

Completing Correlation Matrices of Arbitrary Order by Differential Evolution Method of Global Optimization: A Fortran Program

SK Mishra
Dept. of Economics
North-Eastern Hill University
Shillong (India)

Introduction: A product moment correlation matrix R of order n is a (square) symmetric positive semi-definite matrix such that $r_{ij} = r_{ji} \in R$ lies between -1 and 1 . Moreover, $r_{ii} = 1$. Each r_{ij} is the cosine of angle θ between two variates, say x_i and x_j ; $i, j \in \{1, 2, \dots, n\}$. Such matrices have many applications, particularly in marketing and financial economics as reflected in the works of Chesney and Scott (1989), Heston (1993), Schöbel and Zhu (1999), Tyagi and Das (1999), Xu and Evers (2003), etc. The need to forecast demand for a group of products in order to realize savings by properly managing inventories requires the use of correlation matrices (Budden et al. 2007).

In some cases, the matrix available to the analyst/decision-maker is complete, but it is an invalid (not positive semi-definite) correlation matrix. There could be many reasons that give rise to such invalid matrices (Mishra, 2004). In such cases, the problem is to obtain a positive semi-definite approximate correlation matrix, which, in some sense, is closest to the given invalid matrix. A number of methods have been developed to obtain such nearest correlation matrices. The works of Rebonato and Jäckel (1999), Higham (2002), Anjos et al. (2003), Pietersz and Groenen (2004), Grubisic and Pietersz (2004) and Mishra (2004) are some of them.

In many cases, however, due to paucity of data/information or dynamic nature of the problem at hand, it is not possible to obtain a complete correlation matrix. Some elements of R are unknown. In such cases, the question of validity (positive semi-definiteness) or otherwise (of an incomplete correlation matrix) does not arise. Instead, the problem is to obtain a valid complete correlation matrix. In absence of sufficient side conditions that are often impracticable to specify, this problem cannot be solved uniquely.

Several methods have been suggested to complete a correlation matrix - that is to obtain a valid complete correlation matrix from an incomplete correlation matrix (some of whose elements are unknown). Works of Johnson (1980), Barrett et al. (1989), Helton et al. (1989), Grone et al. (1984), Barrett et al. (1998), Laurent (2001), Kahl and Günther (2005), etc are notable.

In view of non-unique solutions admissible to the problem of completing the correlation matrix, some authors have suggested numerical methods that provide ranges to different unknown elements. Stanley and Wang (1969), Glass and Collins (1970) and

Olkin (1981) have suggested very efficient methods to find such ranges for the unknown elements of very small correlation matrices (of order $n < 4$). Budden et al. (2007) suggest a method to obtain the ranges of missing values of elements of a 4×4 incomplete correlation matrix whose first row elements are known. With the known elements in the first row, the method sets the range for r_{23} and one has to specify its value in that range. Once the value of r_{23} is chosen (within the specified range set for it), the method yields the range in which r_{24} would lie. One has to specify the value of r_{24} within the given range, which yields the range for r_{34} . Thus the matrix is completed. In this procedure it is obvious that the ranges on latter elements are contingent upon the choice of values of former elements. Further, Budden's method is limited to a 4×4 correlation matrix.

Objective of the Present Paper: Our objective in this paper is to suggest a method (and provide a Fortran program) that completes a given incomplete correlation matrix of an arbitrary order. The resulting complete matrices are many in number, but all of them are valid (positive semi-definite – with all non-negative eigenvalues). Additionally, the suggested method does not require any pre-assigned pattern as in case of Budden's method. It allows for holes (unknown elements) in any row and any column. The program that works out such complete matrices does not require any interaction with the user either.

The Method: The method proposed here is based on the Differential Evolution (DE) procedure of global optimization (Storn and Price, 1995). It generates a random population of elements that fit the holes (m in number) in the given incomplete correlation matrix, yielding valid correlation matrices whose eigenvalues are all non-negative summing up to the order of the matrix, which is also the trace of the matrix.

The differential Evolution method is perhaps the fastest evolutionary computational procedure yielding most accurate solutions to continuous global optimization problems. It consists of three basic steps: (i) generation of (large enough) population with individuals in the m -dimensional space, randomly distributed over the entire domain of the function in question and evaluation of the individuals of the so generated population by finding $f(x)$, where x is the decision variable; (ii) replacement of this current population by a better fit new population, and (iii) repetition of this replacement until satisfactory results are obtained or the given criteria of termination are met.

The strength of DE lays on replacement of the current population by a new population that is better fit. Here the meaning of 'better' is in the Pareto improvement sense. A set S_a is better than another set S_b *iff* : (i) *no* $x_i \in S_a$ is inferior to the corresponding member of $x_i \in S_b$; **and** (ii) *at least one* member $x_k \in S_a$ is better than the corresponding member $x_k \in S_b$. Thus, every new population is an improvement over the earlier one. To accomplish this, the DE method generates a candidate individual to replace each current individual in the population. A crossover of the current individual and three other randomly selected individuals obtains the candidate individual from the current population. The crossover itself is probabilistic in nature. Further, if the candidate

individual is better fit than the current individual, it takes the place of the current individual else the current individual passes into the next iteration (Mishra, 2006).

In the present application of DE, the ‘complete correlation problem’ is cast into a minimization problem. It may be noted that the problem has innumerable many minima and we need multiple solutions. Such problems cannot be solved satisfactorily by conventional optimization procedures. A stochastic population method such as DE or PSO (Particle Swarm Optimization) may, therefore, be a suitable choice. In the scheme of DE, a population of N individuals (each represented by an m –dimensional vector, of which each element lies between -1 and 1) is generated by using uniformly distributed random numbers whose each vector provides the candidate values filling in the m number of holes (unknown elements) of the given incomplete matrix. The eigenvalues of the resulting matrices are computed and positive penalties are set if any of them is negative. Minimization of this formulation results into zero penalty, and the solution so obtained yields a valid correlation matrix. Since each individual in the population has gravitational pull to the global optimum, it corresponds to a valid correlation matrix. Thus, we obtain N number of valid correlation matrices.

The Structure of Computer Program and Hints on its Use: The main program (in Fortran) to complete a correlation matrix has eight subroutines. The main program reads the input matrix from a file specified by the user. This file stores the main diagonal and upper diagonal elements of the given matrix. Thus the first row has n elements beginning with 1.0; the second row has $n-1$ elements beginning with 1.0 and so on such that the last (n^{th}) row has only one element (=1.0). In making the input matrix file one has to indicate the known and the unknown elements differently. While the known elements naturally lie *between* -1 and 1 they are put as they really are. However, a number lying *beyond* the range $[-1, 1]$ represents an unknown element. The value could be any number such as 2, -3, 1.5, etc that cannot be a correlation coefficient. For example, if r_{ij} is known to be 0.73, say, it will be put as 0.73, but if r_{ij} is unknown it may be represented by a number, say 2.0 or -1.9 and so on. A number outside the range $[-1,1]$ indicates that it is a hole or an unknown correlation coefficient. When the program runs, it asks for the order of input matrix (*morder*) and the name of input data file in which the input matrix is already stored. The user has to specify them. The program also asks to name the output file in which the final results (valid correlation matrices) would be stored. The user should specify it. Then the program asks for a random number seed. Any 4-digit odd number (say 1271) can be fed as a seed. Subsequently, the program asks for the number of unknown elements (m) in the input matrix. This also has to be given by the user. The main program calls subroutine DE (differential Evolution optimizer). It asks for inputs from the user; the population size (N) and the number of iteration to be performed. The population size determines the number of valid matrices to be obtained as output. It should be normally 100 or so, but for larger problems, this number should be larger. The number of iterations should be specified at 1000 or larger. Then the program needs another random number seed that could be any 4-digit odd number. Once these inputs are given, DE starts running.

Other subroutines in the program are: Normal (generates normally distributed random numbers), Random (generates uniformly distributed random numbers between 0 and 1), Fselect (chooses a function), Func (organizes function calls), Eigen (computes eigenvalues and vectors), Concor (constructs correlation matrices for optimization) and Ncorx (constructs valid correlation matrices and stores them in the output file specified by the user). The output file may be opened in notepad or by any editor program (edit or Microsoft Word of Microsoft Windows) to obtain the results. Directly usable source codes that may be cut or copied and pasted in an editor may be downloaded from the website <http://www1.webng.com/economics/complete-cormat.txt> or alternatively from http://www.freewebs.com/nehu_economics/complete-cormat.txt. A Fortran compiler may be obtained from <http://www.thefreecountry.com/compilers/fortran.shtml> or alternatively from http://www.download.com/Force/3000-2069_4-10233344.html freely. The source codes may be pasted in the Force editor directly. Presently, the dimensions in the program are set to deal with the matrices of order 10 or less. If needed, they may be increased suitably for larger matrices.

An Example: An incomplete matrix of order 7 ($=morder =n$) given in table-1 is used as an example to illustrate an application of the method proposed in this paper. It has 12 ($=m$) holes or unknown elements (colored red). They have been assigned an invalid number (5), outside the permissible range [-1, 1]. Other numbers in the range [-1, 1] are known elements of the matrix. The program is run for population size $N=100$ and it gives N valid correlation matrices. Two sample matrices from the output are given in table-2 and table-3. The program used for computations also gives the eigenvectors for each valid correlation matrix, but they are not presented here.

1.00	-0.50	0.50	-0.50	0.56	0.21	0.34
	1.00	5.00	5.00	5.00	0.30	0.16
		1.00	5.00	5.00	5.00	0.89
			1.00	5.00	5.00	5.00
				1.00	5.00	5.00
					1.00	5.00
						1.00

1.0000000	-0.5000000	0.5000000	-0.5000000	0.5600000	0.2100000	0.3400000
-0.5000000	1.0000000	-0.0285722	0.1840863	-0.0967958	0.3000000	0.1600000
0.5000000	-0.0285722	1.0000000	-0.0249011	0.3674891	0.1476330	0.8900000
-0.5000000	0.1840863	-0.0249011	1.0000000	0.0894851	-0.0430459	-0.0959958
0.5600000	-0.0967958	0.3674891	0.0894851	1.0000000	0.2641564	0.2404028
0.2100000	0.3000000	0.1476330	-0.0430459	0.2641564	1.0000000	0.0415002
0.3400000	0.1600000	0.8900000	-0.0959958	0.2404028	0.0415002	1.0000000
EIGENVALUES, SUM AND PRODUCT OF EIGENVALUES						
2.6161856	1.5874846	1.1577154	1.0120181	0.4686504	0.0975220	0.0604239
7.0000000	0.0134378					

Table-3. Sample Output Correlation Matrix and its Eigenvalues						
1.0000000	-0.5000000	0.5000000	-0.5000000	0.5600000	0.2100000	0.3400000
-0.5000000	1.0000000	0.0784965	-0.0162682	-0.3235212	0.3000000	0.1600000
0.5000000	0.0784965	1.0000000	-0.1237942	0.1573758	0.0478572	0.8900000
-0.5000000	-0.0162682	-0.1237942	1.0000000	-0.0030405	0.0628510	-0.0986661
0.5600000	-0.3235212	0.1573758	-0.0030405	1.0000000	0.0528261	0.0727079
0.2100000	0.3000000	0.0478572	0.0628510	0.0528261	1.0000000	-0.0683791
0.3400000	0.1600000	0.8900000	-0.0986661	0.0727079	-0.0683791	1.0000000
EIGENVALUES, SUM AND PRODUCT OF EIGENVALUES						
2.4707041	1.6609468	1.1648713	1.0422329	0.5538390	0.0926969	0.0147091
7.0000000	0.0037623					

Conclusion: The method given here has an advantage over other algorithms due to its ability to present a scenario of valid correlation matrices that might be obtained from a given incomplete matrix of an arbitrary order. The analyst may choose some particular matrices, most suitable to his purpose, from among those output matrices. Further, unlike other methods, it has no restriction on the distribution of holes over the entire matrix, nor the analyst has to interactively feed elements of the matrix sequentially (as in the scheme of Budden et al.) which might be quite inconvenient for larger matrices. It is flexible and by merely choosing larger population size (N) one might obtain a more exhaustive scenario of valid matrices. As the number of holes increases, the program takes longer time no doubt, but for smaller number of holes it takes a small time even if the input matrix is quite large. This is a special advantage of this method.

References

- Anjos, MF, Higham, NJ, Takouda, PL and Wolkowicz, H (2003) “A Semidefinite Programming Approach for the Nearest Correlation Matrix Problem”, *Preliminary Research Report*, Dept. of Combinatorics & Optimization, Waterloo, Ontario.
- Barrett, WW, Johnson, CR and Lundquist, M (1989). “Determinantal Formulae for Matrix Completions Associated with Chordal Graphs”. *Linear Algebra and its Applications*, 121:265–289.
- Barrett, WW, Johnson, CR and Loewy, R (1998). “Critical Graphs for the Positive Definite Completion Problem”. *SIAM Journal of Matrix Analysis and Applications*, 20:117–130.
- Budden, M, Hadavas, P, Hoffman, L and Pretz, C (2007) “Generating Valid 4 x 4 Correlation Matrices”, *Applied Mathematics E-Notes*, 7:53-59.
- Chesney, M and Scott, L (1989). “Pricing European Currency Options: A Comparison of the Modified Black-Scholes Model and a Random Variance Model”. *Journal of Financial and Quantitative Analysis*, 24:267–284.
- Glass, G and Collins, J (1970) “Geometric Proof of the Restriction on the Possible Values of r_{xy} when r_{xz} and r_{yx} are Fixed”, *Educational and Psychological Measurement*, 30:37-39.
- Grone, R, Johnson, CR, Sá, EM and Wolkowicz, H (1984).” Positive Definite Completions of Partial Hermitian Matrices”. *Linear Algebra and its Applications*, 58:109–124.
- Grubisic, I and Pietersz, R (2004) “Efficient Rank Reduction of Correlation Matrices”, *Working Paper Series*, SSRN, <http://ssrn.com/abstract=518563>
- Helton, JW, Pierce, S and Rodman, L (1989). “The Ranks of Extremal Positive Semidefinite Matrices with given Sparsity Pattern”. *SIAM Journal on Matrix Analysis and its Applications*, 10:407–423.
- Heston, SL (1993). “A Closed-form Solution for Options with stochastic Volatility with Applications to Bond and Currency Options”. *The Review of Financial Studies*, 6:327–343.
- Higham, NJ (2002). “Computing the Nearest Correlation Matrix – A Problem from Finance”, *IMA Journal of Numerical Analysis*, 22, pp. 329-343.
- Johnson, C (1990). “Matrix Completion Problems: A Survey”. *Matrix Theory and Applications*, 40:171–198.
- Kahl, C and Günther, M (2005). “Complete the Correlation Matrix”. <http://www.math.uni-wuppertal.de/~kahl/publications/CompleteTheCorrelationMatrix.pdf>
- Laurent, M (2001). “Matrix Completion Problems”. *The Encyclopedia of Optimization*, 3:221–229.
- Marsaglia, G. and Olkin, I (1984). “Generating Correlation Matrices”. *SIAM Journal on Scientific and Statistical Computing*, 5(2):470-475.
- Mishra, SK (2004) “Optimal Solution of the Nearest Correlation Matrix Problem by Minimization of the Maximum Norm”. <http://ssrn.com/abstract=573241>
- Mishra, SK (2006) “Global Optimization by Differential Evolution and Particle Swarm Methods: Evaluation on Some Benchmark Functions”. <http://ssrn.com/abstract=933827>
- Olkin, I (1981) “Range Restrictions for Product-Moment Correlation Matrices”, *Psychometrika*, 46:469-472.
- Pietersz, R and Groenen, PJF (2004) “Rank Reduction of Correlation Matrices by Majorization”, *Econometric Institute Report EI 2004-11*, Erasmus Univ. Rotterdam.

- Rebonato, R and Jäckel, P (1999) “The Most General Methodology to Create a Valid Correlation Matrix for Risk Management and Option Pricing Purposes”, Quantitative Research Centre, NatWest Group, <http://www.rebonato.com/CorrelationMatrix.pdf>
- Schöbel, R and Zhu, J (1999). “Stochastic Volatility With an Ornstein Uhlenbeck Process: An Extension”. *European Finance Review*, 3:23–46, ssrn.com/abstract=100831.
- Stanley, J and Wang, M (1969) “Restrictions on the Possible Values of r_{12} , given r_{13} and r_{23} ”, *Educational and Psychological Measurement*, 29, pp.579-581.
- Storn, R and Price, K (1995) "Differential Evolution - A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces": *Technical Report, International Computer Science Institute, Berkley*.
- Tyagi, R and Das, C (1999) “Grouping Customers for Better Allocation of Resources to Serve Correlated Demands”, *Computers and Operations Research*, 26:1041-1058.
- Xu, K and Evers, P (2003) “Managing Single Echelon Inventories through Demand Aggregation and the Feasibility of a Correlation Matrix”, *Computers and Operations Research*, 30:297-308.