High Dimensional and Large Span Data Least Square Error: Numerical Stability and Conditionality

Vaclav Skala
Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, CZ 306 14 Plzen, Czech Republic.

Corresponding author. URL: http://www.VaclavSkala.eu

Manuscript submitted May 15, 2017; accepted June 30, 2017.
doi:

Abstract: The Least Square Error (LSE) is widely used method in many engineering and computational problems solution. The LSE method is simple from the "formulation" point of view, offers a simple solution. However, it might lead to wrong or incorrect conclusions, especially if high dimensional or large span data are to be processed or if some variables have significantly smaller influence than the other does, e.g. inconvenient selection of measuring units.

In this paper, we analyze influence of row and column "normalization" inspired by the Gershgorin's theorem. The approach has been experimentally verified on a LSE application for high dimensional and large span data. The proposed approach was tested also on Hilbert's matrix inversion for conditional number change analysis.

Key words: Least square error, conditionality, robustness, Gershgorin's theorem, numerical stability.

1. Introduction

The Least Square Error (LSE) is used, Fig.1, as an approximation method in many areas [1]-[5]. However, it is rotationally and translationally invariant, dependent on variables unit etc. In addition, the LSE method is mostly used for a low dimensionality data, i.e. for a small number of parameters. The LSE is not a convenient method for implicitly defined problems $F(\boldsymbol{x}) = 0$, used especially in computer graphics and geometry fields. In this case the Total Least Square Error (TLSE), Fig.2, is to be used [6][7] which is computationally more expensive. A simple close solution in E^2 is based on a solution of quadratic equation [8]-[10].

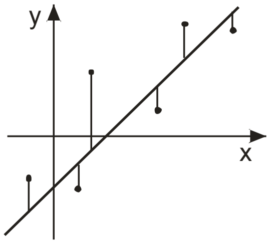


Fig.1: Least Square Error (LSE)

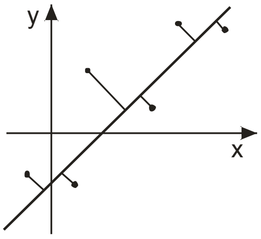


Fig.2: Total Least Square Error (TLSE)

The LSE leads to a solution of a linear system of equations $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$, which leads to problems with numerical stability, which is influenced by eigenvalues of the$\boldsymbol{A}^T\boldsymbol{A}$ matrix.

2. Conditionality and Gershgorin's Theorem

In numerical mathematics conditionality measures how much the output of computation depends on input changes, resp. on errors in measuring, and it is expressed by condition number $\kappa(\boldsymbol{A})$.

2.1. Condition number

The condition number is usually expressed as $\kappa(\boldsymbol{A}) = 10^k$ and it means that approx. k digits in precision of computation can be lost. Let $\boldsymbol{A}$ is a matrix $n \times n$ with a_{ij} elements. Eigenvalues λ_i of the matrix $\boldsymbol{A}$ are determined as

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0 \tag{1}$$

Solving the determinant a polynomial $P_n(\lambda)$ is obtained and roots are eigenvalues. Eigenvalues $\lambda_i \in C^1$are generally complex numbers. The condition number $\kappa(\boldsymbol{A})$ is determined as

$$\kappa(\boldsymbol{A}) = \frac{|\lambda_{max}|}{|\lambda_{min}|} \tag{2}$$

where λ_{max}, resp. λ_{min} are maximal, resp. minimal eigenvalues of the matrix $\boldsymbol{A}$. It is said that the $\boldsymbol{A}$ matrix with a high condition number is *ill-conditioned*.

However, it should be noted that:

- the condition number $\kappa(\boldsymbol{A})$ does not give precise estimation of inaccuracy
- in the case of linear systems of equations, the condition number depends on the order of rows (they can be swapped) and on the order of columns (unknown variables x_i can be re-indexed, as it is only a matter of decision, which physical phenomena will be x_i and which will be x_k)

3. Gershgorin's theorem

The Gershgorin's Theorem (GT) was published in 1931 and gives an estimation of eigenvalues position in the complex plane. It can be described as follows [11]:

Let $\boldsymbol{A}$ is a matrix $n \times n$ with a_{ij} elements. Let $r_i = \sum_{j \neq i}^{n} |a_{ij}|$ for $i = 1, \dots, n$. Let the Gershgorin's disc $D(a_{ii}, r_i)$ is a disc with a radius r_i centered at a_{ii}. Then every eigenvalue of the matrix $\boldsymbol{A}$ lies within one of the Gersgorin's disk $D(a_{ii}, r_i)$. The GT is valid also for the transposed matrix $\boldsymbol{A}$, i.e. $\boldsymbol{A}^T$.

As the conditional number $\kappa(\boldsymbol{A})$ is not fixed in the case of linear system of equations, as row and columns can be swapped depending on re-indexing of variables or on an order of equations. Therefore the Gershgorin's disc $D(a_{ii}, r_i)$ differs for different permutation of rows and columns [11][12].

4. Least Square Error Method

The Least Square Error (LSE) method is widely used in many problems solution, mostly for an optimal approximation of problems formulated in explicit forms, i.e. as $y = f(\boldsymbol{x})$, where $\boldsymbol{x}$ is a vector of known values [13][14]. Let us consider overdetermined system of equations

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{3}$$

where $\boldsymbol{A}$ is a matrix $n \times m$ and $n > m$. An error $\|\boldsymbol{r}\|$ of approximation can be expressed as

$$\|\boldsymbol{r}\| = \|\boldsymbol{Ax} - \boldsymbol{b}\| \tag{4}$$

The LSE method minimizes the square of the error, i.e. minimizes $\|\boldsymbol{r}\|^2$. The condition for an extrema are given as:

$$\frac{\partial \|\boldsymbol{r}\|^2}{\partial \boldsymbol{x}} = \frac{\partial [(\boldsymbol{Ax} - \boldsymbol{b})^T (\boldsymbol{Ax} - \boldsymbol{b})]}{\partial \boldsymbol{x}} = 2(\boldsymbol{A}^T\boldsymbol{Ax} - \boldsymbol{A}^T\boldsymbol{b}) = \boldsymbol{0} \qquad \boldsymbol{A}^T\boldsymbol{Ax} = \boldsymbol{A}^T\boldsymbol{b} \tag{5}$$

Let us consider a non-singular squared matrix $\boldsymbol{B}$ $n \times n$. It can be seen, that the condition number $\kappa(\boldsymbol{B}^T\boldsymbol{B})$ is getting significantly higher as

$$\kappa(\boldsymbol{B}^T\boldsymbol{B}) = \frac{{\lambda_{max}}^2}{{\lambda_{min}}^2} \tag{6}$$

where λ_{max}, resp. λ_{min} are maximal, resp. minimal eigenvalues of the matrix $\boldsymbol{B}$. It means, that if the condition number $\kappa(\boldsymbol{B}) = 10^3$, then the condition number of $\kappa(\boldsymbol{B}^T\boldsymbol{B}) = 10^6$.

It means that the LSE method produces ill-conditioned system of equations in general.

4.1. Matrix normalization

Let us consider linear system of equations $\boldsymbol{Ax} = \boldsymbol{b}$, i.e. the result of the LSE method application. The Gershgorin's Theorem (GT) gives an estimation of eigenvalues position in a complex space. It actually says that if a column $a_{*,j}$ is multiplied by $q_j \neq 0,\ j = 1, \dots, n$ the position and radius for the eigenvalue $\lambda_j \neq 0$ is changed accordingly to the GT. This operation reflects actually the scaling of the variable x_j, i.e. physical unit change in physical applications. On the other hand, a similar process can be made for rows a column $a_{i,*}$ is multiplied by $p_i \neq 0$, $j = 1, \dots, n$ and right hand side must be multiplied by p_i as well.

As the GT depends on the order of rows and columns of the matrix $\boldsymbol{A}$, therefore a small modification is needed, i.e. the diagonal items is not excluded. This resulted to a sequence:

```
for j:=1 to n do
{  q:=sum(|a[*,j]|); /* sum of absolute values in the column including the diagonal one
*/
   A[*,j]:=A[*,j]/q  /* column normalization and scaling of x[j] */
}
for i:=1 to n do
{  p:= sum(|a[i,*]|); /* sum of absolute values in the row including the diagonal one */
   A[i,*]:=A[i,*]/p;  /* row normalization */
   b[i]:=b[i]/p /* right hand side modification */
}
```

The first cycle actually "balances" influence of unknown variables x_j, i.e. scaling their physical units, while the second cycle normalizes the numerical ranges in which each row is to be computed.

It should be noted, that in the real implementation, the division operation should not be used directly, as it changes variable's mantissa. The exponent subtraction operation should be used instead, i.e.

```
exponent(A[*,j]):= exponent(A[*,j]) - exponent(q)
```

as we do not lose the mantissa precision. Also for more precise computation of the $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}^T\boldsymbol{b}$ a special summation algorithms for summation should be used [15], especially in the case of large data spans. It is based on exponent hashing and intended for summation of large data sets with a large span of data

4.2. Practical consequences

Let us consider a data set $\Omega = \{\langle x_i, y_i, f_i \rangle\}_{i=1}^{n}$, i.e. data set containing for x_i and y_i and measured functional value f_i, and we want to find parameters $\boldsymbol{a} = [a, b, c, d]^T$ for optimal fitting function:

$$f(x, y, \boldsymbol{a}) = a + bx + cy + dxy \qquad (7)$$

Minimizing the *vertical* squared distance D, i.e.:

$$D = \min_{a,b,c,d} \sum_{i=1}^{n} \left(f_i - f(x_i, y_i, \boldsymbol{a})\right)^2 \qquad \min_{a,b,c,d} \sum_{i=1}^{n} \left(f_i - (a + bx_i + cy_i + dx_iy_i)\right)^2 \qquad (8)$$

Conditions for an extreme are given as:

$$\frac{\partial f(x, y, \boldsymbol{a})}{\partial \boldsymbol{a}} [1, x, y, xy]^T = 0 \qquad (9)$$

Applying this on the expression of D we obtain

$$\frac{\partial D}{\partial \boldsymbol{a}} = 2 \sum_{i=1}^{n} (f_i - (a + bx_i + cy_i + dx_iy_i)) \frac{\partial f(x, y, \boldsymbol{a})}{\partial \boldsymbol{a}} = 0 \qquad (10)$$

It leads to conditions for $\boldsymbol{a} = (a, b, c, d)$ parameteters in the form of a linear system of equations $\boldsymbol{Ax} = \boldsymbol{b}$:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_iy_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_iy_i & \sum_{i=1}^{n} x_i^2y_i \\ \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_iy_i & \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} x_iy_i^2 \\ \sum_{i=1}^{n} x_iy_i & \sum_{i=1}^{n} x_i^2y_i & \sum_{i=1}^{n} x_iy_i^2 & \sum_{i=1}^{n} x_i^2y_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} f_i \\ \sum_{i=1}^{n} f_ix_i \\ \sum_{i=1}^{n} f_iy_i \\ \sum_{i=1}^{n} f_ix_iy_i \end{bmatrix} \qquad (11)$$

The bilinear form was used to show the LSE method application to a non-linear case; in the case of a linear function, i.e. $f(x, y, \boldsymbol{a}) = a + bx + cy$, the 4th row and column are to be removed. The matrix $\boldsymbol{A}$ is symmetric and the function $f(\boldsymbol{x})$ might be more complex, in general.

Several methods for LSE have been derived [1]-[4], however those methods are sensitive to the vector $\boldsymbol{a}$ orientation and not robust in general as a value of $\sum_{i=1}^{n} x_i^2 y_i^2$ might be too high in comparison with the value n, which has an influence to robustness of a numerical solution. The LSE methods are sensitive to a rotation as they measure vertical distances. Rotational and translation invariance are fundamental requirements especially in geometrically oriented applications [8][9].

The LSE method is usually used for a small size of data, low dimensionality and also a span of a domain is relatively small. However, in some applications the domain span can easily be over several decades, e.g. in the case of Radial Basis Functions (RBF) approximation for GIS applications etc. In this case, the overdetermined system can be difficult to solve [16]

Let us consider a recent simple example again, when points are generated from $(x_i, y_i) \in \langle 10, 10^5 \rangle \times \langle 10, 10^5 \rangle$. It can be found using MATLAB that conditional number $cond(\boldsymbol{A}^T\boldsymbol{A}) \cong 6.10^{10}$, see Fig.3.

Using the approach presented above, the conditional number was decreased significantly to $cond\left(\overline{\boldsymbol{A}^T\boldsymbol{A}}\right) \cong 2.10^6$.

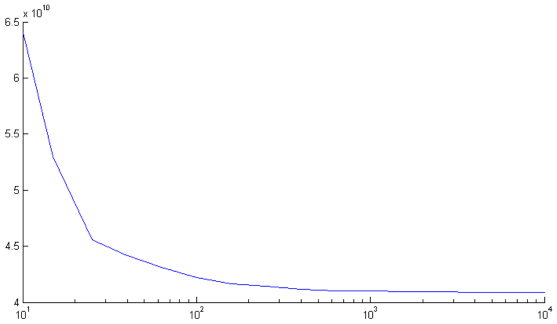


Fig.3: Conditionality of the original matrix depending on data set size, i.e. number of points

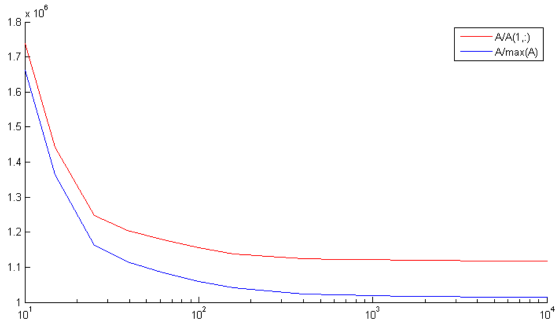


Fig.4: Conditionality of the modified matrix depending on data set size, i.e. number of points

Comparing the condition numbers of the original matrix $\boldsymbol{A}$ and modified matrix $\boldsymbol{A}'$, we can see significant improvement of matrix conditionality as

$$v = {cond(\boldsymbol{A}^T\boldsymbol{A})} \Big/ {cond\left(\overline{\boldsymbol{A}^T\boldsymbol{A}}\right)} \cong \frac{6.10^{10}}{2.10^6} = 3.10^4 \qquad (12)$$

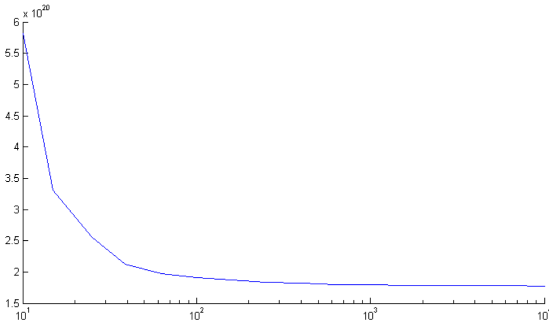


Fig.5: Conditionality of the original matrix depending on data set size, i.e. number of points

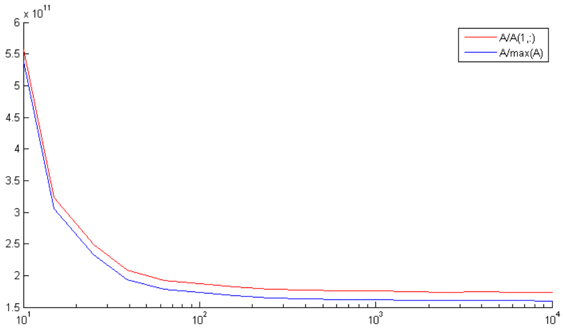


Fig.6: Conditionality of the modified matrix depending on data set size, i.e. number of points

In the case of a little bit more complex function defined by Eq.(7), i.e. $y = a + bx + cy + dxy$ higher condition number and higher improvements are obtained, Fig.5 and Fig.6.

In this case of the LSE defined by Eq.(7) the conditionality improvement is even higher, as

$$\upsilon = {cond(\boldsymbol{A}^T\boldsymbol{A})} \Big/ {cond\left(\overline{\boldsymbol{A}^T\boldsymbol{A}}\right)} \cong \frac{6.10^{20}}{6.10^{11}} = 10^9 \tag{13}$$

It means that better numerical stability is significantly improved by a simple operation. All graphs clearly shows also dependency on a number of points used in the experiments (horizontal axis). As a testing example, the Hilbert's matrix can be used, as it is extremely ill-conditioned matrix.

5. Hilbert's Matrix Conditionality

We should answer a question, how the conditional number of the Hilbert's matrix can be improved if orthogonal basis is used instead of orthonormal one as an experimental test .

A simple experiment can prove that the proposed method does not practically change the conditionality of the Hilbert's matrix $\boldsymbol{H}_n(0,1)$. However, as the LSE approximation is to be used for large span of data, it is reasonable to consider a general case and explore conditionality of the $\boldsymbol{H}_n(a,b)$ matrix, e.g. $\boldsymbol{H}_5(0,b)$.

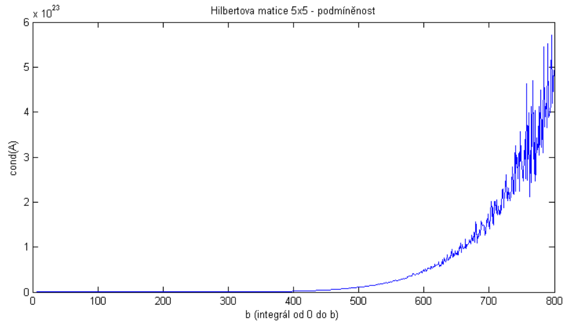


Fig.7: Conditionality of the *original* $\boldsymbol{H}_5(0,b)$ (numerical problems occur for $b > 650$)

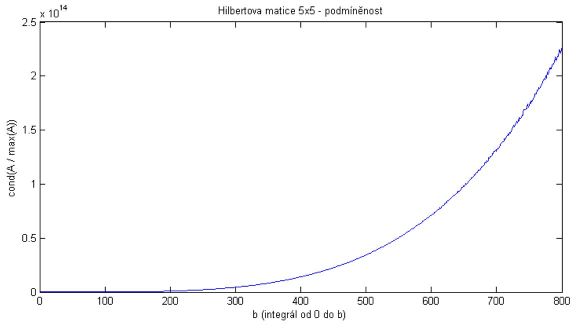


Fig.8: Conditionality of the *modified* $\boldsymbol{H}_5(0,b)$

It can be seen, that $cond(\boldsymbol{H}_5(0{,}800)) = 6.10^{23}$. If the proposed approach is applied $cond\left(\overline{\boldsymbol{H}_5(0{,}800)}\right) = 2{,}5.10^{14}$ for the modified matrix, Fig.7 - Fig.8.

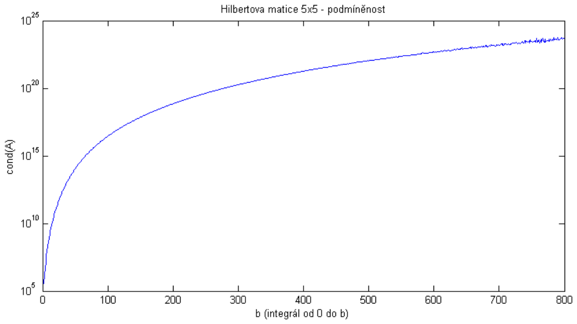


Fig.9: Conditionality of the *original* $\boldsymbol{H}_5(0,b)$ (log scale on vertical axis is used)

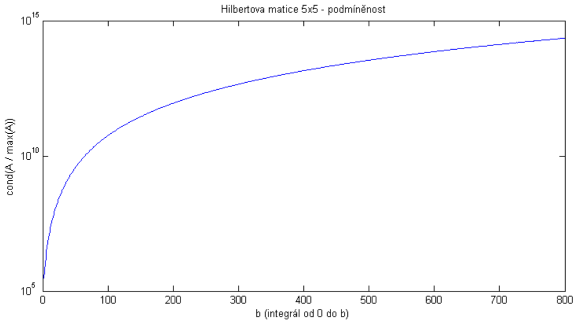


Fig.10: Conditionality of the *modified* $\boldsymbol{H}_5(0,b)$ (log scale on the vertical axis is used)

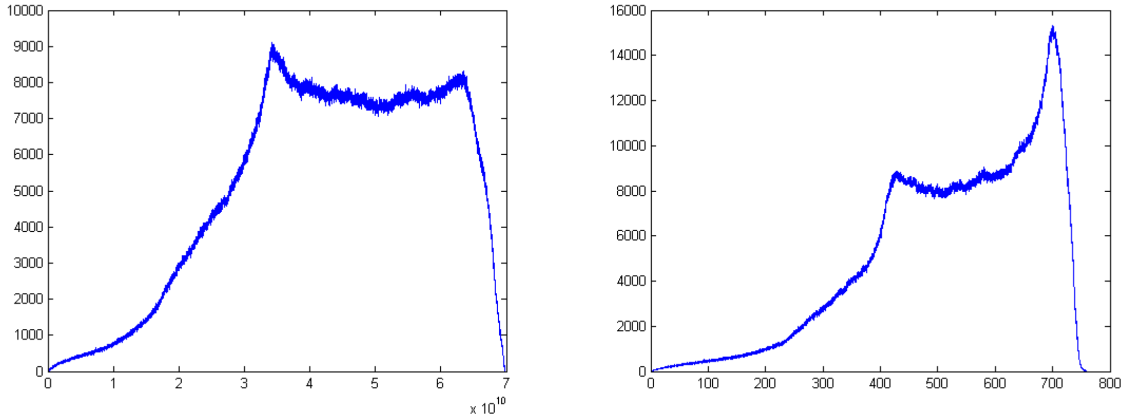
It means that the conditionality improvement

$$\upsilon = \frac{cond\left(\boldsymbol{H}_5(0{,}800)\right)}{cond\left(\overline{\boldsymbol{H}_5(0{,}800)}\right)} \cong \frac{6.10^{23}}{2{,}5.10^{14}} \approx 10^9 \tag{14}$$

This is a similar ratio as for the simple recent examples.

5.1. Experimental evaluation

Let $\boldsymbol{\beta}_i$ is a row the the matrix $\boldsymbol{A}^T\boldsymbol{A}$. It is actually a vector in d-dimensional space. Then the value of the bivector $\|\boldsymbol{\beta}_i \wedge \boldsymbol{\beta}_j\|$ is the size of an oriented area on the plane $\boldsymbol{e}_i\boldsymbol{e}_j$ in that space. It means that if $\boldsymbol{\beta}_i$ vectors are "normalized" a better conditional number is obtained. The proposed approach has been used for St. Helen's volcano RBF approximation by 10 000 points instead of 6 743 176 original points, Fig.12., i.e. 10 000 parameters (weights of the RBFs needed to be determined). This leads to a compression $1{:}\,(67)^2$. In this example of the (RBF) approximation, the polynomial reproduction leads to significant numerical problems analyzed in [16]. A change of the size of the bivectors $\|\boldsymbol{\beta}_i \wedge \boldsymbol{\beta}_j\|$ is a practical result of the application to the RBF approximation, which changes from the interval $\langle eps, 10^{10}\rangle$ to $\langle eps, 8.10^2\rangle$ only, which significantly increases robustness of the RBF approximation, Fig.11.



a) Original matrix (horizontal axis multiplied by 10^{10})

b) Modified matrix (horizontal axis is not multiplied)

Fig.11: Histogram of bivector sizes $\|\boldsymbol{\beta}_i \wedge \boldsymbol{\beta}_j\|$ of the for original matrix and modified one

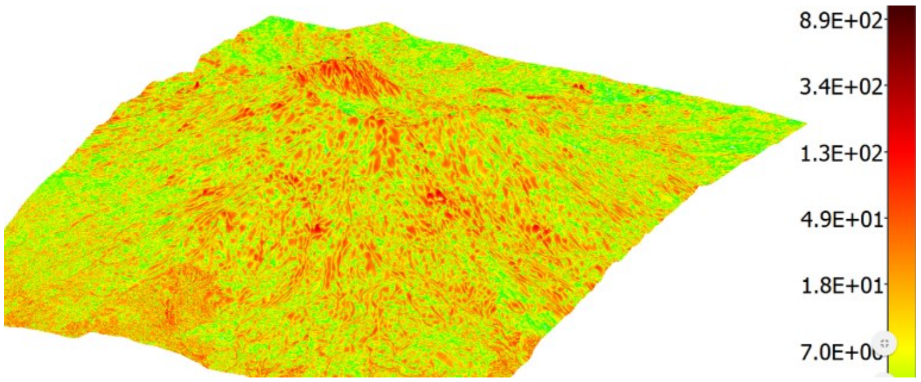


Fig.12: Approximation error of the RBF approximation of the St. Helen's volcano

6. Conclusion

In this paper, we pointed out to the weak points of the Least Square Error (LSE) method application and some relationships with the geometric algebra approach. If the LSE method is used for big data with a large span or units of unknown x_i variables are selected insensitively, the LSE method can produce incorrect results due to very bad conditionality as the LSE matrix is ill-conditioned.

The problem can be partially solved by the row and column "normalization" described in the paper. The proposed method decreases the condition number of a matrix used in the LSE method and increases robustness of a numerical solution especially when domain data range is high. However a special algorithm for precise summation is recommended for high range data sets It can be used also for solving systems of linear equations in general, e.g. if radial basis function interpolation or approximation is used.

Acknowledgment

The author would like to thank to colleagues at the University of West Bohemia and to anonymous reviewers for their comments, which helped to improve the manuscript significantly. Special thanks belong to Zuzana Majdišová and Michal Šmolík for experiments verification, images and generation in MATLAB.

References

(Book)

[1] Charpa,S., Canale,R. (1988). *Numerical methods for Engineers*, McGraw-Hill.

[2] Chatfield,C. (1970). *Statistics for technology*, Penguin Book.

[3] Kryszig,V. (1983). *Advanced engineering mathematics*, John Wiley & Sons.

[4] Levy,D. (2010). *Introduction to Numerical Analysis*, Univ. of Maryland

[5] Jo,S., Kim,S.W. (2005). Consistent normalized least mean square filtering with noisy data matrix. *IEEE Trans. Signal Processing*, Vol. 53, No. 6, 2112–2123.

[6] Van Huffel,S., Lemmerling,P. (2002). *Total least squares and errors-in-variables modeling: Analysis, algorithms and applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers

[7] Lee,S.L. (1994). A note on the total least square fit to coplanar points, *Tech.Rep. ORNL-TM-12852*, Oak Ridge National Laboratory.

[8] Alciatore,D., Miranda,R. (1995). The best least-square line fit, *Graphics Gems V*, 91-97, Academic Press.

[9] Skala,V. (2016). Total least square error computation in E2: A New Simple, Fast and Robust Algorithm, *Proc. CGI 2016*, ACM, pp.1-4, Greece

[10] Skala,V. (2016). A new formulation for total least square error method in d-dimensional space with mapping to a parametric line, *ICNAAM 2015, AIP Conf. Proc.1738*, pp.480106-1-4, Greece

[11] Golub,G.H., Van Loan,C.F. (1980). An analysis of the total least squares problem*. SIAM J. on Numer. Anal.*, 17, 883–893.

[12] Geršgorin,S., (1931). "Über die Abgrenzung der Eigenwerte einer Matrix", Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na, 1931, no. 6, 749–754

[13] Nixon,M.S., Aguado,A.S. (2012). *Feature extraction & image processing for computer vision*, Acad. Press.

[14] DeGroat,R.D., Dowling,E.M. (1993). The data least squares problem and channel equalization. *IEEE Trans. Signal Processing*, Vol. 41, No. 1, 407–411.

[15] Skala,V. (2013). Summation Problem Revisited: More Robust Computation, *17th Int. Conf. on Computers – Recent Advances in Comp. Science CSCC'13*, pp. 56-64, ISBN 978-960-474-311-7, Greece

[16] Skala,V. (2017). RBF Interpolation with CSRBF of Large Data Sets, accepted for publication in *Proc. ICCS 2017*, Procedia Computer Science, Elsevier, pp.2433-2437, Switzerland

Vaclav Skala received his MSc. from the Institute of Technology, Plzen, in 1975, his PhD from the Czech Technical University, Prague, in 1981. In 1988, he was appointed as a Reader at the Institute of Technology, Plzen and was habilitated at the University of West Bohemia in 1989. He was appointed as a full Professor of Computer Science at the University of West Bohemia, Plzen in 1996. Currently he is the Head of the Center of Computer Graphics and Visualization (http://Graphics.zcu.cz).

His research interests are in fundamental algorithms and algorithms for computer graphics, visualization, computational geometry, interpolation and approximation methods, namely radial basis functions. He is the organizer of the WSCG conferences on Computer Graphics, Visualization and Computer Vision (http://www.wscg.eu).

Prof.Vaclav Skala is the Editor-in-Chief of the Journal of WSCG, a Fellow and a member of the Eurographics Association, Advisory Board member of the Computers & Graphics and Machine Graphics & Vision journals. Recently, he was a member of the Computer Graphics Forum, The Visual Computer, The Int. Journal of Virtual Reality, Journal of Object Oriented Technologies editorial boards. He also has served as an IPC member of several international conferences. Project supported by the Czech Science Foundation, project GACR17-05534S