THE STATISTICAL BEHAVIOUR OF CORRELATED DATA

Alessandra Colombo, Luigi Mussio
Politecnico di Milano, Dip. I.I.A.R. sez. Rilevamento
p.za Leonardo da Vinci 32, 20133 Milano (ITALY)

Commission VI, Working Group 3

KEY WORDS: data processing, correlation, estimation, accuracy, reliability

ABSTRACT

Nowadays larger and larger databases are available in order to evaluate parameters in many working problems. So a very intresting item is understanding if larger and larger data sets can always bring better parameter estimations as many cross-correlations are often present among the data in large data sets.

1. INTRODUCTION

Nowadays the modern technologies return very large data sets available for parameters estimations in many kinds of works. The cross-correlation links among data increase dramatically in large data sets and so it is a very important issue to understand and to analyse how the data set size increase, that is always taken by a cross-correlation step increase, can affect and modify the quality of the estimated parameters.

2. WORK AIM

The aim of this work is to analyse the effects of data set cross-correlations on the quality of the estimated parameters in order to understand if it is always reasonable to work with large data sets.
One very important question to begin with, it is if the data set numerousness might be invalidated by its internal interdipendence. In fact the parameter estimations obtained using large data sets, can be exact and reliable only if the cross-correlation level among the data is low. The modern technologies, that return vast amounts of data, do not always warrant low cross-correlation levels and so do not assure high quality for the estimated parameters

3. WORK METHOD

In order to analyse if and how the data set numerousness can be invalidated by its interdipendence some parameters have been estimated beginning from different kinds of data set (different combinations of size and cross-correlation level).
The data sets created for the analysis are characterized by two features:

- the size
- the cross-correlation level

Two different kinds of data set sizes have been fixed:
1) small size (holding 100 elements)
2) large size (holding 10000 elements).

Tree different degrees of data sets cross-correlation levels have been fixed:
1) zero level (cross-correlation step = 1)
2) medium level (cross-correlation step = 10)
3) high level (cross-correlation step = 100)

In this way the work method consists in comparing together the quality of parameter estimations using these following four different kinds of data sets:
I. small and not cross-correlated data set
II. large and not cross-correlated data set
III. large data set and medium cross-correlation level
IV. large data set and high cross-correlation level

4. TEST EXAMPLES

These tests have been implemented using data management, time-optimization and storage-optimization procedures. In such a way it is possible to run huge data sets in very short time legs and using low cost equipments; those benefits are obtained, of course, trought a very sophisticated and advanced coding, involving sparse matrices, gridded data and array algebra.
In photogrammetry, remote sensing, surveying and geodesy many problems are about parameter estimations, in 1D and 2D domains.

In this work some problems of parameter estimations in 1D and 2D have been solved beginning from the four kinds of the above-mentioned data sets in order to make the described analysis.
The problems which are here analysed have been chosen from the most representative problems in 1D and 2D domains.

The objects considered in 1D domain and called time-problems, are:

a) mean value
b) coefficients of a regression line

As long as the 2D domain is concerned, the following items have been considered and called spatial-problems:

c) mean value
d) coefficients of an interpolating plane
e) coefficients of an affine transformation
f) coefficients of a trasformation

Notice that the parameter estimations of the affine transformation repets two times the parameter estimation used for the interpolating plane.
The parameters, of all the above-mentioned six problems, have been valued starting from all the four kinds of data sets.
In such a way the quality analysis for the parameter estimations have been made in many different work combinations, in order to have quite a complete and reliable view of the problem. In such a way the obtained results are pertinent, either to time problems and to spatial problems, and it is possible to have a general view of the statistical behaviour of correlated data sets.

5 THE STOCHASTIC MODEL

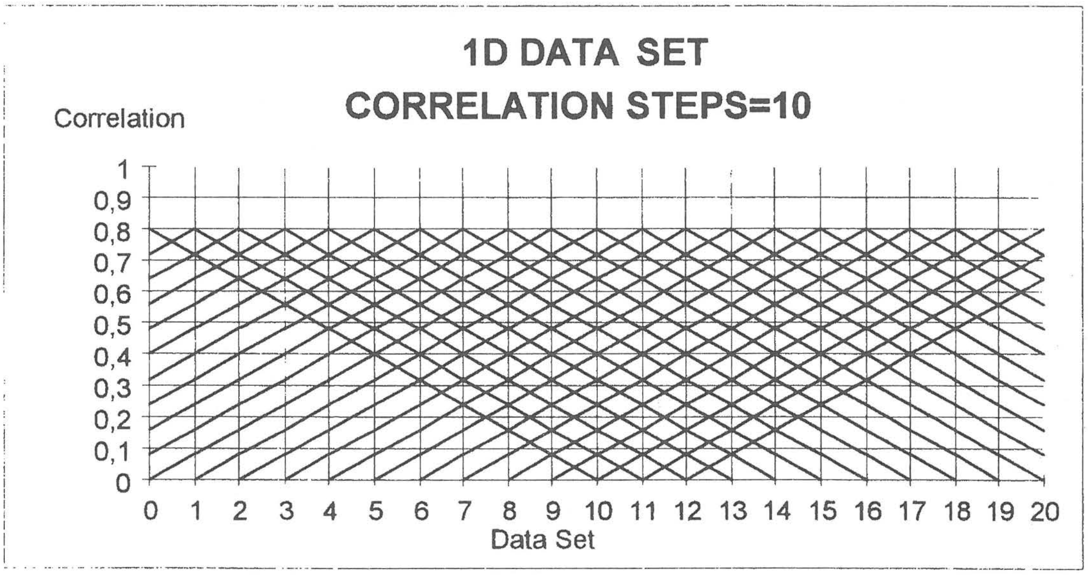
The quality analysis was made comparing the results obtained from non cross-correlated data sets, medium cross-correlated data sets and high cross-correlated data sets.
The data set cross-correlartion level is described by the value of its cross-correlation step which represents the number of data with which one single datum in the data set is correlated.
The cross-correlation links are different in 1D and 2D.
In 1D data sets the following models have been used:

I. Cross-correlation step 1 means that each datum is correlated only with itself (signal to noise ratio equal 0.8) and so there is no cross-correlation among the data.
II. Cross-correlation step 10 means that each datum is correlated with itself (signal to noise ratio equal 0.8) and with 10 other data placed in advance and with 10 more data following in the data set structure (correlation degree linearly decreases from 0.8 to 0.0) and so each datum is cross-correlated with 20 others data.
III. Cross-correlation step 100 means means that each datum is correlated with itself (signal to noise ratio equal 0.8) and with 100 other data placed in advance and 100 more data following in the data set structure (correlation degree linearly decreases from 0.8 to 0.0) and so each datum is cross-correlated with 200 others data.

Picture 1 descrives the cross-correlation links in a data set of 20 elements with cross-correlation step equal to 10.

In 2D data sets the following models have been used:

I. Cross-correlation step 1 means that each datum is correlated only with itself (signal to noise ratio equal 0.8) and so there is no cross-correlation between the data.
II. Cross-correlation step 10 means that each datum is correlated with itself (signal to noise ratio equal 0.8) and with a square of side 21 data placed just around the datum (correlation degree linearly decreases from 0.8 to 0.0) and so each datum is cross-correlated with 440 others data.
III. Cross-correlation step 100 means that each datum is correlated with itself (signal to noise ratio equal 0.8) and with a square of side 201 data placed just around the datum (correlation degree linearly decreases from 0.8 to 0.0) and so each datum is cross-correlated with 40400 others data. The dimensions of the here used data base are 100 and 10000 and so in the second case each datum is cross-correlated with the entire data base.



Picture 1 - Cross-correlation links in a data set of 20 elements with a cross-correlation step 10

6. THE RESULTS OF THE ANALYSIS

The following tables number 1, 2, 3, 4 and 5 show the results obtained in the described analysis.
Each table shows the quality (in terms of mean square deviation) of the estimated parameters, obtained in all the six mentioned problems, starting from the described four different data sets.

I) $\overline{\mathbf{X}} = \mathbf{E}(\mathbf{x})$	σ
data set Numerousness 100 Correlation steps 1	0.10
data set Numerousness 10000 Correlation steps 1	$0.10\ 10^{-1}$
data set Numerousness 10000 Correlation steps 10	$0.29\ 10^{-1}$
data set Numerousness 10000 Correlation steps 100	$0.89\ 10^{-1}$

Table 1 - The quality analysis for 1D mean value

II) Y= aX+b	$\sigma_{1,1}$	$\sigma_{2,2}$
data set Numerousness 100 Correlation steps 1	0.20	$0.35\ 10^{-2}$
data set Numerousness 10000 Correlation steps 1	$0.20\ 10^{-1}$	$0.34\ 10^{-5}$
data set Numerousness 10000 Correlation steps 10	$0.57\ 10^{-1}$	$0.99\ 10^{-5}$
data set Numerousness 10000 Correlation steps 100	0.18	$0.31\ 10^{-4}$

Table 2 - The quality analysis for regression line

III) $\overline{\mathbf{X}} = \mathbf{E}(\mathbf{x})$	σ
data set Numerousness 100 Correlation steps 1	0.10
data set Numerousness 10000 Correlation steps 1	$0.10\ 10^{-1}$
data set Numerousness 10000 Correlation steps 10	$0.77\ 10^{-1}$
data set Numerousness 10000 Correlation steps 100	0.43

Table 3 - The quality analysis for 2D mean value

IV) and V) Z=aX+bY+c	$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{3,3}$
data set Numerousness 100 Correlation steps 1	0.29	$0.35\ 10^{-1}$	$0.35\ 10^{-1}$
data set Numerousness 10000 Correlation steps 1	$0.27\ 10^{-1}$	$0.35\ 10^{-3}$	$0.35\ 10^{-3}$
data set Numerousness 10000 Correlation steps 10	0.19	$0.25\ 10^{-2}$	$0.25\ 10^{-2}$
data set Numerousness 10000 Correlation steps 100	0.75	$0.86\ 10^{-2}$	$0.86\ 10^{-2}$

Table 4 - The quality analysis for an interpolating plane and for a row of an affine trasformation

VI) $\begin{cases} \mathbf{X} = \Delta\mathbf{X} + \mathbf{aX} + \mathbf{bY} \\ \mathbf{Y} = \Delta\mathbf{Y} - \mathbf{bX} + \mathbf{aY} \end{cases}$	$\sigma_{1,1}$	$\sigma_{2,2}$	$\sigma_{3,3}$	$\sigma_{4,4}$
data set Numerousness 100 Correlation steps 1	0.22	0.22	$0.25\ 10^{-1}$	$0.25\ 10^{-1}$
data set Numerousness 10000 Correlation steps 1	$0.20\ 10^{-1}$	$0.20\ 10^{-1}$	$0.24\ 10^{-3}$	$0.24\ 10^{-3}$
data set Numerousness 10000 Correlation steps 10	0.15	0.15	$0.18\ 10^{-2}$	$0.18\ 10^{-2}$
data set Numerousness 10000 Correlation steps 100	0.61	0.61	$0.61\ 10^{-2}$	$0.61\ 10^{-2}$

Table 5 - The quality analysis for a S-transformation

The tables show very clearly that, either in 1D and in 2D domains, the quality of estimated parameters is tightly linked both to the data set size and to the data cross-correletion level. In fact, increasing the data set size (size increase factor 10^2) the quality of the estimated parameters increases (mean square deviation decrease factor between 10^{-1} and 10^{-3}) only if the cross-correlation level does not increase too. If the data set cross-correlation level increases (in the examples increasing factors 10 and 10^2) the quality of the estimated parameters decreases dramatically (mean square deviation increase factor between 10^1 and $1.5\text{x}10^1$).

6. FINAL REMARKS

The results confirm that data set size increase does not always guarantee estimated parameters quality improvements.
The estimated parameters quality can improve on only if new cross-correlation links are not created in the data set while increasing its size because there is a close relationship between the data set correlation increase and its information quality decrease.

References:

1. AKAIKE Hirotugu (1973): Block Toeplitz Matrix Inversion. SIAM J. Appl. Math., vol. 24, n°. 2, pp. 234-241.
2. CUTHILL E., Mc KEE J: Reducing the bandwidth of sparse symmetric matrices, Proc. ACM National Conference , Association for computing machinary, New York, 1969.
3. GIBBS N. E., POOLE W. G., STOCKMEYER P.K. (1976): An Algorithm for reducing the bandwidth and profile of sparse matrix, SIAM Numerical Analysis, vol. 13, n. 2, 1976
4. HESTENES M.R., STIEFEL E.: Methods of Conjugate Gradients for Solving Linear Systems, Journal of Research of national Bureau of Standards, vol. 46, n. 6, december 1952
5. KERSHAW P. S.: The Incomplete Cholesky Conjugate Gradient Method for the Iterative Solution of Systems of Linear Equations, Journal of Computational Physics, n. 26, Gennary, 1978
6. MEIJERING J. A., VAN DER VORST H.A. : An Iterative Solution Method for Linear System of which the Coefficient Matrix is a Symmetric M-Matrix, mathematics of Computation, vol. 31, n. 138, Gennary 1977.
7. TRENCH William F. (1964): An Algorithm for the Inversion of Finite Toeplitz Matrices. J. Soc. Indust. Appl. Math., vol. 12, no 3, pp. 515-522, September 1964.
8. TRENCH William F. (1974): Inversion of Toeplitz Band Matrices. Mathematics of Computation, vol. 28, no. 128, pp. 1089-1095.