Analysis of Arithmetic Algorithms: A Statistical Study

F. Chatelin
IBM-FRANCE
5 Pl. Vendôme
75021 Paris Cedex 01, France

V. Frayssé
IRIT-CERFACS
42 av. Coriolis
31057 Toulouse Cedex, France

Abstract

In order to get an insight on the perturbations generated by running algorithms on a computer, one may simulate them by random perturbations on the data. For linear systems, we find that such a statistical estimation gives results which compare favorably with those given by the backward analysis of Wilkinson and Skeel. We intend to use such a technique mainly for nonlinear problems when no theoretical analysis is available.

1 Stability, backward error and condition number

Consider in the finite dimensional space $\Re^n$ the linear system $Ax = b$. In order to solve this problem, we choose a direct method implemented on a computer. Because of the finite precision arithmetic, the algorithm generates a perturbation, so that the computed solution is not x but x_ϵ, where ϵ is a parameter associated with the arithmetic (ϵ tends to zero means the precision tends to infinity). In exact arithmetic, x_ϵ is the exact solution of the linear system $A_\epsilon x_\epsilon = b_\epsilon$. The algorithm generates a perturbation $(A - A_\epsilon, b - b_\epsilon)$ on the initial problem. It is usually the case that this perturbation is neither unique nor easily computable.
In order to estimate the arithmetic error $x - x_\epsilon$, one needs to have a model for the perturbations generated by the algorithm. This can be done by means of a backward error analysis.

1.1 Backward error

In the context of computational stability analysis, we consider that the computed x_ϵ is the exact solution of a perturbed problem :

$$(A + \Delta A)x_\epsilon = b + \Delta b.$$

where ΔA and Δb belong to the class of perturbations generated by the algorithm. The **backward error** is the minimal relative amplitude ω of the perturbations $(\Delta A, \Delta b)$ of a given class, for which x_ϵ is still a solution of $(A + \Delta A)x_\epsilon = b + \Delta b$.
We recall here the two well-known models of perturbations used to study Gaussian elimination. First, Wilkinson [7] used a global model based on matrix norms and a global backward error of the type :

$$\frac{||\Delta A||}{||A||} + \frac{||\Delta b||}{||b||}.$$

which leads to the **normwise perturbations** $\{||\Delta A|| \leq \omega||A||, ||\Delta b|| \leq \omega||b||\}$. Then, Oettli and Prager [4] introduced a more local model, focused on matrix elements. Skeel [6] studied it in detail. He defines the class of **componentwise perturbations** i.e.

$$\{\Delta A, \Delta b; |\Delta A| \leq \omega E, |\Delta b| \leq \omega f\},$$

where E and f are given as respectively a matrix and a vector of positive elements and where the inequalities are componentwise. For this class of perturbations, the relative backward error is

$$\eta = \min\{\omega; (A + \Delta A)x_\epsilon = b + \Delta b \\ \textit{with } |\Delta A| \leq \omega E, |\Delta b| \leq \omega f\}.$$

E and f are respectively a matrix and a vector to be defined by the user. As noted by Skeel, they may be seen as the maximal uncertainty on the data (tolerance), the special structure of external perturbations ...The choice $(E = |A|, f = |b|)$ seems to be a good model for Gaussian elimination.
Once the backward error is computed, one may want to estimate the error $x - x_\epsilon$ of the solution. This is where the condition number plays a role.

1.2 Condition number

A problem is said to be stable if a "small" variation of the data induces a "small" variation of the solution. Consider a problem (P) with data ξ and solution $\Phi(\xi)$. The stability of this problem (P), when subjected to a certain type of perturbations, can be quantified by means of the relative condition number C:

$$C = \lim_{\bar{\xi} \to \xi} \frac{\text{relative distance from } \Phi(\bar{\xi}) \text{ to } \Phi(\xi)}{\text{relative distance from } \bar{\xi} \text{ to } \xi},$$

if this limit exists, which is always the case for a nonsingular linear system.
The definition of C majorizes, **to the first order**, the relative error:

$$\frac{||\Delta\Phi(\xi)||}{||\Phi(\xi)||} \leq C\eta,$$

CH3015-5/91/0000/0010$01.00 © 1991 IEEE

where η is the relative backward error associated to the chosen type of perturbations. Of course, C depends on the model of the perturbations and on the metric. The condition number also depends on the choice of the data to be perturbed.
Below is a table of the condition numbers (or upper bounds) of a linear system for Wilkinson's global model and Skeel's structured model.

data	global	structured
A, b	$\leq 2\|\|A\|\|\|\|A^{-1}\|\|$	$\frac{\|\|\|A^{-1}\|E\|x\| + \|\|\|A^{-1}\|f\|\|_\infty}{\|\|x\|\|_\infty}$
A	$\leq \|\|A\|\|\|\|A^{-1}\|\|$	$\frac{\|\|\|A^{-1}\|E\|x\|\|\|_\infty}{\|\|x\|\|_\infty}$
b	$\|\|A^{-1}\|\|_\infty \frac{\|\|b\|\|_\infty}{\|\|x\|\|_\infty}$	$\frac{\|\|\|A^{-1}\|f\|\|_\infty}{\|\|x\|\|_\infty}$

Table 1: Some condition numbers or upper bounds for a linear system

2 Statistical estimation

A simple way to get a statistical estimate of a condition number is to perturb the data with random perturbations taken inside the class of perturbations one wants to study. Then, one measures the induced variation on the solution. One can estimate the condition number by computing the ratio of the size of the induced variation and the size of the perturbation [1].
Of course, this method is very costly for linear systems where theoretical tools are already available (explicit formulations of condition numbers). But it allows one to estimate a condition number even when its mathematical formulation is unknown. Therefore, it will be very useful for nonlinear problems where theory has not provided yet such formulations. Then our aim is to test the reliability of the method in the "simpler" linear case first.

2.1 The method and its implementation

$$\text{data } A, b \xrightarrow{G_\epsilon} X \xrightarrow{F} Y = AX - b$$

Perturbing randomly the data of the algorithm G_ϵ, we estimate:

- the relative condition number K_ϵ of F^{-1} by measuring $\frac{\Delta_r(X)}{\Delta_r(Y)}$,
- the relative condition number L_ϵ of G_ϵ by measuring $\frac{\Delta_r(X)}{\Delta_r(data)}$,
- the relative condition number I_ϵ of $\mathcal{I}_\epsilon = F_\epsilon \circ G_\epsilon$ by measuring $\frac{\Delta_r(Y)}{\Delta_r(data)}$

where $\Delta_r(z)$ is a measure of a relative variation around z, with an appropriate norm. One can show that $I_\epsilon \sim \frac{L_\epsilon}{K_\epsilon}$. We call I_ϵ the partial arithmetic stability. If I_ϵ is close to one, then $\mathcal{I}_\epsilon$ is a good approximation of the identity. If I_ϵ is too large, it means that the algorithm is more ill-conditioned than the mathematical problem. I_ϵ is useful to distinguish between the contribution of the algorithm and of the problem itself, to an instability. We introduce two kinds of perturbations:

- type-1 perturbations are global. We define them by $(A_{ij})_{per} = A_{ij} + \alpha||A||t$ and $(b_i)_{per} = b_i + \alpha||b||t$,
- type-2 perturbations are structured. We define them by $(A_{ij})_{per} = A_{ij}(1 + \alpha t)$ and $(b_i)_{per} = b_i(1 + \alpha t)$,

where t controls the amplitude of the perturbations and α is a discrete random variable such as, for example $pb(\alpha = 1) = pb(\alpha = -1) = 1/4$, $pb(\alpha = 0) = 1/2$. From now, when not written explicitely, $||.||$ is $||.||_\infty$
For each type of perturbation, we vary the amplitude of this perturbation from machine precision to 10^{-1} (or more if it is relevant for the problem): this is a way to simulate perturbations of different origins (arithmetic, numerical approximations, physical measurements ...). After applying the algorithm G_ϵ to the randomised data, we collect a sample of computed solutions X and their associated sample of residuals Y. Let x_ϵ be the computed solution of the linear system $Ax = b$.
Let m (resp. ρ) be the mean and σ (resp. v) be the standard deviation of the sample X (resp. Y). Let σ_π be the norm of the relative variation of the data. We define the estimators for the condition numbers following Wilkinson's definition for type-1 perturbations, and Skeel's definition for type-2 perturbations. Tables 2 and 3 present formulae which estimate the condition number and the relative error $\frac{||x - x_\epsilon||}{||x_\epsilon||}$. β takes the value

- $||A||||x_\epsilon|| + ||b||$ when both A and b are perturbed,
- $||A||||x_\epsilon||$ when only A is perturbed,
- $||b||$ when only b is perturbed.

$I_{1\epsilon}(t)$	$\dfrac{\sqrt{\|v\|^2 + \|\rho\|^2}}{\beta\sigma_\pi}$
$L_{1\epsilon}(t)$	$\dfrac{\|\sigma\|}{\|x_\epsilon\|\sigma_\pi}$
$K_{1\epsilon}(t)$	$\dfrac{\|\sigma\|\beta}{\|x_\epsilon\|\|v\|}$
error estimation	$K_{1\epsilon}\dfrac{\|Ax_\epsilon - b\|}{\beta}$

Table 2: Statistical estimators for type-1 perturbations.

perturbation	E, f
$I_{2\epsilon}(t)$	$\dfrac{1}{\sigma_\pi}\max_{1\leq i\leq n}\dfrac{\sqrt{v_i^2+\rho_i^2}}{(E\|x_\epsilon\|+f)_i}$
$L_{2\epsilon}(t)$	$\dfrac{\|\sigma\|}{\|x_\epsilon\|\sigma_\pi}$
$K_{2\epsilon}(t)$	$\dfrac{\|\sigma\|}{\|x_\epsilon\|}\dfrac{1}{\max_{1\leq i\leq n}\frac{\|v_i\|}{(E\|x_\epsilon\|+f)_i}}$
error estimation	$K_{2\epsilon}\max_{1\leq i\leq n}\dfrac{\|r_i\|}{(E\|x_\epsilon\|+f)_i}$

Table 3: Statistical estimators for type-2 perturbations.

For more details on the method and for proofs, see the appendix and [2].

2.2 General behaviour of the condition number

Figure 1 shows the standard behaviour of the condition number estimate for the problem $Ax = b$ when the matrix A is perturbed with our type-1 or type-2 perturbations, and when only the right-hand side is perturbed. t controls the amplitude of the perturbations. For perturbations of A and possibly b, and for $t \leq t_0$, the condition number K is constant, smaller than the classical condition number for type-1 perturbations, and close to Skeel's condition number for type-2 perturbations. For $t > t_0$, K varies like $1/t$ (which appears as a line with slope -1 on the log-log scale of figure 1). For $t > t_0$, the perturbed system is seen as singular by the computer and no estimate is reliably available.
For perturbations of b only, K is constant for all t. This experimentation is very useful because in many cases it allows to make an estimation of the error, by definition of the condition number.

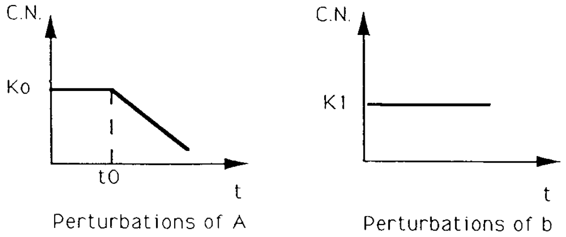

Figure 1: Behaviour of the condition of a linear system with the size of the perturbation (log-log scale).

3 Results and comments

We present here the results when Gaussian elimination, as provided by the LAPACK library, is applied to the linear system $DDx = b$ whose coefficient matrix and exact solution x are given by:

$$\begin{aligned} DD(i,j) &= \frac{i-1}{i+j-1}, i \neq j \\ DD(i,i) &= n \\ x(i) &= \sqrt(i), i = 1, n \end{aligned}$$

For each plot, the value of the parameter t, which controls the amplitude of the perturbation, is on the horizontal axis. On the vertical axis are plotted the relative condition number of the identity I, the relative condition number of the algorithm L, the relative condition number of the mathematical problem K, the exact relative error Eex and the estimated relative error $Eest$.
Classical and componentwise condition numbers of the unscaled matrix are all close to 1 and lead to very good estimations of the error which is around 10^{-16}. All statistical experiments show constant condition numbers (and thus stability) for type-1 and type-2 perturbations. We "descale" this diagonal dominant matrix by multiplying each even row by 10^6 and each odd row by 10^{-6}. We modify the right-hand side in the same way, so that the descaled system has the same solution as the original one. The exact arithmetic error is now around 10^{-14}. This has no influence on row-scaling independant condition numbers (i.e Skeel's condition numbers) but the classical condition numbers are multiplied by 10^{12}. Nevertheless, Skeel's condition numbers still provide a very good error estimation where the classical

condition numbers are far too large.
We observe on figures 2 and 3 that type-2 perturbations do not generate any instability. When the condition number of the identity is one, our estimate of the condition number of the mathematical problem is close to that defined by Skeel and yields a very good estimation of the error (figures 2 and 3).
On the contrary, with type-1 perturbations, the problem is unstable for values of t smaller than 10^{-13} when both A and b are perturbed (figure 4). When only b is perturbed, the condition number of the mathematical problem is constant as expected. But the estimation of the arithmetical error is not good at all: it is largely overestimated. This means that type-1 perturbations are not a good model for the perturbations generated by the LU algorithm: they are too "large" (figure 5). We are actually measuring the error for a different problem which would be a type-1 perturbation of the initial problem. Figure 6 illustrates this point. Let (A, b) be the original "descaled" matrix and vector, and let (A', b') a given type-1 perturbation of (A, b). Let x and x' such that : $Ax = b$ and $A'x' = b'$. If K_0 is the condition number of the linear system $Ax = b$ subjected to type-1 perturbations, then we can make the following estimation:

$$\frac{||x - x'||}{||x||} \sim K_0 \frac{||Ax' - b||}{||A|| ||x'|| + ||b||}.$$

The results shown on figure 6 are now very good.
The conclusion of this example is in threefold:

1. it is very important to have a good model of the perturbations generated by an algorithm if one wants to estimate reliably the arithmetical error,
2. given a model of the perturbation, the statistical method allows one to measure the condition number of a linear system subject to these perturbations and estimate the error generated by this kind of perturbations,
3. the perturbation generated by the LU algorithm seems to be row scaling independant. This is in agreement with the insistence in the literature for building condition numbers which are independent of row scaling.

More extended results, using different matrices are presented in [2].

Appendix

Type-1 perturbations

We perturb A and/or b in the following way:

$$\begin{aligned} (A_{ij})_{per} &= A_{ij} + \alpha ||A|| t, \\ (b_i)_{per} &= b_i + \alpha ||b|| t, \end{aligned}$$

where t controls the size of the perturbations. We have $||\Delta A|| \leq nt||A||$ and $||\Delta b|| \leq t||b||$ where n is the size of the matrix.
We are then simulating a discrete version of the global perturbations of Wilkinson. Therefore we use the appropriate formulation of the relative backward error stated by Rigal and Gaches [5, 3]: if Δy is an **absolute** variation of the residual, $\frac{\Delta y}{||A|| ||x|| + ||b||}$ is the associated **relative** variation when both A and b are perturbed, $\frac{\Delta y}{||A|| ||x||}$ (resp. $\frac{\Delta y}{||b||}$) when only A (resp. b) is perturbed.
We would like to estimate $\Delta_r(x_\epsilon)$ by $\frac{\sqrt{||\sigma||^2 + ||m - x||^2}}{||x_\epsilon||}$ but we dot not know x, the exact solution. That is why we have to take away the term $||m - x||$ called the bias and use $\Delta_r(x_\epsilon) \sim \frac{||\sigma||}{||x_\epsilon||}$. The estimation will be justified when $||\sigma|| \leq ||m - x||$, which will happen as the size of the perturbation grows.
Nevertheless, we know the exact residual which is $y = Ax - b = 0$. We can then estimate $\Delta(y)$ by $\sqrt{||v||^2 + ||\rho||^2}$. That is what we do for computing the condition number of the identity $I_{1\epsilon}(t)$. But since the condition number of the mathematical problem involves both $\Delta_r(x)$ and $\Delta_r(y)$ and that we have to take away the bias for $\Delta_r(x)$, we decide to use $\Delta(y_\epsilon) = ||v||$ in the computation of $K_{1\epsilon}(t)$. This choice is heuristic but we obtained our best results with it. We can also consider that the absolute condition number of the mathematical problem is estimated by $\frac{||\sigma||}{||v||}$.
$I_{1\epsilon}(t)$ is the ratio of $\Delta_r(y_\epsilon)$ and σ_π, $L_{1\epsilon}(t)$ is the ratio of $\Delta_r(x_\epsilon)$ and σ_π, and $K_{1\epsilon}(t)$ is the ratio of $\Delta_r(x_\epsilon)$ and $\Delta_r(y_\epsilon)$. All the estimates were given in table 2.

Type-2 perturbations

We perturb A and/or b in the following way:

$$\begin{aligned} (A_{ij})_{per} &= A_{ij} (1 + \alpha t), \\ (b_i)_{per} &= b_i (1 + \alpha t), \end{aligned}$$

where t controls the size of the perturbations.
We have: $|\Delta A| \leq t|A|$ and $|\Delta b| \leq t|b|$.
In this case, we are simulating a discrete version of the structured perturbations of Skeel [6]. Therefore we use the appropriate formulation of the relative backward error: if Δy is an **absolute** variation of the residual, $\max_{i=1,n} \frac{\Delta y_i}{(|A||x| + |b|)_i}$ is the associated **relative** variation when both A and b are perturbed, and $\max_{i=1,n} \frac{\Delta y_i}{(|A||x|)_i}$ (resp. $\max_{i=1,n} \frac{\Delta y_i}{|b|_i}$) when only A (resp. b) is perturbed.
Like in the previous paragraph, we will estimate $\Delta_r(x_\epsilon)$ by $\frac{||\sigma||}{||x_\epsilon||}$.
When computing the condition number of the identity, we will use $\Delta_r(y) = \max_{1 \leq i \leq n} \frac{\sqrt{v_i^2 + \rho_i^2}}{(A|x_\epsilon| + b)_i}$. When computing the condition number of the mathematical problem, we will use $\max_{1 \leq i \leq n} \frac{|v_i|}{(A|x_\epsilon| + b)_i}$.
$I_{2\epsilon}(t)$ is the ratio of $\Delta_r(y_\epsilon)$ and σ_π, $L_{2\epsilon}(t)$ is the ratio of $\Delta_r(x_\epsilon)$ and σ_π, and $K_{2\epsilon}(t)$ is the ratio of $\Delta_r(x_\epsilon)$ and $\Delta_r(y_\epsilon)$. The estimates were given in table 3.

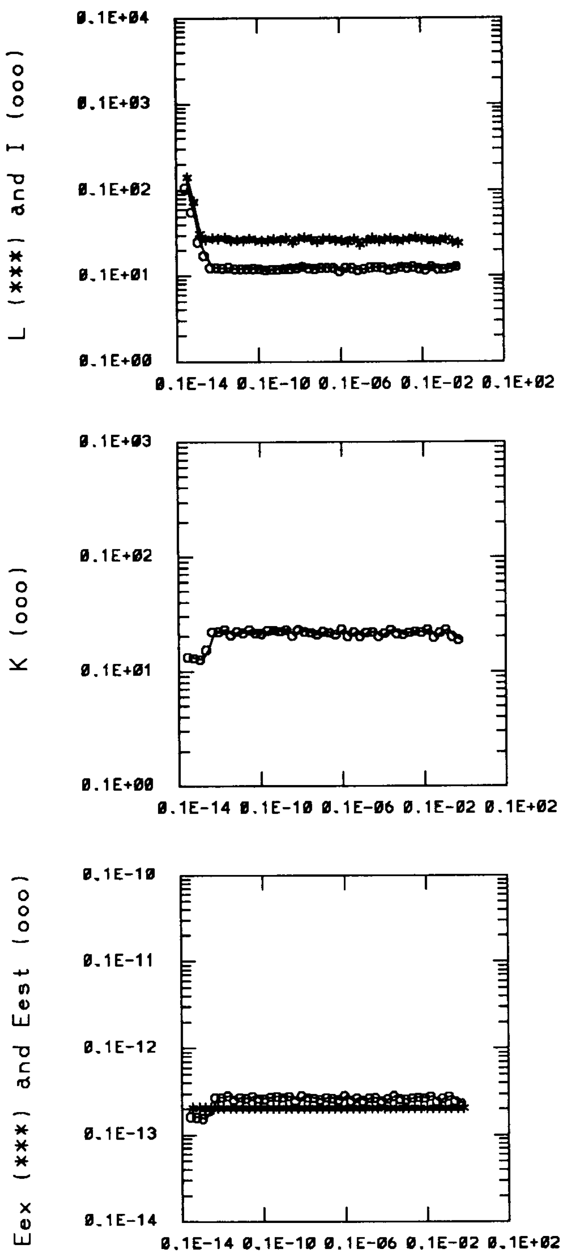

Figure 2: Gaussian elimination. Type-2 perturbations of A and b.

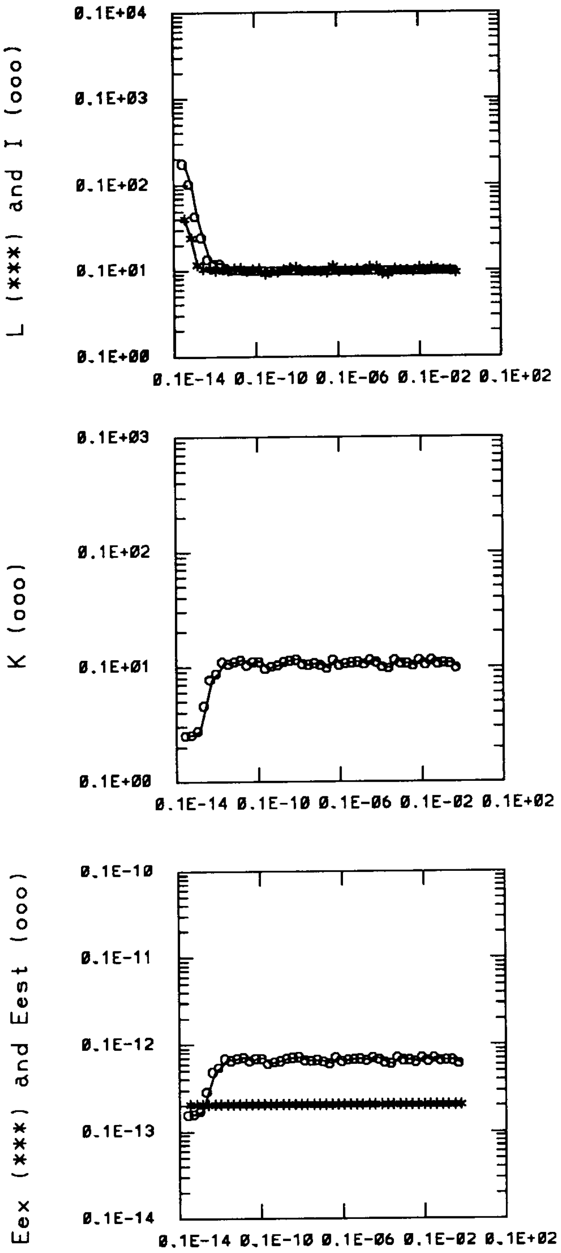

Figure 3: Gaussian elimination. Type-2 perturbations of b.

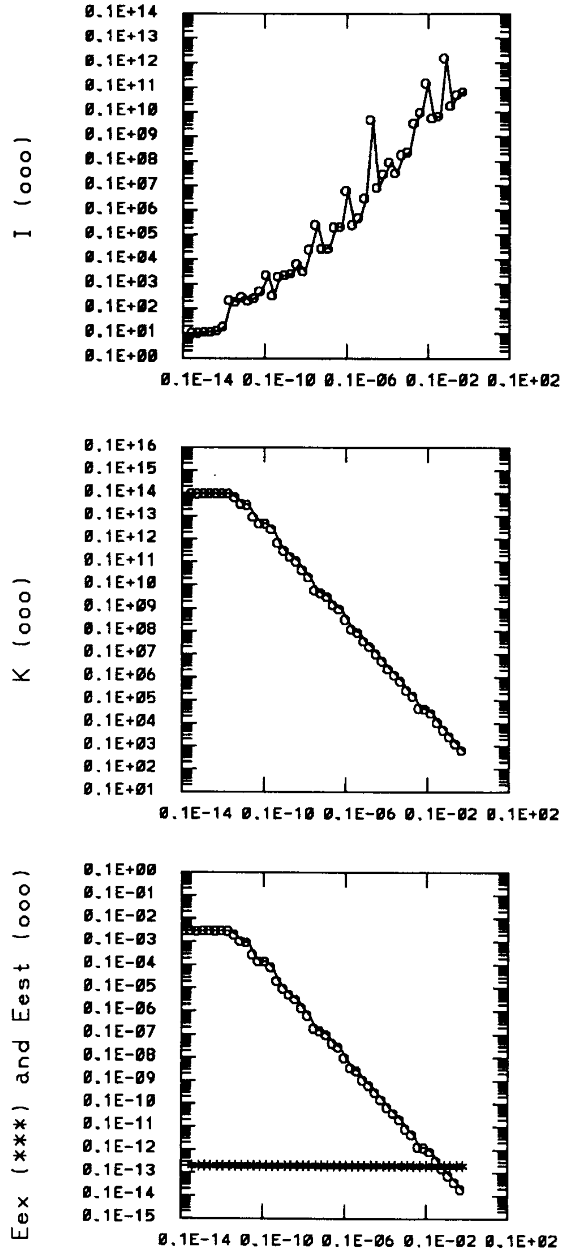

Figure 4: Gaussian elimination. Type-1 perturbations of A and b.

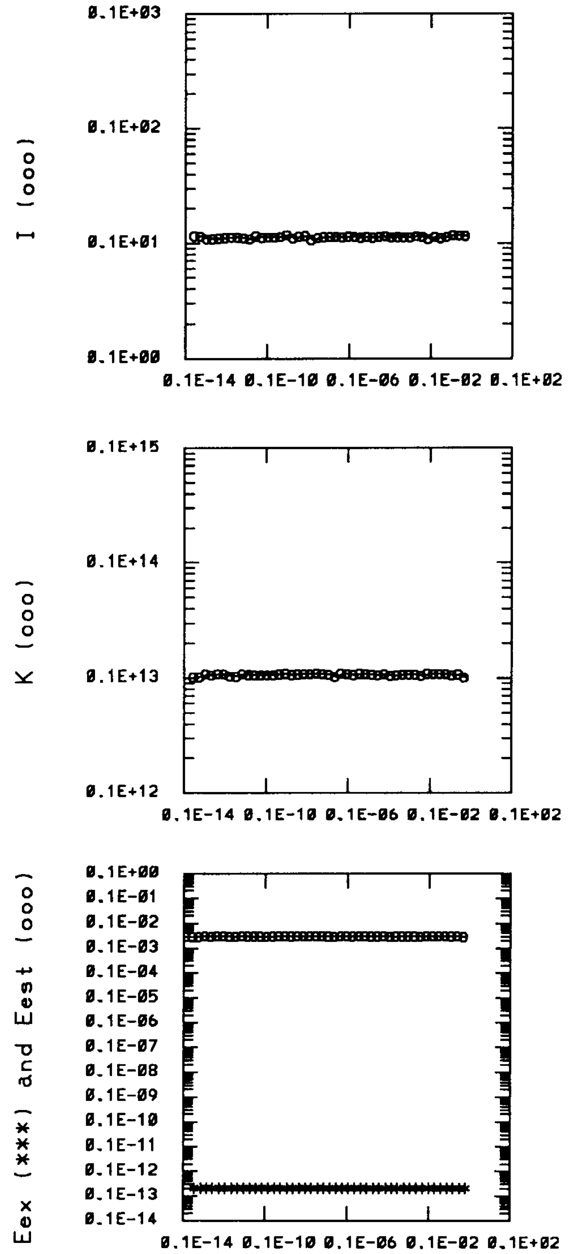

Figure 5: Gaussian elimination. Type-1 perturbations of b.

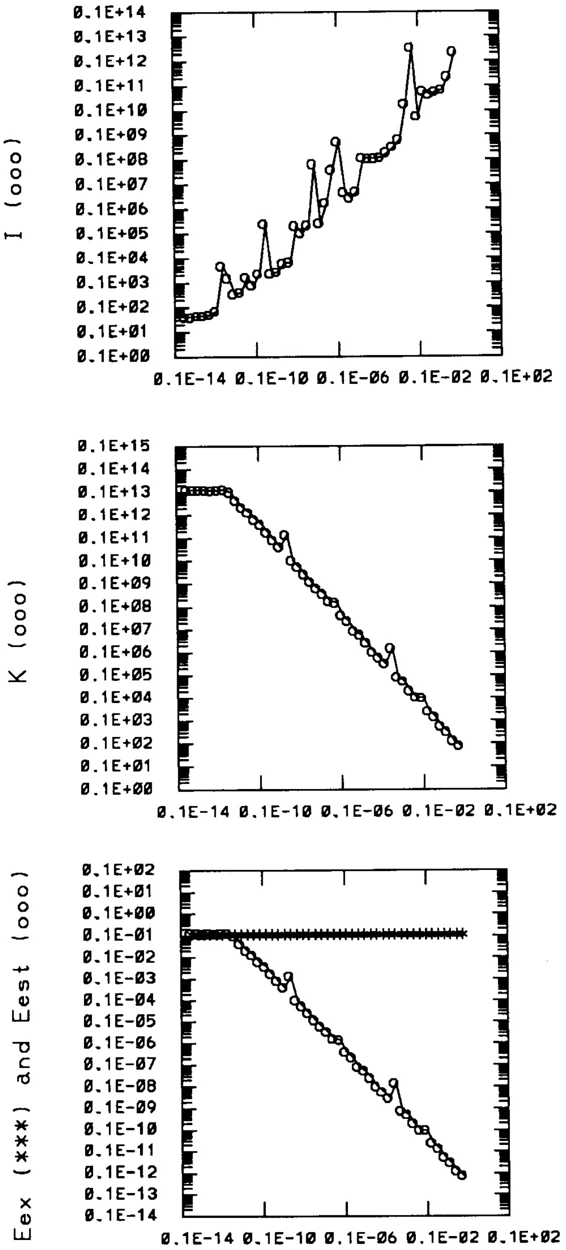

Figure 6: Study of a given type-1 perturbations of A and b. Comparison with the exact error $x - x'$.

References

[1] F. Chatelin. Résolution approchée d'équations sur ordinateur, 1989. notes de DEA, Université Paris Dauphine.

[2] F. Chatelin and V. Frayssé. A statistical study of the stability of linear systems. Technical Report TR/PA/90/43, CERFACS, 1990.

[3] N. J. Higham. How accurate is gaussian elimination? Technical Report TR 89-1024, Cornell University, Ithaca NY 14853-7501, July 1989.

[4] W. Oettli and W. Prager. Compatibily of approximate solution of linear equations with given error bounds for coefficients and right-hand sides. *Numer. Math.*, 6:405–409, 1964.

[5] J.L. Rigal and J. Gaches. On the compatiblity of a given solution with the data of a linear system. *J. Assoc. Comput. Mach.*, 14(3):543–526, July 1967.

[6] R. D. Skeel. Scaling for numerical stability in gaussian elimination. *J. Assoc. Comput. Mach.*, 26(3):494–526, July 1979.

[7] J.H. Wilkinson. Error analysis of direct methods of matrix inversion. *J. Assoc. Comput. Mach.*, 8(3):281–330, July 1961.