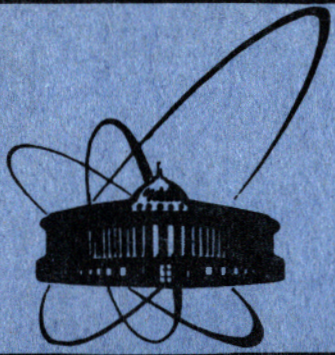

объединенный
институт
ядерных
исследований
дубна

E11-84-692

V.Gadjokov

QUASI-OPTIMUM SCALING OF SYMMETRIC POSITIVE-DEFINITE MATRICES

Submitted to "Computer Physics Communications"

1984

1. INTRODUCTION

Experimental-data processing in physics often requires the numerical solution of operator equations of the type

$$Fx = Y, \tag{1}$$

where $Y \in R^m$ is the vector of quantities measured, $x \in R^n$ is that of unknown parameters and F is the operator (nonlinear or linear) transforming R^n into R^m. This operator corresponds to the model we want to apply to the data processed. The problem formulated is usually overdetermined, i.e., $m > n$ (and often $m \gg n$). When solving numerically equations of this type by means of gradient or Taylor-series methods the necessity arises to deal with matrices built as

$$A = J^T J, \tag{2}$$

where J represents the $(m \times n)$ Jacobi matrix of F. Obviously, A is positive-definite and its condition number is

$$\mathrm{cond}\, A = ||A|| \cdot ||A^{-1}||, \tag{3}$$

the norms being those which suit best the problem considered. In this paper we shall be interested most of Euclidean (spectral) norms $||\cdot||_s$; due to the special property of these, however [1],

$$||\cdot||_s = \inf ||\cdot||, \tag{4}$$

the conclusions remain in force for other norms, too.

For the well-known reasons it is advisable to take certain care of lowering $\mathrm{cond}\, A$ as much as possible. We want to consider the "computationally trivial" [2] operation of scaling

$$A' = DAD \tag{5}$$

as a means of improving $\mathrm{cond}\, A$ and, in particular, the search of such a scaler D' that either leads to

$$\mathrm{cond}(D'AD') = \min_D \mathrm{cond}(DAD) \tag{6}$$

or, at least, is coming close to it.

In section 2 a brief overview of previous results is given; section 3 deals with some geometric properties of matrices J and A and uses such properties for the construction of a simple

scaling approach; then, in section 4 the results of numerical experiments and comparative evaluations are presented which lead to the conclusions of section 5.

2. THEORY AND PRACTICE OF SCALING

Scaling has been considered in the past by many authors and from various points of view (see, e.g., /1-11/). It appears to be generally recognized that the problem of optimum scaling is far from a satisfactory programmable solution. In our view a practicable approach is within reach since a long time but has been used with relative rarity and caution.

To avoid additional round-off errors D is most often built of integer power of the base used /3,5,10,11/ . It is known /3/ that in this case any possible benefits are due to the different selection of pivot elements as compared to the non-scaled matrix manipulation. However, the actual definition of scaling factors in these papers remains without solid theoretical foundations.

If the operator F is a linear superposition of similar functions (e.g., in processing of sectroscopic data), then it may be shown /11/ that ill-condition of A originates primarily from the presence of strongly-varying amplitudes in F. Hence, a remedy may be sought for in the form of a scaling matrix that introduces internal (computational only) measuring units for the amplitudes in such a way as to make them all equal (or, at least, of the same order of magnitude). This sort of scaling may be called "physical". Our experience shows that it is extremely useful. However, it requires non-formal analysis of the problem dealt with and, of course, it lacks generality.

Departing from statistical considerations Marquardt /6/ points out that a suitable scaling matrix D may be constructed from the main diagonal of A,

$$D = \mathrm{diag}(A_{11}^{-1/2}, A_{22}^{-1/2}, \ldots, A_{nn}^{-1/2}) . \qquad (7)$$

This transforms A into the correlation matrix of parameters sought for. It is this approach that we consider neglected, i.e., not used in practical computations as frequently as it should be.

A thourough theoretical treatment of scaling for a variety of absolute norms including all Hölder ones is given by Bauer /2/ whose results are not limited to symmetric matrices. In that paper the nonsingular square matrix A is considered as a metric in dual spaces $\mathfrak{B}$ and $\mathfrak{B}^H$, where norms are introduced with Hölder exponents p and q, respectively, i.e.,

$$y^H A x = (y^H A)x = y^H(Ax) . \qquad (8)$$

Then it is proved that

$$\min_{D_1, D_2} \mathrm{cond}(D_1 A D_2) \leq \pi(\mathcal{P}), \qquad (9)$$

where π is the Perron root (eigenvalue) of the positive matrix $\mathcal{P} = |A| \cdot |A^{-1}|$. Moreover, this minimum is reached for D_1 and D_2 as follows

$$\left.\begin{aligned} D_1 &= Y_1^{1/p} X_1^{-1/q} \\ D_2 &= Y_2^{-1/p} X_2^{1/q} \end{aligned}\right\} , \qquad (10)$$

where X_1, X_2, Y_1, Y_2 are diagonal matrices satisfying the eigenvector equations

$$\left.\begin{aligned} \mathcal{P} X_1 e &= \pi X_1 e \\ e^T Y_1 \mathcal{P} &= \pi e^T Y_1 \\ \mathcal{P}' X_2 e &= \pi X_2 e \\ e^T Y_2 \mathcal{P}' &= \pi e^T Y_2 \end{aligned}\right\} \qquad (11)$$

while $e = \mathrm{col}(1, 1, \ldots, 1)$, $e^T = \mathrm{row}(1, 1, \ldots, 1)$, and

$$\mathcal{P}' = |A^{-1}| \cdot |A| . \qquad (12)$$

This means that the optimum scaling matrices (10) are expressed through the right and left Perron eigenvectors of positive matrices $\mathcal{P}$ and $\mathcal{P}'$. Note that these two are, in general, different, since $|A|$ and $|A^{-1}|$ are not bound to have a zero commutator.

Furthermore, for the cases of maximum ($p = q = \infty$) and sum ($p = q = 1$) norms equality holds in (9). For the maximum norm equations (10) take the form

$$\left.\begin{aligned} D_1 &= X_1^{-1} \\ D_2 &= X_2 \end{aligned}\right\} . \qquad (13)$$

As far as Euclidean ($p = q = 2$) norms are concerned the notion of checkerboard sign distribution is of importance. A matrix G is said to have a checkerboard sign distribution if it can be written in the form

$$G = E_1 |G| E_2 , \qquad (14)$$

where

$$|E_1| = |E_2| = I . \qquad (15)$$

In Euclidean norms (9) is an equality when A has a checkerboard sign distribution and an unequality when A has not.

Note also that for Hermitean matrices $A^H = A$

$$\mathcal{P}' = \mathcal{P}^T, \tag{16}$$

hence, the left (right) eigenvector of $\mathcal{P}$ is also a right (left) eigenvector of $\mathcal{P}'$ and, accordingly, the optimum scaling is reached when $D_1 = D_2 = D$.

It should be clear that these explicit formulae are not suitable for immediate use in numerical calculations for the reason that they express the optimum scalers of A via its inverse A^{-1}. They are precious, however, with the possibility they offer of evaluating the effectiveness of scaling.

3. THE ISOPERIMETRIC SCALING

Let us note that since for any constant $c \neq 0$

$$\mathrm{cond}(cA) = \mathrm{cond}A, \tag{17}$$

we may limit our discussion to the case of such scaling matrices which preserve SpA. For symmetric matrices A built according to (2) it appears convenient to accept the convention

$$\mathrm{Sp}(DAD) = \mathrm{Sp}A = n, \tag{18}$$

i.e.,

$$\sum_{i=1}^{n} d_i^2 a_{ii} = \sum_{i=1}^{n} a_{ii} = n. \tag{18a}$$

When J is a full-rank matrix, A is positive-definite,

$$\det A = \prod_{i=1}^{n} \lambda_i \tag{19}$$

and

$$\mathrm{Sp}A = \sum_{i=1}^{n} \lambda_i . \tag{20}$$

Symmetric scaling is a congruency and does not preserve the eigenvalues $\{\lambda_i\}$; however, in accordance with (18) the one we consider leaves unchanged their sum SpA. Now, applying Cauchy's unequality to the set of eigenvalues we obtain

$$\det A \leq (\mathrm{Sp}A/n)^n = 1, \tag{21}$$

where $\det A = 1$ if and only if all the eigenvalues of A equal unity, i.e., when $\mathrm{cond}A = 1$.

Let A be scale according to (7), i.e.,

$$A_{ii} = 1, \quad i = 1,2,\ldots,n \tag{22}$$

and let its determinant be

$$\det A < 1. \tag{23}$$

If we now carry out another nontrivial scaling which obeys (18) we obtain

$$\det(DAD) = \det A \cdot \prod_{i=1}^{n} d_i^2 . \tag{24}$$

Again, Cauchy's unequality yields

$$\prod_{i=1}^{n} d_i^2 < 1. \tag{25}$$

Therefore, if A has been pre-scaled as in (22) then

$$\det(DAD) < \det A \tag{26}$$

for any non-trivial scaler D.

This relation may be interpreted geometrically. Indeed, due to (2) A may be regarded as the Gram-matrix of the n independent columnvectors of J. Then detA has the meaning of the squared volume of the n-dimensional hyperparallelepiped built over these vectors, while SpA is the sum of their squared Euclidean lengths. The directions of these vectors are not affected by scaling. In view of such an interpretation (26) may be thought of as belonging to the class of isoperimetric unequalities [12], and we propose to call Marquardt's scaling (7) ISOPERIMETRIC.

Naturally, the isoperimetric scaling is different from (10) and may coincide with it in the very exceptional case of $\mathcal{P}$ being symmetric. By the way, it is always so when $n = 2$ but, in this case the optimum scaling is easily solved analytically. At $n > 2$ the isoperimetric scaling appears attractive for at least two reasons: (a) it is indeed "computationally trivial", rather economic ($\sim n^2$ divisions per matrix) and does not require any preliminary knowledge of the inverted matrix; (b) as a greater determinant generally indicates that a matrix is far from degeneration, it may be expected that the condition of isoperimetrically-scaled matrices lie in the vicinity of the theoretical optimum.

4. NUMERICAL EXPERIMENTS

We carried out a series of numerical experiments to evaluate the different condition numbers of matrices scaled isoperimetrically and optimally. All the computations were carried out in single precision (24-bit mantissae) on SM4 which is equivalent to PDP-11/30 at machine-instruction level.

4.1. The first two examples may also be checked analytically and are intended to make clear the technique of comparison.

a) Both types of scaling coincide

$$A = \begin{pmatrix} 1/3 & -\sqrt{2}/3 & 1/3 \\ -\sqrt{2}/3 & 4/3 & 0 \\ 1/3 & 0 & 4/3 \end{pmatrix}, \quad \mathrm{Sp}A = 3, \quad \det A = 4/27 .$$

The isoperimetric scaler D_i is $D_i = \mathrm{diag}(\sqrt{3}, \sqrt{3}/2, \sqrt{3}/2)$ and yields

$$A_i = D_i A D_i = \begin{pmatrix} 1 & -1/\sqrt{2} & 1/2 \\ -1/\sqrt{2} & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \quad \mathrm{Sp}A_i = 3, \quad \det A_i = 1/4 .$$

An attempt to scale A_i further according to (10) with $p = q = 2$ reveals that $\mathcal{P} = |A_i| \cdot |A_i^{-1}|$ is symmetric, hence the scaler $D_{i \to 0} = I$ and no further improvement of condition is attainable. Here $\pi(\mathcal{P}) = 7 + 4\sqrt{3}$, $\mathrm{cond}A = (23 + 5\sqrt{21})/2$, $\mathrm{cond}A_i = \mathrm{cond}A_0 = \pi(\mathcal{P})$ (both A and A^{-1} have a checkerboard sign distribution).

b) Isoperimetric scaling differs from the optimum one (this is the example of /2/ prescaled to satisfy eq. (18))

$$A = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 2/3 & 1 \\ 1/3 & 1 & 2 \end{pmatrix}, \quad \mathrm{Sp}A = 3, \quad \det A = 1/27 .$$

The two scalers are:

isoperimetric $\quad D_i = \mathrm{diag}(\sqrt{3}, \sqrt{3/2}, 1/\sqrt{2})$

optimum $\quad D_0 = \mathrm{diag}(3/2, \sqrt{15}/2\sqrt{2}, 1/\sqrt{2})$

and lead to:

$$A_i = D_i A D_i = \begin{pmatrix} 1 & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{2} & 1 & \sqrt{3}/2 \\ 1/\sqrt{6} & \sqrt{3}/2 & 1 \end{pmatrix}, \quad \mathrm{Sp}A_i = 3, \quad \det A_i = 1/12$$

$$A_0 = D_0 A D_0 = \begin{pmatrix} 3/4 & \sqrt{15}/4\sqrt{2} & 1/2\sqrt{2} \\ \sqrt{15}/4\sqrt{2} & 5/4 & \sqrt{15}/4 \\ 1/2\sqrt{2} & \sqrt{15}/4 & 1 \end{pmatrix}, \mathrm{Sp}A_0 = 3, \quad \det A_0 = 5/64 .$$

The condition numbers are as follows: $\mathrm{cond}A = (4+\sqrt{15})^2 \approx 61.98$, $\mathrm{cond}A_i \approx 39.54$, $\mathrm{cond}A_0 = \pi(\mathcal{P}) = (3+\sqrt{10})^2 \approx 37.97$. Again, the root π is reached because of checkerboard signs in both A and A^{-1}.

4.2. This second series deals with Hilbert matrices of order 2 to 8. These are known to be particularly ill-conditioned, and this fact is not due to non-balanced columns or rows. Nevertheless, it seems instructive to consider the results (see the Table) which show a slight improvement in condition, thus illustrating the limits of scaling.

Since for Hilbert matrices the checkerboard sign distribution holds throughout, optimum scaling is attainable, at least in theory. However, the last two rows in the Table are, obviously, unreliable. Here the initial condA exceeds the inverse machine epsilon; hence round-off errors render the calculation of D_0 practically impossible.

Table

Scaling of Hilbert matrices

Order	Condition numbers			Difference
	Unscaled	Isoperimetric (I)	Optimum (O)	$\frac{I-O}{O}$ %
2	19	14	14	0
3	524	286	254	12
4	15512	7415	5875	26
5	4.76E+5	2.11E+5	1.52E+5	38
6	1.45E+7	5.86E+6	4.04E+6	45
7	3.05E+8	2.37E+8	1.59E+8	49
8	7.56E+8	2.18E+8	1.19E+8	85

4.3. In the third series we generated random matrices J with elements $J_{ij} = a_i x_j$ with $a_i \in (0, 1000]$ and $x_j \in [-1, 1]$, both a_i and x_j having uniform distributions. Then positive-definite matrices were built according to (2) and their condition evaluated with and without scaling for orders $2 \le n \le 40$. A checkerboard sign distribution can appear here by pure chance; therefore, the theory gives an upper bound for cond(DAD). The isoperimetric scaling ensured condition numbers of the same order of magnitude as the optimum one. Differences were computationally

insignificant and lay between 1 and 40% (most often 10-15%) in favour of optimum scaling. The only exception was a difference of 134% which appeared in a matrix with initial cond A $8 \cdot 10^8$. Evidently, this case is unreliable for the same reason as in 4.2.

4.4. The fourth series made use of actual iteration-stop matrices generated during γ-spectra processing by means of KATOK program /11/. In this code a physical scaling is currently used, so there was a possibility of comparison among condition numbers which fell in the following ranges:

unscaled matrices	$10^6 - 10^{14}$
physical scaling	$10^2 - 10^5$
isoperimetric scaling	$10^0 - 10^2$
optimum scaling	$10^0 - 10^2$

The quality of optimum scaling rarely exceeded that of the isoperimetric one by more than a few percent.

4.5. Last but not least, we incorporated the isoperimetric scaling in a computer code which modeled the data processing of continuous β-spectra and, in particular, their high-energy edges. The shape of such spectra is described by a reliable theory. This code was based on the well-known FUMILI minimization program /13/ which makes use of an iteration-step cut-off to avoid divergencies. Apart from introducing the scale-descale procedure nothing was changed in the code as a whole. Despite the fact that this actually meant an increase in the number of operations by n square roots and $n(n+1)$ multiplications/ divisions per iteration, processing time decreased sharply by a factor of 3-4 at the same quality of solution found. We explain this apparent paradox with the better condition of matrices involved which prevented the activation of the cut-off procedure in FUMILI and lead to the same solution in a much lower number of iterations. In other words, a more advantageous trajectory of the iteration process was achieved by means of scaling.

5. COMMENTS AND CONCLUSIONS

It is known /4,14/ that matrix condition in Euclidean norms may be regarded as a quantitative measure of the anisotropy introduced into R^n by the mapping A. When A is built in accordance with (2) it has the structure of a Gram matrix of n independent vectors in R^m. Therefore, cond A depends on these n vectors. One can easily imagine two limit cases of large anisotropy: when the vectors are orthogonal to each other but strongly differ in length and when all the lengths are equal but the space arrangement is very far from orthogonal. These two cases may be termed "linear" and "angular" respectively. In practice we deal with matrices which lie between but, nevertheless, the reasons

for their high condition numbers may be predominantly linear as in 4.3-4.5 or predominantly angular as in the case of Hilbert matrices. Actually, Bauer's theory of optimum scaling is a demonstration of two basic factors: (a) angular and linear causes of high condition numbers are interrelated, optimum scalers being expressed via A^{-1}, where lengths and angles combine together; (b) the angular arrangement puts a lower limit to cond A attainable by scaling (see /2/ Theorems IV-VII).

The results of our numerical experiments show that the reason of ill-condition in physical problems of the class formulated is often predominantly linear. That is why the isoperimetric scaling turns rather close to the theoretical optimum. This accounts for the drastic reduction of condition numbers reported in 4.4 as well as for the improved convergence in 4.5. On the contrary, when matrices treated are Hilbert-like ("predominantly angular"), scaling performs poorly. In such cases other mathematical means should be put into action (orthogonalization, regularization, etc.).

We are convinced that the measuring units in physics are usually selected on the basis of non-computational criteria. Therefore, it is advisable to use isoperimetric scaling whenever a necessity occurs to manipulate numerically matrices of the type considered - so much so that we plan to write a special routine for matrix inversion with built-in scale-descale procedure.

The author acknowledges the assistance of V.Brudanin in the computations described in 4.5.

REFERENCES

1. Parlett B.N. The Symmetric Eigenvalue Problem, Prentice-Hall Inc., Englewood Cliffs, N.J., 1980, ch.1.
2. Bauer F.L. Num.Math., 1963, vol.5, p.73-87.
3. Forsythe G.E., Moler C.B. Computer Solution of Linear Algebraic Systems. Prentice-Hall Inc., Englewood Cliffs, N.J., 1967, ch.11.
4. Forsythe G.E., Straus E.G. Proc.Am.Math.Soc., 1955, vol.6, No.3, p.340-5.
5. McKeeman W.M. Comm.ACM, 1962, vol.5, No.11, p.563-5.
6. Marquardt D.W. J.Soc.Ind.Appl.Math., 1963, vol.11, No.2, p. 431-41.
7. Himmelblau D.M. Applied Nonlinear Programming, McGraw-Hill Book Co., 1972, §3.1.2.
8. Bevington P.R. Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill Book Co., New York-San Francisco-St Louis-Toronto-London-Sydney, 1969, ch.11.

9. Bard Y. Nonlinear Parameter Estimation, Academic Press, New-York-San Francisco-London, 1974, §§5.8 and A.5.
10. Gadjokov V. Instr. and Exper.Technique, 1970, vol.5, p.1338-1341.
11. Gadjokov V. JINR, E10-12352, 12353, 12354, Dubna, 1979.
12. Marshall A.W., Olkin I. Inequalities: Theory of Majorization and Its Applications, Academic Press, New York-London-Toronto-Sydney-San Francisco, 1979, ch. 8-9.
13. Tyapkin A.A. (ed.). Statisticheskie Metody v Eksperimentalnoy Fizike, Atomizdat, Moscow, 1976, Appendix 3, pp. 319-326 (in Russian).
14. Forsythe G.E., Malcolm M.A., Moler C.B. Computer Methods for Mathematical Computations, Prentice-Hall Inc.,Englewood Cliffs, H.J., 1977.

Received by Publishing Department
on October 29, 1984.

В Объединенном институте ядерных исследований начал выходить сборник *"Краткие сообщения ОИЯИ"*. В нем будут помещаться статьи, содержащие оригинальные научные, научно-технические, методические и прикладные результаты, требующие срочной публикации. Будучи частью "Сообщений ОИЯИ", статьи, вошедшие в сборник, имеют, как и другие издания ОИЯИ, статус официальных публикаций.

Сборник "Краткие сообщения ОИЯИ" будет выходить регулярно.

The Joint Institute for Nuclear Research begins publishing a collection of papers entitled *JINR Rapid Communications* which is a section of the JINR Communications and is intended for the accelerated publication of important results on the following subjects:

Physics of elementary particles and atomic nuclei.
Theoretical physics.
Experimental techniques and methods.
Accelerators.
Cryogenics.
Computing mathematics and methods.
Solid state physics. Liquids.
Theory of condenced matter.
Applied researches.

Being a part of the JINR Communications, the articles of new collection like all other publications of the Joint Institute for Nuclear Research have the status of official publications.

JINR Rapid Communications will be issued regularly.

COMMUNICATIONS, JINR RAPID COMMUNICATIONS, PREPRINTS, AND PROCEEDINGS OF THE CONFERENCES PUBLISHED BY THE JOINT INSTITUTE FOR NUCLEAR RESEARCH HAVE THE STATUS OF OFFICIAL PUBLICATIONS.

JINR Communication and Preprint references should contain:

- names and initials of authors,
- abbreviated name of the Institute (JINR) and publication index,
- location of publisher (Dubna),
- year of publication
- page number (if necessary).

For example:

1. *Pervushin V.N. et al. JINR, P2-84-649, Dubna, 1984.*

References to concrete articles, included into the Proceedings, should contain

- names and initials of authors,
- title of Proceedings, introduced by word "In:"
- abbreviated name of the Institute (JINR) and publication index,
- location of publisher (Dubna),
- year of publication,
- page number.

For example:

Kolpakov I.F. In: XI Intern. Symposium on Nuclear Electronics, JINR, D13-84-53, Dubna, 1984, p.26.

Savin I.A., Smirnov G.I. In: JINR Rapid Communications, N2-84, Dubna, 1984, p.3.

Гаджоков В. E11-84-692

Квазиоптимальное масштабирование симметричных положительно определенных матриц

В работе рассмотрены различные подходы к масштабированию в переопределенных физических задачах. Для определенности наложено условие сохранения следа матриц при масштабировании. Показано, что масштабирование к единичной главной диагонали по Марквардту приводит к максимальному определителю. В численных экспериментах проведено сравнение этого метода с теоретически оптимальным. Приведены результаты, указывающие, что отличия в числах обусловленности матриц, масштабированных оптимально и по Марквардту, не существенны для численных расчетов. Рекомендуется более широкое использование масштабирования к единичной главной диагонали при работе с положительно-определенными матрицами.

Работа выполнена в Лаборатории ядерных проблем ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1984

Gadjokov V. E11-84-692

Quasi-Optimum Scaling of Symmetric Positive-Definite Matrices

Various approaches to scaling in overdetermined physical problems are considered. Trace invariance in scaling is imposed for the sake of definiteness. It is then shown that Marquardt's scaling to unit main diagonal leads to a maximum determinant. In a series of numerical experiments this sort of scaling is compared to the theoretical optimum. Results are reported which indicate that differences in condition of matrices scaled optimally and to unit main diagonal are computationally insignificant. A wider use of unit-diagonal scaling is recommended when treating positive-definite matrices.

The investigation has been performed at the Laboratory of Nuclear Problems, JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna 1984