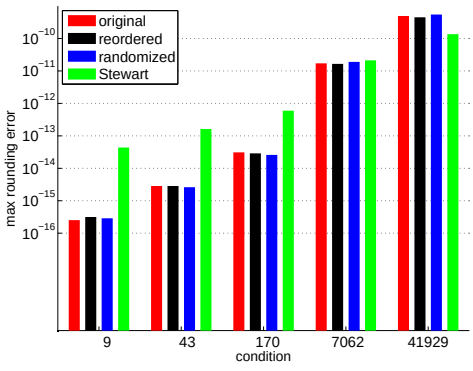

Fig. 5 Results of the simulation for various condition numbers with the three numbering schemes and Stewart's rounding error approximation with c_1 and c_2 of the order unity fixed with 1.

5 Conclusion and discussion

We investigated the hypothesis that reordering induces a clustering of instabilities. The stochastic approach allows for an individual analysis of rounding errors in evaluation processes. As demonstrated here also complex recursive algorithms can be handled and rigorously computed. With respect to our hypothesis it can be stated that the clustering of numerical instabilities caused by reordering strategy has no negative impact on rounding errors.

References

Dahmen, W. and Reusken, A. (2008). Numerics for engineers and natural scientists (German). Springer, Berlin, Heidelberg.

Ernst, A. (2009). Implementation of efficient algorithms to reorder, solve and invert sparse normal equation systems with geodetic applications (German). Master's thesis, Unitversity Bonn, Institute of Geodesy and Geoinformation.

Gaspari, G. and Cohn, S. (1999). Construction of correlation functions in two and three dimensions. Quarterly Journal of the Royal Meteorological Society, 125(554), 723–757.

Gaspari, G., Cohn, S., Guo, J., and Pawson, S. (2006). Construction and application of covariance functions with variable length-fields. *Quarterly Journal of the Royal Meteorological Society*, 132, 1815–1838.

Gibbs, N., Poole, W., and Stockmeyer, P. (1976). An algorithm for reducing the bandwidth and profile of a sparse matrix. SIAM J. Numer. Anal., 13, 236–250.

IEEE Computer Society (2008). IEEE standard for floating point arithmetic. Technical report, The Institute of Electrical and Electronics Engineers, Inc.

Kreyszig, E. (1993). *Advanced engineering mathematics*. 7th edn, Wiley, New York.

Meissl, P. (1980). A priori prediction of roundoff error accumulation in the solution of a super-large geodetic normal equation system. NOAA / National Ocean Survey's National Geodetic Survey (NGS), Rockville, Md., Professional Paper 12.

Moreaux, G. (2008). Compactly supported radial covariance functions. Journal of Geodesy, 82(7), 431–443.

Sansò, F. and Schuh, W.-D. (1987). Finite covariance functions. Bulletin Géodésique, 61, 331–347.

Schuh, W.-D. (1991). Numerical behaviour of covariance matrices and their influence on iterative solution techniques. In R. Rapp and F. Sansò, editors, Determination of the Geoid - Present and Future, volume 106 of IAG Proceedings, pages 432–441, Heidelberg. Springer.

Snay, R. (1976). Reducing the profile of sparse symmetric matrices. NOAA Technical Memorandum, NOS NGS-4.

Stewart, G. (1973). Introduction to Matrix Computations. Academic Press, New York, San Francisco, London.