Numero 4 - 1992

RATIO MATHEMATICA

Atti del Convegno di
Matematica Applicata all'Economia e all'Ingegneria
Ovindoli - Maggio 1991

a cura di

Franco Eugeni e Mario Gionfriddo

Comitato Scientifico

Numero 4 - 1992

RATIO MATHEMATICA

Atti del Convegno di
Matematica Applicata all'Economia e all'Ingegneria
Ovindoli - Maggio 1991

a cura di

Franco Eugeni e Mario Gionfriddo

Comitato Scientifico

Direttore Responsabile: Prof. FRANCO EUGENI

Autorizzazione n. 9/90 del 10- 07-1990

Stampato nel mese di ottobre 1992
da Novastampa srl, Parma, per:

BIBLOS scrl
via dei Peligni, 102 - Pescara
Tel. 085/65920

Ricordiamo affettuosamente

Questi atti e i lavori in essi contenuti sono dedicati alla memoria di due nostri amati Colleghi immaturamente scomparsi dopo il Convegno di Ovindoli. Di questo e del precedente essi erano stati, con noi, animatori e sostenitori. A loro va il nostro affettuoso ricordo: saranno sempre nel nostro cuore!

ILIO ADORISIO

Il prof. Adorisio nelle sue molteplici attività di studioso era andato ben oltre la sua figura di Ingegnere e del suo antico desiderio di essere un Fisico. Era divenuto uno scienziato completo nel senso rinascimentale del termine. Si occupava di fatto di tutte le discipline scientifiche: dalla Matematica all'Economia per l'Ingegneria ma anche di Storia, Scienze Politiche, Antropologia, Psichiatria, e recentemente, non pago di discipline "inquadrate", aveva anche scritto e rappresentato una magnifica ed allegorica commedia, magistralmente interpretata dalla figlia.

Sembrava a noi amici che avesse l'interesse a stupirci sempre, e ci stupiva!.

Il prof. Ilio Adorisio, nato a Cirò (Catanzaro) il 25 aprile 1925, era professore ordinario di Economia Matematica nella Facoltà di Ingegneria della Sapienza. In qualità di consulente della Banca Mondiale dal 1958 al 1974 ha redatto studi economici per conto di paesi del terzo mondo ed ha diretto l'elaborazione di piani di investimento in Sud-America, Africa ed Asia. Oltre che di economia si è occupato di analisi teoriche relative a discipline ingegneristiche.

Il nostro caro Ilio non si può certo dire che fu solo un tecnico: sempre affascinato dai problemi filosofici posti dalla Scienza fu un moderatore-provocatore, forse più provocatore, in tutti i gruppi di ricerca che magistralmente condusse.

GIOVANNI MELZI

Il prof. Melzi era una figura di studioso puro, con interessi multidisciplinari, profondo per lo scibile del passato e attento a tutte le nuove discipline di frontiera, discipline che puntualmente riversava nel suo splendido impegno didattico.

Lo ricordiamo nei Convegni Nazionali della Mathesis, al suo arrivo era tutta una festa: è arrivato Melzi, è arrivato il Melzi, è arrivato Nino!. Nell'ultimo di Cattolica ci parlò con entusiasmo di un Corso che stava facendo in una scuola a fini speciali: "Forse il prossimo anno ne avrò delle belle da raccontarvi". L'anno successivo 1992, a Fermo, non venne: era ricoverato. Noi amici parlammo a lungo di Lui, ci mancava. Pensiamolo ancora a Pandino, in giro con il calesse come soleva fare!.

Il prof. Melzi aveva 60 anni.

Nel 1967 divenne Ordinario di Geometria nella Facoltà di Scienze dell'Università di Milano, successivamente fu Preside della Facoltà di Scienze dell'Università Cattolica di Brescia e anche professore di Logica Matematica prima e Matematica Generale poi.

Dopo gli studi iniziali di Geometria Differenziale si occupò di varie questioni quali ad esempio la sua Teoria delle Neuromacchine che andava dalla Logica alla Cibernetica ed alla Teoria dei Sistemi.

Del convegno di Ovindoli tutti ricordano la magnifica musica ottenuta da un sistema che aveva applicazioni in... economia.

A tutti quelli che hanno avuto il piacere ed il privilegio di conoscerLo rimane il ricordo di un amato Maestro e di un coraggioso ed eclettico Uomo di Scienza.

La sua dotta e pacata voce con cui ci spiegava semplicemente le sue più ardite riflessioni ci manca e ci mancherà!

Indice

ON THE PROFILE REDUCTION OF SPARSE SYMMETRIC MATRICES

Antonio BALDI, Antonio DE PAULIS

Summary. After a short presentation of the profile reduction and band optimization methods for sparse symmetric matrices, a new algorithm is presented. Numerical tests on typical Finite Elements problems are made and the results are discussed.

1. INTRODUCTION

To find the numerical solution to many engineering problems the first step is the discretization of the physical problem. Usually this discretization process leads to algebraic systems of the type

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \tag{1}$$

where the matrix $\mathbf{A}$ is usually symmetric and sparse.
The systems (1) are usually solved using direct methods i.e. the Gauss method or variations such as the Choleskj method. Good results from all these algorithms strongly depend on the matrix structure; some methods for optimizing this structure are outlined below.
Among the principal discretization techniques is the Finite Element Method (F.E.M.).
For Finite Element problems involving a small number of nodes, or a small number of variables x, it is not difficult to order the nodes efficiently; with more complex grids this is often more difficult.
An efficient node ordering is particularly important in non-linear computations where a sequence of equations must be solved many times or when the F.E. meshes are automatically generated and leading to an unnecessarily large band β.
Today there are numerous euristhic methods to minimize approximately the structure of a matrix. These methods can be described with the Graph Theory which provides a succinct means of discussing the various ordering schemes.
A *Graph* $G = (X,E)$ consists of a finite set of *nodes* (or *vertices*) X together with a set E of *edges* , which are unordered pairs of nodes. A mapping of $\{ 1, 2, 3, \ldots , N \}$ onto X , where N denotes the number of nodes of G , is an *ordering* of G

$$G^{\alpha} = G (X^{\alpha}, E)$$

Let $\mathbf{A}$ be an N by N symmetric matrix. An ordered graph G^{A} is one for which the N vertices are numbered from 1 to N and the graph edge $E = \{ x_i, x_j \}$ belongs to E^{A} if and only if $a_{ij} = a_{ji} \neq 0$.
The same graph can be constructed directly by the F.E. mesh; in this case X is the set of the mesh nodes while an edge $E = \{ x_i, x_j \}$ belongs to the graph only if the mesh nodes are connected by one element. Note that in the F.E. method each node usually has more than one degree of freedom, but for our purposes it is sufficient and simpler to consider each of them as having just one degree of freedom. Figure 2 illustrates a simple mesh with its associated graph.
Each graph which is generated in this way can be seen as a *tree*, when one has defined a node as a root. A generical $k + 1$ tree level is defined as the set of nodes which are adjacent to the nodes at k level.
Two nodes x and y are *adjacent* if $\{ x, y \} \in E$.
For a set of node $Y \in X$, the adjacent set, denoted by $Adj(Y)$, is

$$Adj(Y) = \{ x \in X - Y \mid \{ x, y \} \in E \text{ for some } y \in Y \}$$

while the *degree* of Y, denoted by $Deg(Y)$, is the number of members of $Adj(Y)$. Clearly, if Y consists of a single node (y), the adjacent set is formed by the nodes x_i connected to y and its degree is the number of $E_i = \{ x_i, y \}$.

Let i be the i-th row of matrix $\mathbf{A}$. Let

$$f_i(\mathbf{A}) = \max \{ j \mid a_{ij} \neq 0 \},$$

that is the column index of the last non zero element in row i and let

$$\beta_i(\mathbf{A}) = f_i(\mathbf{A}) - i.$$

The *bandwidth* of the matrix $\mathbf{A}$ is defined by

$$\beta(\mathbf{A}) = \max \{ \beta_i(\mathbf{A}) \mid i \leq N \}$$

while the *profile* P is

$$P(\mathbf{A}) = \sum_{i=1}^{N} \beta_i$$

Note that an ordering algorithm which gives the lower profile does not give the smaller bandwidth: in solving a linear system of equations for numerical stabilility purposes is best to have the bandwidth β as small as possible, while to minimize the execution time is best to have a low profile P.

Let x, y be a pair of nodes of a connected graph G. A *path* of length k from node x to y is a set of k vertices $\{ x, x_1, x_2, \ldots, y \}$ such that $x_i \in Adj(x_{i+1})$ for $0 \leq i \leq k - 1$.

The *distance* $d(x,y)$ between x and y is simply the length of the shortest path joining x and y.

If $\lambda(x) = max \{ d(x,y) \mid x,y \in X \}$, then a diameter of the graph G can be given by $\delta(G) = max \{ \lambda(x) \mid x \in X \}$. A node $x \in X$ is *peripheral* if $\lambda(x) = \delta(G)$.

A diameter found by a numerical algorithm is a pseudo-diameter; the first and last nodes of the path defining the pseudo-diameter are pseudo-peripheral nodes. Note that a pseudo-diameter is not guaranteed to be a diameter but only to have a large λ.

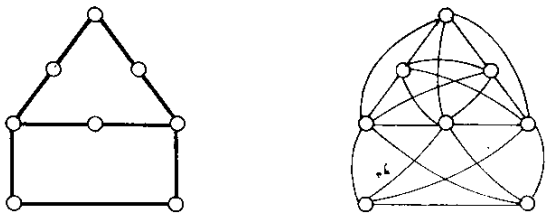


Fig. 1 - A simple F.E.M. mesh and the associated graph.

2. SOME PRACTICAL CONSIDERATIONS

As the algorithms described in subsequent paragraphs operate on Graphs, a computer representation of the adjacency properties of each node, economical in storage and easy to use, is needed.

Let $G(X,E)$ be a graph with N nodes. An *adjacent structure*, i.e. the *list of adjacency* for each $x \in X$, can be implemented quite simply by storing the *adjacency lists* (AL) sequentially in a one dimensional array along with a *index array* containing pointers to the beginning of each list.

Although this method gives the best storage, it is quite complex to use; moreover, it is difficult to construct the connection array when the graph is provided as a list of edges because the final size of each adjacent list is not usually known.

A simpler approach involves a connection table with N rows and $m+1$ columns, where $m = max \{ Deg(x) \mid x \in X \}$.

Each row *i* contains the adjacency list for a node and, in column 0, the node degree. This scheme may be inefficient from a storage point of view (a problem overcome by using the dynamic allocation) but the construction of the adjacent structure can be made easy as, for example, in this Pascal code fragment:

```
for i := 1 to Element_Number do begin
    for j := 1 to Element_Size [Element[i].El_Type]  do begin
        k := 1;  m := Element [i].Node_List[j] ;
         repeat
          if k< >j then begin
          sw := true;
            for i := 1 to AdjT[m,0]  do
              if AdjT[m,i]  = Element [i].Node_List [k]  then
                sw := false;
              if sw then begin
                 AdjT[m,0]  := AdjT[m,0]  + 1;
                 AdjT [m,AdjT[m, 0]]  := Element [i].Node_List[k]
              end
          end;
           k := k+1
        until k > Element_Size [Element[i].El_Type]
end;
```

Here *AdjT* is the adjacency table, *Element_Size* is a table containing the nodes number for each element type and *Element* is an array of records which gives the type (*El_type*) and the list of nodes (*Node_List*) for each of them.

3. ORDERING METHODS

A central concept to many successful ordering algorithms is the *rooted level structure*.
This is a partitioning of the nodes such that each node is assigned to one of the levels $l_1, l_2, \ldots, l_n$ with its distance from a specified *root node* s in the graph. All nodes which lie at the same distance from the root node belong to the same level.
The *depth* of a rooted level structure is simply its total number of levels.
The *width* is the maximum number of nodes which belong to a single level.
The labelling algorithms comprise two distinct steps:

a) the selection of pseudo-peripheral nodes and

b) node labelling.

In step a) a pair of pseudo-peripheral nodes which lie at opposite ends of a pseudo-diameter are found in this way:

1) select a node s with the smallest degree as starting node;
2) generate a level structure rooted at node s;
3) for each node q of the last level generate a level structure; if the height of this structure is higher than the original one, set q as the starting node (s = q) and jump to 1.

Computational experiments indicate that the last level node set is usually quite large and in this set many nodes have the same degree. To reduce the number of nodes in the last level set, nodes for which a rooted level structure must be generated, *shrinking* strategies (which are supposed not to affect the resulting pseudo-peripheral node quality) were adopted by almost all authors. Sloane, for example /4/, creates a list with only one node for each degree; Duff, Reid, Scott /5/ constructed a list with nodes in the last level with ties broken arbitrarily. This strategy led to a slight increase in the computed reduced profile for a class of problems; for other problems they only found a small reduction.

The second step, node labelling, is a field with a wide spectrum of possible choices: usually one starts from the root node and crosses the tree structure in a moving wavefront to end at the tree last level.

Cuthill-McKee's well known algorithm labels, for each tree level node, all the following level unnumbered nodes in increasing order of degree.
Although this method works, it only uses a local parameter, so many authors have presented more sophisticated algorithms. The most effective is the Sloan algorithm, which, during the labelling process, defines active, inactive, pre-active and post-active node states; in this process the current wavefront advancement is controlled by a quantity called the current degree. This label method works dynamically by combining a local parameter (the node degree) with a global one (the node level).
At each stage in the labelling process, the list of eligible nodes (the wavefront) includes those nodes which are either adjacent to a node which has been relabelled or are adjacent to a node which is itself adjacent to a relabelled node.
The next node to be given a new number is the node with the highest priority, where the priority P_i is defined as

$$P_i = - W_1 \, cdeg(i) + W_2 \, d(e,i)$$

Here W_1, W_2 are positive integer weights and the current degree *cdeg(i)* is the number of nodes adjacent to i, not including any node that has been relabelled or is itself adjacent to a relabelled node.
The priorities of the eligible nodes are updated at each step.

To explain the proposed algorithm some new terms are introduced:

in a generic step of the labelling process, all of the previously labelled nodes connected to a node n are *n's fathers*, while all the unlabelled nodes at the next tree level are *n's sons*. The father state is clearly independent of the tree level; nevertheless, the ordering scheme creates an advancing front so that the *fathers* usually belong to the same or former level of node n.

To develop the algorithm an eligible queue has been introduced, which contains all the vertices that can be labelled at the current step. The selection within the queue is made using an index called the *node current priority*, defined as

$$Np(n) = W_F \, N_F - W_S \, N_S$$

where N_F and N_S are, respectively, the number of fathers and sons at the current ordering step ($N_F(n) + N_S(n) \leq Deg(n)$) while W_F and W_S are two positive integer weights.
Numerical tests proved that a good choice is the value 2 for W_F and 1 for W_S , as in the Sloan algorithm, although the two weights refer to parameters with different meanings.
This priority formulation is based on the observation that one must label nodes with few sons first (so nodes with a lot of them are nearest to their sons), but at the same time label as early as possible nodes with many fathers too (Fig. 2).
Numerical tests gave as a good choice the value of 2 for Wf and 1 for Ws.
Although this values are the same as in Sloan algorithm, this is merely a coincidence; in fact they refer to parameters with different meaning.

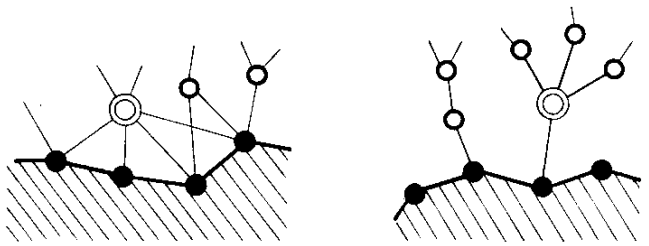


Fig. 2 - The priority scheme.

The algorithm can be summarized in the following steps:

1) look for a starting node *r* using the Gibbs' algorithm or its modifications,

2) construct the tree structure rooted at r.

3) Set start-up priorities: no nodes are labelled: there is no father. All vertices have a current priority $Np(n_i) = -W_s \times N_s(n_i)$ where $N_s(n_i)$ is the number of next level neighbours.

4) For each node set *TRUE* the queue insertion control item (*IcL*); this is an array whose items are *TRUE* if nodes have not been inserted in the *eligibly queue*.

5) Initialize queue:
set $Next\_label \leftarrow 1$, $Queue\_Size \leftarrow 1$, $Queue[1] \leftarrow r$, $IcL(r) \leftarrow false$.

6) Loop: create a list MpL of all the maximum priorities nodes in queue.

In this loop, when there are many nodes with the same priority, the node with minimum distance to the root is selected. This selection method is the same as Sloan used,, but with a different meaning.

a) Select the node k of lower level from MpL;
b) update priorities; for each neighbour i of k
i) set $Np[i] \leftarrow Np[i] + W_F$, (node i gains a father)
ii) if $level[i] > level[k]$ and $Np[i]$ is *TRUE*, insert i in queue and update Np.
iii) if $level[i] < level[k]$ set $Np[i] \leftarrow Np[i] - W_s$, (node i loses a son).
iv) Label k by setting $k \leftarrow Next\_label$,
$Next\_label \leftarrow Next\_label + 1$,
$Queue[k] \leftarrow Queue[Queue\_size]$,
$Queue\_Size \leftarrow Queue\_Size - 1$.
v) Test for termination: if $Queue\_Size > 0$ go to Step 6).

Numerical experiments have proved that the best starting priority setting is with $N_s(n_i)$ (ALG3 b) substituted by $Deg(n_i)$ at step 3) (ALG3 a). This means that not only the next level nodes are considered as being sons but also all the neighbours.

4. NUMERICAL TESTS

To test the performance of the proposed ordering schemes, F.E. meshes were used. They were divided into three groups:

a) Problems that cannot be improved (Test No. 0).

b) Problems for various theoretical configuration (Mostly taken from /2/ : Test Nos. 1, 2, 3, 4, 5).

c) Examples for typical engineering calculations (Test Nos. 6, 7, 8).

Test No. 0 was performed to see how bad the labelling created by an ordering scheme was compared to the original one. Group b) contains small meshes, with particular characteristics. Group c) to see how the algorithms perform in real problems.

Obviously these problems do not cover all the cases and it is easy to find meshes where the ordering scheme performances are quite different from those summarized in Table I; however they are sufficiently reppresentative to extrapolate average algorithm results /4/.

In Table I the results are arranged as follows: each row refers to a different ordering scheme, each column refers to a single test problem. For each problem there is the matrix profile and the ratio obtained by dividing the actual profile by the best result. The original matrix profile is in the first row.

The results are from algorithms RCM, Sloan, ALG3a, ALG3b.

The Table shows that the proposed algorithm nearly always gives good results than the others.

The suggested modification in the start-up priorities generally improves performances, except in test 7.

The Reverse Cuthill McKee method only gave the best performance for test 8 where the exceptionally high and tight graph promotes local against global optimization.

Although the Sloan method only got the best result once with the present tests, it usually performs very well, better than the other algorithms except ALG3b; it got the best result in the quasi optimal test 0.

TABLE I

P =	Test 0 1090	Test 1 403	Test 2 364	Test 3 675	Test 4 156
RCM	1222 (1.121)	370 (1.439)	300 (1.111)	445 (1.093)	136 (1.295)
SLOAN	1199 (1.1)	284 (1.105)	289 (1.070)	480 (1.179)	116 (1.104)
ALG3a	1228 (1.126)	265 (1.031)	389 (1.003)	426 (1.046)	107 (1.019)
ALG3b	1245 (1.142)	257 (1)	270 (1)	407 (1)	105 (1)

P =	Test 5 3218	Test 6 305	Test 7 3764	Test 8 2957
RCM	1251 (1.356)	303 (1.048)	3929 (1.150)	1873 (1)
SLOAN	1080 (1.171)	321 (1.110)	3464 (1.014)	2120 (1.072)
ALG3a	973 (1.055)	298 (1.031)	3415 (1)	2190 (1.169)
ALG3b	922 (1)	289 (1)	3506 (1.026)	1976 (1.054)

APPENDIX: Computer Code for ALG3.

In this Appendix there is the Pascal code to implement ALG3.
Before the listing the definitions of the global variables are reported.

Global routines to use

1 - **Function FindStartNode : integer;**
This function returns a pseudo peripheral node.
2 - **Function CreateTree (root: integer) : integer;**
This function creates a tree structure (in **Nd** see point 3) rooted at the choosen pseudo peripheral node. It returns the tree height (Note that we do not need this information, but it can be useful for other ordering scheme).

Global Variables

3 - **Nd** : array [1..*MAXNODI*, 1..2] of integer;
contains the node label (column 1) and the tree level for each node (column 2).
4 - **AdjT** : array [1..*MAXNODI*, 0..*MAXCON*] of integer;

adjacency Table as described in text; MAXCON is the max number of nodes connected to a node. The nodes are ordered with the degree.

5 - **Nodes** : integer;
is the number of all nodes in the graph.

Global types

6 - **NodeList** = array [1..*MAXNODI*] of integer;
7 - **NodeCtrl** = array [1..*MAXNODI*] of boolean;

The computer codes are

```
procedure ALG3
Const    Ws = 1; Wf = 2;
Var

    Np,Q      : NodeList  { Priority and Queue }
    MaxP      : NodeList  { Max priority List }
    IcL       : NodeCtrl  { Queue insertion control list }
    Ds,Th     : Integer   { Tree root, height }
    qSize     : Integer   { Elements in queue }
    nName     : Integer   { Next node labelled name }
    i,j,k,d,l,n : Integer { Miscellany }
    lvl,q2l   : Integer   { Miscellany }

    procedure UpdatePriority(Wn,Lev:Integer);
    Var i,j : integer;
    begin
        for i := 1 to AdjT [Wn,0] do begin
            j := AdjT [Wn,i];
            Np [j] := Np [j] + Wf;
            if (Nd [j,2] > Lev) and IcL [j] then begin
              qSize := qSize + 1; Q [qSize] := j; IcL [j] := FALSE;
            end ;
            if Nd [j,2] < Lev then Np [j] := Np [j] - Ws;
        end
            end;

begin                    { ALG3 }
    Ds := FindStartNode;                    { Step 1 }
    Th := CreateTree(Ds);                   { Step 2 }

{ ** Init Priorities: use this code fragment for a theoretically consistent algorithm
    for i := 1 to Nodes do begin
        k := Nd [i,2] + 1; l := 0;
        IcL[i] := TRUE;
        for i := 1 to AdjT [i,0] do
        if Nd [AdjT[i,0],2] = k then l := l + 1;
        Np[i] := - Ws * l
    end;                              ** }

    for i := 1 to Nodes do begin
        IcL[i] := TRUE;
        Np[i] := -Ws*AdjT [i,0]
    end;

    qSize := 1; Q [1] := Ds; NName := 1; IcL [Ds] := FALSE; { Step 5 }
    while qSize > 0 do begin    { Main loop }
        d := Np [Q[1]]; n := 1; MaxP [1] := 1;
        for i := 2 to qSize do begin
            k := Np [Q[i]];
            if k = d then begin
                n := n + 1; MaxP [n] := i;
            end else if k > d then begin
```

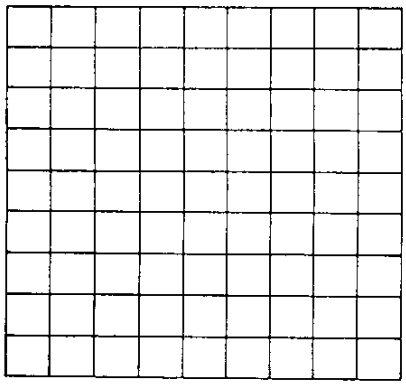
```
                n := 1;  MaxP [1]  := i;  d := k
            end
        end;

        q2l := MaxP [1] ;  lvl := Nd [Q[q2l],2];   { Step 7 }
        for i := 2 to n do
            if Nd [Q[MaxP[i]],2]   < lvl then begin
            q2l := MaxP[i];  lvl := Nd [Q[q2l],2]
            end;

        UpDatePriority(Q [q2l],lvl)                    { Step 8 }
        Nd [Q[q2l],1]  := nName;  nName := nName + 1;   { Step 9 }
        Q [q2l]  := Q [qSize] ;  qSize := qSize - 1;
    end                          { - while loop }
end;
```

REFERENCES

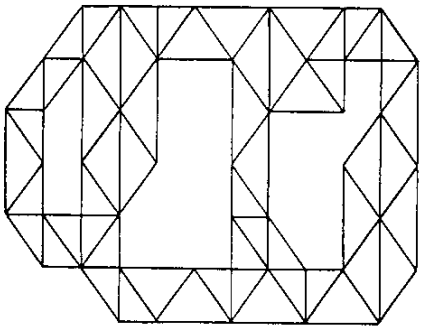
1. A. George, Joseph W.H. Liu
   An Implementation of a Pseudoperipheral Node Finder
   ACM Transactions on Mathematical Software,
   vol. 5, no. 3 (1979), pp 284-295
2. A. George, Joseph W.H. Liu
   Computer Solution of Large Sparse Positive Definite System
   Prentice Hall, Englewood Cliffs, 1981
3. S. W. Sloan
   An Algorithm for Profile and Wavefront Reduction of Sparse Matrices
   Int. J. for Numerical Methods in Engineering, vol 23, 239-251 (1986)
4. S. W. Sloan
   A Fortran Program for Profile and Wavefront Reduction
   Int. J. for Numerical Methods in Engineering, vol 28, 2651-2679 (1989)
5. I.S. Duff, J.K. Reid, J.A. Scott
   The Use of Profile Reduction Algorithms with a Frontal Code
   Int. J. for Numerical Methods in Engineering, vol 28, 2555-2568 (1989)

Author Address

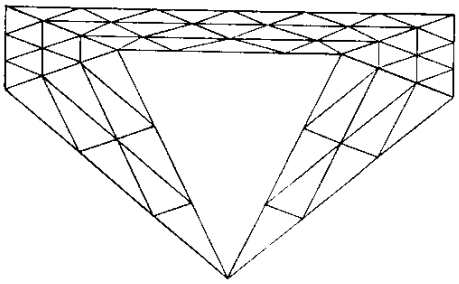
Antonio DE PAULIS
Dipartimento di Energetica
Facolta' di Ingegneria
67040 ROIO POGGIO (AQ) - Italy
FAX number: (0)862-432633



Test 0



Test 1



Test 2

Fig. 3 - F.E. models for Test 0, 1, 2.

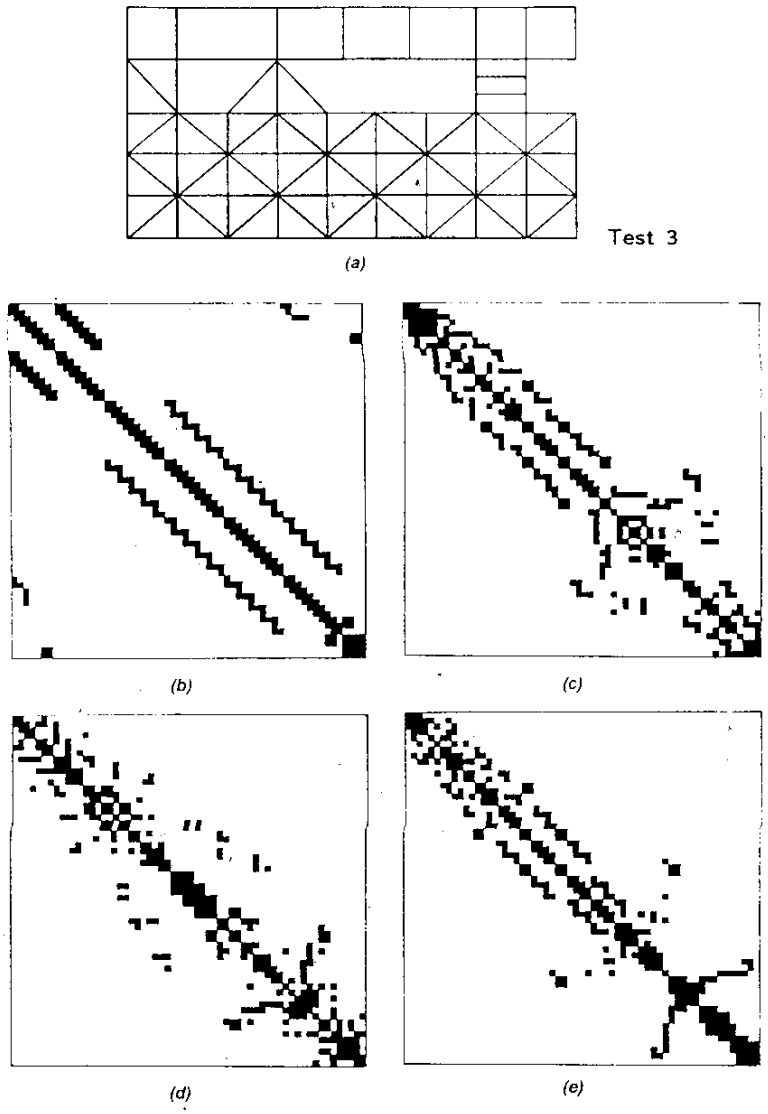


Fig. 4 - Geometrical rappresentation of an ordered matrix. (a) and (b) mesh and original matrix of Test no. 3. The same matrix optimized with (c) Cuthill-Mc Kee, (d) Sloan and (e) ALG3 algorithm.

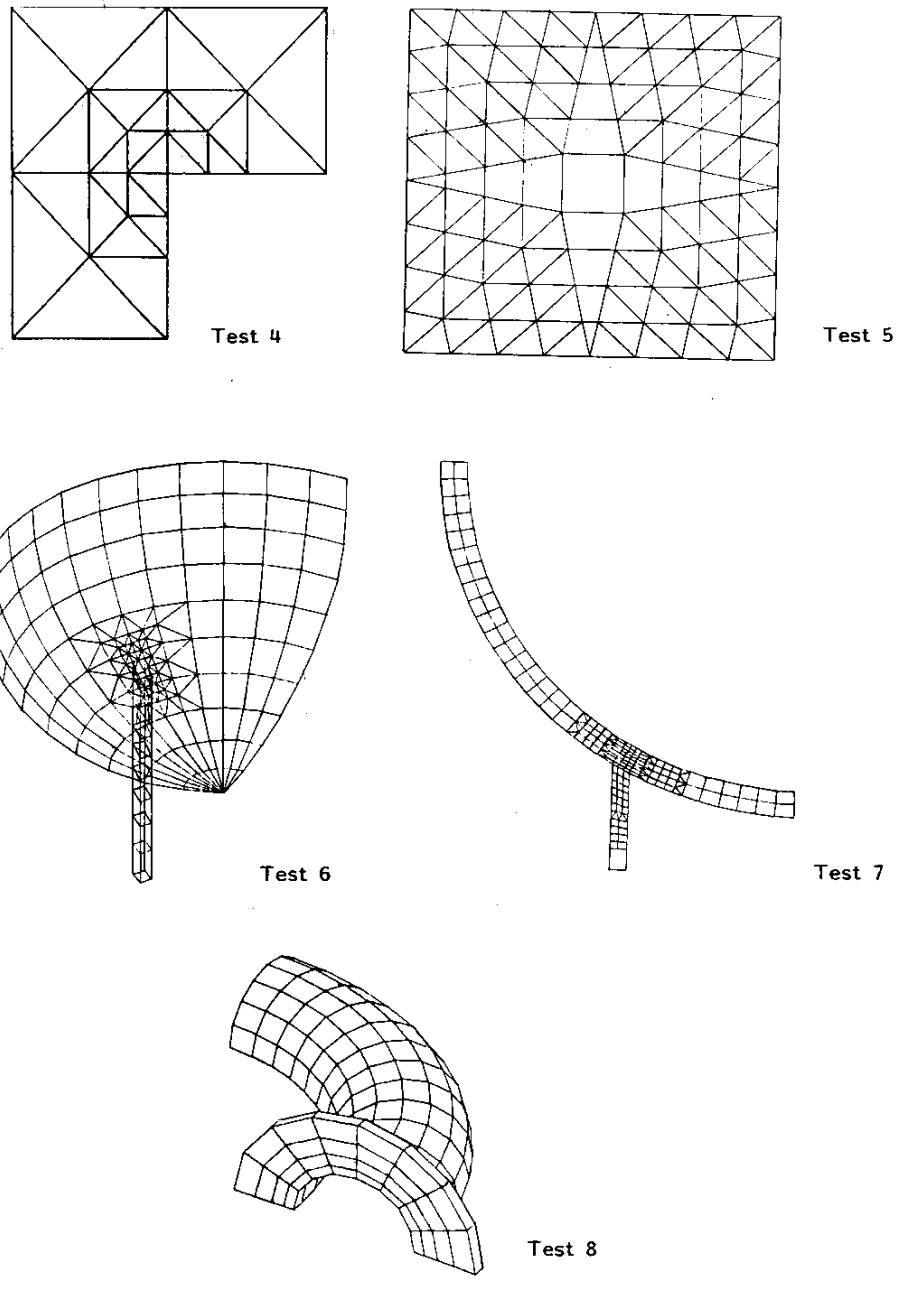


Fig. 5 - F.E. models for Test 4, 5, 6, 7, 8.

A SIMPLE CONSTRUCTIVE PROOF OF VON NEUMANN EQUILIBRIUM

Paolo CARAVANI
Dipartimento di Ingegneria Elettrica, Università degli Studi dell'Aquila, Monteluco di Roio, 67040 L'Aquila, Italy

A proof of existence of general equilibrium in the Von Neumann model of an expanding economy is given based on elementary results of Linear Programming. General equilibrium is computable by solving a sequence of Linear Programs.

1. Introduction

Over the years, existence of the equilibrium in the Von Neumann model has been proved with recourse to analytical tools differing widely in nature as well as in complexity. The original Von Neumann proof made extensive use of game theory. Other proofs rely on fixed point theorems such as Brouwer or Kakutani (for a review, see Burmeister Dobell 1970, p. 207).

In 1960, Charles W. Howe exhibited a proof that did not require game theory or fixed point theorems but used a result on linear inequalities due to A.W. Tucker. Unfortunately though, this result requires independent derivation. To the non-initiated, this derivation can appear lengthy. Murata for instance, derives Tucker' theorem from Farka's Lemma through a sequence of about 10 intermediate theorems (Murata, 1977, pp. 283-288).

Unaware, to my knowledge, of more elementary proofs, I wondered if all this mathematical machinery is really necessary. Here I provide a proof based solely on Linear Programming (henceforth LP). Namely, I employ the only two

notions that 1. if a primal-dual pair of LP problems have feasible solutions, they have optimal solutions with a common value, and 2. these satisfy the so-called complementarity slackness conditions: positive primal variables are associated to active dual constraints and loose primal constraints are associated to zero dual variables. The rest of the proof is elementary algebra.

Besides greater simplicity, the advantage is to characterise solutions through a constructive procedure. This feature is not usually obtained using fixed point theorems or linear inequality results.

2. Model Description

Von Neumann economy comprises m goods and n processes. Process j is characterised by an activity level y_j. When it is operated at unit level ($y_j = 1$) its input demands and output supplies are described by two non-negative vectors

$$\begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix} \quad \begin{bmatrix} b_{1j} \\ \vdots \\ b_{mj} \end{bmatrix}.$$

Thus technology is completely specified by an input matrix A and an output matrix B, both non-negative with m rows (goods) and n columns (processes).

Each good must be demanded by at least one process so there is at least one positive entry in each row of A. Each process must supply at least one good, so there is at least one positive entry in each column of B. We refer to these two assumptions as KMT (from Kemeny Morgenstern Thompson (1956) who formulated them in order to relax those originally made by Von Neumann).

Labour is treated as one amongst the m goods. It is supplied to other processes by a "domestic" sector that produces it using other goods (like food, housing, good TV, etc). So consumption is one of the rows of A and there is no exogenous demand. Labor being paid in nature-goods, there is no wage rate and net income is made up by profits only. These are entirely re-invested in the form of input purchases for next period. Under these conditions an equilibrium is sought so as to satsfy

1. Output at some period By must cover input of next period AY where $Y = (1+g)y$, and g is a (balanced) growth rate. Letting $\alpha = 1+g$

$$\alpha Ay \leq By. \tag{1}$$

2. Revenues $p'B$ cannot exceed production costs $p'A$ plus competitive profits $rp'A$ where r is a (uniform) interest rate. Letting $\beta = 1+r$

$$\beta p'A \geq p'B. \tag{2}$$

3. Goods in excess of next period demand (components of (1) with strict inequality) are worthless

$$p'(B - \alpha A)y = 0. \tag{3}$$

4. Processes generating less than competitive profits (components of (2) with strict inequality) remain idle

$$p'(B - \beta A)y = 0. \tag{4}$$

5. Something of value is produced in the economy $p'By > 0$.

3. Von Neumann Theorem

Theorem. *(Von Neumann, 1937)*

A solution $y \geq 0$ $p \geq 0$ satisfying (1-5) exists if and only if $\alpha = \beta = \gamma$ where

$$\gamma = \max\{\alpha : By - \alpha Ay \geq 0, \quad y \geq 0\}$$

$$\gamma = \min\{\beta : p'B - \beta p'A \leq 0, \quad p \geq 0\}.$$

Proof. *Necessity.* Assume that $y, p \geq 0$ satisfying 1-5 exist. From (3) and (4) it follows $p'By = \alpha p'Ay$ and $p'By = \beta p'Ay$. Since $p'By > 0$, if a solution exists the r.h.s. are positive and $p'Ay > 0$. Hence $\alpha = \beta$.

Furthermore, from (1) and (2)

$$p'By \geq \alpha p'Ay \quad \text{and} \quad -p'By \geq -\beta p'Ay.$$

Adding each side and recalling that $p'Ay > 0$ it follows $\alpha \leq \beta$. Thus there exists only one common value to α and β, call it γ. This necessarily satisfies

$$\max\{\alpha : By - \alpha Ay \geq 0, \quad y \geq 0\} = \gamma = \min\{\beta : p'B - \beta p'A \leq 0, \quad p \geq 0\}.$$

Sufficiency. Define $\gamma = \max\{\delta : By - \delta Ay \geq 0, \quad y \geq 0\}$. The maximum certainly exists, since the set of $y \geq 0$ satisfying (1) is either empty or closed and bounded. Furthermore, $\gamma \geq 0$ due to the non-negativity of A, B, y. Now we have to show that there exist a $y \geq 0$ and a $p \geq 0$ satisfying 1-5, for some α, β. Choose $\alpha = \beta = \gamma$. Letting

$$C = B - \gamma A$$

and introducing scalars $Y_1, Y_2 \geq 0$ and $P_1, P_2 \geq 0$, consider the LP problem

$$\min \quad [0'|1| - 1] \begin{bmatrix} y \\ Y_1 \\ Y_2 \end{bmatrix}$$

$$\begin{bmatrix} C & \mathbf{1} & -\mathbf{1} \\ \mathbf{1}' & 0 & 0 \\ -\mathbf{1}' & 0 & 0 \end{bmatrix} \begin{bmatrix} y \\ Y_1 \\ Y_2 \end{bmatrix} \geq \begin{bmatrix} \mathbf{0} \\ 1 \\ -1 \end{bmatrix}. \tag{6}$$

Its dual is

$$\max \quad [\mathbf{0}'|1|-1] \begin{bmatrix} p \\ P_1 \\ P_2 \end{bmatrix}$$

$$[p'|P_1|P_2] \begin{bmatrix} C & \mathbf{1} & -\mathbf{1} \\ \mathbf{1}' & 0 & 0 \\ -\mathbf{1}' & 0 & 0 \end{bmatrix} \leq [\mathbf{0}'|1|-1]. \tag{7}$$

Rewriting (6,7) as

$$\min \quad Y_1 - Y_2 : \quad Cy + (Y_1 - Y_2)\mathbf{1} \geq 0, \quad \mathbf{1}'y = 1$$

$$\max \quad P_1 - P_2 : \quad p'C + (P_1 - P_2)\mathbf{1}' \leq 0, \quad p'\mathbf{1} = 1$$

it is clear that, for any pair $y, p \geq 0$ with components adding to one, it is always possible to choose $Y_1 - Y_2$ and $P_2 - P_1$ large enough so as to satisfy the constraints. Since there exist feasible solutions there exist optimal solutions, which necessarily satisfy

$$Y_1 - Y_2 = P_1 - P_2.$$

We show now that $Y_1 - Y_2 = P_1 - P_2 = 0$. From the dual constraints we have $P_1 - P_2 \leq 0$. But, if it were $P_1 - P_2 < 0$ then $Y_1 - Y_2 < 0$ and from $Cy + (Y_1 - Y_2)\mathbf{1} \geq 0$ it would follow

$$Cy = [B - \gamma A]y > 0$$

against the assumption that γ is the largest value of α for which (1) has a solution $y \geq 0$. Thus $P_1 - P_2 = 0 = Y_1 - Y_2$ and this shows that there exist two non-zero vectors $y \geq 0$ and $p \geq 0$ satisfying (1) e (2). Conditions 3 and 4, for $\alpha = \beta = \gamma$ are the complementarity slackness conditions of LP and these are obviously satisfied given the optimality of the pair y, p.

We finally prove that 5 holds. Observe first that

$$p'By \geq \bar{p}'\bar{B}\bar{y}$$

where $\bar{p}$ and $\bar{y}$ are vectors obtained from p and y by deleting exactly those zero-components that are associated to active constraints in the respective duals (if there are any) and $\bar{B}$ is the matrix obtained from B by deleting the corresponding rows and columns. Therefore, it is enough to prove 5 under the assumption

$$p + Cy > 0 \qquad y - C'p > 0. \tag{8}$$

Rearranging components if necessary, the optimal solutions can be partitioned as

$$y = \{y_1|y_2\}' \quad p = \{p_1|p_2\}'$$

where the first components of each vector are positive and the remaining components (if there are any) are zero. Partitioning conformally A and B, we get from (3,4) (with $\alpha = \beta = \gamma$)

$$p_1'(B_{11} - \gamma A_{11})y_1 = 0.$$

Now, if $p'By = 0$, then $B_{11} = 0$ (being positive p_1, y_1) and thus also $A_{11} = 0$, and thus also

$$B_{11} - \gamma A_{11} = 0.$$

From (8), if there are components $p_2 = 0$, then the constraints associated to p_2 in the dual of $p'(B - \gamma A) \leq 0$ must be loose, that is

$$(B_{12} - \gamma A_{12})y_1 > 0.$$

But then it is false that γ is the greatest value of α for which (1) has a solution $y \geq 0$. Therefore $p'By > 0$.

If instead there are no components $p_2 = 0$, then there must be components $y_2 = 0$ and the proof is the same, with the roles of y and p interchanged. Q.E.D.

4. Remarks

The proof offered is constructive. Suppose one has an LP algorithm able to compute the optimum value $v = Y_1 - Y_2$ (as well as the solution y) of problem (6). Regard this algorithm as a function

$$f : \gamma \to v.$$

The problem is simply to find the zero of this function. Call y_0 the solution of (6) at $\gamma = 0$. Due to the feasibility of (1) at $\gamma = 0$, we have $v(0) \leq 0$.

Let now $\gamma^* > 0$ be a value of γ for which $B - \gamma A < 0$ (such a value certainly exists due to KMT assumption). Due to infeasibility of (1) at $\gamma = \gamma^*$, we have $v(\gamma^*) > 0$. Since f is a contiuous function, it must have a zero in $[0, \gamma^*]$. But this zero is unique since it is the optimal value of a primal-dual LP pair. So the problem is solvable by a simple line-search.

References

Burmeister E., Dobell A.R., 1970

Mathematical Theories of Economic Growth Macmillan Pub. Co., NY.

Howe C. W., 1960

An Alternative Proof of the Existence of General Equilibrium in a Von Neumann Model Econometrica, 28, 635-639.

Kemeny G., Morgenstern O., Thompson G.L., 1956

A Generalization of the Von Neumann model of an Expanding Economy Econometrica 24, 115-135.

Murata Y., 1977

Mathematics for Stability and Optimization of Economic Systems Academic Press, NY.

Von Neumann J., 1937

Ergebnisse eines Mathematischen Colloquiums, No. 8, 1937 Vienna, translated in *A Model of General Economic Equilibrium,* Review of Economic Studies, 1945-46, 13,33.

Biografia

Paolo Caravani, associato di Modelli e Simulazione nell'Universita' dell'Aquila, si occupa di Economia Matematica, Teoria del Controllo e Teoria dei Giochi. E' autore di circa 40 pubblicazioni nel settore. E' socio della Society for Economic Dynamics and Control.

A Simple Constructive Proof of Von Neumann Equilibrium

P. Caravani

Abstract

A proof of existence of general equilibrium in the Von Neumann model of an expanding economy is given based on elementary results of Linear Programming. General equilibrium is computable by solving a sequence of Linear Programs.

Paolo Caravani, University of Aquila, 67040 L'Aquila, Italy

1. Introduction

Over the years, existence of the equilibrium in the Von Neumann model has been proved with recourse to analytical tools differing widely in nature as well as in complexity. The original Von Neumann proof made extensive use of game theory. Other proofs rely on fixed point theorems such as Brouwer or Kakutani (for a review, see Burmeister Dobell 1970, p. 207).

In 1960, Charles W. Howe exhibited a proof that did not require game theory or fixed point theorems but used a result on linear inequalities due to A.W. Tucker. Unfortunately though, this result requires independent derivation. To the non-initiated, this derivation can appear lengthy. Murata for instance, derives Tucker' theorem from Farka's Lemma through a sequence of about 10 intermediate theorems (Murata, 1977, pp. 283-288).

Unaware, to my knowledge, of more elementary proofs, I wondered if all this mathematical machinery is really necessary. Here I provide a proof based solely on Linear Programming (henceforth LP). Namely, I employ the only two notions that 1. if a primal-dual pair of LP problems have feasible solutions, they have optimal solutions with a common value, and 2. these satisfy the so-called complementarity slackness conditions: positive primal variables are associated to active dual constraints and loose primal constraints are associated to zero dual variables. The rest of the proof is elementary algebra.

Besides greater simplicity, the advantage is to characterise solutions through a constructive procedure. This feature is not usually obtained using fixed point theorems or linear inequality results.

2. Model Description

Von Neumann economy comprises m goods and n processes. Process j is characterised by an activity level y_j. When it is operated at unit level ($y_j = 1$) its input demands and output supplies are described by two non-negative vectors

$$\begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix} \quad \begin{bmatrix} b_{1j} \\ \vdots \\ b_{mj} \end{bmatrix}.$$

Thus technology is completely specified by an input matrix A and an output matrix B, both non-negative with m rows (goods) and n columns (processes).

Each good must be demanded by at least one process so there is at least one positive entry in each row of A. Each process must supply at least one good, so there is at least one positive entry in each column of B. We refer to these two assumptions as KMT (from Kemeny Morgenstern Thompson (1956) who formulated them in order to relax those originally made by Von Neumann).

Labour is treated as one amongst the m goods. It is supplied to other processes by a "domestic" sector that produces it using other goods (like food, housing, good TV, etc). So consumption is one of the rows of A and there is no exogenous demand. Labor being paid in nature-goods, there is no wage rate and net income is made up by profits only. These are entirely re-invested in the form of input purchases for next period. Under these conditions an equilibrium is sought so as to satsfy

1. Output at some period By must cover input of next period AY where $Y = (1+g)y$, and g is a (balanced) growth rate. Letting $\alpha = 1 + g$

$$\alpha Ay \leq By. \tag{1}$$

2. Revenues $p'B$ cannot exceed production costs $p'A$ plus competitive profits $rp'A$ where r is a (uniform) interest rate. Letting $\beta = 1 + r$

$$\beta p'A \geq p'B. \tag{2}$$

3. Goods in excess of next period demand (components of (1) with strict inequality) are worthless

$$p'(B - \alpha A)y = 0. \tag{3}$$

4. Processes generating less than competitive profits (components of (2) with strict inequality) remain idle

$$p'(B - \beta A)y = 0. \tag{4}$$

5. Something of value is produced in the economy $p'By > 0$.

3. Von Neumann Theorem

Theorem. *(Von Neumann, 1937)*

A solution $y \geq 0$ $p \geq 0$ satisfying (1-5) exists if and only if $\alpha = \beta = \gamma$ where

$$\gamma = \max\{\alpha : By - \alpha Ay \geq 0, \quad y \geq 0\}$$

$$\gamma = \min\{\beta : p'B - \beta p'A \leq 0, \quad p \geq 0\}.$$

Proof. *Necessity.* Assume that $y, p \geq 0$ satisfying 1-5 exist. From (3) and (4) it follows $p'By = \alpha p'Ay$ and $p'By = \beta p'Ay$. Since $p'By > 0$, if a solution exists the r.h.s. are positive and $p'Ay > 0$. Hence $\alpha = \beta$.

Furthermore, from (1) and (2)

$$p'By \geq \alpha p'Ay \quad \text{and} \quad -p'By \geq -\beta p'Ay.$$

Adding each side and recalling that $p'Ay > 0$ it follows $\alpha \leq \beta$. Thus there exists only one common value to α and β, call it γ. This necessarily satisfies

$$\max\{\alpha : By - \alpha Ay \geq 0, \quad y \geq 0\} = \gamma = \min\{\beta : p'B - \beta p'A \leq 0, \quad p \geq 0\}.$$

Sufficiency. Assume that $\beta = \alpha = \gamma = \max\{\alpha : By - \alpha Ay \geq 0, \quad y \geq 0\}$. The maximum certainly exists, since the set of $y \geq 0$ satisfying (1) is either empty or closed and bounded. Now we have to show that there exist a $y \geq 0$ and a $p \geq 0$ satisfying 1-5. Letting

$$C = B - \gamma A$$

and introducing scalars $Y_1, Y_2 \geq 0$ and $P_1, P_2 \geq 0$, consider the LP problem

$$\min \quad [0'|1|-1]\begin{bmatrix} y \\ Y_1 \\ Y_2 \end{bmatrix}$$

$$\begin{bmatrix} C & 1 & -1 \\ 1' & 0 & 0 \\ -1' & 0 & 0 \end{bmatrix}\begin{bmatrix} y \\ Y_1 \\ Y_2 \end{bmatrix} \geq \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}. \tag{6}$$

Its dual is

$$\max \quad [0'|1|-1]\begin{bmatrix} p \\ P_1 \\ P_2 \end{bmatrix}$$

$$[p'|P_1|P_2]\begin{bmatrix} C & 1 & -1 \\ 1' & 0 & 0 \\ -1' & 0 & 0 \end{bmatrix} \leq [0'|1|-1]. \tag{7}$$

Rewriting (6,7) as

$$\min \quad Y_1 - Y_2 : \quad Cy + (Y_1 - Y_2)1 \geq 0, \quad 1'y = 1$$

$$\max \quad P_1 - P_2 : \quad p'C + (P_1 - P_2)1' \leq 0, \quad p'1 = 1$$

it is clear that, for any pair $y, p \geq 0$ with components adding to one, it is always possible to choose $Y_1 - Y_2$ and $P_2 - P_1$ large enough so as to satisfy the constraints. Since there exist feasible solutions there exist optimal solutions, which necessarily satisfy

$$Y_1 - Y_2 = P_1 - P_2.$$

We show now that $Y_1 - Y_2 = P_1 - P_2 = 0$. From the dual constraints we have $P_1 - P_2 \leq 0$. But, if it were $P_1 - P_2 < 0$ then $Y_1 - Y_2 < 0$ and from $Cy + (Y_1 - Y_2)\mathbf{1} \geq 0$ it would follow

$$Cy = [B - \gamma A]y > 0$$

against the assumption that γ is the largest value of α for which (1) has a solution $y \geq 0$. Thus $P_1 - P_2 = 0 = Y_1 - Y_2$ and this shows that there exist two non-zero vectors $y \geq 0$ and $p \geq 0$ satisfying (1) e (2). Conditions 3 and 4, for $\alpha = \beta = \gamma$ are the complementarity slackness conditions of LP and these are obviously satisfied given the optimality of the pair y, p.

We finally prove that 5 holds. Observe first that

$$p'By \geq \bar{p}'B\bar{y}$$

where $\bar{p}$ and $\bar{y}$ are vectors obtained from p and y by deleting exactly those zero-components that are associated to active constraints in the respective duals (if there are any) and $\bar{B}$ is the matrix obtained from B by deleting the corresponding rows and columns. Therefore, it is enough to prove 5 under the assumption

$$p + Cy > 0 \qquad y - C'p > 0. \tag{8}$$

Rearranging components if necessary, the optimal solutions can be partitioned as

$$y = \{y_1|y_2\}' \quad p = \{p_1|p_2\}'$$

where the first components of each vector are positive and the remaining components (if there are any) are zero. Partitioning conformally A and B, we get from (3,4) (with $\alpha = \beta = \gamma$)

$$p_1'(B_{11} - \gamma A_{11})y_1 = 0.$$

Now, if $p'By = 0$, then $B_{11} = 0$ (being positive p_1, y_1) and thus also $A_{11} = 0$, and thus also

$$B_{11} - \gamma A_{11} = 0.$$

From (8), if there are components $p_2 = 0$, then the constraints associated to p_2 in the dual of $p'(B - \gamma A) \leq 0$ must be loose, that is

$$(B_{12} - \gamma A_{12})y_1 > 0.$$

But then it is false that γ is the greatest value of α for which (1) has a solution $y \geq 0$. Therefore $p'By > 0$.

If instead there are no components $p_2 = 0$, then there must be components $y_2 = 0$ and the proof is the same, with the roles of y and p interchanged. Q.E.D.

4. Remarks

The proof offered is constructive. Suppose one has an LP computer code able to produce either an optimal solution or an infeasibility message. Consider now the primal LP problem formulated in the proof of the theorem. Starting at $\alpha = 0$ (where a solution certainly exists) choose a non-converging increasing sequence $\{\alpha_k, \quad k = 1, 2, \ldots\}$ and use at each k the LP code. Suppose at $k = m$ an infeasibility message is received for the first time (clearly, $m < \infty$). Then choose a sequence $\{\alpha_h, \quad h = 1, 2, \ldots\}$ converging to α_m with $\alpha_1 = \alpha_{m-1}$ and use again the code at each h. It is obvious that, proceeding in this manner, γ can be approximated to any desired accuracy - and likewise the solution.

A final comment to the KMT assumption. I retained it as a homage to the tradition but it seems that all is needed in the proof is non-negativty and non-nullity of A and B.

eferences

Burmeister E., Dobell A.R., 1970

Mathematical Theories of Economic Growth Macmillan Pub. Co., NY.

Howe C. W., 1960

An Alternative Proof of the Existence of General Equilibrium in a Von Neumann Model Econometrica, 28, 635-639.

Kemeny G., Morgenstern O., Thompson G.L., 1956

A Generalization of the Von Neumann model of an Expanding Economy Econometrica 24, 115-135.

Murata Y., 1977

Mathematics for Stability and Optimization of Economic Systems Academic Press, NY.

Von Neumann J., 1937

Ergebnisse eines Mathematischen Colloquiums, No. 8, 1937 Vienna, translated in *A Model of General Economic Equilibrium,* Review of Economic Studies, 1945-46, 13,33.

IL PROBLEMA DELL'INTEGRAZIONE INDEFINITA

FERDINANDO CASOLARO

SOMMARIO

In quel che segue affronteremo il *"problema di decisione per integrali indefiniti"*, ossia la ricerca di un algoritmo che consenta di decidere se la primitiva di una funzione data sia o meno esprimibile mediante funzioni elementari.

Il problema, come si sa, è insolubile per integrali qualsiasi; però, negli ultimi decenni il matematico americano Risch, sfruttando i risultati di Liouville (1809-1892) del secolo scorso, ha sviluppato due algoritmi che permettono di decidere per alcune classi di funzioni che contengono esponenziali o logaritmi: l'algoritmo per induzione e l'algoritmo parallelo. Qui presenteremo l'algoritmo parallelo.

In precedenza, l'unico algoritmo di decisione noto (già dalla seconda metà del secolo scorso) è riferito alla classe dei *"differenziali binomi"* ed è espresso dal teorema di Tchebichev, da cui trarremo alcune conseguenze che permettono di decidere anche per funzioni trigonometriche.

ABSTRACT

In what follows we deal with the *"decision problem for indefinite integrals"*, nomely the search of an algorithm that lets us of decide if the primitive of a given function can be expressed or not by elementary functions.

As it is know, the problem is irresoluble for integrals of any type; but in the last decades the american mathematical Risch, using the results of Liouville of the previous century, has developed two algorithms that let of decide for some classes of functions which contain exponential or logaritms: the induction algorithm and the parallel algorithm. We present here the parallel algorithm.

Formerly, the only known algorithm of decision (already since the second half of the previous century) is referred to the class of the *"differential binomials"*, and is expressed by the theorem of the Tchebichev, from which we deduce some conseguences that also let of decide for trigonometrical functions.

INTRODUZIONE

Con il termine *funzioni elementari* indicheremo tutte le funzioni complesse di una variabile reale che rientrino nelle seguenti categorie:

a) funzioni razionali;

b) funzioni algebriche sia implicite che esplicite;

c) la funzione esponenziale $exp(x)$;

d) la funzione logaritmica;

e) tutte le funzioni ottenibili mediante una combinazione finita di simboli di funzioni appartenenti alle categorie precedenti.

In particolare, in e) possiamo includere le funzioni goniometriche, le funzioni iperboliche e, rispettivamente, le inverse di esse. Le goniometriche si possono esprimere come combinazioni lineari, nel campo complesso, delle funzioni esponenziali mediante le note formule di Eulero:

a) $sen\, z = \frac{e^{iz} - e^{-iz}}{2i}$; b) $cos\, z = \frac{e^{iz} + e^{-iz}}{2}$; c) $tg\, z = \frac{e^{iz} - e^{-iz}}{i(e^{iz} + e^{-iz})}$

Da esse si ricavano, come combinazioni di funzioni logaritmiche, le inverse:

c) $z = arcsen\, x = \frac{1}{i} \ln\left(\sqrt{1 - x^2} + ix\right)$, con $x = sen\, z$

d) $z = arccos\, x = \frac{1}{i} \ln\left(x + \sqrt{x^2 - 1}\right)$, con $x = cos\, z$

e) $z = arctg\, x = \frac{1}{2i} \ln \frac{1 + ix}{1 - ix}$, con $x = tg\, z$

Analogamente, sussistono le note relazioni per funzioni iperboliche:

$$senh\, z = \frac{e^{z} - e^{-z}}{2} \; ; \quad cosh\, z = \frac{e^{z} + e^{-z}}{2} \; ; \quad tgh\, z = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

e le relative funzioni inverse:

$$settsenh\, x = \ln\left(x + \sqrt{1 + x^2}\right)$$

$$settcosh\, x = \ln\left(x + \sqrt{x^2 - 1}\right)$$

$$sett tgh\, x = \frac{1}{2} \ln\left(\frac{1 + x}{1 - x}\right) .$$

§ 1 - Integrazione delle funzioni razionali.

E' noto che: "*la derivata di una funzione razionale è ancora una funzione razionale*".

Infatti, indicato con $\mathfrak{D}(R)$ l'insieme delle funzioni razionali $R(x)$ sul campo complesso, risulta:

$$R(x) = \frac{P(x)}{D(x)} \Longrightarrow R'(x) = \frac{P'(x)\cdot D(x) - D'(x)\cdot P(x)}{D^2(x)} \in \mathfrak{D}(R) ,$$

con $P(x)$ e $D(x)$ polinomi interi sul campo complesso $\mathbb{C}$.

Di tale proprietà, non vale il viceversa, in quanto la primitiva di una funzione razionale non è, in generale, una funzione razionale.
Sussiste, però, il teorema di Liouville, che, applicato al solo caso delle funzioni razionali, può essere così espresso:

"*Sia* $R(x) = \frac{P(x)}{D(x)}$ *una funzione razionale e sia* n *il grado del polinomio al denominatore* $D(x)$. *Se si conoscono le* n *radici, reali o complesse, di* $D(x)$, *l'integrale*

$$\int R(x)\,dx$$

è esprimibile mediante funzioni elementari e si ha:

$$\int R(x)\,dx = h_o(x) + \sum_i c_i \ln h_i(x) \qquad [1.1]$$

con $h_o(x)$ *funzione razionale di* x, $h_i(x)$ *polinomi irriducibili di primo grado sul campo complesso,* c_i *costanti complesse*".

Verifichiamo tale teorema con un esempio, rimandando, per una trattazione più completa con relativa dimostrazione, all'articolo Pessa-Rizzi "*L'integrazione indefinita delle funzioni trascendenti*"- periodico di matematiche n.2 del 1990 [2].
Esempio -

$$\int \frac{x^5 - x^4 - 8x^3 + 23x - 14}{x^4 - 8x^2 - 8x + 15}\,dx = \int \left[x - 1 + \frac{1}{x^4 - 8x^2 - 8x + 15}\right] dx =$$

$$= \frac{x^2 - 2x}{2} + \int \frac{dx}{(x-3)(x-1)(x^2 + 4x + 5)}\,.$$

Applicando alla funzione integrale il metodo di decomposizione in somma, si ha:

$$\frac{1}{(x-3)(x-1)(x^2+4x+5)} = \frac{A}{x-3} + \frac{B}{x-1} + \frac{Cx + D}{x^2+4x+5}\,,$$

$$A(x-1)(x^2+4x+5) + B(x-3)(x^2+4x+5) + (Cx+D)(x^2-4x+3) = 1$$

per $x = 3$: $52A = 1$; $A = \frac{1}{52}$

per $x = 1$: $-20B = 1$; $B = -\frac{1}{20}$

per $x = 0$: $-5A - 15B + 3D = 1$; $-\frac{5}{52} + \frac{3}{4} + 3D = 1$; $D = \frac{3}{26}$

Annullando il termine di 3° grado si ha:

$$A + B + C = 0\ ;\ \frac{1}{52} - \frac{1}{20} + C = 0\ ;\ C = \frac{2}{65}$$

Pertanto risulta:

$$\int \frac{x^5 - x^4 - 8x^3 + 23x - 14}{x^4 - 8x^2 - 8x + 15}dx = \frac{x^2 - 2x}{2} + \frac{1}{52}\int \frac{dx}{x-3} - \frac{1}{20}\int \frac{dx}{x-1} + \int \frac{\frac{2}{65}x + \frac{3}{26}}{x^2 + 4x + 5}\,dx =$$

$$= \frac{x^2 - 2x}{2} + \frac{1}{52}\ln|x-3| - \frac{1}{20}\ln|x-1| + \frac{1}{65}\int\frac{2x+4}{x^2+4x+5}dx + \frac{7}{130}\int\frac{dx}{x^2+4x+5} =$$

$$= \frac{x^2 - 2x}{2} + \frac{1}{52}\ln|x-3| - \frac{1}{20}\ln|x-1| + \frac{1}{65}\ln|x^2+4x+5| + \frac{7}{130}arctg|x+2| + c_1 .$$

Da: $arctg(x+2) = \frac{1}{2i}\ln\frac{1+i(x+2)}{1-i(x+2)} = \frac{1}{2i}\ln\left[1+i(x+2)\right] - \frac{1}{2i}\ln\left[1-i(x+2)\right] =$

$$= -\frac{1}{2i}\left[\ln\frac{1}{2i}\left(i-x-2\right) - \ln\frac{1}{2i}\left(i+x+2\right)\right] = \frac{1}{2i}\left[\ln\left[i(x+2-i)\right] - \ln\left[-i(x+2+i)\right]\right] =$$

$$= \frac{1}{2i}\ln(x+2-i) - \frac{1}{2i}\ln(x+2+i) + c_2$$

e da: $x^2+4x+5 = (x+2+i)(x+2-i)$, si ha in definitiva:

$$\int\frac{x^5 - x^4 - 8x^3 + 23x - 14}{x^4 - 8x^2 - 8x + 15}dx = \frac{x^2-2x}{2} + \frac{1}{52}\ln|x-3| - \frac{1}{20}\ln|x-1| + \frac{1}{65}\ln(x+2+i) +$$

$$+ \frac{1}{65}\ln(x+2-i) + \frac{7}{260i}\ln(x+2-i) - \frac{7}{260i}\ln(x+2+i) =$$

$$= \frac{x^2-2x}{2} + \frac{1}{52}\ln|x-3| - \frac{1}{20}\ln|x-1| + \left(\frac{1}{65} - \frac{7i}{260}\right)\ln(x+2-i) +$$

$$+ \left(\frac{1}{65} + \frac{7i}{260}\right)\ln(x+2+i) + k \quad , \qquad \text{con } k = c_1 + c_2 \in \mathbb{C}.$$

Come si vede, il risultato è costituito dalla somma della funzione razionale $h_0(x) = \frac{x^2-2x}{2}$ e da una combinazione lineare di logaritmi derivati dai quattro fattori in cui è stato scomposto il polinomio di 4° grado al denominatore.

Da quanto detto, si deduce che il problema della decisione per integrali indefiniti di funzioni razionali è legato alla possibilità di determinare o meno, per via algebrica, le radici del polinomio al denominatore; al di là di tali problemi, gli integrali di funzioni razionali sono sempre esprimibili in termini di funzioni elementari.

§ 2 - Integrali di funzioni trascendenti. Generalizzazione del teorema di Liouville a funzioni che contengono esponenziali e logaritmi di funzioni razionali.

In forma più generale, il teorema di Liouville si estende a funzioni del tipo $F\left(x,\ e^{R(x)},\ log\,S(x)\right)$, con $R(x)$ e $S(x)$ funzioni razionali.

"Sia $f(x)$ una funzione razionale contenente esponenziali e logaritmi di funzioni razionali, cioè del tipo $F\left(x,\ e^{R(x)},\ lg\,S(x)\right)$; se esiste una funzione $g(x)$ tale che:

$$\frac{dg}{dx} = f(x)$$

allora esistono delle funzioni $h_0,\ h_1,\ldots,\ h_n$ delle variabili x, e^x, $lg\,x$,

e delle costanti complesse $c_1, c_2, \ldots, c_n$, *tali che:*

$$g(x) = h_o(x, e^x, lg\, x) + \sum_i c_i\ lg\ h_i(x, e^x, lg\, x) \qquad [2.1]$$

Ovviamente, come per la [1.1], h_o è una funzione razionale di x, e^x, $lg\, x$ e le h_i sono funzioni delle stesse variabili.
In altri termini, dal teorema di Liouville segue che se la $f(x)$ è dotata di primitiva elementare $g(x)$, tale primitiva deve essere espressa dalla [2.1].

Basandosi sui risultati di Liouville, Risch [1] ha sviluppato gli algoritmi che permettono di decidere sulla esistenza o meno di primitive elementari di funzioni del tipo

$$f(x) = F\left[x, e^{R(x)}, lg\, S(x)\right]$$

con R(x) e S(x) funzioni razionali di x.
Noi esporremo *l'algoritmo parallelo* direttamente con alcuni esempi, rimandando, per una più approfondita trattazione teorica, al già citato articolo Pessa-Rizzi [2].

TEOREMA 2.1

L'integrale $\int \frac{e^x}{x}\, dx$ *non è esprimibile mediante funzioni elementari.*

Dim.- Se l'integrale si potesse esprimere mediante funzioni elementari, allora deve esistere una funzione $g(x)$ tale che:

$$\frac{dg}{dx} = \frac{e^x}{x}$$

Per il teorema di Liouville, $g(x)$ dovrebbe essere del tipo:

$$g(x) = h(x, e^x) + c\ ln\ x$$

cioè:
$$\int \frac{e^x}{x}\, dx = h(x, e^x) + c\ ln\ x\ , \qquad [2.2]$$

in cui compare un solo addendo logaritmico in quanto il denominatore presenta l'unico fattore irriducibile x. La funzione $h(x, e^x)$, di cui ipotizziamo l'esistenza, deve essere, ovviamente, una combinazione lineare finita di x ed e^x, per cui la possiamo esprimere nel seguente modo:

$$h(x, e^x) = a_o + a_1 x + a_2 e^x + a_3 x\, e^x + a_4 x^2 + a_5 e^{2x} + \ldots \qquad (*)$$

da cui:

$$h'(x, e^x) = a_1 + a_2 e^x + a_3 e^x + a_3 x\, e^x + 2\, a_4 x + \ldots \qquad (**)$$

Derivando ambo i membri della [2.2], si ottiene:

$$\frac{e^x}{x} = h'(x, e^x) + \frac{c}{x}\ ,$$

cioè:

$$e^x = x\, h'(x, e^x) + c\ .$$

Questa ultima è una semplice equazione differenziale che deve essere verificata dall'espressione di $h'(x, e^x)$ della (* *):

$$e^x - c = a_1 x + (a_2 + a_3) x\, e^x + a_3 x^2 e^x + 2\, a_4 x^2 + \ldots .$$

Tale identità non può essere verificata perchè il termine della sola e^x compare al primo membro con coefficiente 1, ma non al secondo.
Pertanto, l'ipotesi della esistenza della funzione $h(x, e^x)$ verificante la [2.2], conduce ad una contraddizione; l'integrale $\int \frac{e^x}{x}\,dx$ non è esprimibile mediante funzioni elementari.

TEOREMA 2.2

L'integrale $\int e^{x^2} dx$ *non è esprimibile mediante funzioni elementari.*

Dim.- Se l'integrale si potesse esprimere mediante funzioni elementari, esisterebbe una funzione $g(x)$ tale che:

$$g(x) = \int e^{x^2} dx = h\left(x,\ e^{x^2}\right) + c \qquad [2.3]$$

dove non figurano termini logaritmici (la funzione integranda non presenta elementi irriducibili al denominatore). Sviluppando $h\left(x,\ e^{x^2}\right)$ si ha:

$$h\left(x,\ e^{x^2}\right) = a_o + a_1 x + a_2 e^{x^2} + a_3 x\, e^{x^2} + a_4 x^2 + \ldots$$

da cui:

$$h'\left(x,\ e^{x^2}\right) = a_1 + 2\, a_2 x\, e^{x^2} + a_3 e^{x^2} + 2\, a_3 x^2 e^{x^2} + 2\, a_4 x + \ldots$$

Derivando ambo i membri della [2.3], si ottiene:

$$e^{x^2} = h'\left(x,\ e^{x^2}\right)$$

cioè: $$e^{x^2} = a_1 + 2\, a_2 x\, e^{x^2} + a_3 e^{x^2} + 2\, a_3 x^2 e^{x^2} + 2\, a_4 x + \ldots$$

in cui basta osservare che per i coefficienti di e^{x^2} risulta $a_3 = 1$, mentr per i coefficienti di $x^2 e^{x^2}$ risulta $a_3 = 0$, per dedurre la non esistenz dell'ipotetica $h\left(x,\ e^{x^2}\right)$ verificante la [2.3].

§ 3 - Integrali di funzioni irrazionali.

Richiamiamo alcuni noti integrali di funzioni irrazionali:

$$\int \frac{d\ell(\text{æ})}{\sqrt{1 - [\ell(\text{æ})]^2}} = \text{arcsen } \ell(\text{æ}) + c = - \text{ arccos } \ell(\text{æ}) + c \qquad [3.1]$$

$$\int \frac{d\ell(\text{æ})}{\sqrt{1 + [\ell(\text{æ})]^2}} = \text{settsenh } \ell(\text{æ}) + c = \ln\left[\ell(\text{æ}) + \sqrt{1 + [\ell(\text{æ})]^2}\,\right] \qquad [3.2]$$

$$\int \frac{d\ell(\text{æ})}{\sqrt{[\ell(\text{æ})]^2 - 1}} = \text{settcosh } \ell(\text{æ}) + c = \ln\left[\ell(\text{æ}) + \sqrt{[\ell(\text{æ})]^2 - 1}\,\right] \qquad [3.3]$$

L'integrale di una funzione irrazionale riconducibile ad uno dei casi [3.1], [3.2], [3.3], è esprimibile mediante funzioni elementari.
In generale, il problema dell'integrazione di funzioni irrazionali si conduce alla ricerca di opportune sostituzioni che permettono di razionalizzare la funzione integranda. In tal caso il teorema di Liouville assicura l'esistenza di primitive elementari.
Ovviamente, solo in alcuni casi esistono sostituzioni che permettono di razionalizzare la funzione integranda. E' interessante richiamare, a tal proposito, il seguente teorema di Tchebichev che permette di decidere per la particolare classe dei "*differenziali binomi*":

"*L'integrale del differenziale binomio:*

$$\int \text{æ}^p (a\ \text{æ}^r + b)^s \, d\text{æ}$$

con p, r, s *numeri razionali, è esprimibile mediante funzioni elementari se, e solo se, è intero uno dei tre numeri:*
$s\ ;\ \frac{p+1}{r}\ ;\ s + \frac{p+1}{r}$ ".

Una immediata conseguenza di tale proprietà è la seguente:

TEOREMA 3.1

Gli integrali del tipo $\int \sqrt{a \mp \text{æ}^n}\, d\text{æ}$, *con* $a \in \mathbb{R}$, $n \in \mathbb{N}$, *non sono esprimibili mediante funzioni elementari se* $n > 2$.

Dim.- $\sqrt{a \pm \text{æ}^n} = \text{æ}^0 (a \pm \text{æ}^n)^{1/2}$, cioè: $p = 0$; $r = n$; $s = \frac{1}{2}$.
Pertanto: $s \notin \mathbb{Z}$; $\frac{p+1}{r} = \frac{1}{n} \notin \mathbb{Z}$, $\forall\, n \neq 1$;
$s + \frac{p+1}{r} = \frac{1}{2} + \frac{1}{n} < 1$, $\forall\, n > 2$.

Per un'analisi più profonda di altre conseguenze del teorema di Tchebichev

e dei problemi relativi alla integrabilità di funzioni irrazionali vedi: Casolaro-Pessa-Rizzi *"Algoritmi di decisioni per integrali indefiniti"* - Atti del Convegno Nazionale Mathesis 1991 "MATEMATICA MODERNA E INSEGNAMENTO" [3]. Nel prossimo paragrafo trarremo conseguenze relative alla decisione per funzioni goniometriche.

§ 4 - Integrali di funzioni trascendenti contenenti espressioni trigonometriche.

4.1 - Utilizzando le formule di Eulero e l'algoritmo di Risch si può decidere sull'integrabilità di alcune funzioni del tipo: $R(x, \mathit{sen}x)$, $R(x, \mathit{cos}x)$.

TEOREMA 4.1

L'integrale $\int \frac{\mathit{sen}x}{x}\,dx$ *non è esprimibile mediante funzioni elementari.*

Dim. - Risulta:

$$\int \frac{\mathit{sen}x}{x}\,dx = \int \frac{e^{ix} - e^{-ix}}{2ix}\,dx = \frac{1}{2}\int \frac{e^{ix}}{ix}\,dx - \frac{1}{2}\int \frac{e^{-ix}}{ix}\,dx =$$

$$= \frac{1}{2i}\int \frac{e^{ix}}{ix}\,d(ix) - \frac{1}{2i}\int \frac{e^{-ix}}{-ix}\,d(-ix) \qquad [4.1]$$

in cui, dal teorema 2.1, gli integrali all'ultimo membro non sono esprimibili per via elementare.

TEOREMA 4.2

L'integrale $\int \mathit{cos}x^2\,dx$ *non è esprimibile mediante funzioni elementari.*

Dim. - Risulta:

$$\int \mathit{cos}x^2\,dx = \int \frac{e^{ix^2} + e^{-ix^2}}{2}\,dx = \frac{1}{2\sqrt{i}}\int e^{(\sqrt{i}\,x)^2}\,d(\sqrt{i}\,x) + \frac{1}{2}\int e^{-(\sqrt{i}\,x)^2}\,d(x) \qquad [4.2]$$

in cui, ricordando che: $i^2 = -1$, $e^{-ix^2} = e^{i^3x^2}$, risulta:

$$\int \cos x^2\,dx = \frac{1}{2\sqrt{i}}\int e^{(\sqrt{i}\,x)^2}\,d(\sqrt{i}\,x) + \frac{1}{2i\sqrt{i}}\int e^{(i\sqrt{i}\,x)^2}\,d(i\sqrt{i}\,x),$$

per il teorema 2.2, gli integrali all'ultimo membro non sono esprimibili per via elementare.

4.2 - Utilizzando il teorema di Tchebichev si può decidere sugli integrali del tipo: $\int sen^{\alpha}\text{æ}\ cos^{\beta}\text{æ}\ d\text{æ}$ (*)

con α, β numeri razionali.

Posto: $sen\text{æ} = \sqrt{t}$; $cos\text{æ} = \sqrt{1-t}$; $\text{æ} = arcsen\sqrt{t}$; $d\text{æ} = \frac{1}{2\sqrt{t}\cdot\sqrt{1-t}}$

risulta:

$$\int sen^{\alpha}\text{æ}\ cos^{\beta}\text{æ}\ d\text{æ} = \int (\sqrt{t})^{\alpha}(\sqrt{1-t})^{\beta}\ \frac{1}{2\sqrt{t}\cdot\sqrt{1-t}}\ dt =$$

$$= \frac{1}{2}\int t^{(\alpha-1)/2}\ (1-t)^{(\beta-1)/2}\ dt \qquad [4.3]$$

che rappresenta l'integrale del differenziale binomio $\int t^{p}(b + at^{r})^{s}\ dt$, con $p = \frac{\alpha-1}{2}$, $r = 1$, $s = \frac{\beta-1}{2}$.

Dal teorema di Tchebichev, sappiamo che tale integrale è esprimibile mediante funzioni elementari se, e solo se, uno dei numeri:

$$s\ ;\quad \frac{p+1}{r}\ ;\quad s + \frac{p+1}{r}$$

è intero.

Nella [4.3] risulta:

$$s = \frac{\beta-1}{2}\ ;\quad \frac{p+1}{r} = \frac{\alpha+1}{2}\ ;\quad s + \frac{p+1}{r} = \frac{\alpha+\beta}{2}$$

per cui vale il seguente:

Lemma 4.2.1 - *"Gli integrali del tipo* $\int sen^{\alpha}\text{æ}\ cos^{\beta}\text{æ}\ d\text{æ}$ *sono esprimibili mediante funzioni elementari se, e solo se, è intero uno dei numeri:*

$$\frac{\alpha+1}{2}\ ;\quad \frac{\beta-1}{2}\ ;\quad \frac{\alpha+\beta}{2}\ "$$

E' evidente che se α e β sono entrambi interi o se uno dei due è intero dispari e l'altro razionale non intero, l'integrale è esprimibile mediante funzioni elementari. Infatti:

$$\alpha \text{ dispari} \Longrightarrow \frac{\alpha+1}{2} \in \mathbb{Z}\ ;$$

$$\beta \text{ dispari} \Longrightarrow \frac{\beta-1}{2} \in \mathbb{Z}\ ;$$

$$\alpha \text{ e } \beta \text{ pari} \Longrightarrow \frac{\alpha+\beta}{2} \in \mathbb{Z}\ ;$$

Dal Lemma 4.2.1 si deduce facilmente che gli integrali

$$\int \sqrt{\sin x}\,dx\ ,\quad \int \sqrt{\cos x}\,dx\ ,\quad \int \sqrt{\sin x\cos x}\,dx\ ,$$

non sono esprimibili mediante funzioni elementari. Per essi risulta:

$$\int \sqrt{\sin x}\,dx = \int (\cos x)^{0}(\sin x)^{1/2}\,dx\ .$$

$$\frac{\alpha+1}{2} = \frac{3}{4} \notin \mathbf{Z}\ ;\quad \frac{\beta-1}{2} = -\frac{1}{2} \notin \mathbf{Z}\ ;\quad \frac{\alpha+\beta}{2} = \frac{1}{4} \notin \mathbf{Z}\ ;$$

$$\int \sqrt{\cos x}\,dx = \int (\sin x)^{0}(\cos x)^{1/2}\,dx\ .$$

$$\frac{\alpha+1}{2} = \frac{1}{2} \notin \mathbf{Z}\ ;\quad \frac{\beta-1}{2} = -\frac{1}{4} \notin \mathbf{Z}\ ;\quad \frac{\alpha+\beta}{2} = \frac{1}{4} \notin \mathbf{Z}\ ;$$

$$\int \sqrt{\sin x\cos x}\,dx = \int (\sin x)^{1/2}(\cos x)^{1/2}\,dx\ .$$

$$\frac{\alpha+\beta}{2} = \frac{1}{2} \notin \mathbf{Z}\ .$$

In generale, vale il:

TEOREMA 4.3

"Gli integrali del tipo $\int \sqrt[n]{\sin x\cos x}\,dx$ *con* $n \in \mathbf{N}$ *ed* $n > 1$*, non sono esprimibili mediante funzioni elementari."*

Dim. - Risulta:

$$\frac{\alpha+1}{2} = \frac{\frac{1}{n}+1}{2} = \frac{1}{2n}+\frac{1}{2} \notin \mathbf{Z}\ ,\ \forall\, n \neq \pm 1;$$

$$\frac{\beta-1}{2} = \frac{\frac{1}{n}-1}{2} = \frac{1}{2n}-\frac{1}{2} \notin \mathbf{Z}\ ,\ \forall\, n \neq \pm 1;$$

$$\frac{\alpha+\beta}{2} = \frac{\frac{1}{n}+\frac{1}{n}}{2} = \frac{1}{n} \notin \mathbf{Z}\ ,\ \forall\, n \neq \pm 1\ .$$

BIBLIOGRAFIA

[1] R.H. RISCH, *Transaction of the American Mathematical Society*, 1969, 139, 167-189.

[2] PESSA-RIZZI, *"L'integrazione indefinita delle funzioni trascendenti"* - Periodico di matematiche n.2 del 1990.

[3] CASOLARO-PESSA-RIZZI, *"Algoritmi di decisioni per integrali indefiniti"* - Atti del Convegno Nazionale Mathesis 1991 "MATEMATICA MODERNA E INSEGNAMENTO".

CUMULANTI E INVERSIONI DI MOBIUS

Mauro CERASOLI
Università degli Studi dell'Aquila

Sunto. Mediante un'inversione di Möbius nel reticolo delle partizioni di un n-insieme può essere dimostrata la formula che esprime il cumulante n-esimo di una variabile aleatoria in funzione dei suoi momenti. Sono presentati infine alcuni problemi aperti.

1. MOMENTI E CUMULANTI

Sia X una variabile aleatoria su uno spazio di probabilità (Ω,F,P) avente momenti

$$m_k = E(X^k) = \int_\Omega X^k dP \qquad k = 0, 1, 2,...$$

Associata ad m_k consideriamo la successione numerica $K_n(X)$, $n = 1, 2,...$, definita dalla seguente identità:

$$(1) \qquad E(e^{tX}) = \Sigma_{j\geq 0}\, E(X^j)t^j/j! = \exp(\Sigma_{n\geq 1} K_n(X)t^n/n!)$$

dove t è una variabile formale. Il numero $K_n(X)$ viene chiamato <u>cumulante</u> n-esimo (o di ordine n) di X (o della distribuzione di probabilità di X). Introducendo l'operatore derivata D possiamo scrivere in forma più breve

$$K_n(X) = D^n \log E(e^{tX})_{|t=0}.$$

Si può controllare che i primi quattro cumulanti valgono

$K_1 = m_1 = E(X)$

$K_2 = m_2 - m_1^2 = Var(X)$

$K_3 = m_3 - 3m_2m_1 + 2m_1^3$

$K_4 = m_4 - 4m_3m_1 + 12m_2m_1^2 - 6m_1^4 + 3m_2^2$.

I cumulanti hanno notevoli applicazioni nella statistica perché generalizzano il concetto di varianza, ma la loro vera natura ci appare del tutto sconosciuta. Non è difficile dimostrare però che essi godono di tre particolari proprietà che riportiamo nel seguente teorema.

TEOREMA Il primo cumulante $K_1(X)$ coincide con la media m_1 di X e per ogni c reale si ha

a) $K_n(X+c) = K_n(X), \qquad n>1;$

b) $K_n(cX) = c^n K_n(X)$.

c) Se X e Y sono indipendenti allora

$$K_n(X+Y) = K_n(X) + K_n(Y).$$

Dim. Sostituendo X con X+c in (1) e sviluppando risulta

$$\exp(\Sigma_{n\geq 1} K_n(X+c)t^n/n!) = E(e^{t(X+c)}) = e^{tc}E(e^{tX})$$

$$= e^{tc} \exp(\Sigma_{n\geq 1} K_n(X)t^n/n!)$$

$$= \exp[t(c+K_1(X)) + \Sigma_{n>1} K_n(X)t^n/n!]$$

quindi $K_1(X+c) = K_1(X)+c$, ovvero $K_1(X) = m_1$, e $K_n(X+c) = K_n(X)$ per $n>1$.

Anche la seconda proprietà è immediata perché

$$\exp(\Sigma_{n\geq 1} K_n(cX)t^n/n!) = E(e^{t(cX)}) = E(e^{(ct)X})$$

$$= \exp(\Sigma_{n\geq 1} K_n(X)c^n t^n/n!)$$

Infine la terza proprietà, da cui proviene il termine cumulante, è conseguenza del fatto che

$$E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY})$$

se X e Y sono indipendenti.

Riportiamo qui di seguito, per meglio illustrare il concetto di cumulante, quanto è noto a proposito dei cumulanti di tre distribuzioni fondamentali della teoria della probabilità.
a) I primi quattro cumulanti della distribuzione binomiale di parametro p sono rispettivamente

$$K_1 = np, \qquad K_2 = np(1-p),$$

$$K_3 = np(1-p)(1-2p), \qquad K_4 = np(1-p)(1-6p+6p^2).$$

Più in generale si può dimostrare che

$$K_{n+1} = (p-p^2)\frac{dK_n}{dp}$$

e con questa formula si possono calcolare tutti i cumulanti della distribuzione binomiale;
b) la distribuzione di Poisson di parametro α ha tutti i cumulanti uguali ad α;
c) la distribuzione normale ha tutti i cumulanti nulli per $n>2$.

2. LA FORMULA DI CAUCHY

Tenendo presente la (1) consideriamo l'identità

$$(2) \qquad 1+\Sigma_{n\geq 1} a_n t^n/n! = \exp(\Sigma_{n\geq 1} b_n t^n/n!)$$

dove a_n e b_n sono variabili formali. Il problema che si presenta spontaneo è di esprimere a_n in funzione di $b_1, b_2, \ldots, b_n$ e, viceversa, di esprimere b_n in funzione di $a_1, a_2, \ldots, a_n$. Per determinare a_n basta osservare che il secondo membro della (2) si può scrivere nella forma

$$\exp(b_1 t)\exp(b_2 t^2/2!)\ldots\exp(b_n t^n/n!)\ldots =$$

$$= \overline{\left|\ \right|}_{j\geq 1}\ \Sigma_{k\geq 1}[(b_j t^j/j!)^k/k!]$$

$$= \overline{\left|\ \right|}_{j\geq 1}\Sigma_{k\geq 1}[b_j{}^k t^{jk}/(j!{}^k k!)]$$

$$= \Sigma_{n\geq 0} \sum_{i_1,i_2,\ldots,i_n\geq 0} \frac{b_1^{i_1} b_2^{i_2}\ldots b_n^{i_n} t^{i_1} t^{2i_2}\ldots t^{ni_n}}{1!^{i_1} 2!^{i_2}\ldots n!^{i_n} i_1!i_2!\ldots i_n!}$$

e quindi ricaviamo

$$a_n = \sum_{\substack{i_1,i_2,\ldots,i_n\geq 0 \\ i_1+2i_2+\ldots+ni_n=n}} \frac{n!}{1!^{i_1}\ldots n!^{i_n} i_1!\ldots i_n!} b_1^{i_1}\ldots b_n^{i_n}$$

In verità questo non è altro che un caso particolare della formula di Faà di Bruno che dà la derivata n-esima della funzione composta f(g(t)), con f uguale ad exp. I coefficenti a_n sono noti come polinomi di Bell e, ricorrendo ad artifizi di tipo umbrale, si potrebbe ricavare b_n in funzione di $a_1, a_2,\ldots,a_n$ (si veda ad esempio Riordan, pag. 36). Il coefficiente del monomio

$$b_1^{i_1} b_2^{i_2}\ldots b_n^{i_n}$$

ha un significato combinatorio che suggerisce però di risolvere diversamente il problema. Infatti sia S un insieme finito; ricordiamo che una partizione di S è una famiglia $\{A_1,A_2,\ldots,A_r\}$ di sue parti (dette classi o blocchi) non vuote, disgiunte, la cui unione è S. La partizione è detta di tipo

$$(i_1,i_2,\ldots,i_n)$$

se possiede i_k classi di cardinalità k, per ogni k = 1, 2,..., n. Allora il coefficiente del monomio su menzionato è uguale al numero di partizioni di un insieme di n elementi di tipo $(i_1,i_2,\ldots,i_n)$: un risultato che si fa risalire a Cauchy. Questa scoperta

invita a ricavare i coefficienti b_n mediante una inversione di Möbius sul reticolo delle partizioni di un insieme finito. L'idea appare già nel lavoro di Doubilet-Rota-Stanley (p.101) ma, recentemente, studi più approfonditi sono stati affrontati da Speed a partire dal 1983. Vediamo pertanto come il problema proposto viene risolto in quest'altro modo.

3. UN'INVERSIONE DI MOBIUS NEL RETICOLO DELLE PARTIZIONI

Sia $S = \{1,2,\ldots,n\}$; dati una partizione π di S, una classe B di π e le variabili formali $b_1, b_2, \ldots, b_n$, poniamo:

a) $\quad w(B) = b_k \qquad$ se $|B|=k$;

b) $\quad g(\pi) = \prod_{B\in\pi} w(B) = b_1^{i_1} b_2^{i_2} \ldots b_n^{i_n}$

se π è di tipo $(i_1,i_2,\ldots,i_n)$. In particolare se $1\hat{}$ è l'unica partizione di S di tipo $(0,0,\ldots,0,1)$ risulta ovviamente

$g(1\hat{}) = b_n$;

c) $\quad f(\sigma) = \prod_{B\in\sigma} \sum_{\pi\in B^*} g(\pi)$

dove B^* è l'insieme delle partizioni di B. Questa identità equivale a

(3) $\quad f(\sigma) = \sum_{\tau\leq\sigma} g(\tau)$

dove $\leq$ indica l'ordine parziale per raffinamento tra partizioni.

d) Se T è un insieme di k elementi, poniamo

$$a_k = \sum_{\pi\in T^*} g(\pi), \qquad |T| = k.$$

Se a w si dà il significato di funzione peso, possiamo

interpretare a_n come l'enumeratore di S^* (vedi Cerasoli-Eugeni-Protasi, p.131). Dalla (3), con $1\hat{}$ al posto di σ, si ricava

$$f(1\hat{}) = \Sigma_{\tau \leq 1\hat{}}\, g(\tau) = \sum_{\pi \in S^*} g(\pi) = a_n.$$

Applicando il teorema d'inversione di Möbius alla (3) otteniamo

$$(4) \qquad g(\sigma) = \Sigma_{\tau \leq \sigma}\, \mu(\tau,\sigma) f(\tau)$$

dove μ è la funzione di Mobius del reticolo di partizioni di S^*. Per i nostri scopi è necessario conoscere il valore $\mu(\tau, 1\hat{})$ dove $1\hat{}$ è la partizione massima di S. Come è noto, Schützenberger ha dimostrato che

$$\mu(\tau, 1\hat{}) = (-1)^{|\tau|-1}(|\tau|-1)!.$$

Ponendo quindi $1\hat{}$ al posto di σ nel primo membro della (4), si ottiene la formula che esprime b_n in funzione di $a_1, a_2, \ldots, a_n$. Questa, quando scriviamo $K_n = K_n(X)$ per b_n ed $m_k = E(X^k)$ per a_k, diventa:

$$K_n = \sum_{\substack{k=1 \\ i_1+i_2+\ldots+i_n=k \\ i_1+2i_2+\ldots+ni_n=n}}^{n} \frac{(-1)^{k-1}(k-1)!\,n!}{1!^{i_1}\ldots n!^{i_n}\, i_1!\ldots i_n!}\, m_1^{i_1}\ldots m_n^{i_n}$$

In definitiva questa formula esprime i cumulanti di X in funzione dei momenti. Con essa proviamo a ricavare i cumulanti $K_n(X)$ di una variabile aleatoria X quando n = 1, 2, 3, 4, dimostrando le formule date nel primo paragrafo.

a) Per n = 1 risulta ovviamente $K_1(X) = E(X)$;

b) per n = 2 le partizioni di $S = \{1,2\}$ sono $1|2$ e 12 di tipi rispettivamente (2,0) e (0,1). Pertanto il secondo cumulante vale

$$K_2(X) = E(X^2) - E(X)^2 = Var(X)$$

cioè la varianza di X.

c) Le partizioni di (1,2,3) sono:

1|2|3 di tipo (3,0,0),

1|23, 12|3, 13|2 di tipo (1,1,0)

123 di tipo (0,0,1).
Quindi il terzo cumulante è

$$K_3(X) = E(X^3) - 3E(X)E(X^2) + 2E(X)^3.$$

Esso è noto anche in statistica come curtosi.
d) Si può infine dimostrare che

$$K_4(X)=E(X^4)-4E(X)E(X^3)+12E(X^2)E(X)^2-6E(X)^4+3E(X^2)^2$$

e con questo abbiamo ottenuto i quattro cumulanti presentati all'inizio.

Se X è costante, uguale ad 1, i suoi momenti valgono tutti 1 e i cumulanti sono nulli a partire dal secondo. Allora dalla formula che dà i cumulanti otteniamo la semplice identità combinatoria

$$\sum_{k=1}^{n} (-1)^{k-1}(k-1)!S(n,k) = \delta_{n1}$$

dove S(n,k) è il numero di Stirling di seconda specie, cioè il numero di partizioni di un n-insieme in k classi.

4. ALCUNI PROBLEMI APERTI

Come si è detto in precedenza, poco si sa sui cumulanti; elenchiamo pertanto alcuni problemi aperti ed interessanti.
a) Data una successione k_n, n = 1, 2, ..., di numeri reali, determinare sotto quali condizioni esiste una variabile aleatoria il cui n-esimo cumulante è k_n. Il problema è collegato al classico problema dei momenti.

b) Esiste una variabile aleatoria con cumulanti non nulli solo per n>3 ?

c) Tra i coefficienti dello sviluppo in polinomi di Hermite della distribuzione di probabilità di X e i cumulanti di X, esiste una strana relazione: se X è standard (ha media 0 e varianza 1), essi coincidono fino ad n = 5 (si veda Kendall-Stuart, p.158). Studi ulteriori su questo fatto potrebbero illuminare di più il concetto di cumulante.

d) Sia X_n una successione di variabili aleatorie tali che

$$E(X_n) = m, \qquad Var(X_n) = \sigma^2, \qquad n = 1, 2, \ldots$$

Se $K_j(X_n)$ converge a 0 quando $n \to \infty$, per ogni j, la successione X_n converge in legge ad una variabile aleatoria normale di media m e varianza σ^2. Sarebbe interessante conoscere altri teoremi di convergenza simili a questo.

e) Sia $p_n(x)$ una successione di polinomi, di grado n = 0, 1, 2, ... nella variabile reale x. Essa è detta di tipo binomiale, se

$$(5) \qquad p_n(x+y) = \sum_{k=0}^{n} \binom{n}{k} p_k(x) p_{n-k}(y)$$

per ogni n (si veda Rota-Kahaner-Odlyzko). Supponiamo poi che per una variabile aleatoria X siano definiti i momenti binomiali

$$b_n(X) = E(p_n(X)) = \int_{\Omega} p_n(X)\,dP$$

associati alla successione $p_n(x)$. Se X e Y sono indipendenti, allora vale un'identità simile alla (5) con b, X, Y rispettivamente al posto di p, x, y. Sia ora $g_X(t)$ la funzione generatrice esponenziale dei numeri $b_n(X)$,

$$g_X(t) = \Sigma_{n \geq 0} b_n(X) t^n / n!$$

Il problema è di studiare i cumulanti binomiali definiti per analogia dalla identità formale

$$c_n(X) = D^n \log g_X(t)\,|_{t=0}, \qquad n = 1, 2, \ldots$$

Se $p_n(x) = x^n$, i cumulanti binomiali si riducono a quelli ordinari; se $p_n(x) = x(x-1)(x-2)\ldots(x-n+1)$, si riducono invece a quelli fattoriali, ben noti in statistica. Un primo passo in questa direzione potrebbe essere quello di studiare i cumulanti relativi ai polinomi fattoriali crescenti

$$p_n(x) = x(x+1)(x+2)\ldots(x+n-1).$$

BIBLIOGRAFIA

1. M.CERASOLI-F.EUGENI-M.PROTASI, "Elementi di Matematica Discreta", Zanichelli, Bologna, 1988

2. P.DOUBILET-G.C.ROTA-R.STANLEY, "Finite Operator Calculus", The Idea of Generating Function, pp. 83-134 Academic Press, New York, 1975

3. M.G.KENDALL-A.STUART, "The advanced Theory of Statistics", vol.I, Griffin, London, 1963

4. J.RIORDAN, "An Introduction to Combinatorial Analysis", Wiley, New York, 1958

5. G.C.ROTA-D.KAHANER-A.ODLYZKO, "Finite Operator Calculus", (vedi 2.) pp.7-82

6. T.P.SPEED, Cumulants and partitions lattices, Austral. J. Statistics, 25 (1983), no. 2, 378-388

7.-- II. Generalised k-statistics, J. Austral. Math. Soc. Ser. A 40 (1986), no. 1, 34-53

8.-- III. Multipy-indexed arrays, J. Austral. Math. Soc. Ser. A 40 (1986), no. 2, 161-182

9.-- IV. A.s. convergence of generalised k-statistics, J. Austral. Math. Soc. Ser. A 41 (1986), no. 1, 79-94.

THE INVERSE POTENTIAL PROBLEM OF ELETTROCARDIOLOGY IN TERMS OF SOURCES

G. DI COLA, T. MORREA, M. PENNACCHIO

1. INTRODUCTION

The aim of this work is to solve the inverse problem of electrocardiology in terms of sources in a model of cardiac muscle i.e. to identify intracardiac sources from surface conductor potentials.

The solution of this problem is directly related to the localization of the foci of ectopic ventricular beats during surgery.

The oblique dipole layer model of the depolarization wavefront is assumed to describe correctly the potential field at a distance from cardiac sources [1].

A few instants after ectopic cardiac excitation the wavefront is limited to a small region surrounding the ectopic focus. In this case the potential field at a small distance from the wavefront can be approximated by a linear quadrupole centered at the ectopic focus. Without loss of generality we studied the identification of one dipolar source from potential data given on the cardiac muscle surface, because multipolar sources may be represented by superposition of single dipoles.

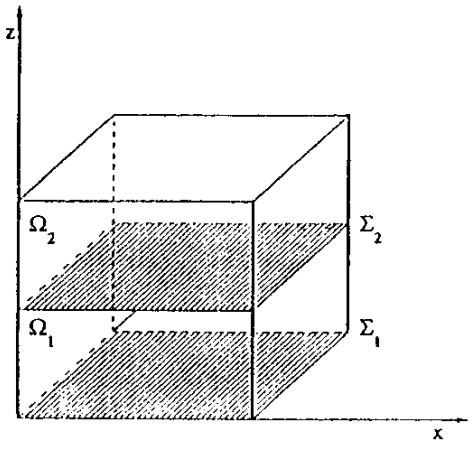
In our model we take into account the influence of cardiac muscle anisotropy due to fiber rotation on generated potential distributions.

The proposed method is being applied to the localization of ectopic intracardiac foci in animal experiments in view of future clinical use.

This work was supported by MURST through 40% grants

2. THE MATHEMATICAL FORMULATION OF THE INVERSE PROBLEM

Let Ω be an insulated conductor volume consisting of the regions Ω_2 homogeneous,representing blood and Ω_1, anisotropic, representing a portion of the myocardium as a set of superimposed layers of parallel fibers with fiber direction rotating from endocardium to epicardium.(fig.1).



$\Omega=\Omega_1 \cup \Omega_2$

$\Gamma=\partial\Omega$

Σ_1, Σ_2 epi- and endocardial surfaces with Σ_2 separating Ω_1 and Ω_2

T_K , k=1,2 conductivity tensor in Ω_K, with $T_2= \sigma I$

T conductivity tensor in Ω

$T=T_K$ in Ω_K

fig. 1

In order to evaluate a dipolar source g from the knowledge of the potential distribution u on the cardiac surface Σ (fig.1), we must solve the following *inverse problem* :

$$
(1)\qquad \begin{cases} -\nabla \cdot (T\nabla u) = g & \text{in} \\ u = z & \text{on } \Sigma=\Sigma_1 \cup \Sigma_2 \\ (T\nabla u)\cdot \underline{n} = 0 & \text{on } \Gamma \\ u_1 = u_2 & \text{on } \Sigma_2 \\ (T_1 \nabla u_1)\cdot \underline{n} = (T_2 \nabla u_2)\cdot \underline{n} & \text{on } \Sigma_2 \end{cases}
$$

$-\nabla\cdot T\nabla$ is unbounded, implying that the solution g does not depend continuously on the data z

The problem is *ill-posed*

We may formulate the problem (1) in the following way:

$$(2) \qquad \begin{cases} \mathcal{A}g = z \\ z = u(g)_{|\Sigma} \end{cases}$$

where $\mathcal{A}$ is called *transfer operator*, which turns out to be

linear, continuous, injective, with unbounded inverse.

Thanks to the properties of the transfer operator, it is possible to utilize the *Tikhonov's regularization method* (Tikhonov et al. 1976) to stabilize the problem, so that we can approximate the solution of the problem (2) with the solution of the following stable problem:

$$\boxed{\inf_{g \in U} \left\{ \| \mathcal{A}g - z \|_F^2 + \lambda \, \| g \|_U^2 \right\}}$$

where λ is called *regularization parameter*, $g \in U$, $z \in F$, U and F are normed spaces. $\forall\ \lambda > 0$, we call g_λ *regularized solution* of the problem. So that we find a class of regularized solutions

$$\left\{ g_\lambda \ : \ \lambda > 0 \right\}.$$

3. DISCRETIZATION AND REGULARIZATION OF THE PROBLEM

In order to find the solution, we assign some dipoles in Ω_1. Each dipole is obtained as a linear combination of three unitary dipoles, oriented along the direction of the orthogonal axes, so we search for a linear combination of them. Let n=3p the number of the unitary dipoles. In the applications z is measured on a finite set of locations P_K, (k=1,...,s) on the surface Σ. For the potential superposition principle, the following equations hold:

$$\sum_{i=1}^{n} \alpha_i U_i(P_k) = z_\delta(P_k) \qquad k=1,...,s$$

where $U_i(P_k)$ is the potential generated by the i-th unitary dipole in

P_k. We know this value by solving the forward problem for every unitary dipole;

$z_\delta(P_k)$ is the observed potential in P_k affected by error related to measurements

α_i is the unknown i-th unitary dipole moment.

In matrix form we have:

(3) $$A_h \underline{\alpha} = \underline{z}_\delta$$

where $A_h = \left(U_i(P_k) \right)_{i=1,\dots,n}^{k=1,\dots,s}$, is the computed transfer matrix with error h: $\|A_h - A\| = h < \infty$

$\underline{z}_\delta$ is an experimental determination of the exact potential $\underline{z}$ such that: $\|\underline{z} - \underline{z}_\delta\| < \delta$.

Since the problem (2) is ill-posed, the problem (4) is *ill-conditioned* ($C(A_h) \gg 1$). We stabilize it by means of Tikhonov's regularization method, so we look for the solution of the following stable problem:

$$\inf_{\underline{\alpha} \in \mathbb{R}^n} \left\{ \frac{1}{s} \| A_h \underline{\alpha} - \underline{z}_\delta \|^2 + \lambda \| \underline{\alpha} \|^2 \right\}$$

4. CRITERIA FOR THE SELECTION OF THE REGULARIZATION PARAMETER

In order to estimate the best value of λ we need some information on the solution.

Gabor and Nelson [3] gave some results on the estimation of the "heart vector" based on the integration of the potential over the bounding surface Γ of the isotropic and homogeneous conductor volume Ω. We have extended these results to the case of inhomogeneous anisotropic media. Let $\mathbf{i}$ be the vector of current density and $s = \nabla \cdot \mathbf{i}$ the source strength, whose sum over the whole body is nil; then the resultant vector dipole moment of such a source system is defined as

$$M = \int_{\Omega} rs \, dv$$

where **r** is the radius vector from an arbitrary origin and dv is the volume element. By Ohm's law, **i** = -A grad u, and from Green's lemma we obtain

$$M_x = \int_{\Gamma} F_x(\underline{\sigma},\vartheta)ud\Gamma \qquad M_y = \int_{\Gamma} F_y(\underline{\sigma},\vartheta)ud\Gamma \qquad M_z = \int_{\Gamma} F_z(\underline{\sigma})ud\Gamma$$

where $F_x(\underline{\sigma},\vartheta)$, $F_y(\underline{\sigma},\vartheta)$ and $F_z(\underline{\sigma})$ are related to the Ω anisotropy.

Then we have information on the moment of the heart vector which is equal to $\|g\|^2$ or in discrete terms equal to $\|\underline{\alpha}\|^2$.

If $\tilde{u}$ is a realization of the potential u on the surface Γ, it is possible to give an estimate of $\|\alpha\|^2$ by means of $\tilde{c}_1^2 = M_x^2 + M_y^2 + M_z^2$.

In order to select the regularization parameter λ (Morozov 1968) we use a criterion for the following auxiliary function

$$\gamma(\lambda) = \|\underline{\alpha}_\lambda\|^2$$

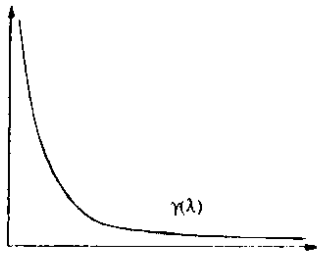


fig. 2

and consider the equation

$$\gamma(\lambda) = c^2$$

If c^2 is a value assumed by the function γ, the solution λ_c of the equation is unique. This method is equivalent to minimize $\|A_h\underline{\alpha}-z_\delta\|^2$ on the compact $\|\underline{\alpha}\|=c$.

5. MATHEMATICAL FORMULATION OF THE FORWARD PROBLEM

The *forward problem* consists in determining the potential u(x) in Ω given the dipolar source. Every dipole is simulated by two opposite electric charges placed at a small distance and the potential must satisfy the following conditions: continuity of the potential and flux across internal boundaries, zero flux on the boundary Γ.

The problem is described by the solution of the following Neumann

problem for the Poisson's equation (G. Di Cola et al 1990) and can be formulated weakly as follows:

find $u \in V$ so that:

$$a(u,v) = \langle g,v \rangle \qquad \forall v \in V$$

where:

$$a(u,v) = \int_\Omega (\nabla v)^T T \nabla u \, d\Omega \quad , \qquad \langle g,v \rangle = \int_\Omega gv \, d\Omega$$

$$H^1(\Omega) \subset V \subset L^2(\Omega)$$

$u = u_k$ in Ω_k $k=1,2$	
q_i	electric charges
x_i', x_i''	charge positions
$g = \sum_{i=1}^{n} q_i \left(\delta(x_i')-\delta(x_i'')\right)$	dipolar source with $\int_\Omega g \, d\Omega=0$
$S = \mathrm{diag}(\sigma_l,\sigma_t,\sigma_t)$	σ_l,σ_t are the conductivity coefficients along a direction respectively parallel and perpendicular to the fiber
$R = \begin{pmatrix} \cos(\vartheta) & -\mathrm{sen}(\vartheta) & 0 \\ \mathrm{sen}(\vartheta) & \cos(\vartheta) & 0 \\ 0 & 0 & 1 \end{pmatrix}$	fiber rotation matrix
$T_1 = RSR^T$	symmetric positive definite matrix related to the conductivity tensor in Ω_1
σ	conductivity in Ω_2
$T_2 = \sigma I$	conductivity matrix of region Ω_2
$T = T_k$ in Ω_k $k=1,2$	conductivity tensor.

Thanks to the Fredholm alternative theorem, the solution u(x) of the problem (1) exists and it is unique if:

$$\int_\Omega u(x) \, d\Omega = 0;$$

the solution $u(x) \in H(\Omega)$ with $H^1(\Omega) \subset H(\Omega) \subset L^2(\Omega)$ and depends continuously on g.

6. NUMERICAL SOLUTION OF THE FORWARD PROBLEM: THE FINITE ELEMENT METHOD

- $\Omega = \bigcup_{K \in \tau_h} K = K_1 \cup K_2 \cup \ldots \cup K_N$
- K parallelepipeds of τ_h
- $Q(k)=\{v \mid v \text{ trilinear on } K\}$
- $V_h=\{v_h \in C^0(\Omega): v_h \in Q(K), \forall K \in \tau_h\}$
- $v_h(x)= \sum_{j=1}^{M} \eta_i \phi_j(x) \qquad \eta_j(x)=v_h(N_j) \qquad x \in \Omega \cup \Gamma$
- N_j interior nodes of the grid
- $\phi_j \in V_h \qquad \phi_j(N_j)=\delta_{ij}=\begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} \quad i,j=1,\ldots,N$

The discrete problem consists in finding $u_h \in V_h$ such that:

$$a(u_h,v_h)=(f,v_h) \qquad \forall\ v_h \in V_h$$

7. THE STUDY OF A SAMPLE PROBLEM

We made the following simulation: we divided the cubic region Ω into the equal regions Ω_1 and Ω_2 with anisotropy coefficients:

$$\sigma=\sigma_x=3\sigma_y=3\sigma_z \qquad \text{and} \qquad \vartheta(z_2)=90°$$

z_2 corresponding to Σ_2

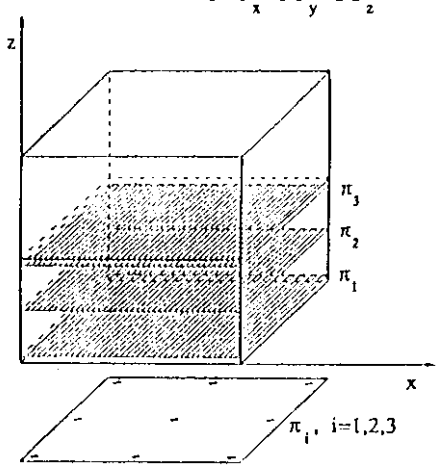


fig. 3

We fixed the possible active dipoles on three different planes in a uniform way as shown in fig.3 and solved the forward problem for every unitary dipole considered, to construct the transfer matrix.

In order to test the method, we studied a sample problem with data obtained by numerical simulation of the potentials generated by a single dipole. A noise of variance $\sigma^2 I$ was added to the surface data. Figures 4a and 5a show some results obtained without regularizing the problem, figures 4b and 5b show the correspondent regularized solutions when we assume $\lambda=\lambda_c$, which identify correctly the assigned dipole. The number of the observations is determined by the realistically available heart surface measurements (a few hundred).

We found the solution by means of a *singular value decomposition* of the matrix $A_h=UWV^T$ (U and V orthogonal matrix, $W=\mathrm{diag}(\sigma_1 \ldots \sigma_s)$) minimizing:

$$F(\underline{\alpha},\lambda)= \frac{1}{s} \quad UWV^T\underline{\alpha}-UU^T\underline{z}_\delta \quad ^2 + \lambda \quad \underline{V}^T\alpha \quad ^2$$

By introducing $\underline{y}=V^T\underline{\alpha}$ and $\underline{\tilde{b}}=U^T\underline{z}_\delta$: $F(\underline{\alpha},\lambda)= \frac{1}{s} \quad W\underline{y} - \underline{\tilde{b}} \quad ^2 + \lambda \quad \underline{y} \quad ^2$,

hence: $$y_\lambda^k = \tilde{b}_k \frac{\sigma_k}{\sigma_k^2+s\lambda} \qquad \text{and} \qquad \underline{\alpha}_\lambda = V\underline{y}_\lambda$$

ACKNOWLEDGMENTS

The research is carried out in cooperation between mathematicians and physiologists of the Department of Biology and General Fisiology, University of Parma, Dipartimento di Informatica e Sistemistica, Università di Pavia, C.V.R.T.I. University of Utah, Salt Lake City, Utah, U.S.A.

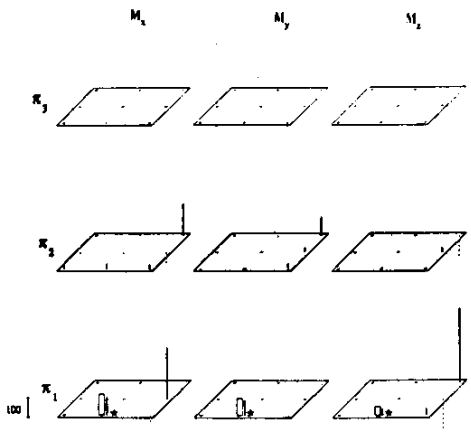


fig.4a

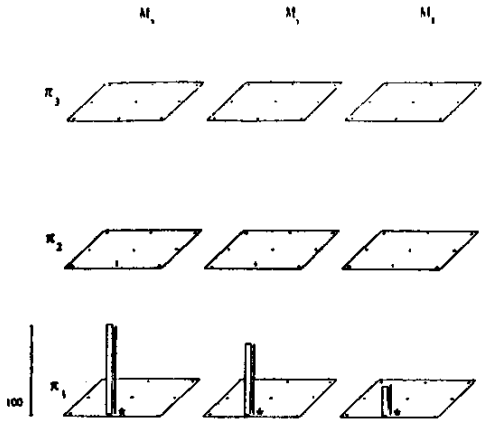


fig.4b

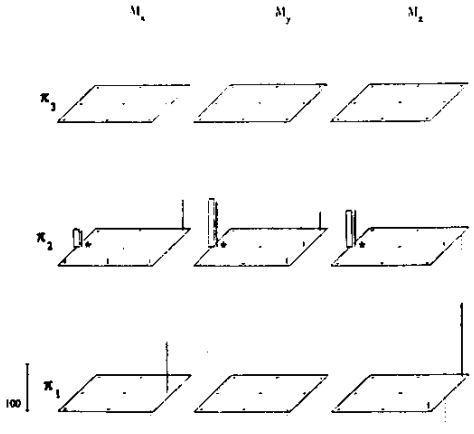


fig.5a

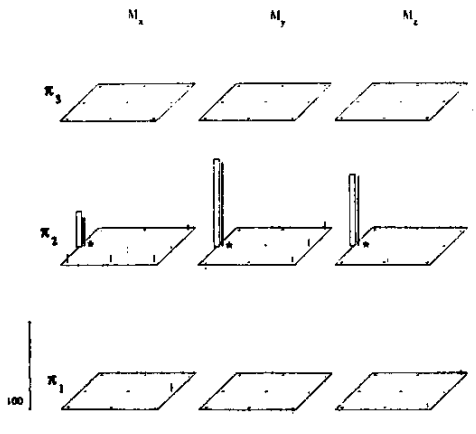


fig.5b

exact solution

computed solution

REFERENCES

[1] P.Colli-Franzone et al. (1982) Potential Fields Generated by Oblique Dipole Layers Modeling Excitation Wavefront in the Anisotropic Myocardium. Comparison with Potential Fields Elicited by Paced Dog Hearts in a Volume Conductor, *Circulation Research,* 51, 330-346

[2] G. Di Cola et al. (1990) Numerical simulation of electric fields generated by dipolar sources in tridimensional anisotropic non homogeneous media 17th International Congress on Electrocardiology Florence 26-29 Set. 1990

[3] D. Gabor, C.V. Nelson (1954). Determination of the resultant dipole of the heart from measurements on the body surface, *Journal of applied physics*,25, 4

[4] V.A. Morozov (1968). The error principle in the solution of operational equations by the regularization method. *Zh. vychisl. Mat. mat. Fiz.*, 8, 2, 295-309.

[5] A. Tikhonov, V. Arsenine (1976). *Méthodes de résolution de problemes mal posés,* Editions MIR-Moscow.

Giulio DI COLA, Tiziana MORREA, Micol PENNACCHIO

Dipartimento di Matematica
Università degli Studi di Parma
via M. D'Azeglio 85, 43100 Parma Italy

UN NUOVO ALGORITMO PER LA SOLUZIONE SIMBOLICA E NUMERICA DI UN' AMPIA CLASSE DI MODELLI FISICI

Marco FACCIO, Giuseppe FERRI

SOMMARIO.

Viene presentato un nuovo metodo che consente la risoluzione di un'ampia classe di modelli fisici attraverso la rapida e diretta determinazione delle grandezze caratteristiche di un particolare tipo di circuito elettrico, la rete a scala, a cui tali modelli vengono ricondotti. L'algoritmo viene illustrato per mezzo di concreti esempi simbolici e numerici.

INTRODUZIONE.

Lo studio del comportamento di alcuni modelli fisici connessi con la propagazione di un segnale in mezzi di varia natura puo' essere ricondotto a quello di un opportuno modello equivalente che si identifica con un particolare circuito elettrico chiamato " *rete a scala* ".

Alcuni esempi di tali modelli fisici sono: smorzatori meccanici, filtri elettrici, reti neuronali analogiche, amplificatori HF di tipo distribuito, linee elettriche di trasmissione dell'informazione, interconnessioni elettriche nei circuiti integrati, e cosi' via.

Nel presente lavoro viene mostrata, attraverso l'uso di alcuni significativi esempi, l'applicazione di un nuovo metodo per la soluzione di tali modelli. Esso consente la scrittura diretta delle grandezze che caratterizzano univocamente la rete a scala evitando il tradizionale ed oneroso utilizzo dei procedimenti di calcolo basati sui noti teoremi circuitali.

L'algoritmo proposto consente inoltre di ottenere le

espressioni delle grandezze sia dal punto di vista simbolico che numerico e di trattare in modo univoco reti di qualsiasi lunghezza (anche con un infinito numero di celle).

Tale metodo, sviluppato presso il Dipartimento di Ingegneria Elettrica dell'Universita' dell'Aquila, permette di esprimere le caratteristiche elettriche della rete per mezzo di un parametro (chiamato fattore di cella) caratterizzante completamente la singola cella e dei coefficienti di due nuovi triangoli numerici, chiamati DFF e DFFz, all'uopo costruiti [1]. Questi ultimi presentano anche interessanti proprieta' matematiche e stretti legami con i numeri di Fibonacci.

A conclusione vengono illustrate le espressioni di alcune grandezze elettriche sia nel caso di rete di lunghezza finita che semi-infinita con alcune applicazioni numeriche.

1. SMORZATORE MECCANICO.

A titolo di esempio si consideri un ricorrente problema legato a misure di spostamenti micro e nanometrici, quali ad esempio misure interferometriche o di correnti di tunnel. Per effettuare un tale tipo di misura e' necessario che i supporti della catena di misura siano immuni da vibrazioni in un ampio campo di frequenze connesso con la misura. E' necessario quindi ottenere opportuni smorzatori che filtrino le inevitabili vibrazioni ambientali di almeno 3 ordini di grandezza (circa 60 dB) in un ampio intorno del range di frequenza di misura. Tali smorzatori possono essere realizzati attraverso la sovrapposizione di piu' stadi ciascuno dei quali copre tutta la banda con una discreta attenuazione.

Il sistema risultante e' presentato in fig.1 :

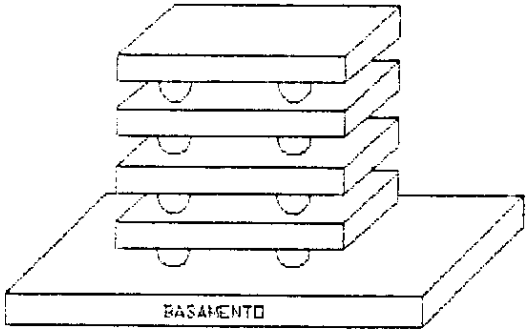


fig. 1

Il modello matematico del sistema meccanico e' espresso dall'equazione del moto armonico smorzato (con attrito) :

$$m\,x'' + b\,x' + K\,x = 0 \qquad (1.1)$$

Attraverso l'equivalenza meccanico-elettrica (tab.1) si ottiene il corrispondente modello circuitale di fig.2 :

molla	(1/K)	<=>	capacita'	C
massa	m	<=>	induttanza	L
attrito (cost.smorzam.)	b	<=>	resistenza	R
spostamento	x	<=>	carica	Q
velocita'	v	<=>	corrente	I

tab. 1

Lo studio della risposta in frequenza del sistema presuppone la determinazione della funzione di trasferimento. Tale determinazione risultera' certamente piu' agevole in ambito elettrico che in quello meccanico, ma, comunque, essa

risulta alquanto articolata, dato il numero di stadi ed il tipo di accoppiamento tra di loro.

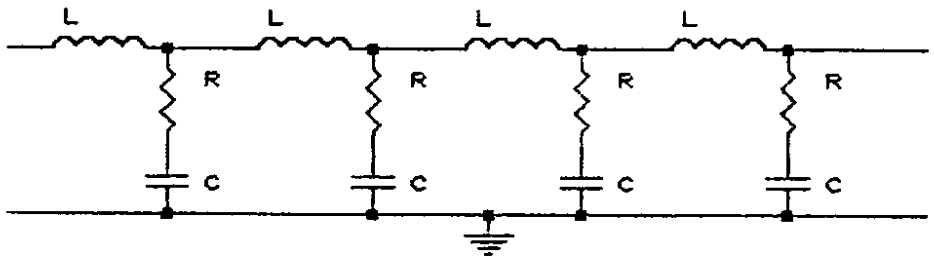


fig. 2

2. RETE NEURONALE ANALOGICA.

Un esempio significativo di elaborazione dell'informazione in modo distribuito con una rete neuronale analogica puo' essere quello che sottende ad un tipo di "visione artificiale" realizzato per mezzo della "retina siliconica".[2]

In fig. 3 e' mostrato un disegno approssimato della sezione verticale di una retina di primati. Si possono riconoscere alcuni blocchi funzionali strategici quali: sensore luminoso (fotorecettore R), alcuni elementi di collegamento orizzontali (H) e le celle gangliari (G) per la trasmissione ai livelli gerarchici superiori.

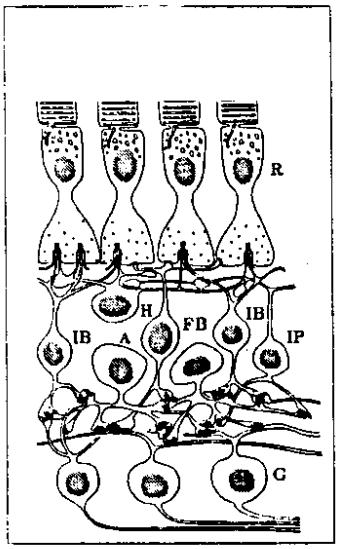


fig. 3

Una proposta di implementazione su silicio con tecnologia VLSI e' mostrata in fig. 4 dove lo strato di celle orizzontali (H) e' rappresentato dalla rete resistiva

con collegamento a celle di tipo esagonale e dove ogni nodo svolge la parte attiva di sensing dell'evento luminoso.

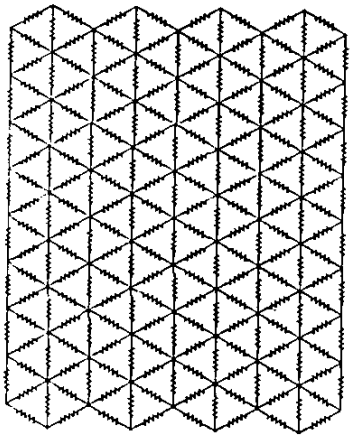


fig. 4

In fig.5 e' mostrato lo schema a blocchi del nodo attivo.
In esso possono riconoscersi un fototransistor (P), realizzato come in fig. 6a, ed alcuni elementi di elaborazione del segnale. Il blocco 1 rappresenta un inseguitore-integratore (capace di non caricare elettricamente la rete resistiva e di eseguire medie temporali), mentre il blocco 2 (fig.6b) rappresenta un amplificatore a transconduttanza il quale fornisce il corrispondente segnale in corrente che, elaborato dagli strati successivi, permettera' la ricostruzione dell'immagine.

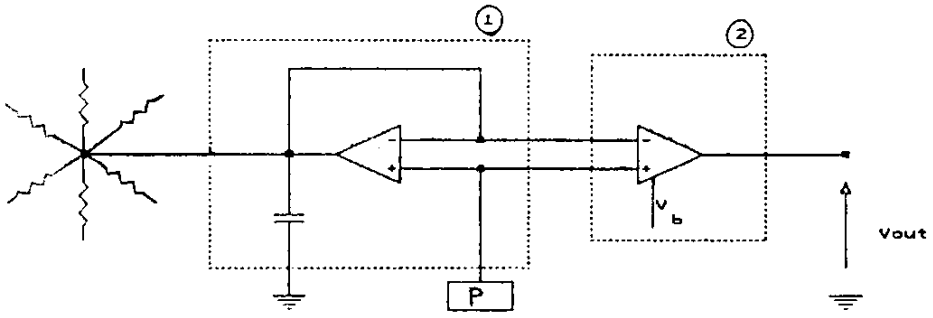
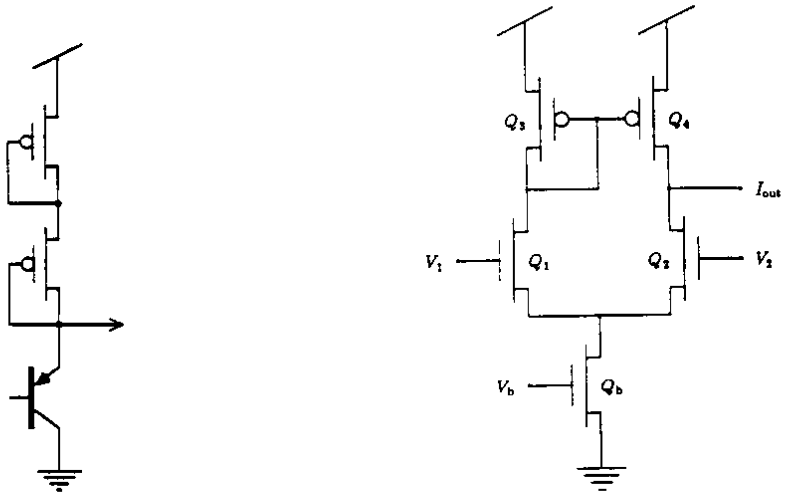


fig.5



a) b)

fig. 6

Nelle reti neuronali di tipo analogico le elaborazioni vengono distribuite nell'architettura del sistema (organizzata in strati gerarchici) e realizzate attraverso opportune soluzioni circuitali peculiari di ogni strato.

Tra le prime elaborazioni da svolgere, particolarmente significa_ tive sono le funzioni di media globale, di media locale e di *smoothing* della consistente massa di segnali disponibili ai fotorecettori.

La media globale su una data area viene ottenuta collegando tra loro le uscite degli inseguitori dei fotorecettori dell'area considerata.

La disposizione circuitale di principio e' mostrata in fig. 7.

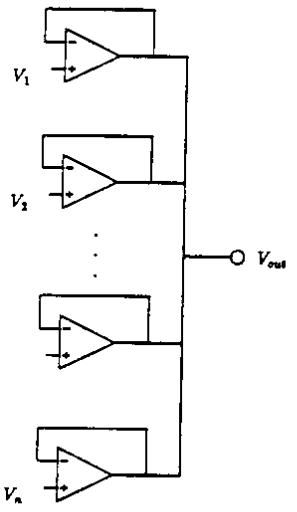


fig. 7

Considerando che ogni amplificatore lavori in zona lineare, l'uscita sara'

$$V_{out} = \frac{\sum_{1}^{n} G_i V_i}{\sum_{1}^{n} G_i} \qquad (2.1)$$

La transconduttanza G rappresenta il peso nella media e puo' essere regolato opportunamente dall'esterno per mezzo della V_b di fig.6b. La media locale (ossia il contributo degli n-1 nodi sul nodo n-esimo) viene ottenuta attraverso i collegamenti resistivi delle celle orizzontali a struttura esagonale (fig.4).

Data l'articolazione della rete orizzontale, l'impulso luminoso presente ad un nodo interessera' tutti gli altri nodi della rete e, considerata la configurazione elettrica di ogni nodo (fig. 5), l'impulso si propaghera' in una rete a scala di tipo resistivo (fig. 8) dove R e' la resistenza di collegamento internodo e G e' la conduttanza verso massa presente ad ogni nodo, coincidente con la transconduttanza dell'amplificatore.

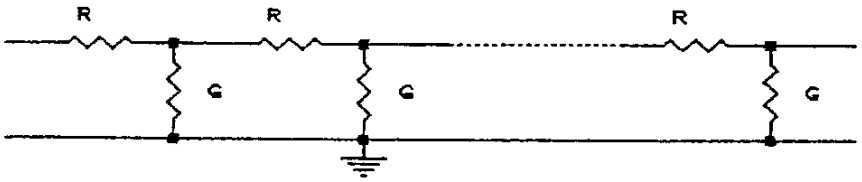


fig. 8

Dato l'elevato numero di nodi e la geometria chiusa della rete di interconnessione, la conseguente rete a scala

risultera' essere una rete discreta di lunghezza semi-infinita.

La propagazione del segnale su tale rete realizza il fenomeno di media locale e, contemporaneamente, quello dello *smoothing* in quanto il segnale si attenuera' lungo la rete con una legge quasi esponenziale.

La determinazione dei potenziali ai nodi e delle correnti nei rami in tale struttura presenta alcune difficolta' dovute alla lunghezza della rete e, di solito, ci si accontenta di soluzioni approssimate, valide sotto restrittive ipotesi di lavoro che impongono vincoli non sempre accettabili sui valori di R e di G.[2]

3. RETE A SCALA.

Le reti a scala (fig. 9) sono reti elettriche formate dalla ripetizione di una cella base costituita da un'impedenza in serie ed una in parallelo.

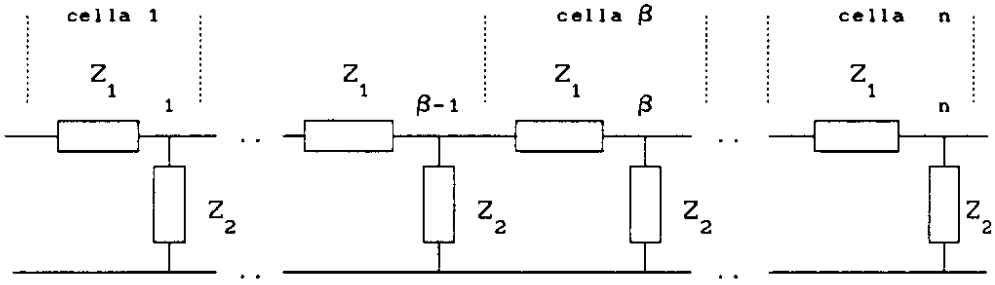


fig.9

In letteratura esistono diversi metodi "ad hoc" che consentono di calcolare le principali caratteristiche elettriche di una rete a scala (funzione di trasferimento, potenziale ai nodi, corrente nei rami, impedenza di ingresso e di uscita, modello equivalente di Thevenin, rapporto

segnale-rumore, matrici di quadripolo e di scattering,...), quali quelli basati sull'applicazione di noti teoremi circuitali, sulla risoluzione di equazioni alle differenze finite o sul metodo della matrice di quadripolo.

Tali metodi tuttavia utilizzano strutture ricorsive che ne limitano la capacita' di elaborazione simbolica o, dal punto di vista numerico, presentano alti tempi di risposta. Inoltre, con tali procedure, risulta pressoche' impossibile generalizzare le espressioni al caso di rete semi-infinita.

Una nuova ed esaustiva caratterizzazione di tale rete [1] consente la scrittura diretta di tali grandezze, noti gli elementi che formano la cella elementare ed il numero delle celle in cascata, attraverso la definizione del fattore di cella K(s) (dove s e' la variabile di Laplace) e l'applicazione di due nuovi triangoli numerici, chiamati DFF e DFFz.

Il metodo proposto consente di riportare ad un'unica trattazione la soluzione di tutte le possibili reti con struttura a scala, sia in base al tipo di elemento elettrico che costituisce la cella elementare (il quale puo' essere anche l'impedenza equivalente di una struttura piu' complessa) che in base alla struttura circuitale della cella stessa (ad L, a T, a Π). Inoltre esso fornisce la soluzione cercata sia dal punto di vista simbolico che numerico.

Le espressioni simboliche vengono ottenute con la scrittura immediata dei polinomi caratteristici in K(s) attraverso i coefficienti estratti da DFF e/o DFFz, mentre alla soluzione numerica si accede direttamente, nel caso di reti RR, LL e CC (dove K e' un numero reale), oppure attraverso il valore assunto, per ogni frequenza di interesse, dal modulo delle stesse caratteristiche elettriche.

Con il metodo proposto i tempi di calcolo per l'elaborazione numerica risultano drasticamente ridotti in quanto i parametri della rete vengono espressi in funzione di due sequenze numeriche di cui sono fornite le espressioni in forma ricorsiva e chiusa.

Quest'ultima consente l'estensione dell'analisi al caso di rete semi-infinita. Se K vale 1, infine, tali sequenze diventano numeri di Fibonacci.

3.1. ILLUSTRAZIONE DEL METODO.

Considerata la rete di fig.9 ed individuata la cella elementare di fig.10, si definisce il fattore di cella K(s) come :

$$K(s) = \frac{Z_1(s)}{Z_2(s)}.$$

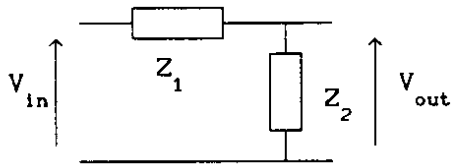


fig.10

Si considerino i seguenti due triangoli numerici :

i \ j	0	1	2	3	4
0	1				
1	1	1			
2	1	3	1		
3	1	6	5	1	
4	1	10	15	7	1
..	.....				

triangolo DFF

i \ j	0	1	2	3	4
0	0				
1	1	0			
2	2	1	0		
3	3	4	1	0	
4	4	10	6	1	0
..	.....				

triangolo DFFz

le cui regole di generazione sono le seguenti :

- triangolo DFF

$$b(i,j) = \begin{cases} 1 & \text{per } j = 0 \\ 0 & \text{per } j > i \\ b(i-1,j) + \sum_{j-1}^{i-1}{}_{k}\ b(k,j-1) & \text{per } j \leq i,\ j \neq 1 \end{cases} \qquad (3.1)$$

dove $b(i,j)$ rappresenta il termine generico.

- triangolo DFFz

$$c(i,j) = \begin{cases} i & \text{per } j = 0 \\ 0 & \text{per } j \geq i \\ c(i-1,j) + \sum_{j-1}^{i-1}{}_{k}\ c(k,j-1) & \text{per } j < i,\ j \neq 1 \end{cases} \qquad (3.2)$$

con $c(i,j)$ termine generico di DFFz.

Si osserva che la somma delle righe del triangolo DFF coincide con la sequenza dei numeri di Fibonacci di indice dispari [3] mentre quella del triangolo DFFz e' uguale alla sequenza dei numeri di Fibonacci di ordine pari [4].

E' possibile dimostrare [1] come tutte le caratteristiche elettriche di una rete a scala formata da n celle elementari accoppiate direttamente possano essere scritte applicando i triangoli DFF e DFFz ed il fattore di cella K(s).

Nel seguito si riportano le espressioni simboliche di alcune di esse, rimandando al lavoro [1] per quelle delle altre caratteristiche.

- Potenziale al generico nodo β :

$$V_\beta = \frac{\sum_{0}^{\alpha}{}_{j}\ b_j\ K^j\ (s)}{\sum_{0}^{n}{}_{i}\ b_i\ K^i\ (s)}\ V_{in} \qquad (3.3)$$

dove : $0 \le \beta \le n$; $\alpha = n - \beta$,

b_j = coefficienti della riga α del triangolo DFF,

b_i = coefficienti della riga n del triangolo DFF.

- Corrente nei rami :

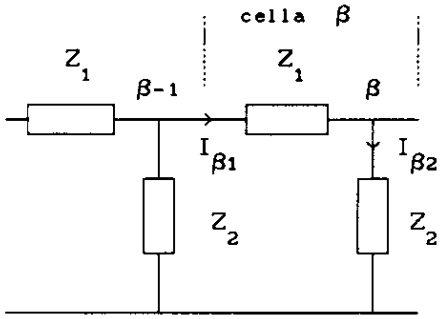


fig. 11

- Corrente che scorre nel ramo serie della cella β :

$$I_{\beta 1} = \frac{1}{Z_1} \frac{\sum_{0}^{\alpha+1} c_j K^{j+1}(s)}{\sum_{0}^{n} b_i K^i(s)} V_{in} \qquad (3.4)$$

dove : c_j = coefficienti della riga $\alpha+1$ del triangolo DFFz,

b_i = coefficienti della riga n del triangolo DFF.

- Corrente che scorre nel ramo parallelo della cella β :

$$I_{\beta 2} = \frac{1}{Z_2} \frac{\sum_{0}^{\alpha} b_j K^j(s)}{\sum_{0}^{n} b_i K^i(s)} V_{in} \qquad (3.5)$$

dove : b_j = coefficienti della riga α del triangolo DFF,

b_i = coefficienti della riga n del triangolo DFF.

- Impedenza di ingresso :

$$Z_{in} = Z_1 \frac{\sum_{0}^{n} b_i K^i(s)}{\sum_{0}^{n} c_i K^{i+1}(s)} = Z_2 \frac{\sum_{0}^{n} b_i K^i(s)}{\sum_{0}^{n} c_i K^i(s)} \qquad (3.6)$$

dove : b_i = coefficienti della riga *n* del triangolo DFF,
c_i = coefficienti della riga *n* del triangolo DFFz.

- Impedenza di uscita :

$$Z_o = Z_1 \frac{\sum_{0}^{n-1} b_j K^j(s)}{\sum_{0}^{n} c_i K^{i+1}(s)} = Z_2 \frac{\sum_{0}^{n-1} b_j K^j(s)}{\sum_{0}^{n} c_i K^i(s)} \qquad (3.7)$$

dove : b_j = coefficienti della riga *n-1* del triangolo DFF,
c_i = coefficienti della riga *n* del triangolo DFFz.

3.2. SOLUZIONE NUMERICA.

La soluzione numerica si ottiene dalle espressioni generali qualificando K come reale. Nel caso di rete formata da elementi dello stesso tipo, K e' un reale e quindi la soluzione e' immediata, mentre se K e' un numero complesso, ne viene considerato il modulo per ogni frequenza di lavoro.

I tempi di calcolo risultanti sono estremamente bassi in quanto tali caratteristiche possono essere immediatamente scritte in termini di opportune sequenze numeriche.

Tali sequenze vengono generate sia in forma ricorsiva, dai triangoli DFF e DFFz, che in forma chiusa, tramite una formula di Binet generalizzata, secondo quanto di seguito esposto.

Sia $\{ S_i(\kappa) \}$ la sequenza generata dalla somma della

i-esima riga del triangolo DFF sommando i coefficienti con peso K^i.

La forma ricorsiva del generico termine $S_i(K)$ appartenente a tale sequenza e':

$$S_i(K) = (2 + K)\, S_{i-1}(K) - S_{i-2}(K) \qquad (i \geq 2) \qquad (3.8)$$

con le condizioni iniziali :

$$S_0(K) = 1 \; ; \; S_1(K) = 1 + K \; .$$

La forma chiusa e' la seguente :

$$S_i(K) = \frac{(K + D)\, A^i - (K - D)\, B^i}{2D} \qquad (3.9)$$

dove : $D = \sqrt{(K + 2)^2 - 4}$;

$$A = \frac{(K + 2) + D}{2} \; ; \; B = \frac{(K + 2) - D}{2} \; ; \; A = 1 \,/\, B \; .$$

Sia inoltre $\left\{ T_i(K) \right\}$ la sequenza generata dalla somma della i-esima riga del triangolo DFFz sommando i coefficienti con peso K^i.

La forma ricorsiva per il generico termine $T_i(K)$ appartenente a tale sequenza e':

$$T_i(K) = (2 + K)\, T_{i-1}(K) - T_{i-2}(K) \qquad (3.10)$$

con $i \geq 2$ e dove $T_0(K) = 0$; $T_1(K) = 1$.

La forma chiusa per $T_i(K)$ vale :

$$T_i(K) = \frac{A^i - B^i}{D} \; .$$

Con tali formule, le espressioni delle caratteristiche elettriche della rete a scala diventano :

- Tensione nodo β :

$$V_\beta = \frac{\sum_0^\alpha{}_j \; b_j \, K^j}{\sum_0^n{}_i \; b_i \, K^i} \, Vin = \frac{S_\alpha(\kappa)}{S_n(\kappa)} \, Vin \qquad (3.11)$$

$$V_\beta = \frac{(K + D)\, A^{n-\beta} - (K - D)\, B^{n-\beta}}{(K + D)\, A^n - (K - D)\, B^n} \, Vin \qquad (3.12)$$

- Corrente ramo serie (cella β) :

$$I_{\beta 1} = \frac{1}{Z_1} \; \frac{\sum_0^{\alpha+1}{}_j \; c_j \, K^{j+1}}{\sum_0^n{}_i \; b_i \, K^i} \, Vin = \frac{1}{Z_1} \frac{S_{\alpha+1}(\kappa) - S_\alpha(\kappa)}{S_n(\kappa)} Vin =$$

$$= \frac{1}{Z_2} \frac{T_{\alpha+1}(\kappa)}{S_n(\kappa)} Vin \qquad (3.13)$$

$$I_{\beta 1} = \frac{1}{Z_2} \; \frac{2\,(\,A^{n-\beta+1} - B^{n-\beta+1})}{(K + D)\, A^n - (K - D)\, B^n} \, Vin \qquad (3.14)$$

- Corrente ramo parallelo (cella β) :

$$I_{\beta 2} = \frac{1}{Z_2} \; \frac{\sum_0^\alpha{}_j \; b_j \, K^j}{\sum_0^n{}_i \; b_i \, K^i} \, Vin = \frac{1}{Z_2} \; \frac{S_\alpha(\kappa)}{S_n(\kappa)} \, Vin \qquad (3.15)$$

$$I_{\beta 2} = \frac{1}{Z_2} \frac{(K + D) A^{n-\beta} - (K - D) B^{n-\beta}}{(K + D) A^{n} - (K - D) B^{n}} V_{in} \qquad (3.16)$$

Formule analoghe possono essere scritte per l'impedenza di ingresso e per quella di uscita.

Nel caso in cui K = 1 ($Z_1 = Z_2 = Z$), le sequenze $\{ S_i(1) \}$, $\{ T_i(1) \}$ diventano numeri di Fibonacci.

Infatti, $\{ S_i(1) \}$ coincide con la sottosequenza dei numeri di Fibonacci di indice dispari, mentre $\{ T_i(1) \}$ e' uguale a quella dei numeri di Fibonacci di indice pari. Cioe':

$$S_i(1) = F_{2i+1} \; ; \qquad T_i(1) = F_{2i} \qquad (3.17)$$

Le espressioni precedenti diventano :

- Potenziale al nodo β : $$V_\beta = V_{in} \frac{F_{2(n-\beta)+1}}{F_{2n+1}} \qquad (3.18)$$

- Corrente serie (cella β) : $$I_{\beta 1} = \frac{1}{Z} \frac{F_{2(n-\beta+1)}}{F_{2n+1}} V_{in} \qquad (3.19)$$

- Corrente parallelo (cella β) : $$I_{\beta 2} = \frac{1}{Z} \frac{F_{2(n-\beta)+1}}{F_{2n+1}} V_{in} \qquad (3.20)$$

- Impedenza di ingresso : $$Z_{in} = Z \frac{F_{2n+1}}{F_{2n}} \qquad (3.21)$$

- Impedenza di uscita :

$$Z_o = Z \frac{F_{2n-1}}{F_{2n}} \qquad (3.22)$$

Nel caso di rete semi-infinita (cioe' con infinito numero di celle), le precedenti espressioni si semplificano in virtu' del fatto che esse non dipendono dal numero delle celle ma solo dal fattore K. Si ha quindi :

- Potenziale al nodo β :

$$\lim_{n \longrightarrow \infty} V_\beta = \lim_{n \longrightarrow \infty} \frac{(K + D)\, A^n\, A^{-\beta} - (K - D)\, B^n\, B^{-\beta}}{(K + D)\, A^n - (K - D)\, B^n} V_{in} =$$

$$= A^{-\beta}\, V_{in}$$

cioe' :

$$V_\beta = V_{in} \left(1 + \frac{K}{2} + \frac{K}{2} \sqrt{1 + \frac{4}{K}} \right)^{-\beta} = \psi^{-\beta}\, V_{in} \qquad (3.23)$$

- Corrente serie (cella β):

$$I_{\beta 1} = V_{in} \frac{1}{Z_1} \left(\psi^{-\beta+1} - \psi^{-\beta} \right) \qquad (3.24)$$

- Corrente parallelo (cella β):

$$I_{\beta 2} = V_{in} \frac{1}{Z_2}\, \psi^{-\beta} \qquad (3.25)$$

- Impedenza di ingresso:

$$\lim_{n \longrightarrow \infty} Z_{in} = Z_2 \frac{(K + D)}{2} = Z_2 \left(A - 1 \right)$$

$$Z_{in} = Z_2 \left(\frac{K}{2} + \frac{1}{2} \sqrt{K^2 + 4K} \right) \qquad (3.26)$$

poiche' $B^n \longrightarrow 0$ dal momento che $B < 1$.

4. ESEMPI.

A conclusione vengono mostrati, a titolo di esempio del metodo presentato, le soluzioni dei problemi posti nei paragrafi 1 e 2 e che si sintetizzano nelle figure 2 e 8.

Per quanto concerne la rete di fig.2, lo smorzamento e' rappresentato dalla funzione di trasferimento del circuito, che puo' essere ricavata dalla (3.3) per n = 4 :

$$\frac{V_{out}}{V_{in}} = \frac{1}{1 + 10\,K + 15\,K^2 + 7\,K^3 + K^4} \qquad (4.1)$$

essendo

$$K = \frac{Z_1}{Z_2} = \frac{sL}{R + 1/sC} = \frac{s^2LC}{1 + sCR} . \qquad (4.2)$$

I valori desunti sperimentalmente per gli elementi di fig.2 sono : R = 10 Ω, L = 60 H, C = 1 mF. Essi garantiscono un'attenuazione minima per f > 1.6 Hz di 70 dB.

Relativamente alla rete di fig. 8, si ha K = RG. La distribuzione dei potenziali ai nodi e' espressa, per n finito, dalla (3.3), mentre nel caso semi-infinito, direttamente ed in modo non approssimato, dalla (3.22).

Nel caso di rete con n finito (es. n = 4) e K=1 (R=1/G), si ha la notevole relazione (3.18) basata sui numeri di Fibonacci. La distribuzione dei potenziali ai nodi sara' :

nodo	0	1	2	3	4
V_β/V_{in}	$\frac{F_9}{F_9} = 1$	$\frac{F_7}{F_9} = \frac{13}{34}$	$\frac{F_5}{F_9} = \frac{5}{34}$	$\frac{F_3}{F_9} = \frac{2}{34}$	$\frac{F_1}{F_9} = \frac{1}{34}$

Nel caso di rete semi-infinita, l'espressione della distribuzione del potenziale ai nodi e' data da:

$$V_\beta = V_{in} \left(1 + \frac{R\,G}{2} + \frac{R\,G}{2} \sqrt{1 + \frac{4}{R\,G}} \right)^{-\beta} \qquad (4.3)$$

e per K=1 si ha:

$$V_\beta = V_{in} \left(\frac{3 + \sqrt{5}}{2} \right)^{-\beta} = V_{in} \left(\frac{1 + \sqrt{5}}{2} \right)^{-2\beta} \qquad (4.4)$$

dove l'ultimo termine tra parentesi coincide con l'espressione del numero di Fibonacci data dalla formula di Binet per n che tende all'infinito.

RINGRAZIAMENTI.

Gli autori desiderano ringraziare il Dr. Piero Filipponi della Fondazione Bordoni, per le chiarificatrici discussioni sulla teoria dei numeri di Fibonacci e la loro storia, ed il Prof. Mauro Cerasoli per gli approfondimenti di analisi combinatoria.

Un particolare e grato ringraziamento va al Prof. Arnaldo D'Amico per il costante ed insostituibile sprone ad approfondire ogni argomento e per il sostegno continuo ed illuminato nell'articolazione dell'attivita' di ricerca.

BIBLIOGRAFIA.

[1] A. D'AMICO, M. FACCIO, G. FERRI , " *Ladder networks characterization and Fibonacci numbers* " , Il Nuovo Cimento, vol. 12 D, pp. 1165-1173, Agosto 1990.

[2] M.A.C. MAHER, S.P. DEWEERTH, M.A. MAHOWALD, C.A. MEAD, " *Implementing neural architectures using analog VLSI circuits* " , IEEE Trans. on Circuits & Systems, Vol.36 N.5, Maggio 1989.

[3] G. FERRI, M. FACCIO, A. D'AMICO , " *A new numerical triangle showing links with Fibonacci numbers* ", Accettato per la pubblicazione su The Fibonacci Quarterly.

[4] G. FERRI, M. FACCIO, A. D'AMICO , " *The DFFz triangle and its links with Fibonacci numbers* ", Accettato per la pubblicazione su The Fibonacci Quarterly.

Marco FACCIO, Giuseppe FERRI
Dipartimento di Ingegneria Elettrica
Universita' degli Studi di L'Aquila
67040 - Poggio di Roio - L'AQUILA

ADDITIONAL MONOTONICITY PROPERTIES AND INEQUALITIES FOR THE ZEROS OF BESSEL FUNCTIONS (*)

C. GIORDANO, Dip. di Matematica dell'Università, v. C. Alberto 10, 10123 Torino, Italy
A. LA FORGIA, Facoltà di Ingegneria, Monteluco di Roio, 67040 L'Aquila, Italy
L.G. RODONO', Dip. di Matematica ed Applicazioni, v. Archirafi 34, 90123 Palermo, Italy

1. INTRODUCTION.

For $\nu \geq 0$ we denote with $j_{\nu k}$ and $c_{\nu k}$ the k-th positive zeros of the Bessel functions $J_\nu(x)$ of the first kind and of the general cylinder function

$$C_\nu(x;\gamma) = C_\nu(x) =: \cos\gamma\, J_\nu(x) - \sin\gamma\, Y_\nu(x) \qquad 0 \leq \gamma < \pi$$

where $Y_\nu(x)$ is the Bessel function of the second kind.

In [2] A. Elbert and A. Laforgia introduced the notation $j_{\nu\kappa} = c_{\nu k}$ where $\kappa = k - \gamma/\pi$ and $k-1 < \kappa < k$. When $\kappa = k$ we get the zeros of the function $J_\nu(x)$. This notation has been used to prove several *monotonicity, concavity, convexity* properties of $j_{\nu\kappa}$ as a function of ν for κ fixed [2,3,4,5,9]. In particular we know that for $\nu \geq 0$ $j_{\nu\kappa}$ is concave for $\kappa \geq 0.344...$ and $j^2_{\nu\kappa}$ is convex for $\kappa \geq 0.7070...$ [5].

We observe that the study of the properties of $j_{\nu\kappa}$ was originated by the paper [19] of Putterman, Kac and Uhlenbeck. They have proposed a quantum mechanical explanation for the origin of the vortex lines produced in superfluid Helium when its container is rotated.

J.T. Lewis and M.E. Muldoon [16] have proved some monotonicity results using the Hellmann- Feynmann theorem of quantum chemistry [8,10,11] that here we recall

Hellmann-Feynmann Theorem. Let's be a pseudo inner product space with a pseudo-inner product $\langle \cdot, \cdot \rangle$. Let $\{H_\nu\}$ be a family of symmetric operators on an inner product space and for $\nu \in (a,b)$ let ψ_ν be an eigenvector (eigenfunction) of H_ν corresponding to an eigenvalue λ_ν.

(*) Work sponsored by Ministero dell' Università e della Ricerca Scientifica e Tecnologica of Italy.

Suppose that for $\mu\epsilon(a,b)$

$$\langle \Psi_\mu , \Psi_\nu \rangle \rightarrow \langle \Psi_\mu , \Psi_\mu \rangle \neq 0$$

as $\mu \rightarrow \nu$ and that

$$\lim_{\mu \rightarrow \nu} \langle \frac{H_\mu - H_\nu}{\mu - \nu} \Psi_\mu , \Psi_\nu \rangle$$

exists.
Moreover we define

$$\langle \frac{\partial H_\nu}{\partial \nu} \Psi_\mu , \Psi_\nu \rangle := \lim_{\mu \rightarrow \nu} \langle \frac{H_\mu - H_\nu}{\mu - \nu} \Psi_\mu , \Psi_\nu \rangle$$

Then

$$\frac{d \lambda_\nu}{d \nu} = \frac{\langle \frac{\partial H_\nu}{\partial \nu} \Psi_\mu , \Psi_\nu \rangle}{\langle \Psi_\mu , \Psi_\nu \rangle}$$

This version of the Theorem is taken from [11]. As a consequence of the Hellmann- Feynmann Theorem, Lewis and Muldoon [16] have proved that $j_{\nu 1}/\nu$ decreases with $\nu > 0$ and that $j^2_{\nu 1}/\nu$ increases with ν, $(3 \leqslant \nu < \infty)$. More recently C. Giordano and L.G. Rodonò proved that for $\nu > 0$ and $\kappa \geqslant 0.7070\ldots$ there exists a value ν_κ such that the function $j^2_{\nu\kappa}/\nu$ decreases on $(0,\nu_\kappa)$ and increases on (ν_κ,∞). Furthermore they showed that $j^2_{\nu\kappa}/\nu$ is convex on $(0,\nu_\kappa)$ [9]. We also recall that M.E.H. Ismail and M.E. Muldoon [12] proved the stronger monotonicity result that $j^2_{\nu 1}/(\nu+1)$ is an increasing function of ν on $(-1, \infty)$.

Further properties for $j_{\nu\kappa}$ and $c_{\nu k}$ can be established by the integral Watson formula [22, p. 508]

$$(1.1) \qquad \frac{d c_{\nu k}}{d \nu} = 2 c_{\nu k} \int_0^\infty K_0\left(2 c_{\nu k} \sinh t\right) e^{-2\nu t} dt , \qquad k=1,2 \ldots$$

where $K_0(x)$ is the standard modified Bessel function.

In section 2 we recall some montonicity properties recently established by A.Laforgia, M.E.Muldoon and L.Lorch and in section 3 we examine their application to the zeros of generalized Airy functions [6,15,17].

2. MONOTONICITY RESULTS.

In [17] L.Lorch has deduced the inequality

$$\frac{\nu}{c_{\nu k}} \cdot \frac{d c_{\nu k}}{d \nu} < 1 , \qquad \text{for} \quad c_{\nu k} \geqslant \nu > 0 ,$$

which implies that, for $k=1,2,\ldots$, $0<\nu<\infty$

$$\frac{\nu}{c_{\nu k}} \qquad \text{increases for} \quad c_{\nu k} \geqslant \nu \ .$$

When $c_{\nu k} > \nu + \pi/4$, there follows the stronger result that [14,17]

$$\frac{\nu+\frac{1}{2}}{c_{\nu k}} \qquad \text{increases for} \quad 0 \leqslant \nu < \infty \ .$$

More generally A.Laforgia and M.E.Muldoon in [14] have showed the inequality

$$c_{\nu k} > \nu + c_{0k} \ , \qquad 0 < \nu < \infty \ , \ k = 2,3,\ldots$$

and

$$(2.1) \qquad \left(\nu + a_k\right) \frac{d\, c_{\nu k}}{d\nu} < c_{\nu k} \ , \quad 0 < \nu < \infty \ , \quad k = 2,3,\ldots$$

where the positive number a_k is defined by

$$a_k^{-1} = c_{0k}^{-1} \left[\frac{d\, c_{\nu k}}{d\nu} \right]_{\nu = 0}$$

The previous results hold also in the case $k=1$, but $0 \leqslant \gamma \leqslant \pi/2$.

Inequality (2.1) can be employed to yield the stronger monotonicity result

$$\frac{\nu + a_k}{c_{\nu k}} \qquad \text{increases for} \quad \nu > 0 \quad \text{when} \quad k = 2,3,\ldots$$

If $0 \leqslant \gamma \leqslant \pi/2$ this monotonicity holds also for $k=1$.

Differentiating we get

$$c_{\nu k} \left(\frac{\nu + a_k}{c_{\nu k}} \right)' = 1 - \frac{\nu + a_k}{c_{\nu k}} \frac{dc_{\nu k}}{d\nu} > 0$$

The previous results can be employed to deduce other monotonicity results as the following ones recently found by A.Laforgia, M.E.Muldoon and L.Lorch [15,17].

We recall now some of them.

If $\nu > 0$ the function

$$\left(\frac{c_{\nu k}}{\nu} \right)^{\nu} \qquad \text{increases for } c_{\nu k} \geqslant \nu + \frac{\pi}{4}$$

and

$$\left(\frac{c_{\nu k}}{2\nu} \right)^{2\nu} \qquad \text{decreases for } 0 < \nu \leqslant c_{\nu k} \leqslant 2\nu$$

The proof is based on the properties [17] of the function

$$\delta_{\nu} = \frac{d}{d\nu} \left\{ \ln \left[\frac{c_{\nu k}}{\beta \nu} \right]^{\nu} \right\} , \qquad \nu > 0 \ , \ \beta > 0$$

in the case β-1, β-2, respectively.

The monotonicity of δ_ν implies that if $\delta_\nu \geqslant 0$, then $[c_{\nu k}/(2\nu)]^{2\nu}$ increases for $0<\nu\leqslant \mu$, while $\delta_\nu \leqslant 0$ implies that this function of ν decreases for all $\nu> \mu$, provided of course that $c_{\nu k}\geqslant \nu+\pi/4$.

The hypothesis that $c_{\nu k}> \nu+\pi/4$ is satisfied when k-2,3,... and for any $c_{\nu 1}$ for which $0\leqslant \gamma\leqslant \pi/2$; in particular for $c_{\nu 1}$- $j_{\nu 1}$.

In the case k-1, $c_{\nu k}$- $j_{\nu 1}$ there exists a value ν - ν_1 such that

(2.2) $$\left(\frac{j_{\nu 1}}{2\nu}\right)^{2\nu} \qquad \text{increases,} \quad 0<\nu \leqslant \nu_1$$

and

(2.3) $$\left(\frac{j_{\nu 1}}{2\nu}\right)^{2\nu} \qquad \text{decreases,} \quad \nu \geqslant \nu_1$$

M.E.Muldoon has calculated some values of $[j_{\nu 1}/(2\nu)]^{2\nu}$ near 1 from which it derives that $1.003 < \nu_1 < 1.006$.

If the order ν is kept constant, but k or γ varies in $c_{\nu k}(\gamma)$ it could be convenient to use the notation introduced by Á.Elbert and A.Laforgia in [2]. They show that $j_{\nu K}$ increases with $k>0$, for fixed ν.

L.Lorch in [17] extends (2.2), (2.3) by implying the existence of unique ν_K such that

(2.4) $$\left(\frac{j_{\nu K}}{2\nu}\right)^{2\nu}$$ increases for $0<\nu \leqslant \nu_K$ and decreases for $\nu \geqslant \nu_K$,

where ν_K is an increasing function of $\kappa>0$.

The behaviour of $[c_{\nu k}/(2\nu)]^{2\nu}$ suggests to consider the character of $[c_{\nu k}/(2\nu)]^{2\nu}/\nu$ [15,17].

First of all A.Laforgia and M.E.Muldoon in [15] have obtained that for some $0<\nu_0\leqslant 1/2$ and some k- 1,2,..., under some hypothesis, the function $[c_{\nu k}/(2\nu)]^{2\nu}/\nu$ decreases as ν increases for $0<\nu\leqslant \nu_0$.

Recently L.Lorch has established in [17] that $[c_{\nu 1}/(2\nu)]^{2\nu}/\nu$ decreases for $0<\nu< \infty$ if $c_{\nu 1}\geqslant \nu$. The proof is divided into four parts. In each subinterval $\lambda\leqslant\nu\leqslant \mu$ he has proved that the expression

$$\frac{1}{2}\,\frac{d}{d\nu}\left\{\ln\left[\left(\frac{c_{\nu 1}}{2\nu}\right)^{2\nu}\cdot\frac{1}{\nu}\right]\right\}$$

is negative , using previous results in [15, 17] .

3. APPLICATION TO THE ZEROS OF GENERALIZED AIRY FUNCTIONS.

The generalized Airy functions are the solutions on $0 \leqslant x < \infty$ of the equation

$$y'' + x^{\alpha} y = 0 \tag{3.1}$$

where α is a positive number. The Airy functions correspond to the case $\alpha = 1$.

The case $\alpha \neq 1$ arises in a fundamental way in the asymptotic solution of certain kinds of differential equations [18,20,21] . A function of this kind also occurs in work concerned with weighted averages of a function at a jump discontinuity [1].

Equation (3.1) is connected to the Bessel equation

$$t^2 \frac{d^2u}{dt^2} + t \frac{du}{dt} + \left(t^2 - \nu^2\right) u = 0$$

by means of the transformations

$$y(x) = x^{\nu/2}\, u(t) \qquad t = 2\nu x^{1/(2\nu)} \tag{3.2}$$

where $\nu = 1/(\alpha+2)$.

So it is possible to use many known results about monotonicity with respect to order of zeros of Bessel functions. We denote $a_{\alpha k}$ the k-th positive zero of a solution of (3.1).

Because of (3.2), it is clear that

$$a_{\alpha k} = \left(\frac{c_{\nu k}}{2\nu}\right)^{2\nu} \qquad \text{where } 0 < \nu = \frac{1}{\alpha + 2} \tag{3.3}$$

Using the connection (3.3) between zeros of cylinder functions and generalized Airy functions, it is possible to obtain results for $a_{\alpha k}$. As Corollary to Theorem 2.1 in [14] there follows that for each k=2,3,... $a_{\alpha k}$ decreases to 1 as α increases , $0 < \alpha < \infty$. If y(0) =0 this decrease holds also for $a_{\alpha 1}$.

Muldoon's bounds for ν_1 permit to assert (2.2) and (2.3) with $1.003 < \nu_1 < 1.006$.

Consequently

$$\begin{aligned} &a_{\alpha 1} \quad \text{increases} \quad -1.00596 < -2 + 1/\nu_1 \leqslant \alpha < \infty \\ &a_{\alpha 1} \quad \text{decreases} \quad -2 < \alpha \leqslant -2 + 1/\nu_1 < -1.00299 \end{aligned} \tag{3.4}$$

Taking into account (2.4), the result (3.4) can be extended to

$$\begin{aligned} &a_{\alpha k} \quad \text{increases} \quad -2 + 1/\nu_k \leqslant \alpha < \infty \\ &a_{\alpha k} \quad \text{decreases} \quad \alpha < -2 + 1/\nu_k \end{aligned} \tag{3.5}$$

This implies that

$$(\alpha+2)\, a_{\alpha,1} \quad \text{increases}, \quad -1.0029 < \alpha < \infty$$

and more generally

$$(\alpha+2)\, a_{\alpha,k} \quad \text{increases}, \quad -2+1/\nu_k < \alpha < \infty$$

for $k = 1,2,\ldots$.

In [6] Á. Elbert and A. Laforgia continued the investigation about the behaviour of $a_{\alpha,k}$ and they proved the following result

Theorem. For each fixed $k = 2,3,\ldots$, let $a_{\alpha,k}$ be the k-th positive zero of (3.1). Then for $\alpha > 0$, the function $\log a_{\alpha,k}$ is convex .

The first important consequence of this theorem is the convexity of $a_{\alpha,k}$, since a positive log-convex function is also convex.

The proof is based on some known properties of $c_{\nu,k}$ and the Watson's formula (1.1).

References

[1] C. Comstock, On weighted averages at a jump discontinuity, Quart. Appl. Mat., 28 (1970), 159-166 .

[2] Á. Elbert and A. Laforgia, On the square of the zeros of Bessel functions, SIAM J. Mat. Anal., 15 (1984), 206- 212.

[3] Á. Elbert and A. Laforgia, Further results on the zeros of Bessel functions, Analysis, 5 (1985), 71-86 .

[4] Á. Elbert and A. Laforgia, On the power of the zeros of Bessel functions, Rend. Circ. Math. Palermo, serie II, Tomo XXXVI (1987), 507-517.

[5] Á. Elbert and A. Laforgia, Some consequences of a lower bound for the second derivative of the zeros of Bessel functions, J. Math. Anal. Appl., (1987), 1-5.

[6] Á. Elbert and A. Laforgia, On the zeros of generalized Airy functions, J. Appl. Math. Phys. (ZAMP) , 42 (1991) 521-526 .

[7] Á. Elbert, A. Laforgia and L. Lorch, Additional monotonicity properties of zeros of Bessel functions, Analysis and Inter. J. (to appear).

[8] R.P. Feynmann, Forces in molecules, Phys. Rev., 56 (1939), 340-343.

[9] C. Giordano and L.G. Rodonò, Further monotonicity and convexity properties of the zeros of cylinder functions, J. Comp. and Appl. Math. (to appear).

[10] H. Hellman, Einfuhrung in die Quantenchemie, Denticke, Vienna 1917.

[11] M.E.H. Ismail and A. Laforgia, Zeros of classical orthogonal polynomials and Bessel functions, Lect. Notes in Math. (to appear).

[12] M.E.H. Ismail and M.E. Muldoon, On the variation with respect to a parameter of zeros of Bessel and q-Bessel functions, J. Math. Anal. and Appl. Vol.135, n.1 (1988).
[13] A. Laforgia and M.E. Muldoon, Inequalities and approximations for zeros of Bessel functions of small order. SIAM J. Mat. Anal., 14 (1983), 383-388.
[14] A. Laforgia and M.E. Muldoon, Monotonicity and concavity properties of zeros of Bessel functions, J. Mat. Anal. Appl., 98 (1984), 470-477.
[15] A. Laforgia and M.E. Muldoon, Monotonicity properties of zeros of generalized Airy functions, J. Appl. Mat. Phys. (ZAMP), 39 (1988), 267-271.
[16] J.T. Lewis and M.E. Muldoon, Monotonicity and convexity of properties of zeros of Bessel functions, SIAM J. Mat. Anal., 8 (1977), 171-178.
[17] L. Lorch, Some monotonicity properties associated with the zeros of Bessel functions, Arch. Mat. (Brno), 26 (1990), 137-146.
[18] L.N. Nosova and S.A. Tumarkin, Tables of generalized Airy functions for the asymptotic solution of the differential equations..., translated by D.E. Brown, Pergamon-McMillan, 1965 (Russian original, Computing Centre of the Academy of Sciences of the USSR, Moscow, (1961).
[19] S.J. Putterman, M. Kac, and G.E. Uhlenbeck, Possible origine of the quantized vortices in He, II, Phys. Rev. Lett. 29 (1972), 546-549.
[20] A.D. Smirnov, Tables of Airy functions and special confluent ypergeometric functions for asymptotic solutions of differential equations of the second order, translated from the Russian by D.G. Fry, Pergamon, Oxford 1960.
[21] C. A. Swanson and V.B. Headley, An extension of Airy's equation, SIAM J. Appl. Mat., 15 (1967), 1400-1412.
[22] G.N. Watson, A treatise on the theory of Bessel functions, 2nd ed., Cambridge University Press (1944)

INEQUALITIES FOR SOME SPECIAL FUNCTIONS (*)

Andrea LA FORGIA
Facoltà di Ingegneria, Monteluco di Roio - 67040 L'Aquila - Italy

Abstract

The present paper is expository. We give a survey of some simple methods for finding bounds for some special functions. Our results are based on the recurrence relations satisfied by these functions and on some integral representations.

AMS (MOS) Subject classification : Primary 33 C 25 ;

Keywords : Bessel functions ; recurrence relations.

(*) Work sponsored by Ministero della Pubblica Istruzione of Italy.

1. **Introduction** . The purpose of the present paper is to describe some elementary methods which can be used to establish inequalities for some special functions. In particular we derive inequalities for the modified Bessel functions $I_\nu(t)$ and $K_\nu(t)$. The bounds are obtained directly from the recurrence relations satisfied by these functions.

Furthermore we describe a method for getting a simple inequality for elliptic integrals K and E of the first and the second kind , respectively.

2. **Bounds for modified Bessel functions** . Let $I_\nu(t)$ and $K_\nu(t)$ be the modified Bessel functions of the first and the third kind, respectively. Several authors studied inequalities for these functions. For example Bordelon [2] and Ross [8] proved the following result

$$e^{x-y}\left(\frac{x}{y}\right)^\nu < \frac{I_\nu(x)}{I_\nu(y)} < e^{y-x}\left(\frac{x}{y}\right)^\nu , \quad \nu > 0 , 0 < x < y$$

and Ifantis and Siafarikas [3] established the upper bound

$$\frac{I_\nu(x)}{I_\nu(y)} < \left(\frac{x}{y}\right)^\nu \left(\frac{x^2 + j_{\nu 1}^2}{y^2 + j_{\nu 1}^2}\right)^{\frac{j_{\nu 1}^2}{4(\nu+1)}} , \quad \nu > -1 , \quad 0 < x < y$$

where $j_{\nu 1}$ denotes the first positive zero of the Bessel function $J_\nu(x)$ of the first kind.

It is possible to derive inequalities for the function $I_\nu(t)$ using the recurrence relations [10, p. 79]

$$t\, I'_\nu(t) - \nu\, I_\nu(t) = t\, I_{\nu+1}(t) \tag{2.1}$$

and

(2.2) $$t\ I'_\nu(t) + \nu\ I_\nu(t) = t\ I_{\nu-1}(t) \quad .$$

Similarly for the function $K_\nu(t)$ we need the recurrence relations [10, p. 79]

(2.3) $$t\ K'_\nu(t) + \nu\ K_\nu(t) = -\ t\ K_{\nu-1}(t)$$

and

(2.4) $$t\ K'_\nu(t) - \nu\ K_\nu(t) = -\ t\ K_{\nu+1}(t) \quad .$$

Theorem 2.1 . For real ν let $I_\nu(t)$ be the modified Bessel function of the first kind . Then the following inequalities

$$\frac{I_\nu(x)}{I_\nu(y)} > e^{x-y}\left(\frac{x}{y}\right)^\nu \quad , \qquad \nu \geq -\frac{1}{2} \ , \ 0 < x < y$$

$$\frac{I_\nu(x)}{I_\nu(y)} < e^{x-y}\left(\frac{y}{x}\right)^\nu \quad , \qquad \nu \geq \frac{1}{2} \ , \ 0 < x < y$$

hold .

Proof . We use the inequality

$$I_\nu(t) > I_{\nu+1}(t) \quad , \qquad \nu \geq -\frac{1}{2}$$

established by Näsell [7] and Soni [9] in the recurrence relation (2.1). So we find

$$\frac{I'_\nu(t)}{I_\nu(t)} < 1 + \frac{\nu}{t} \quad , \qquad \nu \geq -\frac{1}{2} \ .$$

Integrating between x and y we obtain

$$\frac{I_\nu(x)}{I_\nu(y)} > e^{x-y}\left(\frac{x}{y}\right)^\nu \quad , \qquad \nu \geq -\frac{1}{2}$$

which is the desired lower bound.

The proof of the upper bound is similar. Now we refer to the recurrence relation (2.2) and we get

$$\frac{I'_\nu(t)}{I_\nu(t)} > 1 - \frac{\nu}{t} \quad , \qquad \nu \geq \frac{1}{2} \ .$$

Integrating on [x,y] we find

$$\frac{I_\nu(x)}{I_\nu(y)} < e^{x-y}\left(\frac{y}{x}\right)^\nu \quad , \qquad \nu \geq \frac{1}{2} \ .$$

This completes the proof of the theorem.

Theorem 2.2 . For real ν let $K_\nu(t)$ be the modified Bessel function of the third kind. Then the following inequalities

$$\frac{K_\nu(x)}{K_\nu(y)} < e^{y-x}\left(\frac{y}{x}\right)^\nu \quad , \qquad \nu > \frac{1}{2} \quad , \quad 0 < x < y$$

and

$$\frac{K_\nu(x)}{K_\nu(y)} > e^{y-x}\left(\frac{y}{x}\right)^\nu \quad , \qquad 0 < \nu < \frac{1}{2} \quad , \quad 0 < x < y$$

hold.

Proof . The proof is similar to the one of theorem 2.1 . We use the inequality [9]

$$K_{\nu+1}(t) > K_{\nu}(t) \quad , \qquad \nu > -\frac{1}{2}$$

in (2.3) obtaining

$$\frac{K'_{\nu}(t)}{K_{\nu}(t)} = -\frac{\nu}{t} - \frac{K_{\nu-1}(t)}{K_{\nu}(t)} > -\frac{\nu}{t} - 1 \quad , \qquad \nu > \frac{1}{2} \quad ,$$

and integrate on [x,y] , leading to the desired upper bound. The starting point for the proof of the lower bound is the recurrence relation (2.4) and the inequality

$$K_{\nu-1}(t) > K_{\nu}(t) \quad , \qquad 0 < \nu < \frac{1}{2} \quad .$$

This inequality can be checked in several ways. For example it follows immediately from the integral formula [10, p. 181]

$$K_{\nu}(t) = \int_0^{\infty} e^{-t\cosh z} \cosh(\nu z)\, dz \quad .$$

From (2.4) we get

$$\frac{K'_{\nu}(t)}{K_{\nu}(t)} < -\frac{\nu}{t} - 1 \quad , \qquad 0 < \nu < \frac{1}{2}$$

and integrating on [x,y] the desired lower bound follows.

Inequalities similar to ones of theorems 2.1 and 2.2 can be established

for the Bessel functions.

While the proofs of theorems 2.1 and 2.2 are based on the recurrence relations, on the other hand it is possible to establish other more complicated inequalities from the differential equations satisfied by these functions. We recall here a result established by the author and M.L. Mathis [5].

Theorem 2.3. For $\nu > -1/2$ let $I_\nu(t)$ the modified Bessel function of the first kind. Then the following inequalities

$$\frac{I_\nu(x)}{I_\nu(y)} > \left(\frac{x}{y}\right)^{\frac{1}{4}+\nu^2} \exp\left\{\frac{x^2-y^2}{6}\right\}, \qquad 0 < x < y,$$

$$\frac{I_\nu(x)}{I_\nu(y)} < \left(\frac{x}{y}\right)^{\frac{3}{16}+\frac{3}{2}\nu^2-\nu^4} \exp\left\{\frac{x^2-y^2}{36}\left(7-4\nu^2-\frac{x^2+y^2}{5}\right)\right\}$$

$$\text{for } |\nu^2-1/4|^{1/2} \le x < y < t_1,$$

$$\frac{I_\nu(x)}{I_\nu(y)} < \left(\frac{x}{y}\right)^{\frac{1}{2}}, \qquad 0 < x < \nu, \quad \text{for} \quad \nu \ge \frac{1}{2},$$

holds, where t_1 is the first zero of

$$16\,t^4 - 40\,(7-4\nu^2)\,t^2 + 45\,(16\nu^4 - 24\nu^2 - 11).$$

3. **Inequalities for elliptic integrals**. We prove the following result.

Theorem 3.1. For $0 < \kappa < 1$ let

$$E = \int_0^{\frac{\pi}{2}} \left(1 - \kappa^2 \sin^2 t\right)^{\frac{1}{2}} dt \tag{3.1}$$

and

$$K = \int_0^{\frac{\pi}{2}} \left(1 - \kappa^2 \sin^2 t\right)^{-\frac{1}{2}} dt \tag{3.2}$$

be the elliptic integrals of the second and the first kind, respectively. Then

$$2\kappa^2 < (1+\kappa^2)E - (1-\kappa^2)K < \frac{3}{4}\pi\kappa^2 .$$

Proof. Let us consider the function

$$f(\kappa) = \frac{(1+\kappa^2)E - (1-\kappa^2)K}{\kappa^2} . \tag{3.3}$$

We have to prove that

$$2 < f(\kappa) < \frac{3}{4}\pi .$$

Taking into account (3.1) and (3.2) the function $f(\kappa)$ can be written in the form

$$f(\kappa) = \frac{1}{\kappa^2} \int_0^{\frac{\pi}{2}} \left\{ (1+\kappa^2)\left(1-\kappa^2 \sin^2 t\right)^{\frac{1}{2}} - (1-\kappa^2)\left(1-\kappa^2 \sin^2 t\right)^{-\frac{1}{2}} \right\} dt =$$

$$= \int_0^{\frac{\pi}{2}} \frac{\cos^2 t + 1 - \kappa^2 \sin^2 t}{\left(1-\kappa^2 \sin^2 t\right)^{\frac{1}{2}}} \, dt$$

Let $g(\kappa)$ the integrand function. Differentiating with respect to κ we get

$$g'(\kappa) = \frac{\sin^2 t \cdot \kappa}{\left(1-\kappa^2 \sin^2 t\right)^{\frac{3}{2}}} \left[\cos^2 t - \left(1-\kappa^2 \sin^2 t\right)\right] .$$

The term in the brackets is clearly negative, therefore $g(\kappa)$ decreases and also $f(\kappa)$ decreases. To prove the desired bounds we need only to compute $f(0)$ and $f(1)$. We find

$$f(0) = \lim_{\kappa \to 0^+} \frac{(1+\kappa^2)E - (1-\kappa^2)K}{\kappa^2} = \lim_{\kappa \to 0^+} \left[\frac{E - (1-\kappa^2)K}{\kappa^2} + E(0)\right] =$$

$$= \lim_{\kappa \to 0^+} \int_0^{\frac{\pi}{2}} \left(1 - \kappa^2 \sin^2 t\right)^{-\frac{1}{2}} \cos^2 t \, dt + \frac{\pi}{2} =$$

$$= \int_0^{\frac{\pi}{2}} \cos^2 t \, dt + \frac{\pi}{2} = \frac{\pi}{4} + \frac{\pi}{2} = \frac{3}{4}\pi \ .$$

For the value $f(1)$ we have

$$f(1) = 2\,E(1) = 2 \ .$$

This gives

$$2 < f(\kappa) < \frac{3}{4}\pi$$

leading to the desired result . Further inequalities for elliptic integrals will be established in a forthcoming paper with S. Sismondi [6] .

References

1. M. Abramowitz and I. A. Stegun , Handbook of Mathematical Functions , Applied Math. Series , 55, National Bureau of Standards , Washington D.C. , 1964 .

2. D.J. Bordelon , Solution to problem 72-15 , SIAM Rew. , 15 (1973) , 666-668 .

3. E. K. Ifantis and P. D. Siafarikas, Bounds for modified Bessel functions, Rend. Circolo Mat. Palermo, (to appear).

4. A. Laforgia, Bounds for modified Bessel functios, J. Comp. Appl. Math. 1991 (to appear).

5. A. Laforgia and M.L. Mathis, Bounds for Bessel functions, Rend. Circolo Mat. Palermo, XXXVIII (1989) 319-328.

6. A. Laforgia and S. Sismondi, Some inequalities for elliptic integrals, (to appear).

7. I. Näsell, Inequalities for modified Bessel functions, Math. Comp. 28 (1974) 253-256.

8. D. K. Ross, Solution to problem 72-15, SIAM Rew., 15 (1973), 668-670.

9. R. P. Soni, On an inequality for modified Bessel functions, J. Math. Phys., 44 (1965), 406-407.

10. G. N. Watson, A Treatise on the Theory of Bessel Functions, 2nd ed., Cambridge University Press, 1944.

MATEMATICA APPLICATA NELL'ANTICA GRECIA (1)

S. MARACCHIA
Dipartimento di Matematica; Istituto "Guido Castelnuovo"
Università degli Studi di Roma "La Sapienza"
Piazzale Aldo Moro, 2, 00185 Roma

Si legge nella Vita di Marcello scritta da Plutarco come Platone rimproverasse Archita ed Eudosso perché si erano serviti di mezzi meccanici per risolvere problemi altrimenti irresolubili: "Platone, racconta infatti Plutarco, rimase indignato da questo modo di procedere e polemizzò coi due matematici, quasi distruggessero e corrompessero ciò che vi era di buono nella geometria. In tal maniera essa abbandonava infatti i concetti astratti per scendere nel mondo sensibile ed usava anch'essa oggetti che richiedevano ampiamente un grosso lavoro manuale. La meccanica fu così separata e si staccò dalla geometria; per molto tempo la filosofia l'ignorò, ed essa divenne una delle arti militari."[2]

L'influenza di Platone bene si riflette negli Elementi di Euclide nei quali invano cercheremmo formule atte al semplice calcolo dell'area di una figura, di un volume di un solido o qualche indicazione per eseguire anche le più semplici operazioni aritmetiche. In questo modo le opere del filosofo e quelle del matematico collaborarono a creare la convinzione che la matematica in quanto tale doveva rivolgersi solo a problemi astratti, interni alla matematica stessa, ed affrontati con mezzi assolutamente razionali oltretutto limitati al solo uso delle curve perfette quali la retta e la circonferenza. La matematica, inoltre, doveva risultare un godimento in sè e per sè, utile, caso mai, secondo Platone, ad addestrare l'animo all'astrazione e prepararlo alla ricerca della verità attraverso la dialettica[3].

E' noto, d'altra parte, l'episodio raccontato da Stobeo[4] secondo

---

1.Il presente articolo è una rielaborazione di un precedente lavoro presentato al Convegno Nazionale della Mathesis tenuto ad Iseo nell'aprile del 1990.

2.Plutarco,Vita di Marcello, 14, tr.di C.Carena.

3."Senza la matematica -scrive Platone- non può esistere la felicità" (Epinomide 977, c-d) e si tratta ovviamente di una felicità del tutto spirituale.

4.Stobeo, Floril. IV, p. 205, citazione tratta da T.Heath, A History of Greek Mathematics, Clarendon Press, Oxford 1921, I, p. 357.

il quale Euclide fece allontanare un allievo dalla sua scuola "perché aveva pensato di trarre profitto dalla matematica".

Nessuna meraviglia, dunque, se l'opinione corrente attribuisce alla matematica greca un taglio assolutamente astratto, come si vede d'altra parte in quasi tutte le opere di Archimede e nell'opera conservataci di Apollonio sulle coniche.

Però, nonostante le indicazioni di Platone e gli esempi dei grandi matematici greci, non c'è dubbio che applicazioni pratiche della matematica dovevano pur esserci e non soltanto per la comune pratica del misurare e del calcolare necessari per la vita di una civiltà complessa qual era quella greca ["geometria" (γεωμετρια) com'è noto vuol dire appunto "misura della terra" a testimonianza delle origini che i Greci attribuivano ad essa e la "logistica" (λογιστικη) era la comune pratica del calcolo], quanto per vere e proprie applicazioni di risultati teorici a problemi di ordine pratico anche complessi e ad una vera e proprie fisica-matematica del tipo di quella che si trova nelle opere "meccaniche" di Archimede.

D'altra parte i Greci avevano importato dai Babilonesi e dagli Egiziani una matematica di carattere essenzialmente pratico ed è proprio da questa che prese le mosse quella astratta che è una delle loro glorie più fulgide. Lo scopo di questo lavoro, pertanto, è quello di mostrare attraverso qualche esempio che l'aspetto pratico della matematica si conservò durante i secoli, accanto allo straordinario sviluppo di quella teorica.

Nei risultati del primo matematico greco, Talete di Mileto (624-548 a.C. circa) è possibile trovare già questi due aspetti e se non ci si può meravigliare di quello pratico tratto direttamente dalla matematica che aveva appreso durante i suoi viaggi, è l'aspetto razionale, o per meglio dire una certa caratteristica razionalizzante, a meravigliarci poiché ci troviamo per la prima volta difronte a qualcosa di assolutamente nuovo[1]. In seguito accadrà il contrario: sarà la presenza di un aspetto pratico in una matematica tutta volta alla massima razionalizzazione a meravigliarci.

Secondo Proclo, infatti, Talete affrontò la matematica sia nel suo aspetto generale (καθολικωτερον) e sia nel suo

---

1.Molto vasta è la letteratura a questo riguardo, mi limito a dare solo qualche indicazione dalle quali però è possibile trarre altri riferimenti: G.J.Allman, Greek Geometry from Thales to Euclid, Dublino, Londra 1889, p. 7 sgg.; S.Maracchia, Talete nello sviluppo della geometria razionale in "Cultura e Scuola" 1971 nn.1-2; B. Rizzi e T. Viola, Dalla contemplazione ideale delle figure geometriche nell'uomo primitivo a quelle della geometria razionale attraverso l'opera di Talete di Mileto, "Atti dell'Accademia delle Scienze di Torino" vol. 114, 1980.

aspetto sensibile (αισθητικωτερον)[1]. Per il primo aspetto ricordiamo alcune dimostrazioni che gli vengono attribuite dallo stesso Proclo (uguaglianza degli angoli alla base di un triangolo isoscele; esser retto l'angolo inscritto in un semicerchio; divisione in due parti uguali del cerchio per mezzo del diametro; uguaglianza degli angoli opposti al vertice). Per il secondo aspetto ricordiamo la possibilità della misura dell'altezza delle piramidi per mezzo di una proporzione, legata ad un'idea di similitudine e ricordiamo specialmente la possibilità di misurare la distanza di una nave dalla costa attraverso il secondo criterio di uguaglianza (così Proclo,ma probabilmente si tratta anche in questo caso di una applicazione di una similitudine di triangoli).

La misura della distanza delle navi (nemiche?) dalla costa è dunque il primo vero esempio di una matematica applicata ad un avvenimento reale di pratico interesse; il matematico non è soltanto uno studioso astratto, assolutamente distaccato dalla realtà che lo circonda, ma è in grado di applicare la sua scienza al momento opportuno qualora le circostanze lo richiedano.[2] Questa misura eseguita attraverso la similitudine di triangoli rettangoli (questa è la spiegazione del procedimento di Talete che mi sembra più attendibile) oppure attraverso una vera e propria triangolazione che stabilisce la distanza di un punto inaccessibile a partire da due accessibili (ipotesi questa che risponde maggiormente all'indicazione di Proclo ma che risulta più elaborata per quanto riguarda l'applicazione matematica) oppure secondo altre ipotesi ancora, avrebbe spinto Anassimene, secondo quanto scrivono Federigo Enriques e Giorgio De Santillana[3] a tentare una analoga triangolazione col sole in modo da poterne calcolare la distanza dalla terra. Non mi occuperò di questo episodio, non del tutto certo, di matematica applicata; anche un altro allievo di Talete, Anassimandro, aveva tentato di

---

1.Cfr. Proclo,Commento al primo libro degli Elementi di Euclide, Prologo, parte II (Friedlein, 65, 10-11); le successive indicazioni relative ai risultati matematici di Talete si trovano nei commenti alle varie proposizioni. Ricordiamo che dell'opera di Proclo esiste l'edizione curata da Maria Timpanaro Cardini (Giardini, Pisa, 1978) nella quale il brano relativo a Talete si trova a p.71.

2.Viene spontaneo ricordare a questo punto le difese approntate da Archimede in occasione dell'assedio di Siracusa da parte del console Marcello e gli altri episodi che legarono il grande matematico a problemi di carattere pratico: costruzione della sfera armillare; episodio della corona; spostamento di una nave a pieno carico con la forza muscolare di un solo uomo (cfr. ad esempio la citata Vita di Marcello di Plutarco o la Storia di Roma di Tito Livio, XXIV, 34).

3.Cfr. la loro Storia del pensiero scientifico, Zanichelli, Bologna, 1932, pp. 269 e 271.

stabilire la stessa distanza[1], ma vedremo in seguito alcune dimostrazioni di Aristarco di Samo volte alla determinazioni delle dimensioni e delle distanze del sole e della luna.

La spinta per una più consapevole applicazione della matematica avvenne però paradossalmente proprio da parte di chi contribuì in misura maggiore a renderla astratta e razionale: Pitagora di Samo (580-504 a.C. circa). La sua concezione di un universo ordinato e costruito secondo canoni matematici, se da un lato incrementò la ricerca pura volta alla conoscenza dei segreti dell'universo attraverso le proprietà del numero che ne costituiva l'intima essenza, da un altro lato, proprio per questo, indicò all'uomo la possibilità di poter applicare la matematica ad ogni attività umana, pratica o no. La lingua con cui è stato scritto il libro dell'universo è quella matematica, scriverà molti secoli dopo Galileo, e matematici ne sono anche i caratteri[2]. Pitagora aveva detto la stessa cosa, anzi egli era andato ancora più a fondo: la struttura stessa dell'universo (còsmos κοσμος cioè "ordine") non solo era governata dal numero, ma era proprio il numero in tutta la sua magica potenza ordinatrice[3]. E' chiaro che con una concezione siffatta anche qualsiasi problema di ordine pratico deve potersi risolversi nell'ambito della teoria astratta del numero e quindi matematicamente.

Lo stesso Pitagora, d'altra parte, aveva dato un esempio di connessione tra la matematica ed una disciplina apparentemente molto lontana da essa: la musica. Lo studio delle proporzioni tra la lunghezza di una corda vibrante e le note emesse da questa è pur sempre una applicazione della matematica ad una disciplina esterna ad essa. Ma, per ribadire un concetto già espresso, è

---

1. Per quanto riguarda Anassimandro, anche per l'uso scientifico dello "gnomone" all'origine delle principali misurazioni solari, si può leggere l'interessante e denso lavoro di Filippo Franciosi, Le origini scientifiche dell'astronomia greca, L'"Erma" di Bretschneider, Roma, 1990.

2."La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezzi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto." (Il Saggiatore Ed. Naz. VI, 232).

3.Numerose sono le testimonianze relative alla concezione pitagorica; Aristotele e i suoi commentatori se ne sono occupati più volte, cfr. ad esempio La Metafisica 985, b 23 - 986, b 31 e il relativo commento di Alessandro (p. 62 sgg. nei Pitagorici, Testimonianze e Frammenti Fasc. III di Maria Timpanaro Cardini, La Nuova Italia, Firenze, 1964).

nella concezione di un universo matematicamente ordinato, nel sogno di una possibile indagine fisico-matematica della realtà che consiste il contributo maggiore dato da Pitagora ad una applicazione della matematica.

La concezione pitagorica, anche talvolta nei suoi aspetti più esteriori, si diffuse con le fortune della scuola fondata da Pitagora che riscosse grande considerazione in tutta l'Antichità nonostante le varie concezioni filosofiche che la fertile fantasia e capacità speculativa dei Greci elaborarono. Ecco che cosa scrivono a questo proposito F. Enriques e G. De Santillana:[1]

> "Col fiorire della scuola pitagorica la tecnica viene penetrando degli ideali matematici. La ragione geometrica (...) si impone in ogni campo, come norma d'azione (...) ogni cosa deve avere la sua giustificazione nel numero e nella misura. (...) L'entusiasmo matematico si ritrova ovunque. Un tipico rappresentante di questo movimento di idee, Ippodamo di Mileto -che nutrito di filosofia diventò poi grande architetto- è il primo che abbia affrontato i problemi dell'urbanistica con criteri razionali. Fu chiamato dai Rodii a sistemare la loro città; Pericle (...) gli diede a ricostruire il Pireo. Quando la lega Panellenica volle fondare la città simbolica di Turii, gli chiese di farne il progetto, ed egli disegnò quel tipo di città geometrica, dalle strade dritte intersecantisi perpendicolarmente, di cui poi l'antichità face così largo uso (ed anzi abuso: si vedano le piante della città di Priene) e che ritroviamo nel tracciato regolare delle città romane. D'altronde Ippodamo non si era accontentato di portare la geometria all'architettura, ma aveva proposto uno schema di costituzione della polis, in cui i poteri e gli uffici erano tutti organizzati su una base ternaria: idea che si ritroverà nel Tripolitico di Dicearco[2]. dall'architettura si passa alla scultura: anche qui le matematiche si impongono e pretendono portare il

---

1.Op. cit. pp. 487-488.

2.Di Ippodamo (V sec. a.C.) parla Aristotele nella Politica, 1267, b sgg.; la città stato avrebbe dovuto avere solo diecimila uomini divisi in tre classi (agricoltori, artigiani, militari); anche il territorio avrebbe dovuto dividersi in tre tipi (sacro, pubblico, privato) e anche le leggi, infine, avrebbero dovute essere di tre tipi, relative alle tre colpe possibili (oltraggio, danno, omicidio). Ricordo che Ippodamo, cui venne anche attribuita l'idea del terreno a terrazze nel caso di campi scoscesi, viene considerato pitagorico da Stobeo (cfr. H.Diels, I Presocratici. Testimonianze e Frammenti,Laterza, Bari, 1986, I, 442) notizia che in verità M.Timpanaro Cardini considera "tardiva e falsa" (op.cit. Fasc.III, p.370); suo allievo fu Dinocrate di Rodi che iniziò la costruzione di Alessandria a cui poi successe Sostrato di Cnido autore del famoso Faro, una delle sette meraviglie del mondo.

> proprio criterio estetico. Agatarco ateniese verso il 470, applica la stereometria alla Prospettiva (...) le proporzioni geometriche (...) sono condizione indispensabile di bellezza del corpo umano [secondo Policleto]."

In maniera simile, sebbene con diversa valutazione, si esprime Benjamin Farrington allorché scrive:[1] "il pitagorismo col suo culto del numero e della geometria, fu la teoria che influì più profondamente sulla vita dei greci" e, dopo aver osservato che l'esasperazione di questo autoritarismo toglieva agli artisti la freschezza dell'ispirazione, aggiunge: "Vennero di moda i piani regolatori urbani basati su linee geometriche. Ippodamo di Mileto, un fanatico della teoria del numero, preparò nuovi piani per il Pireo, il porto di Atene, per Rodi e per Turi." Anche Farrington accenna al fatto che l'influenza di Ippodamo si fa sentire anche in epoca romana (e oltre!): "per scegliere due esempi fra i tanti, scrive infatti, gli schemi urbani di Pompei e di Timgad costituiscono tarde testimonianze della tradizione pitagorica, e molte città moderne che sono sorte sul modello di accampamenti romani rivelano ancora l'impronta di quella moda. Perfino il nuovo mondo ne è erede. New York, col suo disegno geometrico e con le sue streets e avenues numerate, è una città interamente pitagorica."

Osserviamo ora un chiaro esempio di matematica applicata ad esigenze esclusivamente pratiche. Troviamo in Erodoto questo significativo passo:

> "[Nell'isola di Samo] sono state realizzate le tre opere più importanti che ci siano fra i Greci tutti: in un colle, che si eleva sino a 150 orge, è stata scavata una galleria che comincia proprio ai piedi del colle e va a sboccare sull'altro versante. La galleria è lunga sette stadi, alta otto piedi e larga altrettanti, per tutta la sua lunghezza è stato compiuto un altro scavo, profondo venti cubiti e largo tre piedi, attraverso il quale, incanalata per mezzo di tubi, giunge alla città l'acqua derivata da una grande sorgente. L'architetto di quest'opera di scavo fu il megarese Eupalino, figlio di Naustrofo. Questa è una delle tre meraviglie..."[2].

1.B.Farrington, La scienza nell'antichità, Longanesi, Milano, 1978, pp. 40-41.

2.Erodoto, Le Storie, III, 60 (tr.L.Annibaletto); notiamo che le le altre due meraviglie sono: "il molo che recinge il porto sul mare e raggiunge la profondità perfino di 20 orge [40 metri circa], mentre la sua lunghezza supera i due stadi [355 metri circa]" e "[le dimensioni] di un tempio, il più grande di tutti i templi che noi conosciamo".
Teniamo presente che un "piede" è circa 0,333...metri; un'"orgia" è uguale a sei piedi e quindi a circa due metri e uno "stadio" è

Notiamo innanzi tutto che in misure attuali (v. nota precedente) la collina in questione era alta circa 300 metri, la galleria aveva una lunghezza di circa 1,244 km, un'altezza e un larghezza di circa 2,6 m mentre il secondo scavo era profondo circa 9 m e largo 1 m.

Penso che la meraviglia di Erodoto, che era rivolta presumibilmente solo alle dimensioni del l'opera di Eupalino, si sarebbe accresciuta se egli si fosse soffermato sulla circostanza che Eupalino forò la collina, oggi chiamata Castro, da entrambe le parti in modo che i due tronconi si potessero congiungere circa a metà strada!

Fu proprio questa circostanza che ha dato da pensare agli storici della matematica dopo che nel 1882 l'opera di Eupalino venne riscoperta e successivamente esaminata da varie spedizioni archeologiche. Alla metà del tunnel fu possibile osservare, infatti, un certo sfasamento orizzontale e verticale; ma come era stato possibile programmare la direzione congiungente i due punti di attacco? Evidentemente, prima della realizzazione dell'opera, il problema dovette essere risolto a tavolino, cioè matematicamente ed è questa risoluzione matematica che interessa soprattutto gli storici.

In verità tutto sembrava risolto da un problema presente nella Diottra di Erone (vissuto probabilmente nel 1° sec. d. C. ma che possiamo situare con sicurezza solo tra il 2° sec. a. C. ed il 3° sec. d. C.): nel XV problema dell'opera, infatti, viene affrontato e risolto il problema che si era presentato ad Eupalino e cioè come "Traforare la montagna secondo una retta che congiunga due punti dati sulle sue falde". L'esame sul posto sembra però escludere che Eupalino possa essersi servito del procedimento indicato da Erone che si sviluppa su un ideale piano passante per i due punti da cui iniziare gli scavi. In questa sede non ha interesse esaminare la polemica sorta a questo proposito[1], comunque siano andate le cose, la matematica greca era

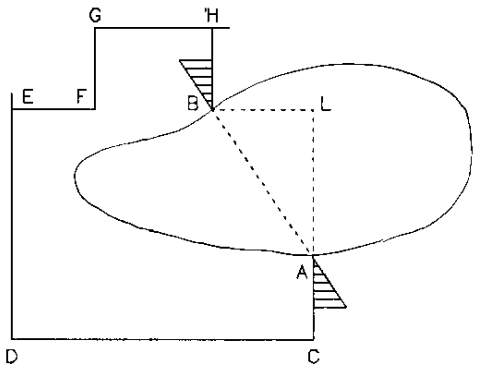
---

...Continua...

uguale a circa 177,70 metri. Un "cubito", infine (questa misura servirà tra poco), ha una lunghezza che oscilla (a seconda dei testi consultati) da 444 mm a 462 mm, cioè poco meno di mezzo metro.

1.Quasi tutti i libri di storia della matematica greca parlano dell'ormai famoso "Tunnel di Eupalino"; riguardo alla polemica relativo al metodo seguito per il congiungimento dei due tronconi, si possono leggere i seguenti articoli nei quali vengono date ulteriori indicazioni bibliografiche: J. Goodfield e S. Toulmin. How was the Tunnel of Eupalinus aligned?, ISIS, 1965, vol. 56, 1; B. L. van der Waerden, Eupalinus and his tunnel, ISIS, 1968; T. Viola, S. Manzoni, M. T. Navale, Problemi geometrici applicati

giunta in qualche modo a risolvere un difficile problema tecnico. Nell'opera di Erone, inoltre, noi possiamo osservarne una soluzione, sia o non sia questa la soluzione di Eupalino.[1]



A e B sono i due punti dai quali si comincia a scavare il tunnel (AB); costruita la poligonale ACDEFGHB con percorso ad angoli retti, è evidente che i cateti AL e BL del triangolo immaginario ABL si possono misurare essendo.

$$AL = DE + FG - (AC + HB)$$
$$BL = DC - (EF + GH)$$

Ma allora, costruendo due triangoli simili al triangolo ABL, come in figura (triangoli tratteggiati), è subito visto che le due ipotenuse sono allineate con AB ed indicano quindi la direzione richiesta.

...Continua...

alle tecniche costruttive e rappresentative in "Imago et mensura mundi", Atti del IX Congresso internazionale di storia della cartografia, Ist. della Encicl. Ital. vol 3.

1.Naturalmente sono state avanzate varie ipotesi relative alla soluzione di Eupalino ma si tratta pur sempre di ipotesi anche se ingegnose; il problema oggi non presenta alcuna difficoltà, ne fanno testimonianza le innumerevoli gallerie che continuamente vengono costruite a partire dalle due parti opposte e con il perfetto congiungimento dei due tronconi e delle maestranze che vi hanno lavorato. A titolo di curiosità, ricordo che nei Problemi di geometria di Lorenzo Mascheroni (Milano, 1832), viene appunto proposto il nostro problema ("continuare la retta AB in C e D [da determinare] al di là dell'ostacolo X che impedisce il traguardo" così lo enuncia Mascheroni) di cui vengono date ben undici soluzioni.

Veniamo ora ad alcune dimostrazioni di Aristarco di Samo (310-230 a. C. circa) su alcune misure astronomiche che si trovano nell'opera Sulla grandezza e sulla distanza del sole e della luna. Ricordiamo, prima di osservare alcuni risultati raggiunti in tale opera, che Aristarco è anche noto per essere stato il primo astronomo greco e forse il primo astronomo in senso assoluto, ad avanzare l'ipotesi eliocentrica. Questa ipotesi, che non si trova nell'opera sopra citata, ma che ricaviamo da una esplicita e attendibile attribuzione presente nell'Arenario di Archimede, non ebbe molti sostenitori: si ha memoria solo di un certo Seleuco che la seguì,[1] dopo di che fu abbandonata, come sappiamo, per molti secoli. Come avverrà per Galileo, anche Aristarco fu attaccato e minacciato per questa ipotesi: riporta Plutarco, infatti, che "Cleante" discepolo di Zenone lo stoico " reputava che Aristarco dovesse essere accusato, davanti ai Greci, di profanazione sacrilega, per aver spostato il focolare del mondo (...) Quest'uomo, infatti, aveva tentato di salvare le apparenze, facendo le ipotesi che il cielo stesse immobile, e che la terra percorresse il cerchio obliquo [l'eclittica] e nello stesso tempo ruotasse intorno al proprio asse"[2]
L'opera di Aristarco, di cui vedremo solo alcuni risultati, è costituita da sei ipotesi e diciotto proposizioni ed è strutturata in forma razionale di tipo euclideo. Aristarco usa diverse proprietà geometriche presenti negli Elementi di Euclide ed inoltre una proprietà che espressa con nostro simbolismo risulta essere:

$$\text{sen}\alpha/\text{sen}\beta < \alpha/\beta < \text{tg}\alpha/\text{tg}\beta \quad (0 < \beta < \alpha < \pi/2)$$[3]

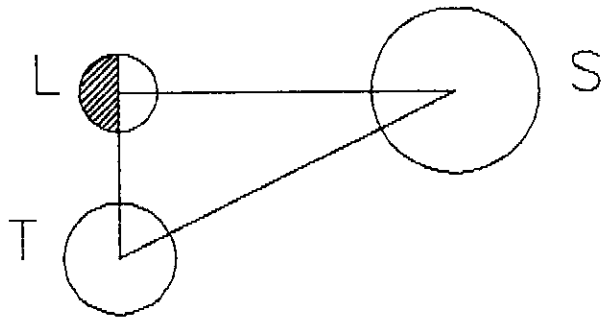
Osserviamo ora le sei Ipotesi su cui Aristarco basa le sue dimostrazioni:

1) La luna riceve la luce dal sole;
2) la terra è come un punto e centro della sfera in cui si muove la luna;
3) quando la luna appare divisa in due parti uguali il [piano del] cerchio massimo che divide la parte illuminata è nella direzione dei nostri occhi;

1.Cfr. T. Heath, op. cit., II, p. 3.

2.Plutarco, De Facie in orbe lunae, cap.6; ho tratto questa citazione dall'opera di Aldo Mieli, Manuale di storia della scienza, ed,Leon. da Vinci, Roma, 1925, p.122.

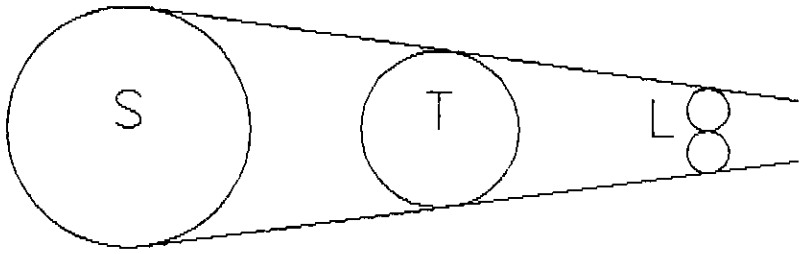
3.Aristarco non sente alcuna necessità di giustificare le due disuguaglianze e da ciò possiamo dedurre che esse dovevavno essere ormai acquisite da tempo; troviamo, infatti, una dimostrazione della seconda nell'ottava proposizione dell'Ottica di Euclide; anche della prima i matematici greci ci hanno tramandato una dimostrazione (Tolomeo, Almagesto, I, 10) anche se questa è successiva ad Aristarco.



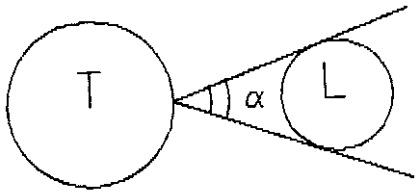
4) quando la luna appare divisa in due parti uguali, la sua distanza dal sole è [vista sotto l'angolo] minore di un quadrante per un trentesimo di quadrante;
(facendo riferimento alla figura precedente si ha cioè LTS = 1R - 1/30 R = 87°)

5) la larghezza dell'ombra [della terra] risulta di due lune;

(in altre parole, il cono d'ombra proiettato dalla terra all'altezza della luna contiene questa esattamente due volte, v.fig.)



6) la luna sottende una quindicesima parte [del segno] dello zodiaco.



$\alpha = 1/15 \cdot (1/12 \cdot 360°) = 2°$

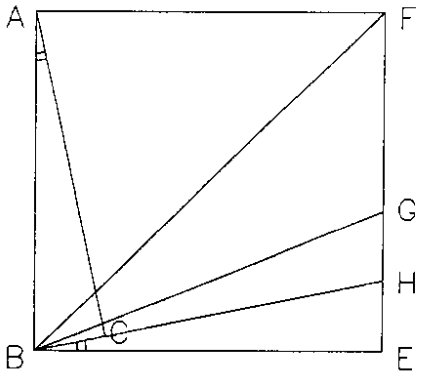
Notiamo che le misure odierne, grazie a strumenti più sofisticati, sono alquanto diverse da quelle di Aristarco; nella quarta ipotesi, ad esempio, a seconda di quando avviene il fenomeno dell'esatta divisione tra ombra e luce nella luna (primo e ultimo quarto), l'angolo LTS assume i valori di 89° 50' 16,8" e 89° 50' 45,6" il che vuol dire che il sole risulterà molto più lontano di quanto verrà ottenuto da Aristarco. Così nell'ipotesi cinque quel valore "due" andrebbe sostituito con 2,606.. ed infine nell'ipotesi sei l'angolo sotto il quale la luna è vista dalla terra oscilla tra 28' 43" e 34' 30" con una media quindi di 31' 36,5". A proposito di quest'ultima misura, notiamo che Aristarco la corresse successivamente in 30', cioè molto vicino a quella di oggi, poiché è questa la misura che gli viene attribuita da Archimede nell'Arenario.

Gli "errori" che si trovano nelle misure di Aristarco, per altro straordinarie ai suoi tempi, nulla però tolgono al rigore del suo ragionamento tanto che sostituissimo alle sue misure quelle di oggi avremmo dei valori abbastanza vicini a quelli che oggi calcoliamo[1]

Vediamo ora, a titolo d'esempio, una delle più importanti proposizioni di Aristarco, la settima, nella quale viene dimostrato che "La distanza che separa il sole dalla terra è maggiore 18 volte, ma minore di 20 volte, della distanza in cui la luna dista dalla terra" In altre parole, indicando con A, B, C rispettivamente i centri del sole, della terra e della luna, Aristarco dimostra che:

1.Ricorda Francesco Pozzo de Somenzi alla voce "Aristarco" dell'Enc. Ital. Treccani), parlando del metodo di Aristarco, che "Keplero nel 1618 ne invocava l'applicazione ad una misura telescopica e Goffredo Wandelin trent'anni più tardi se ne serviva per una determinazione che, pur effetta da errore notevole, è la prima che si approssima alla verità"

$18 < AB/BC < 20$



Si consideri il quadrato di lato AB e sia BG la bisettrice dell'angolo FBE = 1/2·R. Si prolunghi BC sino ad H (v. fig.) e si ricordi che per la quarta ipotesi l'angolo $B\hat{A}C$ (= $H\hat{B}E$) = 1/30·R.

Ricordando la proposizione di Euclide (Ottica,8), si ha:

(*) $GE : HE > G\hat{B}E : H\hat{B}E = 1/4 \cdot R : 1/30 \cdot R = 15/2$

ma per il teorema della bisettrice

(**) $FG : GE = FB : BE > 7/5$[1],

componendo si ha dunque:

(***) $FE : GE > 12/5$.

Dalle (*) e (***), moltiplicando membro a membro, si ha:

$GE/HE \quad FE/GE > 15/2 \cdot 12/5$ cioè $FE/HE > 18$.

ma per la similitudine dei triangoli BHE e ABC, si ha:

---

1. Il rapporto FB : BE è uguale, come scriveremmo noi, a $\sqrt{2}$ che i matematici greci approssimavano con valori d_n / l_n per difetto o per eccesso dedotti dai cosiddetti «numeri diagonali e laterali»:

d_n)	1	3	7	17	41 ...
l_n)	1	2	5	12	29 ...

tenendo presente che si pone inizialmente $d_1 = l_1 = 1$ e si costruiscono i valori successivi con le formule ricorrenti:
$d_n = d_{n-1} + 2\, l_{n-1}$; $l_n = d_{n-1} + l_{n-1}$ (per questa interessante risultato dell'aritmetica pitagorica cfr. «Attraverso la storia della matematica» di Attilio Frajese, Le Monnier, Firenze, 1969, p. 239 sgg.).

$$FE/HE = BE/HE = AC/BC$$

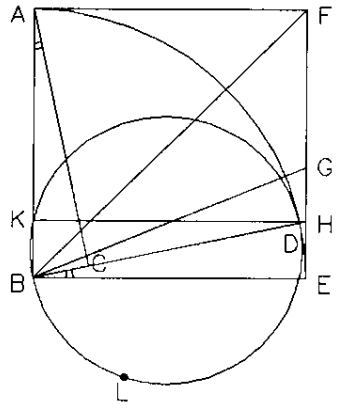
quindi

$$AC/BC > 18$$

e, a maggior ragione:

(1) **$AB/BC > 18$.**

Si consideri ora la parallela per H a BE e sia D il suo punto d'incontro con l'arco di circonferenza di centro B e raggio BA e K quello d'incontro con il lato BA. Poiché l'angolo DKB è retto, la circonferenza di diametro BD passa per K e sia L un punto di tale circonferenza tale che BL sia uguale alla metà di BD e cioè uguale al raggio.



Poiché $\hat{KDB} = \hat{HBE} = \hat{BAC} = 1/30 \cdot R$ e l'angolo al centro dell'arco BL è uguale ad 1/6 di angolo giro e cioè $2/3 \cdot R$ (e dunque quello alla circonferenza è $1/3 \cdot R$), si ha:

(°) $\quad arc\ BK / arc\ BL = 1/10;$

d'altra parte, per la proprietà già menzionata $\alpha/\beta > sen\alpha/sen\beta$ (con $0 < \beta < \alpha < \pi/2$) si ha:

(°°) $\quad arc\ BK / arc\ BL < BK/r.$

Pertanto, dalle (°) e (°°) si ottiene:

$1 : 10 > BK : r \quad$ cioè $\quad r < 10 \cdot BK$

da cui, moltiplicando per due:

(°°°) $BD < 20 \cdot BK$ cioè $BD/BK < 20$

tenendo ora presente che per la similitudine dei triangoli BDK e ABC si ha $BD/BK = AB/BC$, si può concludere che:

(2) **$AB/BC < 20$.** ∎

A titolo di curiosità, osserviamo quali risultati si ottengono con questi calcoli se al posto delle misure di Aristarco si vanno a sostituire quelle di oggi:

Angolo ABC = 89,842° (valore medio tra 89,838° e 89,846°); di conseguenza l'angolo HBE = 0,158 e il rapporta GBE/HBE anziché 15/2 diventa (90/4) : 0,158 = 142,405 (*).
I rapporti FB : BE = $\sqrt{2}$: 1 sono uguali a 1,4142.. cosicché componendo si ha FE/GE = 2,4142 (***).
Moltiplicando (*) e (***) si ottiene dunque (senza ripetere i passaggi già fatti):

(1)' **$AC/BC > 142{,}405 \; 2{,}4142 = 343{,}794$.**

Analogamente, il rapporta tra l'arco BK e l'arco BL diventa (anziché 1/10) 0,158° : 30° = 0,005266.. (°) il cui reciproco é 189,873; dunque si ha $r < 189{,}873$ BK e, in definitiva (moltiplicando per due ecc.):

(°°)' **$AB/BC < 379{,}746$.**

Notiamo infine che il valore esatto (medio) di tale rapporto AB/BC è oggi, con le misure espresse in chilometri, uguale a 149.650.000 / 384.400 = 389,3080[1].

Nelle successive proposizioni Aristarco, indicando con l, t, s i diametri rispettivamente della luna, della terra e del sole e con L e S le distanze dei centri della luna dalla terra e del sole dalla terra, ottiene i risultati:

Prop. 9 $18 < s/l < 20$

Prop. 11 $1/30 < l/L < 2/45$

Prop. 15 $19/13 < s/t < 43/6$

Prop. 17 $108/43 < t/l < 60/19$.

---

1. Si noti che se usare il valore medio dell'angolo BAC ottenuto dai due valori che si possono calcolare al primo o all'ultimo quarto, avessimo usato il più grande (89,846°), avremmo ottenuto per le due limitazioni 352,72.. e 389,61.. ancora più straordinariamente vicine ai valori di oggi.

Ebbene, usando i valori medi, possiamo anche considerare i seguenti risultati:

Prop. 7 S/L = 19

Prop. 9 s/l = 19

Prop. 11 l/L = 7/180 cioè L/l = 180/7

Prop. 15 s/t = 27/4

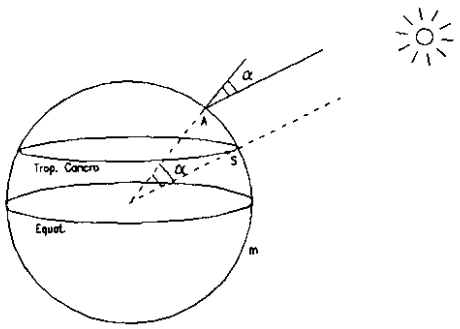
Prop. 17 t/l = 2316/817.

In questo modo è possibile calcolare, in diametri terrestri, la distanza del sole dalla terra:

S/t = L/l S/L l/t = 180/7 19 817/2316 = 172,35...

Questo valore non è naturalmente attendibile e anche se sarà migliorato dai vari astronomi successivi rimarrà sempre lontano da quello che si calcola oggi (S/t = 149.650.000/2 6370 = 11.746,46..).

Per completare le misurazioni di Aristarco (o quelle degli astronomi che lo seguirono) bisognerebbe calcolare dunque il diametro terrestre. A questo penserà Eratostene (276-194 a.C.) con la sua famosa misura del meridiano terrestre che è un altro esempio di matematica applicata.



Nella figura si è indicato con A la città di Alessandria in cui nel giorno di altezza massima del sole (solstizio del 21 giugno) i suoi raggi hanno una inclinazione rispetto alla verticale di circa (7 + 1/7)°; con S la città di Siene (la moderna Assuan) che si trova (quasi) sullo stesso meridiano di Alessandria e sul tropico del cancro il che vuol dire che nello stesso giorno del solstizio considerato (e solo in questo giorno) i raggi cadono perpendicolarmente. Ebbene misurando la distanza delle due città

in 5000 stadi (circa 785 km) forse per mezzo di camminatori dal passo regolare, si ottiene, dall'evidente proporzione (v. fig.):

$$\alpha : AS = 360° : 785$$

cioè, indicando con m la lunghezza in chilometri del meridiano:

$$(7 + 1/7) : 785 = 360 : m$$

da cui si ottiene, con stupefacente approssimazione:

m = 39.564 km ![1]

Silvio Maracchia

---

1. La misura esatta, secondo la tradizionale definizione del metro è 40.000 km; bisogna pensare che gli errori commessi da Eratostene nelle sue misure si siano compensati.

CLUSTERING SU GRAFO

M. MARAVALLE, Facoltà di Economia e Commercio
Università degli Studi dell'Aquila

ABSTRACT

The present paper deals with the following problem: given a connected graph G with n nodes, partition its set of nodes into p classes such that the subgraph induced by each class is connected and a given function, inertia, of the partition is minimized. This paper includes : a heuristic algorithm for general graphs based on the solution of a sequence of restricted problems where G is replaced by one of its spanning tree ; a complete set of real-life applications and a comparison with a hierarchical algorithm.

KEYWORDS : Clustering, constrained clustering.

1. INTRODUZIONE

In molti problemi di classificazione gli oggetti possono essere condizionati a particolari vincoli. I più semplici tipi di vincolo sono legati alla numerosità (ad esempio ciscun gruppo non deve contenere più di un certo numero massimo di soggetti). Un altro tipo di vincolo che spesso si incontra è quello legato alle condizione di contiguità dei membri di un gruppo. In questo caso gli oggetti sono delle località (possono essere città, comuni, quartieri) per le quali l'appartenenza ad uno stesso gruppo richiede che siano geograficamente contigue e che, come si dice in gergo, formino un distretto (particolarmente interessanti per gli studi dei bacini di utenza di servizi quali poste, telefoni - si veda l'esempio di C. Roche 1978 - ecc.). Per poter dare una definizione formale di distretto è necessario introdurre il concetto di grafo, i cui vertici sono le località e laddove esiste uno spigolo fra i due vertici significa che fra queste due località c'è contiguità che in termini pratici indica il collegamento diretto tramite una strada, una ferrovia, un gasdotto ecc. .

La condizione che i siti formino un distretto è espressa dalla condizione che il gruppo dei vertici indotti dalla classificazione sia un sottografo connesso (per la terminologia dei grafi si è utilizzato il testo di Cerasoli, Eugeni e Protasi 1988).

I vincoli di contiguità intervengono anche in altre aree applicative quali la geologia e la botanica, si veda in proposito Gordon 1973 e 1978.

I due principali obiettivi della classificazione sono quelli di massimizzare l'omogeneità all'interno di un gruppo (elementi dello stesso gruppo devono essere simili fra loro) come pure la diversità fra gruppi diversi (elementi di gruppi diversi devono essere differenti fra loro). Questi due obiettivi sono solo apparentemente in contrasto giacché grazie al lemma di Huyghens si può dimostrare che minimizzare la somma delle inerzie interne equivale a massimizzare l'inerzia esterna, si veda Benzécri 1980)

Nel presente articolo è trattato il problema di *classificazione con inerzia minima su un grafo* . A partire da:

- un grafo connesso $G \equiv (V,E)$ con $n=|V|$ vertici ed $m=|E|$ spigoli;
- una funzione vettoriale che associa ad ogni vertice i un vettore di caratteristiche q-dimensionali x(i), cioe' una funzione x: V------> $\Re^q$;
- un intero p , numero dei gruppi, tale che $1 \leq p \leq n$;

si cerca una p-partizione $\pi = \{ C_1,.....,C_p \}$ dei vertici V che sia *ammissibile* nel senso che i sottografi $G(C_k)$ indotti dalla partizione siano tutti connessi; e che sia minima l'inersia interna della partizione

$$h(\pi) = \sum_{k=1}^{p} I(C_k) \qquad (1)$$

dove per inerzia interna del k-mo cluster si intende la quantita' :

$$I(C_k) = \sum_{i \in C_k} \| x(i) - b(C_k) \|^2 ,$$

mentre $\| \cdot \|$ è una norma euclidea in $\Re^p$ ed infine $b(C_k)$ è il baricentro di C_k :

$$b(C_k) = \frac{1}{|C_k|} \sum_{i \in C_k} x(i) \quad .$$

Simili problemi sono stati già trattati nella letteratura scientifica, ed in particolare L. Lebart 1978 ha proposto un algoritmo di tipo gerarchico nel quale ad ogni iterazione non vengono aggregati gli elementi (sottogruppi o semplici individui statistici) più prossimi bensi' quelli che lo sono subordinatamente all'esistenza di uno spigolo che li possa connettere fra loro. Di tale algoritmo esiste anche un programma per computer , AGRAF .

In questo articolo viene presentata nel paragrafo 2 una euristica originale basata su un algoritmo di classificazione non gerarchico concatenato alla soluzione di problemi di ottimo su alberi generati dal grafo stesso. Essendo ben noti i vantaggi/svantaggi di questo tipo di algoritmo rispetto a quelli gerarchici, è stata realizzata, nel paragrafo 3, una comparazione dei risultati di tale euristica comparati con quelli di AGRAF per meglio valutarne le caratteristiche di performance su alcuni problemi reali che coprono un insieme di casi sufficientemente ampio.

2. METODOLOGIA PER LA CLASSIFICAZIONE (EURISTICA A DUE STADI)

Indicato con $\Pi_p(G)$ l'insieme di tutte le p-partizioni di V ammissibili in G; data una partizione π, uno spigolo è chiamato *interno* se i due vertici appartengono entrambi ad uno stesso gruppo indotto dalla partizione π, altrimenti viene indicato come *esterno*. Sia

$$\zeta_p(G) = \min \{ h(\pi) : \pi \epsilon \Pi_p(G) \}, \qquad (2)$$

dove $h(\pi)$ è dato dalla (1).

In Maravalle e Simeone 1992, è stata dimostrata la seguente identità:

$$\zeta_p(G) = \min \zeta_p(T)_{T \epsilon \mathcal{T}} \qquad (3)$$

dove $\mathcal{T}$ è l'insieme di tutti gli alberi di G e per un grafo qualsiasi.

L'identità (3) suggerisce la seguente euristica per il problema (2):

MIDAS (Metodo Iterativo Di Aggregazione Spaziale)

STEP 1 : (*albero iniziale*) Si cerca un albero iniziale "buono" T in G;

STEP 2 : (*partizione iniziale*) Si cerca una partizione iniziale "buona" $\bar{\pi}$ in $\Pi_p(T)$;

STEP 3 : (*ottimizzazione sull'albero*) Partendo dalla partizione $\bar{\pi}$ esegue una ricerca locale per trovare una soluzione quasi-ottimale π^* al problema *ristretto* :

$$\min \{ h(\pi) \mid \pi \epsilon \Pi_p(T) \}$$

STEP 4 : (*modificazione dell'albero*) Si cerca, se possibile, un'altra p-partizione $\bar{\pi}$ ed un altro albero $\bar{T}$ di G tale che :

I) $h(\bar{\pi}) < h(\pi^*)$

II) $\bar{\pi} \notin \Pi_p(T)$

III) $\bar{\pi} \in \Pi_p(\bar{T})$

Se non si trova nessuna coppia $(\bar{\pi}, \bar{T})$ allora il risultato finale è raggiunto con la partizione corrente π^* (poichè π^* è ammissibile in T lo sarà anche in G);

altrimenti sostituisce $\bar{\pi}$ con π^* e si torna allo STEP 3;

END

Discutiamo ora in dettaglio i quattro passi ora descritti

STEP 1 (*albero iniziale*)

Un "buon" albero iniziale può essere trovato nel modo che segue . Definita come $d(i,j)=|| x(i) - x(j)||^2$ la dissimilarità tra i vertici i e j. Si assegna a ciascuno spigolo <i,j> del grafo G una lughezza pari a d(i,j). Successivamente si calcola l'albero di lunghezza minima T di G attraverso l'algoritmo di Kruskal (Lawler, 1976). Questa scelta razionale di T fa si che la lunghezza media di uno spigolo in T sia in generale sufficientemente piccola; quindi è sperabile che, se $\pi=\{C_1,C_2,.....,C_p\}$ è ottimale per T, l'inerzia interna di ogni cluster C_k ed il valore ottimale $\zeta_p(T)$ siano sufficientemente piccoli. I risultati sperimentali del prossimo paragrafo indicano che questa scelta per l'albero iniziale è "buona" nel senso che, partendo da T, MIDAS deve generare pochissimi alberi.

STEP 2 (*partizione iniziale*)

Una buona partizione iniziale $\bar{\pi} \in \Pi_p(T)$ può essere ottenuta attraverso una procedura "ghiotta". In primo luogo si nota che tagliando $p-1$ spigoli, si ottiene una foresta F_p di G consistente in p alberi; si indichino con $C_1,C_2,.....,C_p$ gli insiemi dei vertici di questi alberi. Allora $\pi=\{C_1,C_2,.....,C_p\}$ è una p-partizione ammissibile di T. Eliminando ancora un altro spigolo questo avrà l'effetto di separare un certo cluster C_h in due nuovi clusters $C_{h'}$ e $C_{h''}$. Dopo la separazione il decremento di inerzia interna complessivo è

$$\Delta = \mathbb{I}(C_h) - \mathbb{I}(C_{h'}) - \mathbb{I}(C_{h''}) \geq 0 \ .$$

Si può ora dare una semplice descrizione della procedura "ghiotta".

PROCEDURA GHIOTTA

Passo 1 : Sia $F := T$; $k := 1$;

Passo 2 : Si taglia lo spigolo e di F cui corrisponde la maggiore diminuzione di inerzia interna Δ ; si sostituisce F con $F - e$;

Passo 3 : Se $k = p-1$ allora STOP : si ritorna alla foresta F ed alla partizione associata $\bar{\pi} \in \Pi_p(T)$; altrimenti si incrementa k di un'unità e si torna al Passo 2.

FINE

Il decremento di inerzia interna Δ risultante dal taglio di uno spigolo può essere calcolato con delle semplici formule iterative .

STEP 3 (*ottimizzazione sull'albero*)

L'ottimizzazione (euristica) sull'albero si basa sul fatto esiste una corrispondenza 1-1 fra le partizioni in $\Pi_p(T)$ ed i $(p-1)$-sottoinsiemi di spigoli di T, ed è tale che i $p-1$ spigoli in un sottoinsieme sono precisamente quelli esterni nella corrispondente partizione, cioè sono gli spigoli che devono essere tagliati dalla struttura iniziale per ottenere quella partizione.

Se viene indicata col nome di *base* ogni $(p-1)$-sottoinsieme di spigoli di T, allora uno ***scambio*** (o *pivoting*) consiste nel sostituire uno spigolo nella base con un altro spigolo fuori base.

Un semplice algoritmo di ricerca locale per il problema della classificazione su un albero T è il seguente. Si parte con la base corrispondente alla partizione ammissibile $\overline{\pi}$; ad ogni iterazione verrà indicata con B la base corrente. Se lo spigolo $u \in B$ è sostituito con lo spigolo $v \notin B$, si ottiene una nuova base B'. In corrispondenza ci sarà una variazione $\Delta_v^u = h(\pi) - h(\overline{\pi})$ nella funzione obiettivo, dove π e $\overline{\pi}$ sono le partizioni corrispondenti rispettivamente a B ed ha B'. Se $\Delta_v^u < 0$ si effettua lo scambio e la procedura iterativa continua, se $\Delta_v^u \geq 0$ per tutti gli spigoli $u \in B$ e $v \notin B'$ allora STOP : la partizione corrente π è l'ottimo cercato.

STEP 4 (*modificazione dell'albero*)

Sia T un albero di G e $\pi \in \Pi_p(T)$: si assume che π venga ottenuta attraverso l'algoritmo di scambio descritto precedentemente.

Si selezioni uno spigolo esterno u. I suoi vertici terminali siano i ed j appartenenti a due differenti classi C_h e C_k.

Si considera ora la nuova partizione $\overline{\pi}$ ottenuta partendo dalla π facendo migrare il vertice i dal gruppo C_h a C_k; si dirà in questo caso che $\overline{\pi}$ è ***adiacente*** a π . La migrazione causerà una variazione della funzione obiettivo; se $\overline{\pi}$ è strettamente migliore di π, $\overline{\pi}$ è ammissibile in T : ovvio in quanto π è stata ottenuta attraverso un algoritmo di scambio. La procedura tenta allora di effettuare un intervento di "chirurgia plastica" sull'albero T allo scopo di ottenere un nuovo albero $\overline{T}$ in cui $\overline{\pi}$ sia ammissibile.
È utile, a tal fine, vedere su un piccolo esempio l'essenza di questa procedura.
Prima di descrivere in dettaglio la procedura è necessario introdurre alcune notazioni. Ogni coppia (T,π) induce sul grafo G una partizione degli spigoli in quattro tipologie :

α-*spigolo* : spigolo esterno non appartiene all'albero

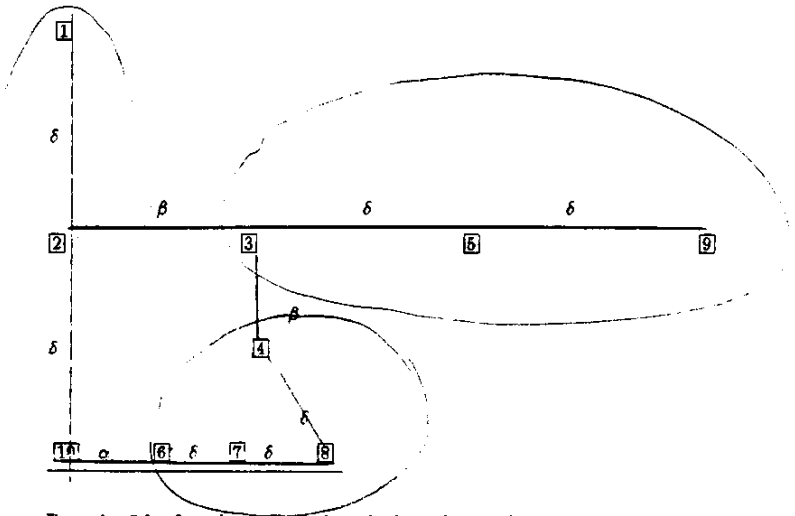
β-*spigolo* : spigolo esterno appartenente all'albero

γ-*spigolo* : spigolo interno non appartenente all'albero

δ-*spigolo* : spigolo interno appartenente all'albero

Per ciascuna tipologia di vertici è definito il suo ...-grado come il numero di ...-spigoli uscenti dal vertice stesso.

Ad esempio nella Fig. 1 è schematizzato un grafo G=(V,E) diviso in tre gruppi e sono indicate le diverse tipologie per gli spigoli :



Il vertice 5 ha δ-grado=2. Il vertice 2 ha b-grado=1 e δ-grado =2.

LEMMA . Se $\bar{\pi}$ è adiacente a π, una condizione necessaria e sufficiente perche' $\bar{\pi}$ sia ammissibile in T è che risulti

a) il δ-grado del vertice x che migra da un cluster all'altro e' 1

b) esiste un β-spigolo che connette x con qualche altro vertice del gruppo C_j in cui x emigra (si noti che il β-spigolo e' necessariamente unico).

Viene ora descritta la procedura di modificazione dell'albero. Come precedentemente indicato , sia x il vertice candidato a migrare da C_i a C_j ed u uno spigolo esterno di G che connette x con qualche altro nodo y di C_j.

MODIFICAZIONE DELL'ALBERO

Step 1 . Se il δ-grado di x è maggiore di 1 passa allo step successivo, altrimenti si va allo Step 3.

Step 2 . Siano $v_1,.....,v_d$ (d>1) i δ-spigoli incidenti x ed S l'albero parziale di T indotto dalla classe C_i. Il grafo $S - x$ è una foresta con d componenti connesse $S_1,S_2,.....,S_d$. Si tratta ora di trovare un γ-spigolo v_0 che connetta S_1 con qualche altro S_k (a tale scopo viene usata una procedura di etichettamento per individuare tale γ-spigolo): Se non si trova alcun γ-spigolo si passa allo Step 9; altrimenti il vertice v_0 viene dichiarato δ-spigolo mentre v_1 γ-spigolo cosi' che il δ-grado di x diminuisce di una unita'. Si ripete lo Step 2 eventualmente fino a che il δ-grado di x non diventi 1. Successivamente si passa allo Step 3.

Step 3 . Sia $v=(x,z)$ l'unico δ-spigolo incidente x. Se il β-grado di x è zero si passa allo Step 6; altrimenti si passa allo step successivo.

Step 4 . Se u è un β-spigolo, v viene dichiarato β-spigolo e si passa allo Step 8. Se u è un α-spigolo e c'è un β-spigolo w che connette x con qualche y' di C_j allora si va allo Step7 , altrimenti passa allo Step 5.

Step 5 . Si indichi con P il cammino (unico) che collega x con y sull'albero T (questo cammino puo' individuarsi facilmente attraverso una procedura di etichettamento). Se P include il δ-spigolo v allora si passa allo Step 6 altrimenti P includerà un β-spigolo w incidente x : in quest'ultimo caso si passa allo Step 7.

Step 6 . Si dichiara u come δ-spigolo, v un α-spigolo e si passa allo Step 8.

Step 7 . Si dichiara u come δ-spigolo, v un β-spigolo , w un γ-spigolo e si passa allo Step 8.

Step 8 . Ciascun γ-spigolo incidente x è dichiarato come α-spigolo; ciascuno α-spigolo differente da u che connette x con qualche $y'' \in C_j$ viene dichiarato come γ-spigolo. La tipologia di ogni altro spigolo incidente x non cambia. Il processo di modificazione dell'albero è completato ed un nuovo albero $\overline{T}$ viene ad essere ammissibile in $\overline{\pi}$.

Step 9 . L'algoritmo non riesce a trovare un albero $\overline{T}$ in cui $\overline{\pi}$ sia ammissibile e la migrazione di x non puo' avvenire.

In quest'ultimo caso, si tenta di verificare la possibilita' di migrare di y da C_j a C_i. Se anche in questo caso l'algoritmo non riesce a modificare l'albero allora si tenta con un altro spigolo esterno u' e si ripete la procedura. Se complessivamente non avviene nessuna migrazione di vertici significa che non c'è nessuna partizione $\overline{\pi}$ che migliora quella di partenza π, che resta percio' la partizione finale.

5 APPLICAZIONI

5.1 Definizione del piano degli esperimenti

Sono stati presi in esame cinque grafi , rispettivamente: a) Grafo TOSCANA b) Grafo ROMA c) Grafo CAMPANIA d) Grafo LAZIO e) Grafo ALTOLAZIO e nella tabella seguente sono fornite sinteticamente le caratteristiche generali dei cinque grafi.

Grafo	numero vertici	numero spigoli	numero caratteristiche	$\rho = \frac{\text{spigoli}}{\text{vertici}}$
TOSCANA	287	773	1	2.69
ROMA	121	302	1	2.50
CAMPANIA	81	165	6	2.04
LAZIO	412	566	1	1.37
ALTOLAZIO	210	331	5	1.58

Allo scopo di avere un elemento di comparazione sono stati messi a confronto i risultati ottenuti con un altro algoritmo di classificazione vincolata AGRAF , di L. Lebart (1978). A differenza di *MIDAS* AGRAF e' un algoritmo ascendente gerarchico.

3.2 Indicatori di performance

Per un confronto fra le diverse partizioni e fra i due algoritmi, MIDAS e AGRAF, sono stati presi in esame i seguenti indici, per ogni predefinito numero di partizioni p (indice che viene per semplicità omesso)

$\%_i = \frac{h(\pi_i)}{h(\pi_0)} \cdot 100$ Percentuale d'inerzia interna corrispondente alla partizione π_i iniziale per l'algoritmo MIDAS [$h(\pi_0)$ è la funzione normalizzante relativa ad un gruppo unico, partizione triviale].

$\%_f = \frac{h(\pi_f)}{h(\pi_0)} \cdot 100$ Percentuale d'inerzia interna corrispondente alla partizione π_f finale . (algoritmo MIDAS o AGRAF).

$\Delta_m = \frac{h(\pi_f)}{h(\pi_i)} \cdot 100$ Guadagno percentuale di inerzia in MIDAS.

$\Delta = \frac{h(\pi_f^a) - h(\pi_f^m)}{h(\pi_f^a)} \cdot 100$ Guadagno percentuale fra le partizioni di MIDAS (π_f^m) e quelle di AGRAF (π_f^a)

Pivot = numero di scambi effettuati nella fase di ottimizzazione sull'albero da MIDAS.

Migr = numero di scambi fra cluster contigui effettuati nella fase di modificazione dell'albero (MIDAS).

Tree = numero di alberi generati (MIDAS).

Tempi di CPU relativi al sistema 1100/20 Univac con sistema opertativo Exec8.

4.3 Analisi e discussione dei risultati

Nell'appendice sono riportate le cinque tabelle relative al confronto puntuale fra le caratteriche di performance dei due algoritmi.

Da questetabelle si possono evincere alcune note generali :

- In tutti i casi contemplati l'algoritmo AGRAF è dominato da MIDAS nel senso che se si prende come criterio di bontà di una partizione π la funzione $h(\pi)$ allora in tutti i casi esaminati, ovviamente a parità di numero di partizioni, risulta

$$h(\pi^a) \geq h(\pi^m)$$

valendo il segno = nel solo caso di p=3 per il grafo CAMPANIA. Un solo caso sui 90 esaminati !

- Si verifica inoltre la circostanza che già l'algoritmo greedy di MIDAS costruisce partizioni migliori, nel senso di cui sopra, rispetto alla partizione finale ottenuta da AGRAF. Nei succitati 90 casi questo si verifica per ben 77 volte. Solo per 14 casi si ha $h(\pi_i)$ di MIDAS inferiore alla $h(\pi_f)$ di AGRAF; rispettivamente per il grafo ALTOLAZIO quattro volte nei casi di p=13, 14, 15 e 16; nel grafo Roma cinque volte nei casi p=3,5,6,7 e 8; nel grafo LAZIO ancora quattro volte per p=4,5,10 e 11 mentre un solo caso, per p=20 , nel grafo della TOSCANA.

La fase di migrazione fra gruppi, relativamente all'algoritmo MIDAS, è sempre significativa per quanto attiene i guadagni d'inerzia, pur se questi sono inferiori in generale a quelli relativi all'algoritmo che la precede. I guadagni medi sono,in ordine crescente, 2.47, 4.99, 5.08, 5.42 e 10.16 rispettivamente per i grafi CAMPANIA,TOSCANA; ALTOLAZIO, LAZIO e ROMA.

-E' noto come negli algoritmi di tipo non-gerarchico si presenti anche il problema della scelta del numero dei gruppi. A tal fine un indicatore interessante è il "guadagno d'inerzia" fra due successive partizioni $v_p = \Delta_m(p\text{-}1) - \Delta_m(p)$. L'andamento di questa funzione del numero dei gruppi (p) mostra abitualmente un forte decremento iniziale per poi presentare uno o piu' massimi relativi per valori di p compresi nell'intervallo $3 \div 20$. Questo fatto "legittima" le motivazioni che hanno spinto a prendere in esame questo intervallo per il numero dei gruppi.

-Δ sembra correlato con ρ. Infatti i guadagni medi, cioè sulle 18 partizioni effettuate per ciascun grafo, sono nell'ordine 6.64, 6.66, 8.04, 22.34 e 66.65 rispettivamente per LAZIO (ρ=1.37), CAMPANIA (ρ=2.04), ALTOLAZIO (ρ=1.58), TOSCANA (ρ=2.69) e ROMA (ρ=2.5). Come si puo' notare, pur

non rispettando fedelmente l'ordine di ρ questo dato medio mette in evidenza una sostanziale correlazione positiva (0.64) tra i due indici.

- il numero di pivot è generalmente basso e sembra essere correlato con il numero di vertici quando il numero di gruppi richiesto è elevato . Per $p \leq 10$ il numero di pivot è molto basso. Se ordiniamo i grafi in ordine crescente del numero dei vertici abbiamo la sequenza CRATL (Campania Roma Altolazio Toscana Lazio) e rispettivamente 0.39 ,0.26 ,0.85 ,0.77 e 0.89 sono le correlazioni tra il numero dei pivot ed il numero delle parti nei quali è ripartito il grafo in questione. I valori medi del numero dei pivot sono, rispettivamente 1.22, 1.44, 2.33, 3.72 e 6.39 in rigoroso ordine con la numerosità dei vertici del grafo.

- il numero delle migrazioni sembra correlato, cioè cresce, al crescere del numero di vertici del grafo. Infatti se ordiniamo i grafi secondo valori del numero di migrazioni crescenti, otteniamo la sequenza dei grafi CRATL (Campania Roma Altolazio Toscana Lazio) che è , a meno di uno scambio fra Toscana e Lazio, lo stesso ordinamento dei grafi per numerosità dei vertici.

3.4 Tempi di calcolo

E' interessante anche il confronto fra i tempi di esecuzione dei due algoritmi, pur tenendo conto della differenza fondamentale che riconosce AGRAF basato su un algoritmo di tipo ascendente gerarchico mentre MIDAS è basato su uno di tipo non-gerarchico. Il confronto è stato fatto utilizzando lo stesso computer, un Univac 110/20. AGRAF fornisce, per il grafo ALTOLAZIO, un tempo complessivo di 17.6 sec di CPU per fornire i risultati relativi a tutte le diciotto partizioni in p classi ($3 \leq p \leq 20$). Per quanto riguarda MIDAS sono stati fatti diciotto "passoggi" diversi essendo l'algoritmo non gerarchico. Nella Tavola 1 sono riportati i tempi di calcolo per i diversi valori di p relativamente al grafo dell'ALTOLAZIO mentre nella Tavola 2 sono iportati quelli relativi al grafo della CAMPANIA .

Tav. 1 Tempi di esecuzione di MIDAS (in secondi). Grafo ALTOLAZIO

numero di classi	tempo di esecuzione	numero di classi	tempo di esecuzione
3	5.89	12	18.03
4	6.61	13	15.08
5	5.20	14	10.39
6	7.23	15	11.76
7	18.54	16	12.15
8	8.45	17	25.63
9	10.64	18	22.80
10	10.90	19	21.85
11	12.11	20	25.13

Tav. 2 Tempi di esecuzione di MIDAS (in secondi). Grafo CAMPANIA

numero di classi	tempo di esecuzione	numero di classi	tempo di esecuzione
3	5.89	12	18.03
4	6.61	13	15.08
5	5.20	14	10.39
6	7.23	15	11.76
7	18.54	16	12.15
8	8.45	17	25.63
9	10.64	18	22.80
10	10.90	19	21.85
11	12.11	20	25.13

BIBLIOGRAFIA

Benzécri J.-P. & F. (1980) *Pratique de L'Analyse Des Données*. Vol.1. Dunod - Paris.

Cerasoli M.,Eugeni, F., Protasi M. (1988) *Elementi di Matematica Discreta*. Zanichelli - Bologna.

Christofides, N. (1975) : *Graphe Theory: an algorithmic approach*. Academic Press, New York.

Gordon A.D. (1973) Classification in the presence of constraints, *Biometrics* - 29 -821 ÷ 827.

Gordon A.D. (1987) Classification and Assignment in Soli Science, *Soil Use and Management*-3-3 ÷ 8.

Lawler E. (1976) . *Combinatorial Optimization: Network and Matroid,* Holt, Rinehart & Winston.

Lebart L. (1978) . Programme d'aggrégation avec contraintes. *Les Cahiers de l'Analyse des Données,* vol. 3 -275 ÷ 287. Dunod, Paris.

Roche C. (1978) . Exemple de classification hiérarchique avec contraintes de contiguité . *Les Cahiers de l'Analyse des Données,* vol. 3 -289 ÷ 305. Dunod, Paris.

Maravalle M. e Simeone B. (1992) - A two-stage heuristic for optimal graph partitioning. Tecnical Report .

Murtagh F. (1985) A Survey of Algorithms for contiguity-constrained clustering and related problems. *The Computer Journal* - 28 - 82 ÷ 88.

APPENDICE

GRAFO CAMPANIA

caratteristiche generali: 81 vertici ; 165 spigoli ; 6 caratteristiche

	M I D A S						AGRAF	
p	l_i	l_f	Δ_m	pivot	migr	tree	l_f	Δ
3	67.418	67.418	0.00	0	0	1	67.418	0.00
4	61.006	61.006	0.00	0	0	1	62.120	1.83
5	55.786	55.786	0.00	0	0	1	56.567	1.40
6	50.825	50.825	0.00	0	0	1	51.104	0.55
7	46.446	44.865	3.52	1	15	2	47.230	5.28
8	43.609	43.442	0.38	0	1	2	44.495	2.42
9	40.407	40.241	0.41	0	1	2	40.672	1.07
10	36.234	36.118	0.32	4	3	2	37.90	4.93
11	33.915	31.708	6.96	3	6	2	35.617	12.33
12	32.037	29.950	6.97	2	8	2	33.48	11.82
13	30.185	28.380	6.36	1	13	2	31.67	11.59
14	28.662	27.963	2.50	2	11	2	30.122	7.73
15	27.444	26.743	2.62	3	11	2	29.770	11.33
16	26.296	25.525	3.02	2	14	2	27.600	8.15
17	25.260	24.488	3.15	1	15	2	27.09	10.62
18	24.362	23.739	2.62	1	15	2	25.96	9.35
19	23.452	22.828	2.83	1	17	3	25.09	9.95
20	22.609	21.987	2.83	1	15	3	24.08	9.55

GRAFO ROMA

caratteristiche generali: 121 vertici ; 302 spigoli ; 1 caratteristica

	MIDAS						AGRAF	
				performance				
p	t_i	t_f	Δ_m	pivot	migr	tree	t_f	Δ
3	24.453	21.950	20.51	0	30	2	22.05	0.456
4	15.702	13.344	17.67	0	18	2	15.882	19.04
5	13.938	12.366	12.71	0	5	2	12.937	4.61
6	11.866	10.307	15.12	2	5	2	11.20	8.63
7	10.388	8.886	16.90	3	15	2	9.27	4.39
8	8.516	7.161	18.92	3	5	2	8.50	18.71
9	7.038	5.740	22.61	3	15	2	7.21	25.61
10	6.080	4.827	25.95	2	26	3	6.79	40.58
11	4.890	4.816	1.5	1	21	2	6.63	37.77
12	4.270	4.258	0.28	0	27	2	6.28	47.42
13	3.918	3.905	0.33	0	27	2	5.47	40.25
14	3.573	3.561	0.33	1	25	2	5.78	62.36
15	3.222	3.210	0.37	0	27	2	5.62	75.08
16	2.957	2.945	0.41	0	19	2	5.51	87.41
17	2.733	2.721	0.44	0	17	2	5.40	98.53
18	2.161	1.904	13.5	5	12	2	5.33	179.94
19	1.937	1.680	15.29	2	11	2	5.31	216.07
20	1.583	1.583	0.00	4	1	2	5.26	232.91

GRAFO ALTOLAZIO

caratteristiche generali: 210 vertici ; 331 spigoli ; 5 caratteristiche

	MIDAS						AGRAF	
				performance				
p	t_i	t_f	Δ_m	pivot	migr	tree	t_f	Δ
3	82.823	80.702	2.63	1	15	2	83.21	3.10
4	77.900	76.071	2.40	1	14	2	81.40	7.00
5	73.307	71.509	2.51	0	11	2	77.34	8.15
6	69.811	66.278	5.33	0	13	2	75.86	14.45
7	66.236	62.192	6.50	3	33	4	71.688	15.27
8	63.648	60.344	5.47	0	24	2	68.558	13.62
9	61.079	57.972	5.36	1	21	2	65.04	12.196
10	58.907	57.580	2.30	1	15	2	61.842	7.40
11	56.974	55.625	2.43	1	15	2	58.912	5.915
12	55.273	52.802	4.68	2	26	3	56.	6.06
13	53.605	50.168	6.85	2	19	3	53.	5.64
14	51.978	47.469	9.66	3	26	2	50.	5.33
15	50.359	47.455	6.24	4	30	2	49.37	4.03
16	48.680	45.720	6.47	5	30	2	48.08	5.16
17	46.997	44.279	6.14	5	29	3	47.505	7.27
18	45.350	42.864	5.8	4	25	3	46.212	7.82
19	43.853	41.370	6.00	4	25	3	44.98	8.73
20	42.532	40.648	4.63	5	18	3	43.762	7.65

GRAFO LAZIO

caratteristiche generali: 412 vertici ; 566 spigoli ; 1 caratteristica

	MIDAS						AGRAF	
				performance				
p	l_i	l_f	Δ_m	pivot	migr	tree	l_f	Δ
3	63.667	61.448	3.61	1	12	2	64.05	4.25
4	58.830	57.022	3.17	1	18	2	57.27	0.43
5	53.225	51.665	3.02	1	27	2	52.08	0.83
6	48.392	47.063	2.82	1	23	2	49.09	4.30
7	45.219	44.037	2.68	1	18	2	46.52	5.6
8	42.301	41.094	2.94	1	24	2	42.7	3.9
9	33.404	30.695	8.82	9	62	2	33.80	10.10
10	31.225	29.144	7.14	9	72	2	30.64	5.3
11	30.541	28.613	6.74	8	62	2	29.73	3.9
12	28.312	26.430	7.12	8	82	3	28.8	8.9
13	27.622	26.243	5.25	7	46	2	28.1	7.00
14	26.502	25.064	5.73	7	49	2	27.28	8.8
15	25.514	24.337	4.83	8	96	3	26.43	8.6
16	24.560	23.175	5.97	8	103	3	25.62	10.5
17	23.677	22.178	6.76	12	102	3	24.18	9.02
18	23.008	21.529	6.87	10	105	3	23.47	9.01
19	22.395	20.943	6.93	11	99	3	22.92	9.4
20	21.789	20.337	7.14	12	105	3	22.29	9.6

GRAFO TOSCANA

caratteristiche generali: 287 vertici ; 773 spigoli ; 1 caratteristica

	MIDAS						AGRAF	
	performance							
p	t_i	t_f	Δ_m	pivot	migr	tree	t_f	Δ
3	60.858	60.839	0.03	0	3	2	75.9143	24.77
4	53.588	50.716	5.66	0	54	2	67.1956	32.50
5	47.433	44.535	6.51	0	54	2	58.7072	31.84
6	40.981	40.885	0.23	2	11	2	51.3704	25.66
7	37.7	36.872	2.24	2	32	2	48.3964	31.245
8	33.897	32.118	5.54	2	44	2	44.9969	40.01
9	31.392	29.617	5.99	2	44	2	41.6173	40.51
10	28.828	27.680	4.15	2	32	2	35.6539	28.79
11	26.866	25.717	4.47	2	32	2	32.5527	26.55
12	25.310	24.116	4.95	2	32	2	29.4923	22.01
13	23.921	22.783	4.99	2	32	2	27.0615	18.79
14	22.686	21.555	5.25	4	52	3	25.0418	16.19
15	21.451	20.442	4.93	3	52	3	23.1666	13.36
16	20.265	19.345	4.75	6	75	4	21.4699	11.01
17	19.097	17.331	10.19	13	139	4	19.8044	14.25
18	17.545	16.347	7.33	7	144	3	18.322	12.05
19	16.353	15.380	6.32	13	137	3	16.9157	9.99
20	16.175	15.225	6.24	5	92	3	15.6133	2.55

SU ALCUNI METODI PER RAZIONALIZZARE LE SCELTE FRA PIU' ALTERNATIVE VALUTATE CON CRITERI MULTIPLI QUANTITATIVI

Antonio MATURO, Giuseppina VARONE
Dip. di Scienze, Storia dell'Architettura e Restauro
Università "G. D'Annunzio", viale Pindaro, 42 - Pescara

SOMMARIO Si presentano alcuni metodi per scegliere quale alternativa, fra quelle appartenenti ad un dato insieme finito, soddisfa maggiormente ad un dato obiettivo.

A tale scopo ogni alternativa A viene valutata in base a piú criteri, qualitativi o quantitativi, che forniscono dei punteggi i quali individuano, rispettivamente, delle misure o delle relazioni di preordine fra le alternative.

Vengono poi presentati degli algoritmi che, in funzione dei punteggi ottenuti, permettono di assegnare ad ogni alternativa A un numero reale, compreso fra 0 e 1, il quale indica in che misura l'alternativa A deve essere considerata preferibile alle altre.

1. POSIZIONE DEL PROBLEMA

Vi sono n possibilitá di scelta, dette *alternative* o *oggetti*, e k *obiettivi* spesso in contrasto fra loro.

Il raggiungimento o meno di ciascuno degli obiettivi è valutato con uno o piú *criteri di scelta* o *variabili*, assegnando, per ogni criterio, un punteggio ad ogni alternativa che indica il " grado " di soddisfacimento dell'obiettivo da parte dell'oggetto.

I punteggi possono rappresentare una valutazione

(a) di tipo quantitativo se indicano una *misura*;

(b) di tipo qualitativo se sono utilizzati esclusivamente per stabilire una *relazione di preordine* fra gli oggetti, ossia una relazione riflessiva e transitiva che indichiamo con $\leq\cdot$.

Per ogni criterio supponiamo che la relazione sia *totale.*

Siano

$\mathcal{A} = \{ A_1, A_2,\ldots,A_n \}$ l'insieme delle alternative,

$\mathcal{V} = \{ C_1, C_2,\ldots,C_m \}$ l'insieme dei criteri qualitativi,

$\mathcal{W} = \{ D_1, D_2,\ldots,D_q \}$ quello dei criteri quantitativi.

Diciamo *punteggio* associato al criterio C_j, *compatibile* con $\leq\cdot$, ogni funzione $P_j: \mathcal{A} \to \mathbb{N}$ tale che, $\forall r,s \in \{ 1,2, \ldots.n \}$,

$$A_r \leq\cdot A_s \Rightarrow P_j(A_r) \leq P_j(A_s).$$

Il punteggio si dice *strettamente compatibile* con $\leq\cdot$ se $\Rightarrow$ è sostituito con $\Leftrightarrow$. Il problema di scelta risulta *implicitamente definito* da una matrice bipartita con n righe ed m+q colonne del tipo

$$
M = \begin{array}{c|c|c}
C_1 \; C_2 \ldots C_j \ldots C_m & D_1 \; D_2 \ldots D_r \ldots D_q & \\
\hline
e_{11}\; e_{12} \ldots e_{1j} \ldots e_{1m} & d_{11}\; d_{12} \ldots d_{1r} \ldots d_{1q} & A_1 \\
e_{21}\; e_{22} \ldots e_{2j} \ldots e_{2m} & d_{21}\; d_{22} \ldots d_{2r} \ldots d_{rq} & A_2 \\
\ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots & \vdots \\
e_{i1}\; e_{i2} \ldots e_{ij} \ldots e_{im} & d_{i1}\; d_{i2} \ldots d_{ir} \ldots d_{iq} & A_i \\
\ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots & \vdots \\
e_{n1}\; e_{n2} \ldots e_{nj} \ldots e_{nm} & d_{n1}\; d_{n2} \ldots d_{nr} \ldots d_{nq} & A_n \\
\hline
P_1 \; P_2 \ldots P_j \ldots P_m & Q_1 \; Q_2 \ldots Q_r \ldots Q_q &
\end{array}
$$

Diciamo *punteggio* associato al criterio quantitativo D_r ogni funzione

$$Q_r : \mathcal{A} \to [0, +\infty)$$

esprimente le misure attribuite alle varie alternative con il criterio D_r.

Nella matrice M risulta $e_{ij} = P_j(A_i)$, $d_{ir} = Q_r(A_i)$.

2. METODO DEL REGIME

Limitiamoci, per il momento, allo studio di problemi di scelta basati su criteri solo di tipo qualitativo.

La matrice M per q=0 si riduce alla seguente

$$E = \begin{array}{c|cccc|c} & C_1 & C_2 \ldots C_j \ldots & C_m & & \\ \hline & e_{11} & e_{12} \ldots e_{1j} \ldots & e_{1m} & | & A_1 \\ & e_{21} & e_{22} \ldots e_{2j} \ldots & e_{2m} & | & A_2 \\ & \ldots\ldots\ldots\ldots\ldots & & & | & \vdots \\ & e_{i1} & e_{i2} \ldots e_{ij} \ldots & e_{im} & | & A_i \\ & \ldots\ldots\ldots\ldots\ldots & & & | & \vdots \\ & e_{n1} & e_{n2} \ldots e_{nj} \ldots & e_{nm} & | & A_n \\ \hline & P_1 & P_2 \ldots P_j \ldots & P_m & & \end{array} \quad , \quad e_{ij} = P_j(A_i).$$

Siano

$\underline{a}_i = (e_{i1}, e_{i2}, \ldots, e_{im})^t$, (detto vettore associato ad A_i),

$\underline{c}_j = (e_{1j}, e_{2j}, \ldots, e_{nj})^t$, (detto vettore associato a C_j).

Poniamo, $\forall a \in \mathbb{R}$, $s(a) = \begin{cases} 0 \text{ per } a=0 \\ \dfrac{a}{|a|} \text{ per } a \neq 0. \end{cases}$

Per ogni $\underline{a} = (a_1, a_2, \ldots, a_r)^t \in \mathbb{R}^r$ poniamo

$$s(\underline{a}) = \left(s(a_1), s(a_2), \ldots\ldots, s(a_r) \right)^t.$$

Siano inoltre $\underline{d}_{hk} = \underline{a}_h - \underline{a}_k$ (detto vettore delle differenze per la coppia (A_h, A_k)) e $\underline{r}_{hk} = s(\underline{d}_{hk})$ (detto vettore di regime per la coppia (A_h, A_k)).

Chiamiamo *matrice delle differenze* quella avente per colonne i vettori delle differenze $\underline{d}_{hk}$ con $h \neq k$, ossia la matrice

$$= \left[\underline{d}_{12}\ \underline{d}_{13}\cdots\underline{d}_{1n} \mid \underline{d}_{21}\ \underline{d}_{23}\cdots\underline{d}_{2n} \mid \cdots\cdots \mid \underline{d}_{11}\ \underline{d}_{12}\ \cdots\cdots \right.$$

$$\left. \underline{d}_{1(1-1)}\ \underline{d}_{1(1+1)}\cdots\underline{d}_{1n} \mid \cdots \underline{d}_{n1}\ \underline{d}_{n2}\cdots\underline{d}_{n(n-1)} \right].$$

Chiamiamo *matrice di regime* quella avente per colonne i vettori di regime $\underline{r}_{hk}$ con $h \neq k$, ossia la matrice

$$R = \left[\underline{r}_{12}\ \underline{r}_{13}\cdots\underline{r}_{1n} \mid \underline{r}_{21}\ \underline{r}_{23}\cdots\underline{r}_{2n} \mid \cdots\cdots \mid \underline{r}_{11}\ \underline{r}_{12}\cdots\cdots \right.$$

$$\left. \underline{r}_{1(1-1)}\ \underline{r}_{1(1+1)}\cdots\underline{r}_{1n} \mid \cdots \underline{r}_{n1}\ \underline{r}_{n2}\cdots\underline{r}_{n(n-1)} \right].$$

Diciamo *peso* ogni funzione

$$W : \left\{ C_1, C_2, \ldots C_m \right\} \rightarrow [0,1]$$

tale che $W(C_i) \leq W(C_j) \Leftrightarrow$ il criterio C_i è considerato " non più importante" di C_j.

Poniamo $W_i = W(C_i)$, $\underline{W} = (W_1, W_2, \ldots, W_m)^t$.

Il concetto di "non piú importante" viene precisato se è assegnata una relazione $\preceq$ di preordine fra i criteri. In tal caso consideriamo il peso come un vettore aleatorio con supporto $[0,1]^m$ tale che $C_i \preceq C_j \Leftrightarrow W_i \leq W_j$.

I pesi si possono ottenere nel seguente modo.

Si considera una variabile casuale X con supporto [0,1] ed un campione casuale di ampiezza m, della X, $\underline{X} = (X_1, X_2, \ldots, X_m)$.

Si pone W_i uguale alla statistica ordinale di ordine m-i+1.

In particolare $W_1 \geq W_2 \geq W_3 \geq \ldots\ldots \geq W_m$.

Si ha che, se F(x) è la funzione di ripartizione della X, allora la W_i ha funzione di ripartizione

$$G_i(y) = \sum_{\ell=m-i+1}^{m} \binom{m}{\ell} \left[F(x)\right]^{\ell} \cdot \left[1-F(y)\right]^{m-\ell}.$$

Se la X ha distribuzione UC [0,1] allora, derivando, si vede che la W_i ha densitá

$$g_i(y) = \binom{m}{m-i+1} (\, m-i+1 \,) \cdot y^{m-i} (1-y)^{i-1},$$

ossia W_i è una $\beta(m-i+1,\ i)$, variabile casuale beta di parametri $(m-i+1,\ i)$, dove

$$\beta\ (r,s) = \int_0^1 x^{r-1}(1-x)^{s-1}\, dx = \frac{\Gamma(r)\ \Gamma(s)}{\Gamma(r+s)}$$

è l'integrale euleriano di prima specie.

Sia $\underline{r}_{is} = (\, \sigma_{is1}\ ,\ \sigma_{is2}\ ,\ \ldots\ldots, \sigma_{ism}\ )^t$.

Si ammette che la scelta fra le alternative A_i e A_s sia in base al valore assunto dalla funzione aleatoria

$$V_{is} = \sum_{j=1}^{m} \sigma_{isj}\ W_j$$

esprimente il "grado di soddisfacimento" nella scelta di A_i rispetto ad A_s.

La matrice V, di termine generale V_{ls}, si ottiene dalla matrice R, di regime, e dal vettore casuale W ponendo

$$V = W^t \cdot R.$$

La matrice V è un vettore riga con n(n-1) colonne.

Sia (A_l, A_s) una coppia di alternative distinte.

Diciamo *probabilita' per la dominanza* di A_l rispetto ad A_s il numero

$$p_{ls} = \text{prob}\ (V_{ls} > 0).$$

Diciamo matrice delle *probabilita' di dominanza* la seguente

$$p = \begin{array}{c|cccccc} & A_1 & A_2 & \cdots & A_s & \cdots & A_n \\ \hline A_1 & 0 & p_{12} & \cdots & p_{1s} & \cdots & p_{1n} \\ A_2 & p_{21} & 0 & \cdots & p_{2s} & \cdots & p_{2n} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_l & p_{l1} & p_{l2} & \cdots & p_{ls} & \cdots & p_{ln} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_j & p_{j1} & p_{j2} & \cdots & p_{js} & \cdots & p_{jn} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_n & p_{n1} & p_{n2} & \cdots & p_{ns} & \cdots & 0 \end{array}$$

L'idea di base è di considerare una alternativa A_l preferibile all'alternativa A_j se il vettore riga i della matrice P ha gli elementi, disposti in ordine decrescente, rispettivamente maggiori o uguali a quelli, anch'essi disposti in ordine decrescente, della riga j.

In generale, in questa maniera si ottiene una relazione d'ordine non totale.

Per ottenere un ordinamento totale si puó sostituire ad ogni vettore riga una media che puó essere la media aritmetica, la media geometrica (se le probabilitá sono tutte positive), la media quadratica o la mediana.

La media piú utilizzata è la media aritmetica, ossia ad ogni alternativa A_i si associa il numero

$$p_i = \frac{1}{n-1} \sum_{j=1}^{n} p_{ij}$$

3. CRITERI DI VALUTAZIONE DELLE p_{ls}

Per calcolare la p_{ls} occorre determinare, in $\mathbb{R}^m$, la zona A soddisfacente il sistema di disequazioni

$$\begin{cases} \sum_{j=1}^{m} \sigma_{lsj} \cdot W_j \geq 0 \\ W_j \geq 0, \ \forall j \in \{ 1,2,\ldots,m \} \\ W_j \geq W_{j+1}, \ \forall j \in \{ 1,2,\ldots,m-1 \} \end{cases}$$

In corrispondenza si ottiene

$$p_{ls} = \int_A g(w_1, w_2, \ldots, w_m) \cdot dw_1 \cdot dw_2 \ldots dw_m$$

con g funzione di densità di probabilità della variabile casuale m-dimensionale (W_1, W_2, ...,W_m).

Nel caso in cui $\underline{X}$ = (X_1, X_2,....,X_m) ha distribuzione uniforme la g (W_1, W_2,.....W_m) è uniforme nella zona S soddisfacente le disequazioni

$$W_j \geq 0, \forall j \in \{ 1,2,....,m \}$$

$$W_j \geq W_{j+1}, \forall j \in \{ 1,2,....,m-1 \}$$

per cui, dette m(A) e m(S) le misure di A ed S, risulta

$$p_{ls} = \frac{m(A)}{m(S)} .$$

I calcoli risultano alquanto elaborati già per m≥4.

Per m elevato, in pratica, è necessario calcolare le p_{ls} con il metodo di Montecarlo.

Si genera una successione di lunghezza N di campioni casuali empirici

$$\underline{x}^{(r)} = \left(x_1^{(r)}, x_2^{(r)},.....,x_m^{(r)} \right), \quad r=1,2,...N$$

determinazioni indipendenti della variabile casuale X.

Se $\underline{y}^{(r)} = \left(y_1^{(r)}, y_2^{(r)},.....,y_m^{(r)} \right)$ è la permutazione ottenuta

da $\underline{x}^{(r)}$ tale che $y_i \geq y_{i+1}$, $\forall i<m$, allora si pone

$$v_{is}^{(r)} = \sum_{j=1}^{m} \sigma_{isj}\, y_j^{(r)}.$$

Posto $N_p = \left| \left\{ r \in \{1,2,\ldots,N\} : v_{is}^{(r)} > 0 \right\} \right|$, si assume

$$p_{is} = \text{prob}\left(V_{is} > 0 \right) \cong \frac{N_p}{N}.$$

4. UN ESEMPIO

Antonio vuole acquistare una macchina scegliendo una delle seguenti

A1) ALFA 33

A2) AUDI 80

A3) PEUGEOT 405

I criteri per la scelta sono i seguenti

C_1) ripresa

C_2) comoditá

C_3) costo

C_4) estetica

Valutando i dati esposti da riviste specializzate, ad esempio QUATTRORUOTE, GENTE MOTORI, L'AUTOMOBILE etc..., si arriva stabilire la seguente matrice dei punteggi

$$E = \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & C_3 & C_4 \end{array} \\ \left[\begin{array}{cccc} 4 & 1 & 3 & 1 \\ 2 & 2 & 1 & 3 \\ 1 & 5 & 4 & 2 \end{array}\right] \end{array} \begin{array}{c} A_1 \\ A_2 \\ A_3 \end{array}$$

Supponiamo di avere la seguente relazione di preordine

$$C_1 > C_2 > C_3 > C_4.$$

Sia $W_1 = W(C_1)$. Allora $W_1 > W_2 > W_3 > W_4$ e W_i è uguale alla statistica ordinale di ordine 5–i.

La matrice di regime è la seguente

$$R = \begin{array}{c} \begin{array}{cccccc} r_{12} & r_{13} & r_{21} & r_{23} & r_{31} & r_{32} \end{array} \\ \left[\begin{array}{cccccc} 1 & 1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & -1 \end{array}\right] \end{array}.$$

Il vettore $V = W^t \cdot R$ è

$$V = \left[V_{12} = W_1 - W_2 + W_3 - W_4,\ V_{13} = W_1 - W_2 - W_3 - W_4,\ V_{21} = -V_{12},\right.$$

$$\left. V_{23} = W_1 - W_2 - W_3 + W_4,\quad V_{31} = -V_{31},\quad V_{32} = -V_{23} \right].$$

Le probabilitá di dominanza sono

$$p_{12} = \text{prob}\left(V_{12} > 0 \right) = \text{prob}\left(W_1 - W_2 + W_3 - W_4 > 0 \right),$$

$$p_{13} = \text{prob}\left(V_{13} > 0 \right) = \text{prob}\left(W_1 - W_2 - W_3 - W_4 > 0 \right),$$

$$p_{23} = \text{prob}\left(V_{23} > 0 \right) = \text{prob}\left(W_1 - W_2 - W_3 + W_4 > 0 \right).$$

Assumiamo l'ipotesi che la variabile casuale X abbia distribuzione uniforme continua in [0,1].

Allora la densitá di probabilitá congiunta è uniforme nella zona S soddisfacente le disequazioni

$$W_j \geq 0, \quad \forall j \in \{1,2,3,4\}$$

$$W_j \geq W_{j+1}, \quad \forall j \in \{1,2,3\}.$$

Poniamo $W^*_{j-1} = W_j / W_1$. Si puó verificare che la terna (W^*_1, W^*_2, W^*_3) ha distribuzione uniforme nella zona S^* soddisfacente le disequazioni

$$W^*_1 \geq 0, \; W^*_2 \geq 0, \; W^*_3 \geq 0$$

$$W^*_1 > W^*_2 > W^*_3.$$

Come si puó verificare con semplici operazioni geometriche, il volume della zona S^* è 1/6.

Le probabilitá di dominanza si possono esprimere come

$$p_{12} = \text{prob}\left(1 - W_1^* + W_2^* - W_3^* > 0 \right),$$

$$p_{13} = \text{prob}\left(1 - W_1^* - W_2^* - W_3^* > 0 \right),$$

$$p_{23} = \text{prob}\left(1 - W_1^* - W_2^* + W_3^* > 0 \right).$$

Poichè $W_1^* < 1$, $W_3^* < W_2^*$, l'espressione $1 - W_1^* + W_2^* - W_3^*$ è positiva, $\forall (x,y,z) \in S^*$ e quindi $p_{12}=1$.

La zona D intersezione di S^* con il semispazio di equazione $1 - W_1^* - W_2^* - W_3^* > 0$ è un tetraedro avente volume 1/36.

Poichè la distribuzione delle terne (W_1^*, W_2^*, W_3^*) è uniforme in S^* risulta

$$p_{13} = \frac{\text{volume}(D)}{\text{volume}(S^*)} = 1/6.$$

La zona H intersezione di S^* con il semispazio $1 - W_1^* - W_2^* + W_3^* > 0$ è un tetraedro avente volume 1/12. Segue che

$$p_{23} = \frac{\text{volume}(H)}{\text{volume}(S^*)} = 1/2.$$

La matrice delle probabilitá di dominanza è la seguente

	A_1	A_2	A_3
A_1	0	1	1/6
A_2	0	0	1/2
A_3	5/6	1/2	0

Utilizzando il criterio della media aritmetica i numeri

associati alle alternative A_1, A_2, A_3 sono rispettivamente

$$p_1 = 1/12, \qquad p_2 = 1/4, \qquad p_3 = 2/3,$$

per cui le alternative vanno ordinate nel seguente modo

$$A_3 > A_1 > A_2.$$

La macchina scelta è la PEUGEOT 405.

BIBLIOGRAFIA

1. Albers, L.H. *Het gewichtloze gewogen, cultuurhistorische betekenis van landgoederen geevalueerd met behulp van multicriteria analyse.* Delftse Universitaire Pers,Delft,1987.

2. Albers, L., P. Nijkamp. "*Multimensional analysis for plan or project evalutation; how to fit the right method to the right problem*". Procedures of the conference at Capri, Napoli, April 1988.

3. Dall'Aglio G. "*Calcolo delle probabilitá*", Zanichelli, Bologna.

4. Delft, A. van and Nijkamp, P. (1977). *Multicriteria Analysis and Regional Decision-Making,* The Hague/Boston: Martinus Nijhoff.

5. Hinloopen, E., Nijkamp, P. and Rietveld, P. (1983). "Qualitative discrete multiple criteria choice models in regional planning", *Regional Science and Urban Economics* 13: pp 77-102.

6.Hinloopen, E. (1985). *De Regime Methode,* M.A. Thesis, Interfaculty Actuariat and Econometrics, Free University, Amsterdam (minimeographed)

7. Hinloopen, E. and Smyth, A.W. (1985). "A description of

principles of a new multicriteria evalutation tecnique, The Regime Method", in *Proceedings Colloquium Vervoerplanologisch Speurwerk,* Delft, pp. 422-431.

8. Keeney, R. and Raiffa, H. (1976). *Decision with Multiple Objectives, Preferences and Value Tradeoffs.* Neww York: Wiley.

9. Kmietowicz, Z.W. and Pearman, A.D. (1981). *decision Theory and Incomplete Knowledge.* Aldershot: Gower.

10. Lancaster, K. (1971). *Consumer Demand.* New York: Columbia University Press.

11. Lootsma, T.A. (1980). "Saaty's Priority Theory and the nomination of a senior professor in operation research", *European Journal of Operational Research* 4: pp.380-388.

12. Mastenbroek, P and Paelinck, J.H.P. (1977). "Qualitative multicriteria analysia-applications to airport location", *environment and Planning* A 9(8): pp. 883-895.

13. Nijkamp, p. and Voogd, H. (1981). "New multicriteria methods for physical planning by means of multimensional scaling tecniques", pp.19-30 in Haimes, Y. and Kindler, J. (eds.), *Water and Related Land Resource System.* Oxford: Pergamon Press.

14. Nijkamp, P., Leitner H. and Wrigley N. (eds.) (1985). *Measuring the Unmeasurable,* Dordrecht: Martinus Nijhoff.

15. Rietveld, P. (1980) *Multi Objective Methods and Regional Planning,* Amsterdam: North-Holland Publ.Co.

16. Saaty, T.l. (1977). "A scaling method for priorities in hierarchical structures", *journal of Mathematical Psychology* 15: pp.234-281

17. Taha, H. A. (1976). *Operations Research,* New York: Macmillan.

18. Voogd, H. (1983). *Multicriteria Evalutation for Urban and Regional Planning.* London: Pion

VIRUS INFORMATICI: NATURA E RIMEDI

Giorgio MESSI

Si parla spesso di protezione delle informazioni da attacchi ispirati da interessi economici. La minaccia costituita dai virus, essendo il risultato di motivazioni diverse, merita una trattazione a sé stante. Si inizierà con il delineare uno scenario generale per arrivare, in conclusione, a consigliare alcuni metodi di protezione.

1. Introduzione

L'impiego di personal computer per lo svolgimento delle più svariate attività (commerciali, di ricerca, industriali, etc.) si allarga ogni giorno ad un nuovo settore. Questo diffondersi della cultura informatica comporta un flusso continuo di nuovi utenti; come ben ricorda chiunque usi un computer, l'approccio iniziale con la macchina ha aspetti traumatici; domina la paura (dopo i primi inevitabili errori) di perdere i dati, di "toccare il tasto sbagliato". Via via che si apprendono le nozioni basilari questa paura si ridimensiona, salvo riemergere quando si è di fronte a qualcosa che non si conosce.

I virus informatici hanno tutte le caratteristiche necessarie per suscitare timore, anzi direi quasi terrore, in coloro che fanno uso di personal computer.

Viene spontaneo il parallelismo con l'AIDS, come un qualcosa di cui non si sa cos'è, come si diffonde, cosa fa di preciso.

Succede allora che se il monitor video si surriscalda qualcuno pensa possa essere colpa di un virus.

Fino a pochi anni fa se un computer non funzionava le possibilità erano due: o la macchina era guasta, o c'era un errore umano nell'immissione dei dati. Adesso invece viene spontanea in mente una terza possibilità: il virus.

2 La storia

Si può risalire alle origini del termine "virus" in campo informatico: esso fu attribuito per primo nel 1983 da Fred Cohen, ricercatore americano, ad un tipo di software capace di autoriprodursi senza alcun comando da parte dell'utilizzatore.

Nell'agosto dell'anno successivo, alla National Computer Security Conference, Cohen illustrò le proprie intuizioni ed insistette perché fosse aperto un pubblico dibattito sulla materia.

Non tutti i partecipanti alla conferenza erano però d'accordo riguardo questa pubblicità: si temeva la diffusione di idee pericolose. Purtroppo però queste idee erano già nell'aria, e di lì a breve si iniziarono a registrare i primi casi di "infezione" (è un termine che useremo anche in seguito) da virus informatico.

La definizione di Cohen si basava su concetti conosciuti già da tempo, anche se con canoni diversi. Infatti già intorno ai primi anni 70 Victor Vyssotsky, impiegato nei Laboratori Bell, aveva inventato un gioco di nome Darwin, dal nome dello scienziato della teoria dell'evoluzione. Questo gioco simulava l'evoluzione di programmi in un ambiente ostile; vinceva il concorrente capace di continuare a riprodursi sbaragliando i suoi avversari.

Nel 1984 A.K. Dewdney ha introdotto un gioco basato sullo stesso principio con un nuovo nome "CORE WARS" ed esiste tuttora un'associazione internazionale che organizza tornei annuali tra programmi "gladiatori" che si affrontano all'ultimo bit.

Purtroppo la facile adattabilità di queste idee dal campo ludico a quello del normale impiego professionale, assieme alla diffusione di massa del personal computer, hanno portato, in breve tempo, al verificarsi di numerosi casi di crimine informatico. In una società in cui atti di esibizionismo e fenomeni vandalici fanno parte della cronaca di tutti i giorni, i virus altro non sono che l'equivalente in campo informatico.

3 Elementi fondamentali di un virus

Scendiamo per un momento nei dettagli tecnici, per spiegare COS'E' un virus. Per virus si intende, in senso stretto, un programma con due caratteristiche fondamentali: la prima è la capacità di *riprodursi*, ossia di attaccarsi ad un altro programma e propagarsi da esso; la seconda è relativa all'*azione* attraverso cui il virus si manifesta, azione che comporta di solito un danno.

Esistono programmi che difettano di una di queste due proprietà, ma che vengono comunemente definiti virus: programmi che compiono danni senza però sapersi riprodurre (Cavalli di Troia, Catene di lettere, Bombe a tempo), o all'inverso programmi (worms = vermi) che si riproducono soltanto senza compiere alcuna azione dannosa (in effetti poi il danno deriva dalle risorse del sistema progressivamente saturate da questa riproduzione).

Analizziamo ora COSA FA un virus. Va' premesso che data la grande varietà di virus esistenti (diverse centinaia) non tutti compiono esattamente le stesse azioni; è però possibile schematizzare alcune linee di comportamento comuni.

Un virus arriva solitamente attraverso un supporto magnetico: può trattarsi di un dischetto contenente programmi acquistato da terzi, oppure dato in prestito da un nostro amico che non sa di essere già "infetto" : può inoltre essere trasmesso direttamente se il computer fa parte di una rete locale, o si collega ad altri computer attraverso le linee telefoniche mediante un modem.

La prima fondamentale attività di un virus è la stessa di ogni programma al momento di andare in esecuzione: **caricarsi nella RAM.**

Perché ciò avvenga il virus deve essere letto o all'inizio di un programma che lo contiene (che viene normalmente definito infetto) o in particolari zone (BOOT SECTOR) del disco usato per caricare i files di sistema al momento dell'accensione.

Una volta caricato in memoria il virus può restarvi fino allo spegnimento della macchina come un qualsiasi programma TSR[1] oppure restarvi solo fino al compimento delle azioni descritte in seguito.

La seconda caratteristica è la **diffusione**: alcune istruzioni del programma indicano le modalità di duplicazione, che interessano -come già detto- alcune zone fisiche dei supporti magnetici oppure i files eseguibili (COM EXE); in casi più rari anche i file di Overlay (OVL) i SYS (drivers) e particolari tipi di dati (DBF)(DB3)[2].

Non si hanno notizie di files di dati diversi da quelli nominati che siano stati oggetto di "infezione" (ad esempio documenti redatti con un word processor).

Terza caratteristica è la verifica di una **condizione** (di solito controllo della data o contatore incrementato ad ogni utilizzo). Ciò permette al virus di esistere in un sistema, senza manifestarsi, per un certo periodo di tempo, durante il quale ne approfitta per diffondersi.

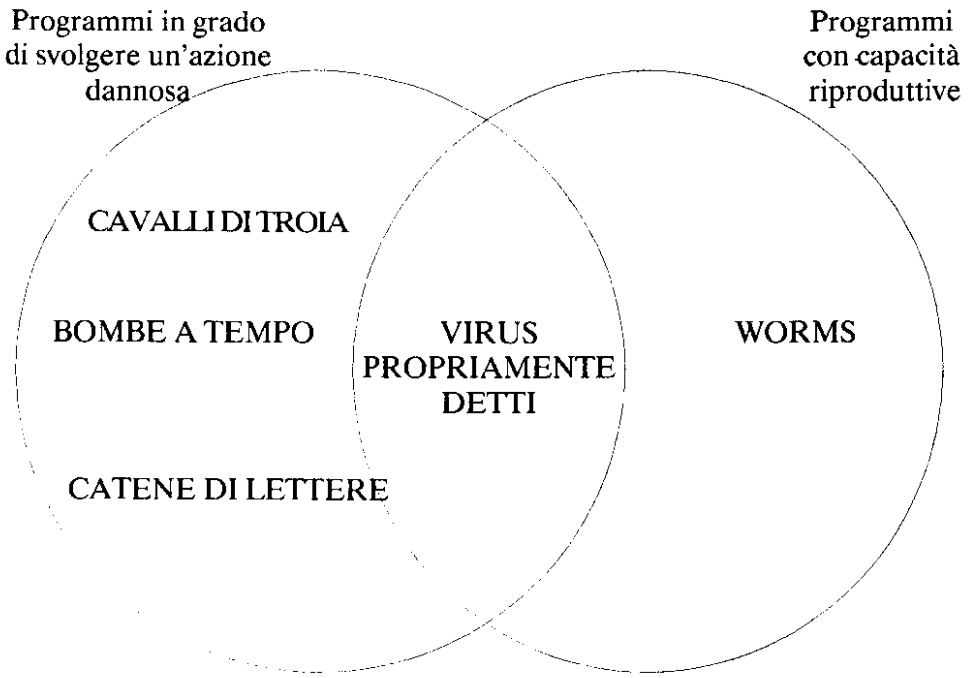
Soddisfatta la condizione di cui si è parlato il virus compie un'**azione**, di solito distruttiva e comunque dannosa: fastidiosi messaggi che vanno e vengono, blocco di programmi durante l'esecuzione, perdita di dati e, in alcuni casi, la cancellazione totale: ciò equivale a dire che il massimo danno di un virus equivale al comando "FORMAT" del DOS.

Finora si è descritto il problema: una volta compreso appieno, è possibile studiarne le soluzioni. Lo schema che segue riassumerà sinteticamente quanto detto :

(1) Sono così definiti i programmi che, una volta eseguiti e terminato il loro impiego, lasciano nella memoria volatile una porzione di sé stessi che non può essere eliminata. Ad esempio: il file di configurazione della tastiera (keybit.com), il Sidekick, etc.

(2) In questo contesto si attribuisce alla voce "dati" il significato più ristretto, cioè il risultato del lavoro di un utente, mentre ove non altrimenti specificato si intenderà per "dati" tutto ciò che è contenuto nell'elaboratore

Classificazione dei programmi "virali"



4 Prevenzione e rimedi

Il primo consiglio non è una novità: effettuare backup dei dati a diversi livelli (su scala mensile; su scala settimanale; su scala giornaliera; naturalmente a seconda dei ritmi di lavoro) ed inoltre non usare, se si lavora su dati di una certa importanza, software non originale o di provenienza sconosciuta.

Un altro metodo di prevenzione, in presenza di ambienti multiutente (reti), può essere l'attenta scelta dei diritti di accesso (di ogni singolo utente) alle risorse comuni. Le reti di comunicazione sono organizzate di solito in maniera gerarchica ed è perciò facoltà del supervisore prevenire "infezioni" più o meno dolose: in un'azienda ad esempio, in cui sia installata una rete di PC, si possono regolamentare i diritti del singolo utilizzatore, impedendogli di usare floppy disk personali che potrebbero contenere virus. Tipicamente ciò avviene non consentendo l'esecuzione di programmi da dischetto, permettendo soltanto la lettura-scrittura di dati[3]. Nelle reti collegate via modem, dove si scambiano quotidianamente programmi di libero dominio, si pone in atto un accorgimento: i programmi che i singoli "scaricano" sull'elaboratore centrale subiscono una quarantena, un periodo in cui vengono testati per verificarne l'"immunità".

Si è già detto che il virus, nelle diverse fasi della sua "vita", modifica in qualche modo i dati contenuti nel computer. Un modo per scoprirlo può essere quindi confrontare, ad intervalli periodici, il contenuto del disco rigido (in particolare i file eseguibili) con copie di riserva effettuate al momento dell'installazione, per questo sicuramente immuni. Tale confronto dovrebbe riguardare sia le dimensioni -e ciò risulta relativamente facile- sia il contenuto dei file, visto che alcuni virus evoluti riescono a conservare immutata la lunghezza dei file da loro infetti. Quest'ultimo tipo di controllo risulta meno agevole; è però possibile, per accelerare entrambi, l'impiego della crittografia. Per chi non conosca il significato di tale termine, si esporranno in breve i chiarimenti necessari alla comprensione di quanto segue, tralasciando una trattazione più completa dell'argomento che esula dagli scopi del presente lavoro.

(3) Per il significato del termine "dati" vedere la nota (2) al paragrafo 3

Crittografia deriva da due termini greci, criptos (nascosto) e grafia (scrittura), significa quindi scrittura segreta, incomprensibile a chi non abbia la chiave per decifrarla. La crittografia nel corso dei secoli ha avuto importanza soprattutto in ambito militare, o come passatempo per una ristretta elite; dalla fine della seconda guerra mondiale ad oggi si è invece sviluppata come una scienza, che cammina di pari passo con la matematica. Agli originali metodi di impiego, che consistevano nel "camuffare" un messaggio per non permetterne la comprensione se non al destinatario, se ne sono aggiunti di nuovi ed è di questi che ci interesseremo.

L'innovazione sta nel poter "marchiare" un messaggio, in modo che non sia possibile a terzi cambiarne un solo bit a nostra insaputa. Operando su di un programma, in questo modo, delle operazioni crittografiche otteniamo come risultato una stringa alfanumerica di controllo. E' assai facile a questo punto memorizzarla per un successivo riscontro .

La prevenzione dai virus si ottiene perciò controllando il più frequentemente possibile (l'ideale sarebbe perdere qualche minuto ad ogni accensione) l'hard disk; in caso di "infezione" viene subito ravvisata la modifica del file (o BOOT SECTOR) ed è possibile limitare al massimo i danni, eliminando la fase di diffusione latente del virus. Esiste già in commercio software che effettua controlli di questo tipo: produce per ogni file, al momento dell'installazione, una chiave di identificazione che viene confrontata, ogni volta, con un valore calcolato in quel momento con uguale procedura.

Anche il tradizionale uso della crittografia può essere d'aiuto nel difendersi dall'attacco di virus. E'necessario però a tale fine conservare i propri dati sotto forma cifrata: quando il virus tenta di "infettare" il programma, cercando di aggiungersi alle istruzioni in esso contenute, non riesce a leggere tali istruzioni e fallisce perciò nel suo intento. Ai vantaggi che un simile sistema di protezione può comportare si contrappongono però degli svantaggi, soprattutto in termini di tempo: per ogni operazione di lettura-scrittura di dati è necessaria infatti la relativa cifratura; procedure di questo tipo risultano perciò proponibili solo in caso si lavori con informazioni di carattere altamente riservato.

La ricerca dei virus può avvenire anche attraverso particolari programmi che ricercano, in ogni file, le istruzioni che contraddistinguono i virus già conosciuti. Tale ricerca non ha naturalmente esito, se il programma non conosce il virus in questione perchè "nuovo di zecca" (per questo vengono rilasciate nuove versioni del programma su scala mensile). Occorre inoltre fare attenzione: sono già due volte che versioni del programma devirus più

diffuso si rivelano "infette".

Al termine di questa trattazione è giunto il momento di parlare dei rimedi: se esistono dei dubbi sulla presenza di virus è opportuno cercare di identificarlo con appositi software (magari non usando l'ultimissima versione appena arrivata da una BBS americana per non correre i rischi di cui si è appena detto), quindi usare un software di rimozione virus per eliminare il problema.

Nella maggioranza dei casi, qualsiasi utente che abbia una discreta conoscenza dell'MS-DOS è in grado, utilizzando tali programmi con le accluse istruzioni, di cavarsela da solo. E' comunque buona norma, se possibile, la creazione di un backup prima dello svolgimento di un tale lavoro, ed è inoltre opportuno rivolgersi a personale specializzato se i dati compromessi sono di valore e l'utente ha scarsa esperienza.

BIBLIOGRAFIA

Estratto da: "La crittografia: aspetti teorici ed applicativi della matematica per il trattamento di dati riservati" Tesi di Laurea, 1991.

CALCULATION OF THE POLARIZATION PARAMETERS FOR ELECTROMAGNETIC FIELDS

Tommaso SCOZZAFAVA
Dipartimento di Ingegneria Elettrica
Università degli Studi dell'Aquila

SUMMARY: A new method utilizing the "affinity", a concept of the projective geometry, is exposed in order to identify the polarization ellipse of an assigned electromagnetic field.

1. INTRODUCTION

In the study of the electromagnetic fields the problem may arise of calculating the parameters of the so-called "polarization ellipse", namely the lengths and directions of the semiaxes of such ellipse, with respect to an assigned reference frame. To clarify this problem, first we will recall how that ellipse originates.

From now on, we indicate the complex numbers by a dot typed over their symbols; because the phasors and their related operators are time-invariable, no confusion is possible with the time derivative operator, needless in the present work. For the physic vectors we assume boldface symbols. As the polarization phenomenon involves both electric and magnetic fields, is sufficient to consider only one of them, e.g. the electric one.

2. THE POLARIZATION ELLIPSE

We define "harmonic field" an electric field whose cartesian coordinates are sinusoidal function of the time, with the same angular frequency ω. The general expression of such a field is

$$\mathbf{e}(t) = \mathbf{u}_x\, e_x(t) + \mathbf{u}_y\, e_y(t) + \mathbf{u}_z\, e_z(t) \tag{1}$$

where $\mathbf{u}_k$ (with $k = x,y,z$) are unit vectors and

$$e_k(t) = E_k \cos(\omega t + \alpha_k) \tag{2}$$

are the components, whose peak values are indicated by E_k and the phases by α_k.

If we introduce the phasors

$$\dot{E}_k = E_k\, e^{j\alpha_k} \qquad (k = x,y,z)$$

corresponding to the scalar fields (2), in the frequency domain the harmonic vector (1) becomes a *complex vector*

$$\dot{\mathbf{E}} = \mathbf{u}_x \dot{E}_x + \mathbf{u}_y \dot{E}_y + \mathbf{u}_z \dot{E}_z \tag{3}$$

from which, as it is well-known, the physical real vector (1) is obtainable by the algorithm

$$\mathbf{e}(t) = Re\left(\dot{\mathbf{E}}\, e^{j\omega t}\right) . \tag{4}$$

It is also well-known [1] that another representation is possible for the complex vector (3): if we expand the cosines in eqn. (2) and let

$$E_{Rk} = Re\,(\dot{E}_k), \qquad E_{Jk} = Im\,(\dot{E}_k),$$

we may define two *constant and real* vectors

$$\begin{aligned} \mathbf{E}_R &= \mathbf{u}_x E_{Rx} + \mathbf{u}_y E_{Ry} + \mathbf{u}_z E_{Rz} \\ \mathbf{E}_J &= \mathbf{u}_x E_{Jx} + \mathbf{u}_y E_{Jy} + \mathbf{u}_z E_{Jz} \end{aligned} \tag{5}$$

by means, after some algebraic manipulation, the vector (1) may assume the expression

$$\mathbf{e}(t) = \mathbf{E}_R \cos \omega t + \mathbf{E}_J \cos(\omega t + \pi/2) \tag{6}$$

whose representation in the frequency domain is:

$$\dot{\mathbf{E}} = \mathbf{E}_R + j\, \mathbf{E}_J . \tag{7}$$

The relation (6) is more interesting than its equivalent (1) and (2) because it points out that the vector $\mathbf{e}(t)$ moves always lying on a stationary plane defined by the two stationary vectors (5); this important result is not easy to draw directly from eqns. (1) and (2). Besides, it is also apparent that the vector $\mathbf{e}(t)$, defined as sum of three harmonic vectors, coincides also with the sum of the following only two harmonic vectors

(8) $\mathbf{e}_R(t) = \mathbf{E}_R \cos \omega t, \qquad \mathbf{e}_J(t) = \mathbf{E}_J \cos(\omega t + \pi/2),$

whose spatial directions form the angle

(9) $$\vartheta = \arccos \frac{\mathbf{E}_R \cdot \mathbf{E}_J}{E_R \; E_J} ;$$

it is noteworthy that, whatsoever their spatial angle ϑ is, their representation on the complex plane always form a right angle, as $\mathbf{e}_R(t)$ lags $\pi/2$ behind $\mathbf{e}_J(t)$.

The trajectory described by the tip of the vector $\mathbf{e}(t)$ is an ellipse, because it is the composition of the harmonic motions of the two tips of vectors (8); the field $\mathbf{e}(t)$ is said to be "elliptically polarized". This ellipse becomes a circle if $E_R=E_J$ and $\vartheta=\pi/2$ (field "circularly polarized"), or may degenerate into a segment if $\vartheta=0$, or $E_R=0$, or $E_J=0$ (field "linearly polarized").

The phenomenon of the elliptical polarization is very important because an electric (or magnetic) polarized field does stress the medium in a more complicate way than a variable field with constant direction does. In many problems, especially if the medium is anisotropic, it may be necessary to determine the maximum and the minimum stresses (and their respective directions) caused by a polarized field.

3. THE GEOMETRICAL PROBLEM

Generally speaking, a couple of stationary vectors like (5) completely defines a polarized field by means of the eqn. (6) and identifies the corresponding ellipse. The maximum and minimum values of the vector magnitude $e(t)$ coincide with the semiaxes of this ellipse, whose equation is readily obtainable in an "affine plane" eliminating the variable ωt from the moduli of eqns. (8):

(10) $$\frac{e_R^2}{E_R^2} + \frac{e_J^2}{E_J^2} = 1 ;$$

from this equation it is easy to recognize that: the ellipse is centered on the origin of the frame, whose x and y axes, not necessarily orthogonal, form the angle ϑ of eqn. (9) and contain,

respectively, the vectors $\mathbf{E}_R$ and $\mathbf{E}_J$ of eqns. (5). These vectors are a couple of conjugated semidiameters, that is, the polus at infinity of a diameter identifies the direction of the other: in other words, if we consider a parallelogram circumscribing the ellipse, each couple of opposite sides is parallel to the diameter passing through the tangential points of the other couple of sides.

After all, the physical problem of calculating the maximum and minimum values (which we will name e_{max} and e_{min}) of modulus of $\mathbf{e}(t)$ coincides with the geometrical problem of calculating the semiaxes of an ellipse whose only a generic couple of conjugated semidiameters is known. This problem is classically solved by resorting to the "metric" property of the ellipse [2,3]; this approach, however, leads to intricate mathematical developments, very tedious to follow.

4. THE PROJECTIVE GEOMETRY APPROACH

The flexibility of projective geometry enables us to choose as reference frame a cartesian plane, which is an affinity plane with the x and y axes forming a more conventional angle of $\pi/2$. Let us suppose we know the vectors (5) whose origin is the same of the reference frame; in order to simplify the notation, let us indicate the vectors (5) by $\mathbf{M}$ and $\mathbf{N}$ and the cartesian coordinates by x and y, rather than e_x and e_y suggested by the physical problem.

If the aforesaid vectors are known, also the implicit equations of their lines are known, that is:

$$(11) \qquad f_1(x,y) = y - mx = 0, \qquad f_2(x,y) = y - nx = 0.$$

whose angular coefficients are, respectively, m and n. The ellipse we are looking for belongs to the sheaf of conics identified by the double lines of the conjugated diameters involution [4]; then their parametric equation has the form

$$(12) \qquad f_1^2 + h\, f_2^2 + k = 0$$

where h and k are two constants to determine by imposing to the ellipse to have $\mathbf{M}$ and $\mathbf{N}$ as semidiameters. Afterwards, let us write

eqn. (12) in the canonical form

$$a_{11}x^2 + 2\,a_{12}x\,y + a_{22}y^2 + a_{33} = 0 \quad (13)$$

then the following bilinear transformation holds for the angular coefficients of conjugated diameters:

$$n = -\frac{a_{11} + a_{12}\,m}{a_{12} + a_{22}\,m} \quad (14)$$

Now we impose the orthogonality of the two diameters:

$$m\,n = -1 \quad (15)$$

which means that these two generic diameters must become the two axes; combining (14) with (15) it easy to obtain the solutions:

$$\mu,\nu = \frac{a_{22} - a_{11} \pm \sqrt{(a_{11} - a_{22})^2 + 4\,a_{12}^2}}{2\,a_{12}} \quad (16)$$

The two lines having equations

$$y = \mu\,x \quad \text{and} \quad y = \nu\,x \quad (17)$$

allow us to obtain the values $e_{max}=|\mathbf{e}_{max}|$ and $e_{min}=|\mathbf{e}_{min}|$ as distances of their intersections with the ellipse (13) from the origin of coordinates.

5. A NUMERICAL EXAMPLE

In order to make clear the plainness of the proposed approach, we will apply them to a practical case.

Let us calculate the maximum and minimum values of a polarized field identified by two vectors **M** and **N**, having their origin O placed in the origin of a cartesian frame and their tips placed in the points C and D whose coordinates are, respectively, (2.5; 1) and (-1; 1.5); as consequence, the lines (11) on which the vectors lay have, respectively, equations $f_1=y-0.4x$ and $f_2=y+1.5x$, so the eqn. (12) becomes

$$(y - 0.4\,x)^2 + h\,(y + 1.5\,x)^2 + k = 0 \;;$$

by imposing the passage for the points C and D we obtain h=0.16 and k=-3.61, so that the parametric equation (13) of the ellipse becomes

$$0.52\,x^2 - 0.32\,xy + 1.16\,y^2 - 3.61 = 0 \tag{18}$$

with $a_{11}=0.52$, $a_{12}=-0.16$ and $a_{22}=1.16$. Then, resorting to (16), we get $\mu=-4.236$ and $\nu=0.236$, that is, the lines on which the semiaxes lay have equations:

$$y = 0.236\,x \tag{19}$$

$$y = -4.236\,x\ . \tag{20}$$

Let AO and BO the semiaxes; introducing in eqn.(18), respectively, the eqns.(19) and (20) we obtain the coordinates $x_A= 2.663$, $y_A= 0.629$ of the point A and $x_B= -0.399$, $y_B= 1.690$ of the point B; as consequence, the semiaxes lengths are:

$$e_{max}= \sqrt{x_A^2+y_A^2} = 2.736; \qquad e_{min}= \sqrt{x_B^2+y_B^2} = 1.736\ .$$

Finally, we have obtained the moduli of the maximum and minimum stresses acting, respectively, along the directions (19) and (20); the corresponding angles they form with the positive versus of the x axis are:

$$\text{arc tg } 0.236 = 0.232 \text{ rad} \cong 13.3 \text{ deg};$$

$$\text{arc tg } (-4.236) + \pi = 1.803 \text{ rad} \cong 103.3 \text{ deg}.$$

BIBLIOGRAPHY

[1] F.Barozzi, F.Gasparini: *Fondamenti di Elettrotecnica: Elettromagnetismo;* U.T.E.T., Torino, 1989.

[2] L.M.Magid: *Electromagnetic Fields, Energy and Waves:* John Wiley & Sons, New York, 1972.

[3] H.Mott: *Polarization in Antennas and Radar;* John Wiley & Sons, New York, 1986.

[4] O.Chisini: *Lezioni di Geometria Analitica e Proiettiva;* Zanichelli, Bologna, 1970.

SUI CODICI UNIDIREZIONALI

Luca TALLINI

Oltre ad una breve descrizione della problematica sui codici Unidirezionali è quì data la dimostrazione di una congettura di S. Al-Bassan e B. Bose che interviene nella teoria dei suddetti autori, sulla costruzione di codici ottimali nella classe dei codici bilanciati che si ottengono con il metodo della complementazione di Knuth.

1 INTRODUZIONE.

Durante la trasmissione dei dati su un canale è inevitabile che avvengano degli errori. Nel caso binario si possono classificare tre tipi di errori:

1. **errori simmetrici**: il tipo di errore è detto simmetrico quando durante la trasmissione si possono avere sia errori $0 \rightarrow 1$ che errori $1 \rightarrow 0$. I canali caratterizzati da questo tipo di errori sono i Canali Binari Simmetrici (BSC).

2. **errori asimmetrici**: il tipo di errore è detto asimmetrico quando durante la trasmissione si possono avere solo errori $1(0) \rightarrow 0(1)$ e mai errori $0(1) \rightarrow 1(0)$. I canali caratterizzati da questo tipo di errori sono i Canali Asimmetrici ($ZC(\overline{Z}C)$).

3. **errori unidirezionali**: il tipo di errore è detto unidirezionale quando durante la trasmissione si possono avere sia errori $0 \rightarrow 1$

che errori $1 \rightarrow 0$, ma mai simultaneamente nella stessa parola di dati. I canali caratterizzati da questo tipo di errori sono i Canali Unidirezionali (UC).

Mentre sono ben noti alla letteratura esempi di BSC, esempi di canali fisici che obbediscono propriamente al modello di errore asimmetrico e/o unidirezionale sono: le Fibre Ottiche, i Dischi ottici, le PROMs (Programmable Read Only Memories), le Semiconductor memories, i circuiti e le memorie VLSI dei computers.

Il problema è quello di ottenere una trasmissione affidabile di dati sui canali su menzionati, codificando l'informazione con codici efficienti che correggano o almeno riconoscano errori, ovvero garantiscano la correttezza del dato ricevuto.

Piú precisamente ci si propone di risolvere i seguenti due tipi di problemi:

1. **Riconoscimento degli errori (Error Detection)**: questo problema si propone di stabilire se, nella parola di dati ricevuta sono stati commessi o meno errori.

2. **Correzione degli errori (Error Correction)**: questo problema si propone di correggere gli eventuali errori della parola di dati ricevuta.

I codici efficienti che offrono i servizi di correzione e riconoscimento di errori sui canali Asimmetrici e/o Unidirezionali sono detti codici Unidirezionali.

In seguito parleremo di codici binari a lunghezza fissa.

Nell'ambito dei codici riconoscitori di errori su UC e/o ZC, fino ad oggi, sono state studiate le seguenti quatro classi di codici:

1. **AUED** (All Unidirectional Error Detecting): è l'insieme dei codici che incondizionatamente riconoscono se la parola di dati in output al canale è affetta da errori unidirezionali o meno.

2. **t-UED** (t-Unidirectional Error Detecting): è l'insieme dei codici che riconoscono se la parola di dati in output al canale è affetta da errori unidirezionali, sotto l'ipotesi che il numero di bit errati è minore od uguale a t [4].

3. **b-BUED** (b-Burst Unidirectional Error Detecting): è l'insieme dei codici che riconoscono se nella parola di dati in output al canale è avvenuto un burst di errori unidirezionali di lunghezza minore od uguale a b, *sotto l'ipotesi che sia avvenuto un burst di errori* unidirezionali di lunghezza minore od uguale a b (Data una parola di dati in output X, si dice che è avvenuto un burst di errori di lunghezza b, se i bit erronei di X sono confinati in b bit consecutivi dei quali il primo e l'ultimo sono in errore) [2].

4. **t-UBED** (t-Unidirectional Byte Error Detecting): Si dice che sono avvenuti degli errori unidirezionali di byte in una parola di m byte di b bit se sono avvenuti degli errori unidirezionali in alcuni byte della parola. t-UBED è l'insieme dei codici che riconoscono se nella parola di dati in output al canale sono avvenuti degli errori unidirezionali di byte, sotto l'ipotesi che il numero di byte errati è minore od uguale a t [5].

2 METODO DI KNUTH PER I CODICI BILANCIATI.

Diamo la seguente:

Definizione 2.1 *Un codice $\mathcal{C}$ si dice bilanciato se ogni parola di $\mathcal{C}$ ha lunghezza n e peso $\left\lceil \frac{n}{2} \right\rceil \left(\left\lfloor \frac{n}{2} \right\rfloor \right)$. $\mathcal{C}$ si dice bilanciato con k bit di informazione e r bit di controllo se:*

1. $|\mathcal{C}| = 2^k$,
2. *Ogni parola di $\mathcal{C}$ ha lunghezza $k+r$ e peso $\left\lceil \frac{k+r}{2} \right\rceil \left(\left\lfloor \frac{k+r}{2} \right\rfloor \right)$.*

I codici bilanciati sono particolari codici AUED. Infatti, se una parola di codice è affetta da errori unidirezionali, il peso necessariamente deve cambiare, di modo che, se il codice è bilanciato, il decodificatore, contando il numero degli 1 della parola ricevuta, riesce a stabilire se sono stati commessi errori o meno.

Per quanto segue abbiamo bisogno delle seguenti notazioni:

1. $\mathcal{I} \stackrel{\text{def}}{=} \mathbb{Z}_2^k$ è l'insieme delle parole di informazione.
2. $\mathcal{A} \stackrel{\text{def}}{=} \mathbb{Z}_2^r$ è l'insieme delle parole di controllo.
3. Data $X \in \mathbb{Z}_2^k$, $w(X)$ è il peso di X.
4. Data $X \in \mathcal{I}$, $X^{(i)}$ è la parola X complementata dei primi i bit.

Se $X \in \mathcal{I}$, l'insieme:

$$\{(0, w(X^{(0)})), (1, w(X^{(1)})), (2, w(X^{(2)})), \ldots, (k, w(X^{(k)}))\},$$

descrive una passeggiata aleatoria da $w(X)$ a $k - w(X)$. Poiche:

$$\forall X \in \mathcal{I} \quad \left\lceil \frac{k}{2} \right\rceil \left(\left\lfloor \frac{k}{2} \right\rfloor \right) \in [w(X), k - w(X)],$$

per ogni X esiste sicuramente un indice i tale che:

$$w(X^{(i)}) = \left\lceil \frac{k}{2} \right\rceil \left(\left\lfloor \frac{k}{2} \right\rfloor \right).$$

A partire da questa osservazione è possibile codificare la generica parola $X \in \mathcal{I}$ al seguente modo. Sia $i \in [0, k]$ il piú piccolo indice per cui $w(X^{(i)}) = \left\lfloor \frac{k}{2} \right\rfloor$, codificando i con una parola di controllo Y bilanciata, ovvero tale che $w(Y) = \left\lceil \frac{r}{2} \right\rceil$, possiamo far corrispondere a X la parola bilanciata $YX^{(i)}$.

Il codice appena proposto è un semplice esempio, certamente non ottimale, di come applicare il metodo di Knuth [6]. In generale si dovrà colmare lo sbilanciamento di $X^{(i)}$ con un opportuno sbilanciamento del check Y.

3 TEORIA DI AL-BASSAN E BOSE.

Nel 1989 Al-Bassam e Bose in [1] hanno costruito una teoria generale per lo schema su proposto, proggettando un codice bilanciato ottimale, nella classe di tutti i codici bilanciati che si ottengono con il metodo della complementazione di Knuth e hanno decodifica sequenziale (la decodifica si dice parallela se basta un solo confronto del check Y per decodificare, altrimenti si dice sequenziale). Cominciamo ad esporre la teoria generale su detta.

Sia $\mathcal{S}_i$ l'insieme di tutte le stringhe binarie di lunghezza k e peso i. Dati $Y \in \mathcal{A}$, $w_1, w_2, \ldots, w_q, v \in \{0, \ldots, k\}$, con la notazione:

$$Y :: \mathcal{S}_{w_1} \cup \mathcal{S}_{w_2} \cup \ldots \cup \mathcal{S}_{w_q} \longrightarrow \mathcal{S}_v, \tag{3.1}$$

intenderemo che ogni parola di peso $w_1, w_2 \ldots$, o w_q (ovvero ogni parola in $\mathcal{S}_{w_1} \cup \mathcal{S}_{w_2} \cup \ldots \cup \mathcal{S}_{w_q}$), è codificata, complementando un bit alla volta, fino a che il peso sia v, e quindi appendendo, ad essa, la parola di controllo Y. Il codice che così si ottiene è bilanciato se $v = \left\lceil \frac{k+r}{2} \right\rceil - w(Y)$, per ogni parola di controllo $Y \in \mathcal{A}$. Il decodificatore, leggendo Y, il quale codifica $w_1, w_2 \ldots, w_q$, complementa il resto della parola di codice un bit alla volta, fino a che è ottenuto uno dei pesi $w_1, w_2 \ldots, w_q$. Ovviamente la mappa (3.1) deve essere ben definita e iniettiva affinché lo schema sia corretto.

D'ora in poi, in (3.1), supporremo, senza perdita di generalità, che $w_1 < w_2 < \ldots < w_q$.

Se $q = 1$, chiameremo la funzione (3.1), mappa singola, se $q = 2$, mappa doppia, e così via. Il seguente Teorema da delle condizioni necessarie e sufficienti perche una mappa singola sia ben definita e iniettiva.

Teorema 3.1 *La mappa singola* $Y :: \mathcal{S}_a \longrightarrow \mathcal{S}_v$ *è ben definita e iniettiva se, e solo se* $\min\{a, k-a\} \leq v \leq \max\{a, k-a\}$.

Per riferirci ad una mappa doppia, scriveremo:

$$Y :: \mathcal{S}_a \cup \mathcal{S}_b \longrightarrow \mathcal{S}_v,$$

essendo $b > a$. Si ha il seguente:

Teorema 3.2 *La mappa doppia* $Y :: \mathcal{S}_a \cup \mathcal{S}_b \longrightarrow \mathcal{S}_v$ *(dove* $b > a$*) è ben definita e iniettiva se, e solo se:*

$$b - a \geq \max\{v, k-v\}.$$

Per quanto riguarda le mappe triple, quadruple, ecc., notiamo che dal Teorema appena enunciato segue che se una mappa (3.1), per $q \geq 3$, è ben

definita e iniettiva allora

$$\forall i, j \in [1, q] \qquad |w_i - w_j| > \max\{v, k - v\} \geq \left\lceil \frac{k}{2} \right\rceil,$$

che, dovendo risultare $0 \leq w_i \leq k$, è impossibile.

Il problema è quindi ricondotto a trovare la miglior combinazione di mappe singole soddisfacenti il Teorema 3.1 e di mappe doppie soddisfacenti il Teorema 3.2.

Ci sono 2^r mappe corrispondenti alle 2^r parole di controllo disponibili. Se d parole di controllo sono usate per le mappe singole e $2^r - d$ per le mappe doppie, il numero di pesi differenti che possiamo riconoscere negli algoritmi di codifica e decodifica è $2(2^r - d) + d = 2^{r+1} - d$. Siccome dobbiamo riconoscere $k + 1$ pesi differenti, si deve avere:

$$k + 1 = 2^{r+1} - d \iff k = 2^{r+1} - d - 1.$$

È quindi chiaro che per massimizzare k dobbiamo minimizzare il numero delle mappe singole d. Nella costruzione di Bose, data in [3] d era uguale a $r + 1$, per cui in seguito supporremo $d \leq r + 1$.

Cerchiamo di caratterizzare qual'è la scelta ottimale per la combinazione di mappe. Diamo il seguente:

Teorema 3.3 *Dato $d \leq r + 1$, se esiste una combinazione di mappe:*

$$\begin{aligned}
&Y_1 :: S_{a_1} \longrightarrow S_{v_1}, \\
&Y_2 :: S_{a_2} \longrightarrow S_{v_2}, \\
&\vdots \\
&Y_d :: S_{a_d} \longrightarrow S_{v_d}, \\
&Y_{d+1} :: S_{a_{d+1}} \cup S_{b_{d+1}} \longrightarrow S_{v_{d+1}}, \\
&Y_{d+2} :: S_{a_{d+2}} \cup S_{b_{d+2}} \longrightarrow S_{v_{d+2}}, \\
&\vdots \\
&Y_{2^r} :: S_{a_{2^r}} \cup S_{b_{2^r}} \longrightarrow S_{v_{2^r}},
\end{aligned} \tag{3.2}$$

con $b_i > a_i$, per $i = d+1, \ldots, 2^r$, tale che:

$$\begin{gathered} \min\{a_i, k-a_i\} \leq v_i \leq \max\{a_i, k-a_i\} \quad \forall i \in [1,d], \\ b_i - a_i > \max\{v_i, k-v_i\} \quad \forall i \in [d+1, 2^r], \\ \{a_1 \ldots, a_{2^r}, b_{d+1}, \ldots, b_{2^r}\} = \{0, \ldots, k\}, \\ w(Y_i) = \left\lceil \tfrac{k+r}{2} \right\rceil - v_i \quad \forall i \in [1, 2^r], \end{gathered} \tag{3.3}$$

allora esiste anche una combinazione di mappe per cui valgono le (3.3) e:

$$\begin{cases} a_i \geq \left\lceil \frac{k}{2} \right\rceil \Longrightarrow v_i \geq \left\lceil \frac{k}{2} \right\rceil \\ a_i \leq \left\lfloor \frac{k}{2} \right\rfloor \Longrightarrow v_i \leq \left\lfloor \frac{k}{2} \right\rfloor \end{cases} \quad \forall i \in [1,d], \tag{3.4}$$

oppure:

$$\begin{cases} \begin{cases} a_i \geq \left\lceil \frac{k}{2} \right\rceil \Longrightarrow v_i \geq \left\lceil \frac{k}{2} \right\rceil \\ a_i \leq \left\lfloor \frac{k}{2} \right\rfloor \Longrightarrow v_i \leq \left\lfloor \frac{k}{2} \right\rfloor \end{cases} & \text{se } a_i \neq \left\lceil \frac{k+r}{2} \right\rceil + 1, \\ v_i = \left\lceil \frac{k-r}{2} \right\rceil & \text{se } a_i = \left\lceil \frac{k+r}{2} \right\rceil + 1. \end{cases} \tag{3.5}$$

Dimostrazione: Posto:

$$\begin{cases} v' \stackrel{\text{def}}{=} \left\lceil \frac{k-r}{2} \right\rceil \\ v'' \stackrel{\text{def}}{=} \left\lceil \frac{k+r}{2} \right\rceil \end{cases},$$

notiamo che, poiché:

$$\forall v, w \in [0,k] \quad \left| \left\lfloor \frac{k}{2} \right\rfloor - w \right| \leq \left\lceil \frac{k}{2} \right\rceil \leq \max\{v, k-v\},$$

e:

$$\forall v, w \in [0,k] \quad \left| \left\lceil \frac{k}{2} \right\rceil - w \right| \leq \left\lceil \frac{k}{2} \right\rceil \leq \max\{v, k-v\},$$

nella combinazione (3.2), dovendo valere le (3.3), si ha:

$$\left\lfloor \frac{k}{2} \right\rfloor, \left\lceil \frac{k}{2} \right\rceil \in \{a_1, \ldots, a_d\}.$$

Poiché:

$$|\{w \in \{a_i, b_i\}_{1=d+1,\ldots,2^r} : w < \left\lfloor \frac{k}{2} \right\rfloor\}| =$$

$$|\{w \in \{a_i, b_i\}_{1=d+1,\ldots,2^r} : w > \left\lceil \frac{k}{2} \right\rceil\}| = 2^r - d,$$

si ha:

$$|\{a \in \{a_i\}_{1=1,\ldots,d} : a < \left\lfloor \frac{k}{2} \right\rfloor\}| = |\{a \in \{a_i\}_{1=1,\ldots,d} : a > \left\lceil \frac{k}{2} \right\rceil\}| = \left\lfloor \frac{d-1}{2} \right\rfloor.$$

Notiamo inoltre che, se è possibile la mappa singola:

$$Y :: S_a \longrightarrow S_v,$$

poiché vale la prima relazione delle (3.3), si ha:

$$\begin{aligned} &\min\{a, k-a\} = \min\{k-a, a\} \leq \min\{v, k-v\} \leq \\ &\max\{v, k-v\} \leq \max\{k-a, a\} = \max\{a, k-a\}, \end{aligned} \tag{3.6}$$

e allora sono possibili anche le mappe:

$$\begin{aligned} &Y' :: S_a \longrightarrow S_{k-v}, \\ &Y'' :: S_{k-a} \longrightarrow S_v, \\ &Y''' :: S_{k-a} \longrightarrow S_{k-v}, \end{aligned}$$

se invece è possibile la mappa doppia:

$$Y :: S_a \cup S_b \longrightarrow S_v,$$

poiché vale la seconda relazione delle (3.3), si ha:

$$(k-a) - (k-b) = b - a > \max\{v, k-v\} = \max\{k-v, v\}, \tag{3.7}$$

e allora sono possibili anche le mappe:

$$\begin{aligned} &Y' :: S_a \cup S_b \longrightarrow S_{k-v}, \\ &Y'' :: S_{k-b} \cup S_{k-a} \longrightarrow S_v, \\ &Y''' :: S_{k-b} \cup S_{k-a} \longrightarrow S_{k-v}. \end{aligned}$$

Effettuiamo degli scambi, nelle mappe (3.2), tra elementi dell'insieme $\{S_{a_1}, \ldots, S_{a_{2^r}}, S_{b_{d+1}}, \ldots, S_{2^r}\}$, al modo che segue. Per $i \in [1,d]$, se $Y_i :: S_{a_i} \longrightarrow S_{v_i}$ è tale che $a_i \geq \left\lceil \frac{k}{2} \right\rceil$ distinguiamo i seguenti casi:

1. $v_i \in \left[\left\lceil \frac{k}{2} \right\rceil, v'' \right]$: Non effettuiamo nessuno scambio.

2. $v_i \in \left[v', \left\lfloor \frac{k}{2} \right\rfloor \right]$ ed esiste una mappa singola $Y_j :: S_{a_j} \longrightarrow S_{v_j} = S_{k-v_i}$ con $a_j \leq \left\lfloor \frac{k}{2} \right\rfloor$: In questo caso poniamo:

$$Y_i :: S_{a_j} \longrightarrow S_{v_i} \text{ e } Y_j :: S_{a_i} \longrightarrow S_{v_j}.$$

   Per la (3.6) è possibile fare ciò.

3. $v_i \in \left[v', \left\lfloor \frac{k}{2} \right\rfloor \right]$ e non esiste nessuna mappa come in 2.: Si distinguono i seguenti sottocasi:

   (a) $v_i \neq v'$: In questo caso, poiché $d \leq r+1$, esiste sicuramente una mappa doppia $Y_j :: S_{a_j} \cup S_{b_j} \longrightarrow S_{v_j} = S_{k-v_i}$. Allora poniamo:

$$Y_i :: S_{a_j} \cup S_{b_j} \longrightarrow S_{v_i} \text{ e } Y_j :: S_{a_i} \longrightarrow S_{v_j}.$$

   Per le (3.6) e (3.7) è possibile fare ciò.

   (b) $v_i = v'$: In questo caso, se è possibile rifarsi ai casi precedenti, allora è possibile far si che:

$$a_i \geq \left\lceil \frac{k}{2} \right\rceil \Longrightarrow v_i \geq \left\lceil \frac{k}{2} \right\rceil \text{ con } i \in [1,d].$$

   Se invece l'unica mappa $Y_j :: S_{a_j} \longrightarrow S_{v_j}$ con $v_j = k - v_i = v''$ è tale che $a_j \geq \left\lceil \frac{k}{2} \right\rceil$, allora è possibile solo fare in modo che:

$$\begin{cases} a_i \geq \left\lceil \frac{k}{2} \right\rceil \Longrightarrow v_i \geq \left\lceil \frac{k}{2} \right\rceil & \text{se } a_i \neq v'' + 1 \\ v_i = v' & \text{se } a_i = v'' + 1 \end{cases} \quad \text{con } i \in [1,d].$$

Se $a_i \leq \left\lceil \frac{k}{2} \right\rceil$ con ragionamenti analoghi si perviene ad analoghi risultati. Per avere il Teorema basta notare che se esiste una combinazione di mappe del tipo (3.2) per cui valgono le (3.3), poiché $\binom{r}{i} = \binom{r}{r-i}$ e valgono le (3.6) e (3.7), esiste anche la combinazione di mappe simmetrica, cioè quella che si ottiene scambiando tutti i pesi da w a $k - w$. ∎

Posto, $\forall v \in [v', v'']$:

$$h_v \stackrel{\text{def}}{=} \max\{v, k - v\} - (2^r - 1),$$

$$\bar{x}_v \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } v \notin [2^r - d, 2^r - 1] \\ 1 & \text{se } v \in [2^r - d, 2^r - 1], \end{cases} \tag{3.8}$$

e:

$$g_v = \binom{r}{v'' - v},$$

sia:

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} \stackrel{\text{def}}{=} \sum_{v=v'}^{v''} (g_v - \bar{x}_v) h_v.$$

Si ha il seguente:

Teorema 3.4 *Dato $d \leq r + 1$, per ogni combinazione di mappe (3.2) per cui valgono le (3.3), e la (3.4) o la (3.5) si ha:*

$$\sum_{i=d+1}^{2^r} [\max\{v_i, k - v_i\} + 1 - (b_i - a_i)] \geq (\mathbf{g} - \bar{\mathbf{x}})\mathbf{h}, \tag{3.9}$$

e l'uguale vale se, e solo se:

$$a_i = v_i \in [v', v''] \quad \forall i \in [i, d]. \tag{3.10}$$

Dimostrazione: Notiamo che per ogni $\mathbf{x} = (x_{v'}, \ldots, x_{v''}) \in \mathbb{N}^{r+1}$, tale che $\sum_{v=v'}^{v''} x_v = d$, si ha:

$$(\mathbf{g} - \mathbf{x})\mathbf{h} = \sum_{v=v'}^{v''} (g_v - x_v) h_v = \sum_{v=v'}^{v''} (g_v - x_v)(\max\{v, k - v\} + 1 - 2^r) =$$

$$\sum_{v=v'}^{v''}(g_v - x_v)(\max\{v, k-v\}+1) - 2^r \sum_{v=v'}^{v''}(g_v - x_v) =$$

$$\sum_{v=v'}^{v''}(g_v - x_v)(\max\{v, k-v\}+1) - 2^r(2^r - d),$$

ovvero:

$$(\mathbf{g} - \mathbf{x})\mathbf{h} = \sum_{v=v'}^{v''}(g_v - x_v)(\max\{v, k-v\}+1) - 2^r(2^r - d). \qquad (3.11)$$

Si ha:

$$\sum_{i=d+1}^{2^r} b_i + \sum_{a_i \geq \lceil \frac{k}{2} \rceil : i \in [1,d]} a_i = \sum_{v=\lceil \frac{k}{2} \rceil}^{k} v = \sum_{v=2^r}^{k} v + \sum_{v=\lceil \frac{k}{2} \rceil}^{2^r-1} v,$$

e:

$$\sum_{i=d+1}^{2^r} a_i + \sum_{a_i \leq \lceil \frac{k}{2} \rceil - 1 : i \in [1,d]} a_i = \sum_{v=0}^{\lceil \frac{k}{2} \rceil - 1} v = \sum_{v=0}^{2^r-d-1} v + \sum_{v=2^r-d}^{\lceil \frac{k}{2} \rceil - 1} v,$$

da cui rispettivamente segue che:

$$\sum_{i=d+1}^{2^r} b_i = \sum_{v=2^r}^{k} v - \Big(\sum_{a_i \geq \lceil \frac{k}{2} \rceil : i \in [1,d]} a_i - \sum_{v=\lceil \frac{k}{2} \rceil}^{2^r-1} v \Big) = \sum_{v=2^r}^{k} v - \Delta_b \qquad (3.12)$$

e:

$$\sum_{i=d+1}^{2^r} a_i = \sum_{v=0}^{2^r-d-1} v + \Big(\sum_{v=2^r-d}^{\lceil \frac{k}{2} \rceil - 1} v - \sum_{a_i \leq \lceil \frac{k}{2} \rceil - 1 : i \in [1,d]} a_i \Big) = \sum_{v=0}^{2^r-d-1} v + \Delta_a, \qquad (3.13)$$

dove:

$$\Delta_b \stackrel{\text{def}}{=} \sum_{a_i \geq \lceil \frac{k}{2} \rceil : i \in [1,d]} a_i - \sum_{v=\lceil \frac{k}{2} \rceil}^{2^r-1} v \geq 0$$

e:

$$\Delta_a \stackrel{\text{def}}{=} \sum_{v=2^r-d}^{\lceil \frac{k}{2} \rceil - 1} v - \sum_{a_i \leq \lceil \frac{k}{2} \rceil - 1 : i \in [1,d]} a_i \geq 0.$$

Dalle (3.12) e (3.13), segue che:

$$\sum_{i=d+1}^{2^r} b_i - a_i = \sum_{v=2^r}^{k} v - \sum_{v=0}^{2^r-d-1} v - \Delta_b - \Delta_a = 2^r(2^r-d) - \Delta_b - \Delta_a. \quad (3.14)$$

Posto:

$$x_v \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } v \notin \{v_i\}_{i\in[1,d]} \\ 1 & \text{se } v \in \{v_i\}_{i\in[1,d]}, \end{cases} \quad (3.15)$$

si ha:

$$\sum_{i=d+1}^{2^r} [\max\{v_i, k-v_i\} + 1 - (b_i - a_i)] =$$

$$\sum_{i=d+1}^{2^r} (\max\{v_i, k-v_i\} + 1) - \sum_{i=d+1}^{2^r} (b_i - a_i) =$$

$$\sum_{v=v'}^{v''} (g_v - x_v)(\max\{v, k-v\} + 1) - \sum_{i=d+1}^{2^r} (b_i - a_i) \stackrel{(3.11)}{=}$$

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} - \sum_{v=v'}^{v''} (g_v - \bar{x}_v)(\max\{v, k-v\} + 1) + 2^r(2^r - d) +$$

$$\sum_{v=v'}^{v''} (g_v - x_v)(\max\{v, k-v\} + 1) - \sum_{i=d+1}^{2^r} (b_i - a_i) \stackrel{(3.14)}{=}$$

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} + \sum_{v=v'}^{v''} (g_v - x_v - g_v + \bar{x}_v)(\max\{v, k-v\} + 1) + \Delta_b + \Delta_a =$$

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} + \sum_{v=v'}^{v''} (\bar{x}_v - x_v) \max\{v, k-v\} + \not{g} - \not{g} + \Delta_b + \Delta_a =$$

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} + \sum_{v=v'}^{\lceil \frac{k}{2} \rceil - 1} (\bar{x}_v - x_v)(k-v) + \sum_{v=\lceil \frac{k}{2} \rceil}^{v''} (\bar{x}_v - x_v)v + \Delta_b + \Delta_a =$$

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} + \sum_{v=v'}^{\lceil \frac{k}{2} \rceil - 1} \bar{x}_v(k-v) - \sum_{v=v'}^{\lceil \frac{k}{2} \rceil - 1} x_v(k-v) + \sum_{v=\lceil \frac{k}{2} \rceil}^{v''} \bar{x}_v v - \sum_{v=\lceil \frac{k}{2} \rceil}^{v''} x_v v + \Delta_b + \Delta_a =$$

$$(\mathbf{g}-\bar{\mathbf{x}})\mathbf{h}+\left\{\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}\bar{x}_v k-\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}x_v k\right\}+\left\{\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}x_v v-\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}\bar{x}_v v\right\}+$$

$$\left\{\sum_{v=\lceil\frac{k}{2}\rceil}^{v''}\bar{x}_v v-\sum_{v=\lceil\frac{k}{2}\rceil}^{v''}x_v v\right\}+\Delta_b+\Delta_a \overset{(3.8),(3.15)}{=}$$

$$(\mathbf{g}-\bar{\mathbf{x}})\mathbf{h}+\left\{\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}\bar{x}_v k-\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}x_v k\right\}+\left\{\sum_{v_i\le\lceil\frac{k}{2}\rceil-1:i\in[1,d]}v_i-\sum_{v=2^r-d}^{\lceil\frac{k}{2}\rceil-1}v\right\}+$$

$$\left\{\sum_{v=\lceil\frac{k}{2}\rceil}^{2^r-1}v-\sum_{v_i\ge\lceil\frac{k}{2}\rceil:i\in[1,d]}v_i\right\}+\Delta_b+\Delta_a=$$

$$(\mathbf{g}-\bar{\mathbf{x}})\mathbf{h}+\left\{\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}\bar{x}_v k-\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}x_v k\right\}+\left\{\sum_{a_i\le\lceil\frac{k}{2}\rceil-1:i\in[1,d]}a_i-\sum_{v=2^r-d}^{\lceil\frac{k}{2}\rceil-1}v\right\}+$$

$$\left\{\sum_{v_i\le\lceil\frac{k}{2}\rceil-1:i\in[1,d]}v_i-\sum_{a_i\le\lceil\frac{k}{2}\rceil-1:i\in[1,d]}a_i\right\}+\left\{\sum_{v=\lceil\frac{k}{2}\rceil}^{2^r-1}v-\sum_{a_i\ge\lceil\frac{k}{2}\rceil:i\in[1,d]}a_i\right\}+$$

$$\left\{\sum_{a_i\ge\lceil\frac{k}{2}\rceil:i\in[1,d]}a_i-\sum_{v_i\ge\lceil\frac{k}{2}\rceil:i\in[1,d]}v_i\right\}++\Delta_b+\Delta_a.$$

Se vale la (3.4), poiché:

$$\left\{\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}\bar{x}_v k-\sum_{v=v'}^{\lceil\frac{k}{2}\rceil-1}x_v k\right\}=0,$$

$$p_a \overset{\text{def}}{=} \left\{\sum_{v_i\le\lceil\frac{k}{2}\rceil-1:i\in[1,d]}v_i-\sum_{a_i\le\lceil\frac{k}{2}\rceil-1:i\in[1,d]}a_i\right\}\ge 0,$$

e:

$$p_b \overset{\text{def}}{=} \left\{\sum_{a_i\ge\lceil\frac{k}{2}\rceil:i\in[1,d]}a_i-\sum_{v_i\ge\lceil\frac{k}{2}\rceil:i\in[1,d]}v_i\right\}\ge 0,$$

si ha:

$$\sum_{i=d+1}^{2^r} [\max\{v_i, k - v_i\} + 1 - (b_i - a_i)] =$$

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} - \not\Delta_a + p_a - \not\Delta_b + p_b + \not\Delta_b + \not\Delta_a \geq (\mathbf{g} - \bar{\mathbf{x}})\mathbf{h},$$

e il segno di uguaglianza vale se, e solo se:

$$p_a, p_b = 0 \iff (3.10).$$

Se vale la (3.5), poiché:

$$\left\{ \sum_{v=v'}^{\lceil \frac{k}{2} \rceil - 1} \bar{x}_v k - \sum_{v=v'}^{\lceil \frac{k}{2} \rceil - 1} x_v k \right\} = -k,$$

$$p'_a \stackrel{\text{def}}{=} \left\{ \left[\sum_{v_i \leq \lceil \frac{k}{2} \rceil - 1 : i \in [1,d]} v_i - v' \right] - \sum_{a_i \leq \lceil \frac{k}{2} \rceil - 1 : i \in [1,d]} a_i \right\} \geq 0,$$

e:

$$p'_b \stackrel{\text{def}}{=} \left\{ \left[\sum_{a_i \geq \lceil \frac{k}{2} \rceil : i \in [1,d]} a_i - (v'' + 1) \right] - \sum_{v_i \geq \lceil \frac{k}{2} \rceil : i \in [1,d]} v_i \right\} \geq 0,$$

si ha:

$$\sum_{i=d+1}^{2^r} [\max\{v_i, k - v_i\} + 1 - (b_i - a_i)] =$$

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} - k - \not\Delta_a + p'_a + v' - \not\Delta_b + p'_b + v'' + 1 + \not\Delta_b + \not\Delta_a =$$

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} + (v' + v'' + 1 - k) + p'_a + p'_b > (\mathbf{g} - \bar{\mathbf{x}})\mathbf{h}.$$

Onde l'asserto. ∎

Diamo ancora il seguente:

Teorema 3.5 *Dato $d \leq r+1$, condizione necessaria e sufficiente affinché esista una combinazione di mappe (3.2) per cui valgono le (3.3) e risulti:*

$$Y_i :: \mathcal{S}_{v_i} \longrightarrow \mathcal{S}_{v_i} \quad \forall i \in [1, d], \tag{3.16}$$

con i v_i scelti in mezzo all'intervallo $[0,k]$, ovvero tali che:

$$v_i \in [2^r - d, 2^r - 1] \iff v_i \stackrel{\text{def}}{=} 2^r - d - 1 + i \quad \forall i \in [1,d], \qquad (3.17)$$

è che sia:

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} \leq 0.$$

Dimostrazione: Se le (3.2) sono una combinazone di mappe per cui valgono le (3.3), la (3.16) e la (3.17), allora, per il Teorema precedente e le (3.3), si ha:

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} = \sum_{i=d+1}^{2^r} [\max\{v_i, k - v_i\} + 1 - (b_i - a_i)] \leq 0.$$

Per quanto riguarda la restante parte della dimostrazione, notiamo che se disponessimo di tutti check Y_i tali che $v_i = \left\lceil \frac{k+r}{2} \right\rceil - w(Y_i) \leq 2^r - 1$ allora definire le mappe doppie sarebbe semplice, in quanto per ogni $i = d+1, \ldots, 2^r$ si potrebbe porre $Y_i :: \mathcal{S}_{i-(d+1)} \cup \mathcal{S}_{i-(d+1)+2^r} \longrightarrow \mathcal{S}_{v_i}$. Ma così non è. $\forall t \in [1, -\min_{v \in [v',v'']} h_v]$, è però possibile associare a t mappe di check Y_i tali che $h_{v_i} > 0$, h_{v_i} mappe di check Y_j per cui $h_{v_j} = -t$.

Si ha:

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} = \sum_{v=v'}^{v''} (g_v - \bar{x}_v)h_v = \sum_{h_v > 0} (g_v - \bar{x}_v)h_v + \sum_{h_v < 0} (g_v - \bar{x}_v)h_v \leq 0,$$

per cui:

$$\sum_{h_v > 0} (g_v - \bar{x}_v)h_v \leq \sum_{h_v < 0} (g_v - \bar{x}_v)|h_v|,$$

e quindi, posto $M = max_{v \in [v',v'']} h_v$ e $m = -min_{v \in [v',v'']} h_v$, ponendo:

$$\sum_{h_v > 0} (g_v - \bar{x}_v)h_v \stackrel{\text{def}}{=} \sum_{h=1}^{M} t_h^+ h$$

e:

$$\sum_{h_v < 0} (g_v - \bar{x}_v)|h_v| \stackrel{\text{def}}{=} \sum_{h=1}^{m} t_h^- h,$$

si ha:

$$\sum_{h=1}^{M} t_h^+ h \leq \sum_{h=1}^{m} t_h^- h = t_1^- 1 + t_2^- 2 + \ldots t_m^- m.$$

Suddividiamo la prima sommatoria della precedente relazione, al seguente modo:

$$\sum_{h=1}^{M} t_h^+ h = (\sum_{h=h_1+1}^{M} t_h^+ h + t_{h_1}^{+'} h_1) + (t_{h_1}^{+''} h_1 + \sum_{h=h_2+1}^{h_1-1} t_h^+ h + t_{h_2}^{+'} h_2) + \ldots$$

$$\ldots + (t_{h_{m-1}}^{+''} h_{m-1} + \sum_{h=1}^{h_{m-1}-1} t_h^+ h),$$

in modo che si abbia:

$$(\sum_{h=h_1+1}^{M} t_h^+ h + t_{h_1}^{+'} h_1) = t_1^- 1,$$

$$\forall l = 2, \ldots, m-1 \quad (t_{h_{l-1}}^{+''} h_{l-1} + \sum_{h=h_l+1}^{h_{l-1}-1} t_h^+ h + t_{h_l}^{+'} h_l) = t_l^- l,$$

e:

$$(t_{h_{m-1}}^{+''} h_{m-1} + \sum_{h=1}^{h_{m-1}-1} t_h^+ h) \leq t_m^- m,$$

ovvero:

$$(\sum_{h=h_1+1}^{M} t_h^+ h + t_{h_1}^{+'} h_1) = t_1^-,$$

$$\forall l = 2, \ldots, m-1 \quad \left(\frac{t_{h_{l-1}}^{+''} h_{l-1}}{l} + \sum_{h=h_l+1}^{h_{l-1}-1} \frac{t_h^+ h}{l} + \frac{t_{h_l}^{+'} h_l}{l}\right) = t_l^-,$$

e:

$$\left(\frac{t_{h_{m-1}}^{+''} h_{m-1}}{m} + \sum_{h=1}^{h_{m-1}-1} \frac{t_h^+ h}{m}\right) \leq t_m^-.$$

Sfruttando le ultime tre relazioni scritte è possibile, associando a l mappe di livello $h > 0$ h mappe di livello $-l$, costruire una combinazione di mappe soddisfacente le (3.3).

Ad esempio per $d = 6$ e $r = 27$, $k = 2^{r+1} - d - 1 = 268435449$, posto $* = 1342177$, si ha $v' = *11$, $v'' = *38$, $\left\lfloor \frac{k}{2} \right\rfloor = *24$, $\left\lceil \frac{k}{2} \right\rceil = *25$, $2^r - d = *22$ e $2^r - 1 = *27$. Si ha $M = 11$, $m = 2$ e:

$$t_{11}^+ = \binom{27}{0} + \binom{27}{27} = 2 \cdot 1 = 2,$$
$$t_{10}^+ = \binom{27}{1} + \binom{27}{26} = 2 \cdot 27 = 54,$$
$$t_9^+ = \binom{27}{2} + \binom{27}{25} = 2 \cdot 351 = 702,$$
$$t_8^+ = \binom{27}{3} + \binom{27}{24} = 2 \cdot 2925 = 5850,$$
$$t_7^+ = \binom{27}{4} + \binom{27}{23} = 2 \cdot 17550 = 35100,$$
$$t_6^+ = \binom{27}{5} + \binom{27}{22} = 2 \cdot 80730 = 161460,$$
$$t_5^+ = \binom{27}{6} + \binom{27}{21} = 2 \cdot 296010 = 592020,$$
$$t_4^+ = \binom{27}{7} + \binom{27}{20} = 2 \cdot 888030 = 1776060,$$
$$t_3^+ = \binom{27}{8} + \binom{27}{19} = 2 \cdot 2220075 = 4440150,$$
$$t_2^+ = \binom{27}{9} + \binom{27}{18} = 2 \cdot 4686825 = 9373650,$$
$$t_1^+ = \binom{27}{10} + \binom{27}{17} = 2 \cdot 8436285 = 16872570,$$

$$t_1^- = \binom{27}{12} + \binom{27}{15} - 2 = 2 \cdot 17383859 = 34767718,$$
$$t_2^- = \binom{27}{13} + \binom{27}{14} - 2 = 2 \cdot 20058299 = 40116598.$$

Si ha $t_2^{+'} = \frac{t_1^- - \sum_{h=3}^{11} t_h^+ h}{2} = 5057394$ e $t_2^{+''} = t_2^+ - t_2^{+'} = 4316256$, per cui una combinazione di mappe può essere costruita associando ad una mappa di livello h, h mappe di livello -1, ciò per ogni $h = 11, \ldots, 2$, per $\sum_{h=3}^{11} t_h^+ + t_2^{+'}$ volte, dopodiché si associano a due mappe di livello h, h mappe di livello -2, ciò per $h = 1, 2$, per $t_2^{+''} + t_1^+$ volte. Infine si definiscono le mappe doppie che accoppiano i pesi q e $q + 2^r$, per ogni q, fino ad esaurire i check. Ciò si può fare perché i check rimasti sono tali che $\max\{v, k - v\} \leq 2^r - 1$.

È facile vedere che quanto abbiamo fatto si può fare, con qualche accortezza, in generale. ∎

La caratterizzazione della scelta ottimale per la combinazione di mappe è data dal seguente:

Teorema 3.6 *Il piú piccolo d per cui esiste una combinazione di mappe (3.2) per cui valgono le (3.3), è il piú piccolo d per cui risulta:*

$$(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} \leq 0.$$

Per tale scelta di d è possibile supporre che:

$$Y_i :: \mathcal{S}_{v_i} \longrightarrow \mathcal{S}_{v_i} \quad \forall i \in [1, d],$$

con i v_i scelti in mezzo all'intervallo $[0, k]$, ovvero tali che:

$$v_i \in [2^r - d, 2^r - 1] \quad \Longleftrightarrow \quad v_i \stackrel{\text{def}}{=} 2^r - d - 1 + i \quad \forall i \in [1, d],$$

Inoltre si ha:

$$d \geq 1 + 0.8\sqrt{r},$$

e quindi:

$$k \simeq 2^{r+1} - 0.8\sqrt{r} - 2.$$

Dimostrazione: Sia d il piú piccolo intero tale che (3.2) è una combinazione di mappe per cui valgono le (3.3), allora:

1. Per il Teorema 3.3 esiste una combinazione di mappe (3.2) per cui valgono le (3.3) e la (3.4) o la (3.5).
2. Per il Teorema 3.4 si ha $(\mathbf{g} - \bar{\mathbf{x}})\mathbf{h} \leq 0$.
3. Per il Teorema 3.5 esiste una combinazione di mappe (3.2) per cui valgono le (3.3), la (3.16) e la (3.17).

Per la restante parte del Teorema rimandiamo a [1]. ∎

Bibliografia

[1] S. Al-Bassam and B. Bose, *Design of Efficient Balanced Codes*, in Proc. IEEE 19th Int. Symp. Fault Tolerant Computing, June 1989.

[2] B. Bose, *Burst Unidirectional Error-Detecting Codes*, IEEE Trans. Comput. vol. c-35, pp. 350–353, April 1986.

[3] B. Bose, *On Unordered Codes*, in Proc. IEEE 17th Int. Symp. Fault Tolerant Computing, pp. 102–107, July 1987

[4] B. Bose and D. J. Lin, *Systematic Unidirectional Error-Detecting Codes*, IEEE Trans. Comput. vol. c-34, pp.1026–1032, Nov. 1985.

[5] L. A. Dunnin, G. Dial and M. R. Varnasi, *Unidirectional Byte Error Detecting Codes for computer Memory System*, IEEE Trans. Comput, vol. c-39, pp. 592–595, April 1990.

[6] D. E. Knuth, *Efficient Balanced Codes*, IEEE Trans. Inform. Theory, vol. IT-32, pp. 51–53, Jan. 1986.

Luca TALLINI
Viale Ippocrate n. 97
Roma, ITALIA,
cap. 00161.

DETERMINAZIONE DEL PIANO SPERIMENTALE DI UNA SPERIMENTAZIONE FATTORIALE A 2 E 3 LIVELLI CON CONFUSIONE E FRAZIONATA

L. TORO, F. VEGLIO'
Dipartimento di Chimica, Ingegneria Chimica e Materiali: Facoltà di Ingegneria - Monteluco di Roio - L'Aquila

SOMMARIO

Nella sperimentazione industriale spesso vengono usati dei *designs* di prove realizzati secondo i principi della sperimentazione fattoriale.
Quando il numero dei fattori da analizzare e' molto elevato il lavoro sperimentale puo' risultare troppo oneroso.
In questi casi possono essere usati dei *designs* sperimentali estratti dalle prove del fattoriale completo (fattoriale frazionato).
Nel presente lavoro e' stata effettuata una rassegna dei principi della sperimentazione fattoriale e delle metodologie che permettono di individuare in maniera sistematica piani sperimentali frazionati a due ed a tre livelli.

1. INTRODUZIONE

Nella sperimentazione industriale una delle fasi piu' importanti e' quella della ricerca di un modello matematico che descriva un determinato fenomeno.
Il problema in genere e' quello di trovare come una determinata variabile dipendente (superfice di risposta) sia legata ad una serie di variabili assunte come indipendenti (fattori), che possono assumere vari valori (livelli), arbitrariamente selezionati dallo sperimentatore.
La combinazione dei vari livelli dei vari fattori viene definita trattamento, ed in corrispondenza di questo viene determinata sperimentalmente la superfice di risposta.
Il problema viene risolto ricercando il modello che descrive il fenomeno in esame e cio' viene realizzato impiegando varie tecniche a seconda della tipologia del modello (algebrico o differenziale, lineare o non lineare) [1] [3] [4] [5].
In tal senso una delle fasi piu' delicate che occorre bene tener presente e' la pianificazione delle prove: in definitiva occorre scegliere i livelli dei vari fattori che influenzano il fenomeno in esame.
Cioe' in definitiva occorre che il piano sperimentale sia sufficentemente esplorato in quanto una cattiva scelta dei livelli dei fattori si riperquote sulla stima dei parametri del modello [1], [4].
Risulta evidente che se il numero dei fattori e dei livelli risulta elevato il numero di prove da realizzare puo' risultare molto oneroso sia in termini di tempo impiegato nella sperimentazione e sia in termini di costo; si puo' aggiungere infine che a volte possono subentrare problemi di scala che rendono improponibile una sperimentazione con molte prove da realizzare.
Si consideri che se si vogliono studiare 5 fattori a 3 livelli il numero delle prove da realizzare mediante un reticolo fattoriale e' pari a 243.
Il problema e' pertanto quello di individuare una metodologia che permetta di realizzare una sperimentazione contenente il minimo numero di prove con il massimo contenuto di informazione.
Piani sperimentali di questo tipo sono i fattoriali frazionati che costituiscono un sottoinsieme

opportunamente definito dell'insieme contenente tutti i trattamenti di una sperimentazione fattoriale completa [2].
In Fig.1 viene mostrato un reticolo completo di prove di un fattoriale a tre livelli con 3 fattori (27 trattamenti), mentre in Fig.2 viene mostrato un fattoriale frazionato in cui i trattamenti sono stati opportunamente estratti dal fattoriale completo (9 trattamenti).
Dall'analisi della Fig.2 e' possibile constatare che guardando il cubo da qualsiasi lato, la visione che si avra' sara' sempre quella di 9 trattamenti distinti non sovrapposti. In altre parole il design sperimentale a due dimensioni risulta ortogonale.
In questo lavoro viene descritta la metodologia della ricerca sistematica del fattoriale frazionato, in sperimentazioni fattoriali a 2 ed a 3 livelli;tale procedura e' stata utilizzata per la realizzazione di un programma di calcolo per la definizione di un piano sperimentale frazionato.

2.SPERIMENTAZIONE FATTORIALE

La sperimentazione fattoriale rappresenta un modo molto efficente per realizzare delle prove sperimentali [2].
In pratica si tratta di realizzare le prove mediante un cosiddetto reticolo ortogonale.
I fattori analizzati vengono studiati a vari livelli ed in corrispondenza di un determinato trattamento si ottiene la superfice di risposta .
In Fig.1 e' mostrato il reticolo di una sperimentazione con 3 fattori a tre livelli ciascuno.
Le prove da realizzare sono 27 e sono costituite dai nodi del cubo di Fig.1, che costituiscono i trattamenti della sperimentazione.
Occorre fare qualche precisazione sul simbolismo che sara' adottato nel seguito.
Nella sperimentazione 2^n (fattoriale con n fattori a due livelli), un generico trattamento viene indicato con delle lettere minuscole (notazione di Yates) [2].
Ad es. il trattamento ab di un fattoriale 2^3 (con i fattori A, B, C) rappresenta il trattamento in cui i

fattori A e B sono al livello piu' alto, mentre C e' al livello piu' basso.
Il trattamento denotato con 1 indica la prova con tutti i fattori al livello piu' basso.
Nella sperimentazione 3^n (fattoriale con n fattori a 3 livelli), per definire il generico trattamento verranno usati una serie di numeri interi (0,1,2).
Ad es. il trattamento 012 indica una prova in cui il fattore A si trova al livello piu' basso, B a quello intermedio e C a quello piu' alto.
E' chiaro che quest' ultima notazione risulta essere di piu' facile estendibilita' per sperimentazioni fattoriali a qualsiasi numero di livelli e pertanto nel seguito si fara' riferimento prevalentemente a quest'ultima notazione.
Mediante la tecnica dell'analisi della varianza e' possibile definire quali sono i contributi dei vari fattori sulla superfice di risposta e definirne la loro significativita' mediante degli F-tests [2].
Dall'analisi del fattoriale generalmente e' possibile determinare il valore degli effetti e delle interazioni tra fattori; questi rappresentano il contributo dei singoli fattori o delle loro interazioni, sulla superfice di risposta.
Un effetto o una interazione e' in generale una combinazione lineare dei trattamenti.
Queste funzioni lineari godono della proprieta' di ortogonalita' e di indipendenza, definite maniera usuale [5].

3.RELAZIONE TRA CONFUSIONE E FATTORIALI FRAZIONATI

Esiste una profonda relazione tra fattoriale frazionato e fattoriale confuso in quanto e' possibile mostrare che il frazionato puo' essere derivato da un fattoriale completo confuso [2].
Per questa ragione nel seguito si fara riferimento ai concetti di confusione nella sperimentazione fattoriale.
La confusione viene applicata nella sperimentazione fattoriale quando non e' possibile garantire che le prove vengano condotte in condizioni uniformi a prescindere dai fattori che invece vengono deliberatamente variati.

Infatti in un fattoriale completo la dimensione del blocco dei trattamenti tende ad essere grande all'aumentare del numero dei fattori e dei livelli; all'interno del blocco pero' possono esistere sorgenti di variazione, non imputabili all'errore sperimentale, che incrementano l'errore stesso diminuendo in tal modo l'efficenza dell'esperimento.
Il problema e' legato al fatto che quando si vuole studiare l'effetto di un certo fattore sulla superfice di risposta e' auspicabile che questa stima non sia inquinata da altre sorgenti di variazione nascoste che ne inficerebbero tale confronto.
Ad esempio se si vuole studiare l'effetto della temperatura sulla resa di un certo processo chimico che tratta una determinata partita di materiale base si desidera che tale partita sia di caratteristiche costanti durante tutto il corso della prove.
Questo puo' non essere sempre possibile e in tal modo si potrebbe verificare che l'effetto del fattore in esame venga confuso con la variazione tra partite di materiale.
Per rendere la stima degli effetti piu' importanti da stimare, non contaminata da tale variazione si ricorre ad un piano sperimentale confondendo una o piu' *high interaction* con la variazione tra le partite [2].
Questo metodo suggerito da Yates si basa sulla scarsa inportanza delle interazioni con ordine elevato per realizzare un'esperimento piu' efficente.
Verranno realizzati dei blocchi di prove effettuate con una certa partita di materiale.
Un blocco e' rappresentato da un gruppo di trattamenti realizzato in condizioni praticamente omogenee.
Il fattoriale frazionato e' rappresentato da uno dei blocchi che costituiscono un determinato sistema di confusione [2].
Tale blocco conserva la importante proprieta' del fattoriale completo, cioe' la condizione di ortogonalita' degli effetti stimati anche se a causa del minor numero di prove possono verificarsi delle confusioni tra effetti ed interazioni.
In definitiva con il fattoriale completo ogni effetto proprio in virtu' di tale proprieta' , puo' essere stimato indipendentemente dagli altri effetti.
Quando si estraggono dei trattamenti dal fattoriale completo per ridurre il numero delle prove non e'

detto che tale proprieta' venga conservata, rendendo cosi' la stima degli effetti non indipendente tra loro.
Lo scopo del presente lavoro e' quello descrivere una procedura sistematica per l'individuazione di un fattoriale ridotto che conservi la proprieta' di ortogonalita' (fattoriale frazionato) tenendo conto comunque che in questo caso puo' originarsi una certa confusione nella stima di effetti e di interazioni.

4.DETERMINAZIONE DEL FATTORIALE FRAZIONATO

E' possibile determinare un fattoriale frazionato passando per i concetti della confusione e partendo da due punti di vista:

a) dalla costruzione di quadrati, cubi ed ipercubi latini [2],
b) utilizzando la definizione di ortogonalita' [2].

Nel primo caso si costruisce il fattoriale completo avente il numero di prove che si desidera realizzare ; si determinano le espressioni degli effetti ed interazioni.
Dalla definizione di questi effetti o interazioni e' possibile dedurre il piano sperimentale confondendo uno di questi con l' effetto di un nuovo fattore [2].
Questo modo di procedere potrebbe definirsi come "ricerca in avanti".
Con il secondo modo di procedere si parte dal fattoriale completo contente tutti i fattori che si vogliono studiare e si estrae da tutti i trattamenti un sottogruppo (blocco principale) dal quale viene derivato il fattoriale frazionato.
Questo modo di procedere puo' essere definito come "ricerca all'indietro" e sara' oggetto del presente lavoro.

5. APPLICAZIONE DEL CONCETTO DI GRUPPO ALLA DEFINIZIONE DELLA SPERIMENTAZIONE FATTORIALE

Gli effetti ed i trattamenti di una sperimentazione fattoriale completa possono essere determinati tenendo conto della definizione di gruppo, cioe' di un insieme di elementi definito con una determinata legge di composizione [2].
Gli elementi del gruppo possono essere definiti assegnando un gruppo di generatori ad esempio del tipo A, B, C, tenendo presente che:

$$A^2 = B^2 = C^2 = 1$$

Nello stesso modo e' possibile ottenere gli elementi dello stesso gruppo considerando come generatori gli elementi A, B, BC, dove gli stessi rappresentano l'effetto di A, di B e l'interazione BC.
La sola condizione richiesta per i generatori e' che siano indipendenti: ogni elemento non deve risultare combinazione degli altri. Ad esempio il set A, B, AB non puo' essere preso come gruppo generatore del fattoriale completo.
In Tab. 1 e' stato determinato il gruppo del fattoriale completo 2^3 considerando come generatori A, B, BC.

Nota 1.

Un reticolo fattoriale puo' essere considerato come uno spazio fattoriale n-dimensionale (n=numero dei fattori) i cui elementi o vettori hanno componenti discrete.
Nel caso di fattoriali a due livelli le componenti sono Booleane; ad esempio se n=3 (fattori A, B, C) i vettori
$V_1 = (0,1,0)$ e $V_2 = (1,1,0)$ rappresentano gli elementi b ed ab del gruppo dei trattamenti.
In questa maniera tutti i trattamenti del fattoriale a due livelli con 3 fattori possono essere determinati, sommando con mod.2, ad es. i seguenti 3 vettori:

$V_1 = (1,0,0)$; $V_2 = (0,1,0)$; $V_3 = (0,0,1)$;

In tal caso il trattamento abc sara' dato da $V_1 + V_2 + V_3$

e cosi' via.
Il discorso e' analogo se i livelli sono 3 (mod.3), 4 (mod.4) etc..

6.DEFINING CONTRAST (D.C.)

Operando in un sistema di confusione con un certo numero di blocchi, necessariamente alcuni effetti verranno di conseguenza confusi.
Un set di effetti confusi in un qualsiasi sistema di confusione viene definito *defining contrast* (DC). Ad es. in una sperimentazione 2^3 in cui si vuole confondere l'interazione ABC con la variazione tra due blocchi il DC sara' costituito dagli elementi I e ABC (dove I e' l'elemento neutro del gruppo degli effetti ed interazioni).
In generale dato un fattoriale 2^n, questo puo' essere confuso in 2^p blocchi (p=1,2..,n); in tal caso vengono a costituirsi 2^{p-1} DC di cui p sono indipendenti, mentre i rimanenti 2^p-p-1 sono combinazione dei p elementi.
I DC con l'elemento neutro costituiscono un gruppo finito chiuso rispetto all'operazione di combinazione (che chiameremo d'ora in poi moltiplicazione).
Ad es. si consideri un fattoriale 2^5 in cui si vogliono realizzare 8 blocchi di 4 osservazioni (p=3); possono essere arbitrariamente scelti 3 DC mentre i rimanenti 4 verranno determinati impiegando la legge di composizione del gruppo [2].
Ponendo ad es. come DC indipendenti i tre elementi CDE, ACE e ABDE si ottengono con la moltiplicazione di questi elementi tra loro, il gruppo completo dei DC, costituito dai seguenti 8 elementi (compreso l'elemento neutro):

I, AD, AB, BE, CDE, ACE, BCD, ABDE

Questo gruppo puo' essere impiegato per la determinazione degli effetti confusi nel particolare sistema di confusione.
E' ovvio che in una sperimentazione fattoriale si desidera confondere le *high-interactions*, facendo in modo da rendere puliti gli effetti che si vogliono stimare con una maggiore precisione come quelli

principali e le *two-interactions* [2].

Nota 2.

Due o piu' effetti si dicono confusi tra loro, quando per la loro determinazione si usa lo stesso confronto. Ad es. si consideri un fattoriale 2^3 (A,B,C). Se si confonde l'interazione ABC con un nuovo fattore, ad es. D, si avra' che:

D=ABC

In definitiva i termini per determinare ABC, vengono usati per valutare D.
In questo caso il DC sara' dato da:

I, ABCD.

Moltiplicando questo DC per i vari effetti si ottengono i fattori che vengono confusi; ad es. se si moltiplica per A si otterra' che l'effetto A si confonde con BCD (perche' $A^2=1$).
In questa maniera verranno definiti tutti i fattori confusi.

Nota 3.

La determinazione dei trattamenti di una sperimentazione fattoriale puo' essere derivata applicando gli stessi concetti visti sopra, selezionando un determinato gruppo di generatori costituito questa volta da trattamenti.
Selezionando come gruppi generatore il gruppo 1, a, b, c, e' possibile derivare il fattoriale completo 2^3 o gruppo dei trattamenti.

7.BLOCCO PRINCIPALE

Si definisce blocco principale, il gruppo dei trattamenti di un sistema di confusione che contiene l'elemento neutro del gruppo dei trattamenti.
Definito un gruppo di DC puo' essere mostrato che il gruppo principale puo' essere derivato da questo,

cercando i trattamenti che risultano ortogonali ai DC. Un effetto di un DC puo' essere scritto nella forma:

$$A^p B^q C^r D^s \qquad (1)$$

mentre un trattamento come:

$$a^w b^x c^y d^z \qquad (2)$$

Nel caso di fattoriale a due livelli gli esponenti possono assumere valori pari a 0 ed 1.
Il trattamento (2) risulta ortogonale all'effetto (1) se:

$$pw+qx+ry+sz = 0 \ (\text{mod}.2)$$

Quindi in una sperimentazione fattoriale a due livelli 2^n in cui e' necessario realizzare un sistema di confusione con 2^p blocchi, una volta scelto opportunamente il set di effetti facenti parte del DC, puo' essere individuato, applicando i concetti di ortogonalita', il blocco principale.
Questo puo' essere realizzato per individuare gli altri blocchi da realizzare, nel sistema di confusione assegnato [2].

Nota 4.
La condizione di ortogonalita' e' definita come il prodotto scalare di due vettori dello spazio vettoriale.
Ad es. in una sperimentazione 2^3 (A,B,C) si desideri confondere ABC con la variazione tra i blocchi; l'interazione ABC e' data dal seguente confronto:

$ABC = 1/4\ [(abc+a+b+c)-(bc+ac+ab+1)]$

Il gruppo bc, ac, ab, 1, definito anche come:

(0,1,1) (1,0,1) (1,1,0) (0,0,0)

e' un blocco principale e risulta ortogonale ad ABC definito anche (1,1,1).
Il blocco principale rappresenta un gruppo chiuso rispetto alla moltiplicazione.
Da questo possono essere derivati gli altri blocchi

del sistema di confusione considerato (in questo caso ABC confuso); se viene usata la notazione con le lettere, occorre moltiplicare il blocco principale per un trattamento non contenuto nel blocco stesso; se viene usata la notazione vettoriale occorre sommare ai vettori che costituiscono il blocco principale un vettore che non e' compreso in esso.

8.CONFUSIONE IN UN 3^n DESIGN

Analogamente a quanto detto precedentemente per i 2^n designs, e' possibile applicare gli stessi concetti visti sopra ai 3^n designs operando solo alcune modifiche.
I generici effetti e trattamenti espressi tramite la (1) e la (2) saranno validi anche in questo caso tenendo presente pero' che gli esponenti possono assumere valori di 0, 1 e 2, essendo tre i livelli.
Inoltre l'elemento neutro, sia dei trattamenti che degli effetti, e' dato da:

$$A^3 = B^3 = C^3 = D^3 = A^0 = \ldots = D^0 = I$$

$$a^3 = b^3 = c^3 = d^3 = a^0 = \ldots = d^0 = 1$$

In definitiva facendo assumere ad (1) ed a (2) tutte le possibili combinazioni di 0, 1 e 2, viene generato il fattoriale completo, sia come trattamenti che come effetti.
La condizione di ortogonalita' tra due trattamenti o tra due effetti o tra un trattamento ed un effetto e' data da:

$$pw+qx+ry+sz = 0 \ (\mathrm{mod}.3)$$

Nota 5.
Occorre considerare che in un fattoriale a tre livelli il concetto di effetto risulta di piu' difficile comprensione ed in ogni caso si puo' sempre parlare della significativita' dei fattori.

Nel caso di fattoriali a tre livelli e' necessario

fare alcune considerazioni particolari sugli effetti e sui DC; viene considerato come esempio un fattoriale 3^3. Dal generico effetto-interazione $A^a B^b C^c$, come e' stato gia' detto, e' possibile ottenere il gruppo degli effetti; questi saranno costituiti dai 27 termini mostrati in Tab.2.
Questo gruppo, come nel caso di un 2^n design, e' chiuso rispetto all'operazione di moltiplicazione.
Da questi elementi e' possibile estrarre dei sottogruppi di ordine 3 che hanno la proprieta' di essere chiusi rispetto alla moltiplicazione.
Ad es. costituiscono un sottogruppo gli elementi I, A, A^2, oppure I, B, B^2.
Si consideri il sottogruppo I, A, A^2; sara' possibile trovare un gruppo di 9 trattamenti che risultano ortogonali al sottogruppo; questi trattamenti sono mostrati in Tab.3.
Nel seguito per indicare un generico trattamento verranno indicati solo gli esponenti; ad es. $a^0 b^1 c^2$ = 012.
Se si moltiplica per a ed a^2 questo gruppo di trattamenti si ottengono i tre gruppi di trattamenti del sistema di confusione che confonde la variazione tra i blocchi con l'effetto di A; in altre parole il sottogruppo I, A, A^2 individua l'effetto di A.
Procedendo analogamente per gli altri sottogruppi si ottiene la Tab.4 in cui vengono riportati le corrispondenze tra sottogruppi ed effetti determinati.
Nel caso di una sperimentazione 3^n tali sottogruppi individuano i DC che possono essere selezionati per individuare il sistema di confusione voluto.
Un sistema 3^3 puo' essere confuso o con 3 blocchi da 9 o con 9 blocchi da 3 trattamenti ciascuno.
Nel primo caso il DC puo' essere costituito ad es. da:

$$I,\ ABC,\ A^2B^2C^2$$

Nel secondo caso occorre confondere due effetti; se si confonde per esempio [2]:

$$I,\ AB,\ A^2B^2 \quad [\ F(AB)\ ]$$
$$I,\ BC,\ B^2C^2 \quad [\ F(BC)\ ]$$

saranno confusi anche:

AB^2C, A^2BC^2 e AC^2, A^2C,

che si ottengono dalla legge di moltiplicazione dei due effetti confusi inizialmente.
Il DC complessivo sara' dato pertanto dal seguente gruppo chiuso:

I, AB, A^2B^2, BC, B^2C^2, AB^2C, A^2BC^2, AC^2, A^2C,

che corrisponde a confondere [2]:

F(AB), F(BC), I(AC), X(ABC).

Come nel 2^n *design,* una volta noto il gruppo dei DC, e'possibile individuare il blocco principale del sistema di confusione in base alla condizione di ortogonalita' (3).
Noto tale blocco possono essere derivati i rimanenti moltiplicando per un certo set di trattamenti: ad es. per (1,0,0) e (2,0,0).

9.CONCLUSIONI

Nel presente lavoro e' stata descritta una metodologia sistematica per la definizione di un piano sperimentale frazionato a 2 ed a 3 livelli.
Tale procedure comunque risulta essere generalizzabile.
In Fig.2 viene riportato il diagramma di flusso di tale procedura.
In definitiva per la determinazione delle prove da realizzare occorre :

1. individuare con quante prove si vogliono studiare i fattori in esame,
2. fissare il sistema di confusione assegnando un opportuno DC,
2. determinare il blocco principale applicando la condizione di ortogonalita'.
3. ottenere gli altri blocchi moltiplicando per un certo set di trattamenti,
4. scegliere tra questi il blocco che costituira' il frazionato.

Questo piano sperimentale puo' essere applicato oltre che nello studio della significativita' dei vari fattori, anche nella pianificazione preliminare delle prove per la ricerca di un modello.
Infatti come e' noto nelle fasi successive si seguono le regole della "programmazione della sperimentazione" per la ricerca delle prove da realizzare per una migliore stima dei parametri del modello stesso [4].

BIBLIOGRAFIA.

[1] D.H. Himmelblau, " Process Analysis by Statistical Methods", John Wiley & Sons, 1970.
[2] O.L. Davies et al., " The Design and Analysis of Industrial Experiments", ICI, IIª edizione, 1979.
[3] O.L. Davies and P.L. Goldsmith, "Statistical Methods in Research and Production", ICI, IVª edizione, 1972.
[4] G.B. Ferraris, "Analisi ed identificazione di modelli", CLUP, Milano 1981.
[5] R.C. Bose, "Combinatorial Problems of Experimental esign II: Factorial Design", in Annals of Discrete Mathematics 6 (1980).

TAB.1 - Generazione fattoriale completo.

GENERATORI	A, B, BC
PRODOTTO DUE ALLA VOLTA	AB, ABC, C
PRODOTTO TRE ALLA VOLTA	AC
PRODOTTO CON STESSO	I

TAB.2 - Gruppo del fattoriale completo a tre fattori e tre livelli.

$A, A^2, B, B^2, C, C^2, A^2B, AB^2, A^2C, AC^2, B^2C, BC^2$
$BC, B^2C^2, A^2BC, AB^2C^2, A^2BC, AB^2C^2, AB^2C, A^2BC^2$
$ABC^2, A^2B^2C, ABC, A^2B^2C^2, I$

TAB.3 - Gruppo principale di un fattoriale a tre livelli e tre fattori.

$a^0b^0c^0$	$a^0b^0c^1$	$a^0b^0c^2$
$a^0b^1c^0$	$a^0b^1c^1$	$a^0b^1c^2$
$a^0b^2c^0$	$a^0b^2c^1$	$a^0b^2c^2$

TAB. 1

I, A^2, A	A
I, B^2, B	B
I, C^2, C	C
I, A^2B, AB^2	I(AB)
I, AB^2, A^2B	F(AB)
I, A^2C, AC^2	I(AC)
I, AC, A^2C^2	F(AC)
I, B^2C, BC^2	I(BC)
I, BC, B^2C^2	F(BC)
I, A^2BC, AB^2C^2	W(ABC)
I, AB^2C, A^2BC^2	X(ABC)
I, ABC^2, A^2B^2C	Y(ABC)
$I, ABC, A^2B^2C^2$	Z(ABC)

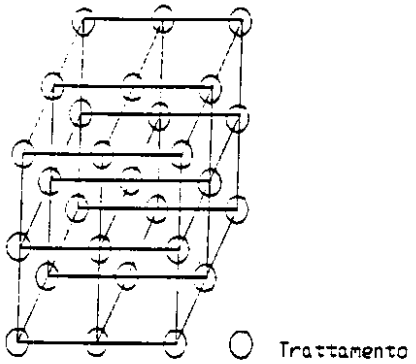


fig 1

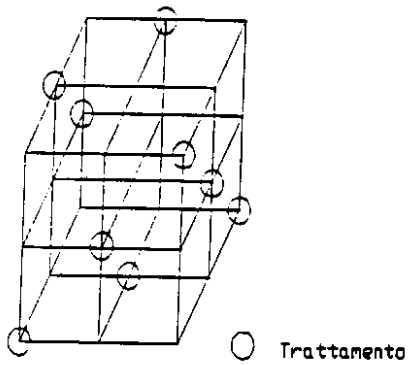


fig 2

SULL'AUTOCORRELAZIONE SPAZIALE DI FENOMENI QUALITATIVI

Giuliano VISINI, Dip. di Scienze, Storia dell'Architettura e Restauro
Facoltà di Architettura - Università "G. D'Annunzio" di Chieti

ABSTRACT

It is well known that Moran (1948), Krishna Iyer (1949), Dacey (1965), Cliff (1969) and Cliff-Ord (1973, 1981) determined the first steps of the join count statistic.

Nevertheless these steps, on account of the parameters (A, D, etc.) by which they are characterized, can be referred to different contiguity hypothesis.

On the other end, the problem concerning the determination of the join count statistics still remains open for a finite subdivision of a region $\{a_i\}_{i\in N_n}$, which can be defined through small values of n and generally only through some hypothesis of contiguity.

In this note the casual variable $N_{BW}(V)$ of the contiguous diversities type BW will be determined, limitedly referred to the case of dicotomical observations (type {B, W}) and independent on n regional unity, in the particular case of the contiguity operator V defined for each r.u. a_i as $V(a_i)=\{a_j \mid j\in N_n-\{i\}\}$.

Thus, taken for equivalent the operators of simmetrical contiguities of the same A parameter (A stands for the bonds

number), $A = \sum_{i=1}^{n} L''_i = \sum_{i=1}^{n} L'_i$, it will be determined the c.v. $N_{BW}(2A)$ of the contiguous diversities type BW, these being pointed out for the equivalent class $[L]$ identifiable in the parameter 2A.

Eventuay, being observed that, on account of the symmetry bond of the operator, the c.v. $N_{BW}(2A)$ is similar to the c.v. $N_{BWvWB}(A)$ of the contiguous diversities type BW or WB, when associated to the class of non-symmetrical operators L'' (L') and L limitations, those symmetricall operators L with $\underline{L}''=(L''_i)_{i\in N_n}$ (or $\underline{L}'=((L'_i)_{i\in N_n})$) are considered equivalent.

Limitedly referring to the previous hypothesis (dichotomy and independence) and to the V'' operator, limitation of V, c.v. is defined as $\underline{N}_{BWvWB}((V''_i)_{i\in N_n})$ which has, as determination of n successions, values whose i^{th} element stands for the amount of diversities exisiting between the r.u. a_i, and those to follow and from this the c.v. $N_{BWvWB}([(L''_i)_{i\in N_n}])$ through the use of a suitable algoritm.

Key-words: two colour maps; contiguity; finite populations; join-count; distributions.

SOMMARIO

E' noto come Moran (1948), Krishna Iyer (1949), Dacey (1965), Cliff (1969) e Cliff-Ord (1973, 1981) abbiano determinato i primi momenti della statistica join count.

Tuttavia questi momenti per i parametri (A, D, etc.) che li caratterizzano possono essere riferiti ad ipotesi diverse di contiguita'.

D'altra parte, la determinazione della distribuzione della statistica join count, per una suddivisione finita di un territorio $\{a_i\}_{i\in N_n}$, definibile per piccoli valori di n ed in generale solo per alcune ipotesi di contiguita', resta ancora un

problema aperto.

In questa nota, limitatamente al caso di osservazioni dicotomiche (di tipo {B, W}) ed indipendenti sulle n unita' territoriali, si determina, nel caso particolare dell'operatore di contiguita' V, definito per ogni u.t. a_i come $V(a_i)=\{a_j | j\in N_n-\{i\}\}$, la variabile casuale $N_{BW}(V)$ delle diversita' contigue di tipo BW.

Poi, considerati equivalenti gli operatori di contiguita' simmetrici L di medesimo parametro A (numero dei legami), $A = \sum_{i=1}^{n} L''_i = \sum_{i=1}^{n} L'_i$, si determina la v.c. $N_{BW}(2A)$ delle diversita' contigue di tipo BW rilevabili per la classe d'equivalenza $[L]$ identificabile nel parametro 2A.

Successivamente, osservato che per il vincolo di simmetria degli operatori L, la v.c. $N_{BW}(2A)$ e' uguale alla v.c. $N_{BWvWB}(A)$ delle diversita' contigue del tipo BW o WB, associata alla classe degli operatori non simmetrici L'' (L') limitazioni di L, si considerano equivalenti quegli operatori simmetrici L aventi medesima n-pla $\underline{L}'' = (L''_i)_{i\in N_n}$ (o $\underline{L}' = (L'_i)_{i\in N_n}$).

Limitatamente alle precedenti ipotesi (dicotomia ed indipendenza) e all'operatore V'', limitazione di V, si definisce la v.c. $\underline{N}_{BWvWB}((V''_i)_{i\in N_n})$, che ha come determinazioni delle successioni di n valori il cui elemento i-esimo rappresenta il numero delle diversita' esistenti tra l'u.t. a_i e le successive, e da questa la v.c. $N_{BWvWB}([(L''_i)_{i\in N_n}])$ attraverso l'introduzione di un opportuno algoritmo.

Parole chiavi: mappe a due colori; contiguita'; popolazioni finite; join-count; distribuzioni.

1. INTRODUZIONE

Sia $N_n = \{1,2,\ldots,n\}$ ed $n \geq 2$; se $\{a_i \mid i \in N_n\}$ indica una suddivisione territoriale finita di un territorio, un criterio di contiguita' spaziale tra le unita' territoriali e' definito da L,

$$L: \{a_i \mid i \in N_n\} \longrightarrow \{a_i \mid i \in N_n\},$$

che esprime una corrispondenza generalmente multivoca e non ovunque definita tale che $L(a_i) \subset \{a_j \mid j \in N_n,\ j \neq i\}$. Nel caso valga l'implicazione

$$a_j \in L(a_i) \Longrightarrow a_i \in L(a_j) \qquad (j \in N_n),$$

l'operatore di associazione L lo diremo simmetrico.

La funzione I,indicatrice delle unita' territoriali contigue,

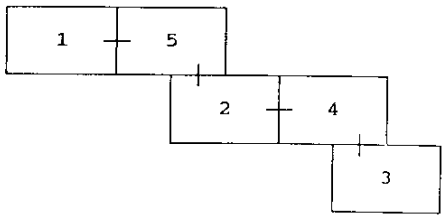
$$I: \{a_i \mid i \in N_n\} \times \{a_i \mid i \in N_n\} \longrightarrow \{0,1\}$$

tale che

$$I(a_p,a_q)=1 \Longleftrightarrow a_q \in L(a_p),$$

individua una matrice $\mathbb{L}$, detta matrice delle contiguita', che e' in corrispondenza biunivoca con l'operatore L.

Ad es., se si e' indicata con —— la contiguita' spaziale



all'operatore L resta associato la matrice $\mathbb{L}$:

$$\begin{vmatrix} 00001 \\ 00011 \\ 00010 \\ 01100 \\ 11000 \end{vmatrix}$$

Ad $\mathbb{L}$ sono legati dei parametri descrittivi, gia' noti in letteratura in quanto impiegati nella ricerca dei momenti degli indici di Moran, quali:

$$L_1=|L(a_1)|,\quad \sum_{i=1}^{n} L_i=2A,\quad \sum_{i=1}^{n} L_i(L_i-1)=2D,\text{etc..}$$

A questi aggiungiamo in particolare:

$$L'_i = L'(a_i) \qquad \text{ed} \qquad L''_i = L''(a_i) \qquad (i\in N_n)$$

per i quali rispettivamente L' indica la limitazione di L tale che

$$L'(a_i) = L(a_i) \cap \{a_j |\ j\in N_n,\ j<i\} \qquad (i\in N_n)$$

ed L'' la limitazione per cui

$$L''(a_i) = L(a_i) \cap \{a_j |\ j\in N_n,\ j>i\} \qquad (i\in N_n).$$

Tra tali parametri intercorrono le relazioni:

$$L_i = L'_i + L''_i \qquad (i\in N_n)$$

con $L'_1 = L''_n = 0$ e $0\le L'_i\le i-1$, $0\le L''_i\le n-i$.

$$\mathbb{L}'=\begin{vmatrix}00000\\00000\\00000\\01100\\11000\end{vmatrix} \quad \underline{L}'=\begin{pmatrix}0\\0\\0\\2\\2\end{pmatrix} \qquad \mathbb{L}''=\begin{vmatrix}00001\\00011\\00010\\00000\\00000\end{vmatrix} \quad \underline{L}''=\begin{pmatrix}1\\2\\1\\0\\0\end{pmatrix}$$

Sia $E=\{E_1,E_2,\ldots,E_m\}$ un sistema completo di m eventi con probabilita' $P(E_1)=p_1$, $P(E_2)=p_2$, ..., $P(E_m)=p_m$ e $\sum_{i=1}^{m} p_i=1$; se si assume che una successione di n eventi $\underline{X}=(X_i)_{i\in N_n}$ sia il risultato di n prove indipendenti per le quali si sia verificato k_1 volte l'evento E_1, k_2 volte l'evento E_2, ..., k_m volte l'evento E_m ($k_1+k_2+\ldots+k_m=n$) allora all'espressione $\underline{X}^T\mathbb{L}\underline{X}$ compete la probabilita':

$$P(\underline{X})=p_1^{k_1}\cdot p_2^{k_2}\cdot\ldots\cdot p_m^{k_m}.$$

Dall'uguaglianza

$$\underline{X}^T\mathbb{L}\underline{X} = E^T\mathbb{F}_L(\underline{X})E, \tag{1.1}$$

ove $\mathbb{F}_L(\underline{X})$ e' la matrice quadrata di ordine m costituita dalle frequenze assolute degli eventi $\{E_1E_1,\ E_1E_2,\ldots,\ E_1E_m,\ E_2E_1,\ldots,\ E_mE_m\}$, si ha che le espressioni $\underline{X}^T\mathbb{L}\underline{X}$ e $E^T\mathbb{F}_L(\underline{X})E$ si realizzano con la medesima probabilita' $P(\underline{X})$, come anche ogni elemento dello sviluppo di $E^T\mathbb{F}_L(\underline{X})E$, ogni singolo elemento di $\mathbb{F}_L(\underline{X})$, etc..

La forma algebrica $\sum_{i=1}^{n} \mathbb{F}_{L(a_i)}(\underline{X})$, equivalente della $E^T\mathbb{F}_L(\underline{X})E$, in cui $\mathbb{F}_{L(a_i)}(\underline{X})$ indica la matrice di ordine m i cui elementi $f_{rsL(a_i)}(\underline{X})$ sono le frequenze della modalita' E_rE_s di cui la prima E_r rilevabile sull'unita' territoriale a_i e la seconda E_s sull'unita' $a_j \in L(a_i)$, porta a riconsiderare l'espressione $\underline{X}^T\mathbb{L}\underline{X}$ riproponendola in notazione vettoriale-matriciale $\left(\mathbb{F}_{L(a_i)}(\underline{X})\right)_{i\in N_n}$ piu' efficace per le considerazioni che seguono. In particolare, riferendosi alla modalita' E_rE_s si hanno le n-ple di frequenze $\underline{f}_{rsL}(\underline{X})=(f_{rsL(a_i)}(\underline{X}))_{i\in N_n}$.

Nel precedente es. se $E=\{E_1, E_2, E_3\}$ ed $\underline{X}^T=(E_1,E_2,E_1,E_3,E_2)$ si ha

		a_1	a_2	a_3	a_4	a_5
		$E_1E_2E_3$	$E_1E_2E_3$	$E_1E_2E_3$	$E_1E_2E_3$	$E_1E_2E_3$
a_1	E_1	0 0 0	0 0 0	0 0 0	0 0 0	0 1 0
	E_2	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_3	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
a_2	E_1	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_2	0 0 0	0 0 0	0 0 0	0 0 1	0 1 0
	E_3	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
a_3	E_1	0 0 0	0 0 0	0 0 0	0 0 1	0 0 0
	E_2	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_3	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
a_4	E_1	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_2	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_3	0 0 0	0 1 0	1 0 0	0 0 0	0 0 0
a_5	E_1	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_2	1 0 0	0 1 0	0 0 0	0 0 0	0 0 0
	E_3	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0

Le componenti di $\left(\mathbb{F}_{L(a_i)}(\underline{X})\right)_{i\in N_n}$ sono:

$$\mathbb{F}_{L(a_1)}(\underline{X}) = \begin{vmatrix} 010 \\ 000 \\ 000 \end{vmatrix} \qquad \mathbb{F}_{L(a_2)}(\underline{X}) = \begin{vmatrix} 000 \\ 011 \\ 000 \end{vmatrix} \qquad \mathbb{F}_{L(a_3)}(\underline{X}) = \begin{vmatrix} 001 \\ 000 \\ 000 \end{vmatrix}$$

$$\mathbb{F}_{L(a_4)}(\underline{X}) = \begin{vmatrix} 000 \\ 000 \\ 110 \end{vmatrix} \qquad \mathbb{F}_{L(a_5)}(\underline{X}) = \begin{vmatrix} 000 \\ 110 \\ 000 \end{vmatrix}$$

ed $\mathbb{F}_L(\underline{X}) = \sum_{i=1}^{5} \mathbb{F}_{L(a_i)}(\underline{X}) = \begin{vmatrix} 011 \\ 121 \\ 110 \end{vmatrix}$.

Le n-ple delle frequenze delle modalita' $E_r E_s$ con riferimento all'es. sono:

$\underline{f}^T_{11L}(\underline{X}) = (00000)$ $\quad \underline{f}^T_{12L}(\underline{X}) = (10000)$ $\quad \underline{f}^T_{13L}(\underline{X}) = (00100)$

$\underline{f}^T_{21L}(\underline{X}) = (00001)$ $\quad \underline{f}^T_{22L}(\underline{X}) = (01001)$ $\quad \underline{f}^T_{23L}(\underline{X}) = (01000)$

$\underline{f}^T_{31L}(\underline{X}) = (00010)$ $\quad \underline{f}^T_{32L}(\underline{X}) = (00010)$ $\quad \underline{f}^T_{33L}(\underline{X}) = (00000)$.

2. LA VARIABILE CASUALE $N_{BW}(2A)$.

Sia E la variabile casuale $\begin{pmatrix} B & W \\ p & q \end{pmatrix}$ osservabile su ogni unita' territoriale a_i e sia $\underline{X}=(X_i)_{i\in N_n}$ una successione di n eventi indipendenti osservabile sulla suddivisione territoriale finita.

Al variare di $\underline{X}$ resta definita la v.c. $N_{BW}(L)$ delle diversita' contigue BW, dipendente dal particolare operatore di associazione L, le cui determinazioni η altro non sono che frequenze $f_{BWL}(\underline{X})$.

In particolare, per l'operatore V, la cui matrice delle contiguita' $\mathbb{V}$ ha generico elemento $v_{ij} = \begin{cases} 1 & \text{per } i \neq j \\ 0 & \text{per } i=j \end{cases}$, la v.c. $N_{BW}(V)$ ha la seguente distribuzione (2.1):

ν	0	n-1	...	$k_1 k_2$	...
$P(\nu)$	p^n+q^n	$n(pq^{n-1}+p^{n-1}q)$	...	$\frac{n!}{k_1!k_2!}\left(p^{k_1}q^{k_2}+p^{k_2}q^{k_1}\right)$	...

...	[n/2][(n+1)/2]
...	.....

$$\begin{cases} k_i \geq 0 \ (i \in N_2) \\ k_1+k_2=n \end{cases} .$$

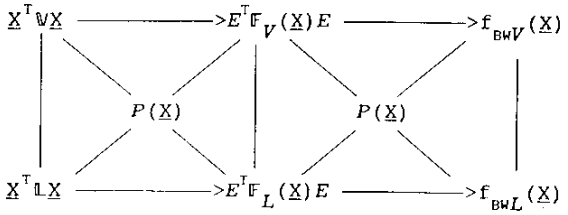
Per attribuire all'evento [n/2][(n+1)/2] la sua probabilita', si distinguono due casi a_0) e a_1) rispettivamente riferiti ad n pari e n dispari:

a_0) $P(n^2/4) = \binom{n}{n/2}(pq)^{n/2}$,

a_1) $P([n/2][(n+1)/2]) =$

$$= \binom{n}{[n/2]}\left(p^{[n/2]}\, q^{[(n+1)/2]} + p^{[(n+1)/2]}\, q^{[n/2]}\right).$$

Ove si consideri un generico operatore simmetrico L, la corrispondente variabile casuale $N_{BW}(L)$, le cui determinazioni $\nu(L)$ sono le $f_{BWL}(\underline{X})$, resta definita attraverso la commutativita' dei diagrammi ottenuti al variare delle n-ple osservabili $\underline{X}$:



Si noti tuttavia che la cardinalita' 2^n dell'insieme delle

n-ple $\underline{X}$ rende eccessivamente oneroso il calcolo per i valori di n riscontrabili nei casi concreti.

Se introduciamo nell'insieme degli operatori di associazione $\{L\}$ la relazione di equivalenza ~ cosi' definita:

$$L_1 \sim L_2 \Longleftrightarrow \sum_{i=1}^{n} L_{1i} = \sum_{i=1}^{n} L_{2i} = 2A,$$

ed indichiamo con 2A la classe degli operatori equivalenti $[L]$ allora ad ogni n-pla ordinata $\underline{X}$ corrisponde una $\left(\binom{n}{2}\right)$-pla non ordinata $\left(f_{BW_L}(\underline{X})\right)_{L\in 2A}$.

Quindi, indicata con $N_{BW}(2A)$ la variabile casuale le cui determinazioni $\eta(2A)$ sono le $f_{BW}(L,\underline{X})$ al variare di L nella classe 2A e di $\underline{X}$, la probabilita' di $\eta(2A)$ e' somma di probabilita' composte di eventi indipendenti di cui l'uno fa riferimento ad una distribuzione ipergeometrica e l'altro alla distribuzione binomiale (2.1):

$$(2.2) \qquad P(\eta(2A)) = \sum_{\nu\in N_{BW}(V)} \frac{\binom{\nu}{\eta(2A)}\binom{\binom{n}{2}-\nu}{A-\eta(2A)}}{\binom{\binom{n}{2}}{A}} P(\nu)$$

con le condizioni $\begin{cases} 1\le A\le \binom{n}{2} \\ \nu-\eta(2A)\ge 0 \\ \binom{n}{2}-A-(\nu-\eta(2A))\ge 0 \end{cases}$

In particolare per $A=\binom{n}{2}$ si ha $N_{BW}\left\{2\binom{n}{2}\right\} = N_{BW}(V)$ mentre per A=1 la v.c. $N_{BW}(2)$ ha distribuzione:

η	0	1
$P(\eta)$	p^2+q^2	2pq

.

Facendo riferimento all'es. per p =.5 la v.c. $N_{BW}(8)$ ha la seguente distribuzione:

η	0	1	2	3	4
$P(\eta)$	.0878	.1905	.4018	.2738	.0461

.

3.LA VARIABILE CASUALE $N_{BWvWB}(\underline{L}")$.

Se $V"$ e' l'operatore limitazione di V, la cui matrice W" ha elementi $v"_{ij}=\begin{cases} 1 \text{ per } i<j \\ 0 \text{ per } i\geq j \end{cases}$, allora la v.c. $N_{BWvWB}(V")$, che ha per determinazioni il numero complessivo delle diversita' tipo BW o WB, e' esattamente la v.c. $N_{BW}(V)$ (2.1) e, piu' in generale, per la simmetria e' anche $N_{BWvWB}(L") = N_{BW}(L)$ come anche $N_{BW}(2A) = N_{BWvWB}(A)$, essendo A la classe degli operatori $L"$ tali che $\sum_{i=1}^{n} L"_i=A$.

Ora gli $\left(\binom{n}{2}\right)_A$ operatori equivalenti di classe $\dot{A} = [L"]$ possono essere ripartiti in sottoclassi $[\underline{L}"] = [(L"_i)_{i\in N_n}]$ attraverso la nuova relazione d'equivalenza:

$$L"_1 \approx L"_2 \iff (L"_{1i})_{i\in N_n}=(L"_{2i})_{i\in N_n}.$$

In particolare gli insieme quozienti $\{L"\}/-$ e $\{L"\}/\approx$ hanno rispettivamente la classe $A=\binom{n}{2}$ e la classe $[\underline{L}"]=[(n-i)_{i\in N_n}]$ costituita dal solo operatore $V"$.

Al fine di definire la v.c $N_{BWvWB}([\underline{L}"])$, che ha per determinazioni il numero complessivo delle diversita' di tipo BW o

WB al variare dell'operatore nella classe [$\underline{L}$"], introduciamo la v.c. $\underline{N}_{BWvWB}(\underline{V}")$, che ha come determinazioni delle successioni di n valori il cui elemento i-esimo e' il numero delle diversita'(o BW o WB) esistenti tra la u.t. a_i e le successive u.t. a_j con $j\in(N_n-N_1)$.

E' possibile rappresentare il processo di generazione delle diverse determinazioni di $\underline{N}_{BWvWB}(\underline{V}")$ con il seguente schema (3.1), in cui distinguiamo:

a) due tipi di frecce ———> e - - ->, le quali esprimono rispettivamente il realizzarsi dell'evento B e dell'evento W;

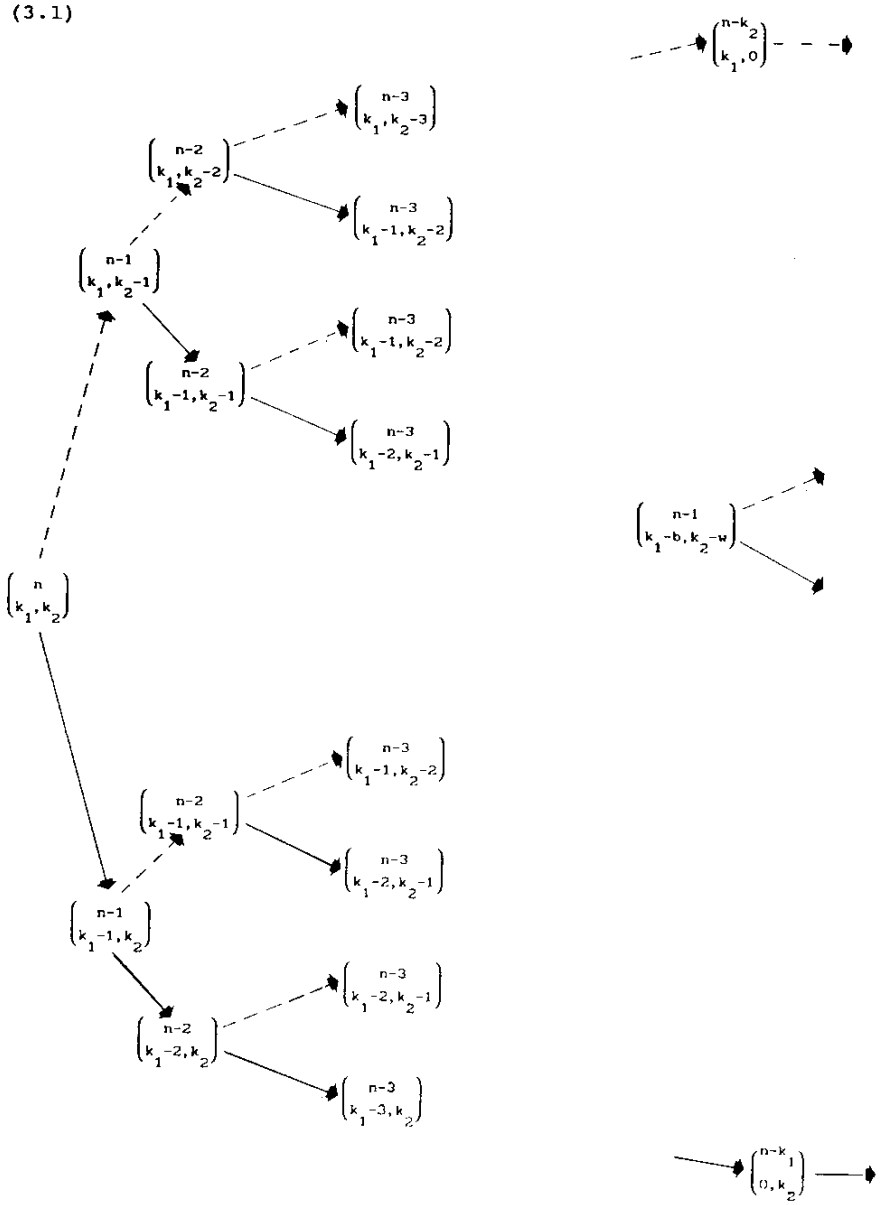
b) un vertice iniziale $\binom{n}{k_1,k_2}$ al variare di k_1 e k_2 con $k_1+k_2=n$;

c) vertici individuati dalle terne (i,b,w) con $1\le i\le n$ in cui i e' l'ordine dell'unita' territoriale di riferimento mentre b e w sono rispettivamente il numero degli eventi B e W che si sono realizzati dalla prima u.t. alla i-esima u.t. per cui $b+w=i$ con le limitazioni $0\le b\le k_1$ e $0\le w\le k_2$;

e dove la biforcazione dall'elemento generico $\binom{n-1}{k_1-b,k_2-w}$ e' eseguibile se il suo valore e' maggiore di 1.

Infatti, $\binom{n}{k_1,k_2}$ per ogni k_1, k_2 rappresenta il numero delle osservazioni $\underline{X}$ per le quali si e' realizzato k_1 volte l'evento B e k_2 volte l'evento W e se questo valore e' maggiore di 1, significa che per la prima u.t. puo' realizzarsi sia l'evento B che l'evento W e si analizzano i due casi si vede che $\binom{n}{k_1,k_2}$ si scinde in $\binom{n-1}{k_1,k_2-1}$ e $\binom{n-1}{k_1-1,k_2}$ la cui somma e' $\binom{n}{k_1,k_2}$ e, se uno dei due valori trovati e' maggiore di 1, significa che nella seconda u.t. e' di nuovo possibile la realizzazione sia dell'evento B che dell'evento W, cosi' proseguendo.

(3.1)


$\binom{n}{k_1,k_2}$
$\binom{n-1}{k_1,k_2-1}$
$\binom{n-2}{k_1,k_2-2}$
$\binom{n-3}{k_1,k_2-3}$
$\binom{n-3}{k_1-1,k_2-2}$
$\binom{n-2}{k_1-1,k_2-1}$
$\binom{n-3}{k_1-1,k_2-2}$
$\binom{n-3}{k_1-2,k_2-1}$
$\binom{n-1}{k_1-1,k_2}$
$\binom{n-2}{k_1-1,k_2-1}$
$\binom{n-3}{k_1-1,k_2-2}$
$\binom{n-3}{k_1-2,k_2-1}$
$\binom{n-2}{k_1-2,k_2}$
$\binom{n-3}{k_1-2,k_2-1}$
$\binom{n-3}{k_1-3,k_2}$
$\binom{n-k_2}{k_1,0}$
$\binom{n-1}{k_1-b,k_2-w}$
$\binom{n-k_1}{0,k_2}$

Inoltre l'elemento $\binom{n-1}{k_1-b,k_2-w}$, con $1 \le i \le n$, esprime in sintesi e le componenti i-esime delle frequenze $\underline{\nu} = \underline{f}_{BWvWB|V''}(\underline{X})$ attraverso i valori (k_1-b) o (k_2-w), secondo che si realizzi il w-esimo evento W o il b-esimo evento B sulla i-esima u.t., ed i modi di osservare tali valori nella quantita' data dallo stesso binomiale e le probabilita' $p^{k_1}q^{k_2}$ per ciascuna delle $\binom{n}{k_1,k_2}$ successioni differenti $\underline{X}$ di k_1 eventi B e k_2 eventi W.

Facendo riferimento allo schema (3.1) ed all'esempio, in cui e' n=5, la v.c. $\underline{N}_{BWvWB}(V'')$ e':

$\underline{\nu}$	0 0 0 0 0	41111 03111 00211 00011 00000	3333222222 3111222222 0211211112 0011011110 0000000000	2222223333 2222221113 2111121120 0111101100 0000000000	11114 11130 11200 11000 00000	0 0 0 0 0
$P(\underline{\nu})$	p^5	pq^4	p^2q^3	p^3q^2	p^4q	q^5

che ridotta a forma normale diventa:

ν	0	4			
$\underline{\nu}$	0 0 0 0 0	4 0 0 0 0	1 3 0 0 0	1 1 2 0 0	1 1 1 1 0
$P(\underline{\nu})$	p^5+q^5	p^4q+pq^4	p^4q+pq^4	p^4q+pq^4	$2p^4q+2pq^4$

6				
3 3 0 0 0	3 1 2 0 0	3 1 1 1 0	2 2 2 0 0	2 2 1 1 0
$p^3q^2+p^2q^3$	$p^3q^2+p^2q^3$	$2p^3q^2+2p^2q^3$	$2p^3q^2+2p^2q^3$	$4p^3q^2+4p^2q^3$

Se ora facciamo riferimento ad una specifica u.t., ad es. la i-esima, attraverso lo schema (3.1) possiamo determinare la v.c. $N_{BWvWB}(V_i'')$ (3.2), numero delle diversita' BW o WB riferite alla stessa u.t., le cui determinazioni $\nu_i \in [0,n-i]$ e le probabilita' $P(\nu_i)$ si ottengono per riduzione, secondo il principio delle probabilita' totali, dalle $P(\underline{\nu})$.

(3.2)

i \ ν_i	0	1	2	3	4
1	p^5+q^5	$4(p^4q+pq^4)$	$6p^2q^2$	$4p^2q^2$	p^4q+pq^4
2	p^4+q^4	$3(p^3q+pq^3)$	$6p^2q^2$	p^3q+pq^3	
3	p^3+q^3	$2pq$	pq		
4	p^2+q^2	$2pq$			
5	1				

Pertanto la v.c. $N_{BWvWB}([L_i''])$ ha come determinazioni i valori $\eta([L_i'']) \in [0,L_i'']$ con probabilita' (3.3)

$$P(\eta([L_i'])) = \sum_{\nu_i=\eta([L_i''])}^{n-i} \frac{\binom{\nu_i}{\eta([L_i''])}\binom{n-i-\nu_i}{L_i''-\eta([L_i''])}}{\binom{n-i}{L_i''}} P(\nu_i)$$

sotto le condizioni $\begin{cases} 0 \le L_i'' \le n-i \\ (n-i)-L_i''-(\nu_i-\eta([L_i'])) \ge 0 \end{cases}$.

Piu' in generale, per la v.c. $\underline{N}_{BWvWB}([\underline{L}''])$ le determinazioni $\underline{\eta}([\underline{L}''])$ sono costituite dalle successioni di n elementi, con n-esimo elemento 0, tali che queste abbiano almeno una successione $\underline{\nu}$ tra le determinazioni della v.c. $\underline{N}_{BWvWB}(\underline{V}'')$ per cui $\eta_i([\underline{L}'']) \le \nu_i$ ($i \in N_n$); allora per le $\underline{\eta}([\underline{L}''])$, sotto l'ipotesi di indipendenza, le

probabilita' sono espresse da (3.4):

$$P(\underline{\eta}([\underline{L}'']))=\sum_{\underline{\nu}\in \underline{N}_{BWvWB}(V'')}\left\{\prod_{i=1}^{n-1}\frac{\binom{\nu_i}{\eta_i([\underline{L}''])}\binom{n-i-\nu_i}{L_i''-\eta_i([\underline{L}''])}}{\binom{n-i}{L_i''}}\right\}P(\underline{\nu})$$

con le condizioni $\begin{cases}\nu_i-\eta_i([\underline{L}''])\geq 0\\ 0\leq L_i''\leq n-i\\ (n-i)-L_i''-(\nu_i-\eta_i([\underline{L}'']))\geq 0\end{cases}$

Infine, si determina la v.c. $N_{BWvWB}([\underline{L}''])$ per mezzo della v.c. $\underline{N}_{BWvWB}([\underline{L}''])$ ponendo $\eta([\underline{L}''])=\sum_{i=1}^{n}\eta_i([\underline{L}''])$ con $\eta_i([\underline{L}])$ componente i-esima di $\underline{\eta}([\underline{L}])$.

Per l'es. avendosi $(\underline{L}'')^T=(1,2,1,0,0)$ per p=.5 si hanno:

a) Le v.c. $N_{BWvWB}([L_i''])$

i \ $\eta([L_i''])$	0	1	2
1	.5	.5	
2	.25	.5	.25
3	.5	.5	
4	1		
5	1		

Si osservi, attraverso l'esempio, che $N_{BWvWB}([L_i''])=N_{BWvWB}(V_j'')$ per $j=n-L_i''$.

$\underline{L}''=\begin{vmatrix}00001\\00011\\00010\\00000\\00000\end{vmatrix}$, $(\underline{L}'')^T=(1,2,1,0,0)$ ed A=4, per p=.5 si hanno le distribuzioni:

η / P	0	1	2	3	4
$P(\eta(A))$	.0878	.1905	.4018	.2738	.0461
$P(\eta([\underline{L}'']))$	.0833	.2083	.3750	.2917	.0417

rispettivamente con media e varianza:

$E[N_{BWvWB}(A)]=2$ $\quad$ $E[N_{BWvWB}([\underline{L}''])]=2$

$Var[N_{BWvWB}(A)]=1$ $\quad$ $Var[N_{BWvWB}([\underline{L}''])]=1$

valori coincidenti con quelli ricavabili utilizzando le formule proposte da P.A.P. Moran (1948) per il calcolo dei momenti della statistica N_{BW}.

Ringrazio il Sig. F. Critani per la collaborazione prestata nella stesura dei programmi mirati alla realizzazione di questa nota e per i suggerimenti propositivi ai fini di una generalizzazione dei risultati sin qui ottenuti.

BIBLIOGRAFIA

Badaloni M.-Vinci E.,1988-"Contributi all'analisi dell'autocorrelazione spaziale", Metron vol.XLVI, n.1-4, pp.119-140.

Cliff A.D.-Hagget P.-Ord J.K.-Basset K.A.-Davies R.B.-"Elements of spatial structure"- Cambridge University Press, London, 1975.

Cliff A.D.-Ord J.K.- "Spatial autocorrelation" - Pion Limited, London, 1973.

Cliff A.D.-Ord J.K.- "Spatial process" - Pion Limited, London, 1981.

Coli M., 1982- "Sulla previsione dei dati territoriali: la

b) La v.c. $N_{BWvWB}([\underline{L}''])$

η	0	1			2			
$\underline{\eta}$	0 0 0 0 0	1 0 0 0 0	0 1 0 0 0	0 0 1 0 0	1 1 0 0 0	1 0 1 0 0	0 1 1 0 0	0 2 0 0 0
$P(\underline{\eta})$	.0833	.0833	.0833	.0417	.0833	.0417	.1667	.0833

3			4
1 1 1 0 0	1 2 0 0 0	0 2 1 0 0	1 2 1 0 0
.1667	.0833	.0417	.0417

c) La v.c. $N_{BWvWB}([\underline{L}''])$

η	0	1	2	3	4
$P(\eta)$	.0833	.2083	.3750	.2917	.0417

4. CONCLUSIONI

La ricerca presentata in questa nota evidenzia come attraverso l'acquisizione di informazioni quali la simmetria, la numerosita' dei legami e la distribuzione degli stessi per unita' territoriale, si possano considerare distribuzioni di probabilita' che "tendono" alla distribuzione $N_{BW}(L)$ del numero delle diversita' BW per lo specifico operatore di contiguita' L.

Infatti riferendosi all'esempio proposto per il quale:

funzione di autocorrelazione spazio-temporale per l'identificazione e la stima di modelli della classe STARMA" - Atti della XXXI Riunione Scientifica S.I.S., Torino.

Dacey, M.F., 1965- "A review on measures of contiguity for two and k-color maps", Technical Report N.2, Spatial Diffusion Study, Department of Geography Northwestern University, Evanston, Illinois.

Krishna Iyer, P.V.A., 1949- "The first and second moments of some probability distributions arising from points on a lattice, and their applications", Biometrika 36,135-141.

Moran P.A.P., 1948- "The interpretation of statistical maps" - J.R.S.S., Series B,10,243-251.

Rényi A.- "Calcul des probabilités" - Dunod, Paris, 1966.

Visini G.-Sclocco T.,1981-"Sull'indice di autocorrelazione spaziale C di Geary",Atti del V Convegno A.M.A.S.E.S., pp.277-298, Perugia 22-24 ottobre 1981.

Visini G.-Sclocco T.,1982-"Sui primi due momenti della statistica "join count" di Moran",Atti del VI Convegno A.M.A.S.E.S., pp.621-642, Marina di Campo (Isola d'Elba) 1-3 ottobre 1982.

Visini G.,1985-"Sull'analisi di carte tematiche bicromatiche attraverso l'autocorrelazione spazio-temporale",Atti Giornate di Lavoro A.I.R.O. 1985,pp.211-229, Venezia 30 settembre-2 ottobre 1985, Ed. Tecnoprint Bologna.

Visini G.,1989-"Considerazioni sul test z per gli indici di Moran",Pescara, accettata per la pubblicazione sugli Atti del I Convegno di Matematica Applicata all'Economia e alla Ingegneria,26-28 gennaio 1989,Pescara.

Visini G.,1990-"Considerazioni sulla statistica join count e l'informazione di Shannon",Convegno Giornate di Lavoro A.I.R.O. 1990, Sorrento 3-5 ottobre 1990, pp.853-865, Ed.Fast Print, Napoli.

Visini G.,1990-"Considerazioni sull'indice di correlazione spaziale",Giornate di Studio "Analisi dei dati spaziali e classificazione dei dati", Pescara 11-12 ottobre 1990, pp.477-488, Ed. Solfanelli M., Chieti.

SULL'AUTOCORRELAZIONE SPAZIALE DI FENOMENI QUALITATIVI

Giuliano VISINI, Dip. di Scienze, Storia dell'Architettura e Restauro
Facoltà di Architettura - Università "G. D'Annunzio" di Chieti

ABSTRACT

It is well known that Moran (1948), Krishna Iyer (1949), Dacey (1965), Cliff (1969) and Cliff-Ord (1973, 1981) determined the first steps of the join count statistic.

Nevertheless these steps, on account of the parameters (A, D, etc.) by which they are characterized, can be referred to different contiguity hypothesis.

On the other end, the problem concerning the determination of the join count statistics still remains open for a finite subdivision of a region $\{a_i\}_{i\in N_n}$, which can be defined through small values of n and generally only through some hypothesis of contiguity.

In this note the casual variable $N_{BW}(V)$ of the contiguous diversities type BW will be determined, limitedly referred to the case of dicotomical observations (type {B, W}) and independent on n regional unity, in the particular case of the contiguity operator V defined for each r.u. a_i as $V(a_i)=\{a_j \mid j\in N_n-\{i\}\}$.

Thus, taken for equivalent the operators of simmetrical contiguities of the same A parameter (A stands for the bonds

number), $A = \sum_{i=1}^{n} L''_i = \sum_{i=1}^{n} L'_i$, it will be determined the c.v. $N_{BW}(2A)$ of the contiguous diversities type BW, these being pointed out for the equivalent class $[L]$ identifiable in the parameter 2A.

Eventuay, being observed that, on account of the symmetry bond of the operator, the c.v. $N_{BW}(2A)$ is similar to the c.v. $N_{BWvWB}(A)$ of the contiguous diversities type BW or WB, when associated to the class of non-symmetrical operators L'' (L') and L limitations, those symmetricall operators L with $\underline{L}''=(L''_i)_{i\in N_n}$ (or $\underline{L}'=((L'_i)_{i\in N_n})$) are considered equivalent.

Limitedly referring to the previous hypothesis (dichotomy and independence) and to the V'' operator, limitation of V, c.v. is defined as $\underline{N}_{BWvWB}((V''_i)_{i\in N_n})$ which has, as determination of n successions, values whose i^{th} element stands for the amount of diversities exisiting between the r.u. a_i, and those to follow and from this the c.v. $N_{BWvWB}([(L''_i)_{i\in N_n}])$ through the use of a suitable algoritm.

Key-words: two colour maps; contiguity; finite populations; join-count; distributions.

SOMMARIO

E' noto come Moran (1948), Krishna Iyer (1949), Dacey (1965), Cliff (1969) e Cliff-Ord (1973, 1981) abbiano determinato i primi momenti della statistica join count.

Tuttavia questi momenti per i parametri (A, D, etc.) che li caratterizzano possono essere riferiti ad ipotesi diverse di contiguita'.

D'altra parte, la determinazione della distribuzione della statistica join count, per una suddivisione finita di un territorio $\{a_i\}_{i\in N_n}$, definibile per piccoli valori di n ed in generale solo per alcune ipotesi di contiguita', resta ancora un

problema aperto.

In questa nota, limitatamente al caso di osservazioni dicotomiche (di tipo {B, W}) ed indipendenti sulle n unita' territoriali, si determina, nel caso particolare dell'operatore di contiguita' V, definito per ogni u.t. a_i come $V(a_i)=\{a_j | j\in N_n-\{i\}\}$, la variabile casuale $N_{BW}(V)$ delle diversita' contigue di tipo BW.

Poi, considerati equivalenti gli operatori di contiguita' simmetrici L di medesimo parametro A (numero dei legami), $A = \sum_{i=1}^{n} L''_i = \sum_{i=1}^{n} L'_i$, si determina la v.c. $N_{BW}(2A)$ delle diversita' contigue di tipo BW rilevabili per la classe d'equivalenza $[L]$ identificabile nel parametro 2A.

Successivamente, osservato che per il vincolo di simmetria degli operatori L, la v.c. $N_{BW}(2A)$ e' uguale alla v.c. $N_{BWvWB}(A)$ delle diversita' contigue del tipo BW o WB, associata alla classe degli operatori non simmetrici L'' (L') limitazioni di L, si considerano equivalenti quegli operatori simmetrici L aventi medesima n-pla $\underline{L}'' = (L''_i)_{i\in N_n}$ (o $\underline{L}' = (L'_i)_{i\in N_n}$).

Limitatamente alle precedenti ipotesi (dicotomia ed indipendenza) e all'operatore V'', limitazione di V, si definisce la v.c. $\underline{N}_{BWvWB}((V''_i)_{i\in N_n})$, che ha come determinazioni delle successioni di n valori il cui elemento i-esimo rappresenta il numero delle diversita' esistenti tra l'u.t. a_i e le successive, e da questa la v.c. $N_{BWvWB}([(L''_i)_{i\in N_n}])$ attraverso l'introduzione di un opportuno algoritmo.

Parole chiavi: mappe a due colori; contiguita'; popolazioni finite; join-count; distribuzioni.

1. INTRODUZIONE

Sia $N_n = \{1,2,\ldots,n\}$ ed $n \geq 2$; se $\{a_i \mid i \in N_n\}$ indica una suddivisione territoriale finita di un territorio, un criterio di contiguita' spaziale tra le unita' territoriali e' definito da L,

$$L: \{a_i \mid i \in N_n\} \longrightarrow \{a_i \mid i \in N_n\},$$

che esprime una corrispondenza generalmente multivoca e non ovunque definita tale che $L(a_i) \subset \{a_j \mid j \in N_n,\ j \neq i\}$. Nel caso valga l'implicazione

$$a_j \in L(a_i) \Longrightarrow a_i \in L(a_j) \qquad (j \in N_n),$$

l'operatore di associazione L lo diremo simmetrico.

La funzione I,indicatrice delle unita' territoriali contigue,

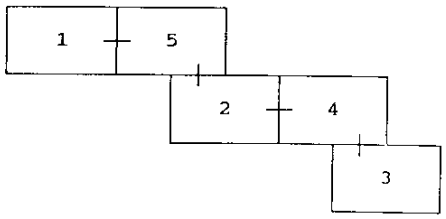
$$I: \{a_i \mid i \in N_n\} \times \{a_i \mid i \in N_n\} \longrightarrow \{0,1\}$$

tale che

$$I(a_p,a_q)=1 \Longleftrightarrow a_q \in L(a_p),$$

individua una matrice $\mathbb{L}$, detta matrice delle contiguita', che e' in corrispondenza biunivoca con l'operatore L.

Ad es., se si e' indicata con —— la contiguita' spaziale



all'operatore L resta associato la matrice $\mathbb{L}$:

$$\begin{vmatrix} 00001 \\ 00011 \\ 00010 \\ 01100 \\ 11000 \end{vmatrix}$$

Ad $\mathbb{L}$ sono legati dei parametri descrittivi, gia' noti in letteratura in quanto impiegati nella ricerca dei momenti degli indici di Moran, quali:

$$L_1=|L(a_1)|,\quad \sum_{i=1}^{n} L_i=2A,\quad \sum_{i=1}^{n} L_i(L_i-1)=2D,\text{etc..}$$

A questi aggiungiamo in particolare:

$$L'_i = L'(a_i) \qquad \text{ed} \qquad L''_i = L''(a_i) \qquad (i\in N_n)$$

per i quali rispettivamente L' indica la limitazione di L tale che

$$L'(a_i) = L(a_i) \cap \{a_j |\ j\in N_n,\ j<i\} \qquad (i\in N_n)$$

ed L'' la limitazione per cui

$$L''(a_i) = L(a_i) \cap \{a_j |\ j\in N_n,\ j>i\} \qquad (i\in N_n).$$

Tra tali parametri intercorrono le relazioni:

$$L_i = L'_i + L''_i \qquad (i\in N_n)$$

con $L'_1 = L''_n = 0$ e $0\le L'_i\le i-1$, $0\le L''_i\le n-i$.

$$\mathbb{L}'=\begin{vmatrix}00000\\00000\\00000\\01100\\11000\end{vmatrix} \quad \underline{L}'=\begin{pmatrix}0\\0\\0\\2\\2\end{pmatrix} \qquad \mathbb{L}''=\begin{vmatrix}00001\\00011\\00010\\00000\\00000\end{vmatrix} \quad \underline{L}''=\begin{pmatrix}1\\2\\1\\0\\0\end{pmatrix}$$

Sia $E=\{E_1,E_2,\ldots,E_m\}$ un sistema completo di m eventi con probabilita' $P(E_1)=p_1$, $P(E_2)=p_2$, ..., $P(E_m)=p_m$ e $\sum_{i=1}^{m} p_i=1$; se si assume che una successione di n eventi $\underline{X}=(X_i)_{i\in N_n}$ sia il risultato di n prove indipendenti per le quali si sia verificato k_1 volte l'evento E_1, k_2 volte l'evento E_2, ..., k_m volte l'evento E_m ($k_1+k_2+\ldots+k_m=n$) allora all'espressione $\underline{X}^T\mathbb{L}\underline{X}$ compete la probabilita':

$$P(\underline{X})=p_1^{k_1}\cdot p_2^{k_2}\cdot\ldots\cdot p_m^{k_m}.$$

Dall'uguaglianza

$$\underline{X}^T\mathbb{L}\underline{X} = E^T\mathbb{F}_L(\underline{X})E, \tag{1.1}$$

ove $\mathbb{F}_L(\underline{X})$ e' la matrice quadrata di ordine m costituita dalle frequenze assolute degli eventi $\{E_1E_1,\ E_1E_2,\ldots,\ E_1E_m,\ E_2E_1,\ldots,\ E_mE_m\}$, si ha che le espressioni $\underline{X}^T\mathbb{L}\underline{X}$ e $E^T\mathbb{F}_L(\underline{X})E$ si realizzano con la medesima probabilita' $P(\underline{X})$, come anche ogni elemento dello sviluppo di $E^T\mathbb{F}_L(\underline{X})E$, ogni singolo elemento di $\mathbb{F}_L(\underline{X})$, etc..

La forma algebrica $\sum_{i=1}^{n} \mathbb{F}_{L(a_i)}(\underline{X})$, equivalente della $E^T\mathbb{F}_L(\underline{X})E$, in cui $\mathbb{F}_{L(a_i)}(\underline{X})$ indica la matrice di ordine m i cui elementi $f_{rsL(a_i)}(\underline{X})$ sono le frequenze della modalita' E_rE_s di cui la prima E_r rilevabile sull'unita' territoriale a_i e la seconda E_s sull'unita' $a_j \in L(a_i)$, porta a riconsiderare l'espressione $\underline{X}^T\mathbb{L}\underline{X}$ riproponendola in notazione vettoriale-matriciale $\left(\mathbb{F}_{L(a_i)}(\underline{X})\right)_{i\in N_n}$ piu' efficace per le considerazioni che seguono. In particolare, riferendosi alla modalita' E_rE_s si hanno le n-ple di frequenze $\underline{f}_{rsL}(\underline{X})=(f_{rsL(a_i)}(\underline{X}))_{i\in N_n}$.

Nel precedente es. se $E=\{E_1, E_2, E_3\}$ ed $\underline{X}^T=(E_1,E_2,E_1,E_3,E_2)$ si ha

		a_1	a_2	a_3	a_4	a_5
		$E_1E_2E_3$	$E_1E_2E_3$	$E_1E_2E_3$	$E_1E_2E_3$	$E_1E_2E_3$
a_1	E_1	0 0 0	0 0 0	0 0 0	0 0 0	0 1 0
	E_2	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_3	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
a_2	E_1	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_2	0 0 0	0 0 0	0 0 0	0 0 1	0 1 0
	E_3	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
a_3	E_1	0 0 0	0 0 0	0 0 0	0 0 1	0 0 0
	E_2	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_3	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
a_4	E_1	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_2	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_3	0 0 0	0 1 0	1 0 0	0 0 0	0 0 0
a_5	E_1	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
	E_2	1 0 0	0 1 0	0 0 0	0 0 0	0 0 0
	E_3	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0

Le componenti di $\left(\mathbb{F}_{L(a_i)}(\underline{X})\right)_{i\in N_n}$ sono:

$$\mathbb{F}_{L(a_1)}(\underline{X}) = \begin{vmatrix} 010 \\ 000 \\ 000 \end{vmatrix} \qquad \mathbb{F}_{L(a_2)}(\underline{X}) = \begin{vmatrix} 000 \\ 011 \\ 000 \end{vmatrix} \qquad \mathbb{F}_{L(a_3)}(\underline{X}) = \begin{vmatrix} 001 \\ 000 \\ 000 \end{vmatrix}$$

$$\mathbb{F}_{L(a_4)}(\underline{X}) = \begin{vmatrix} 000 \\ 000 \\ 110 \end{vmatrix} \qquad \mathbb{F}_{L(a_5)}(\underline{X}) = \begin{vmatrix} 000 \\ 110 \\ 000 \end{vmatrix}$$

ed $\mathbb{F}_L(\underline{X}) = \sum_{i=1}^{5} \mathbb{F}_{L(a_i)}(\underline{X}) = \begin{vmatrix} 011 \\ 121 \\ 110 \end{vmatrix}$.

Le n-ple delle frequenze delle modalita' $E_r E_s$ con riferimento all'es. sono:

$\underline{f}^T_{11L}(\underline{X}) = (00000)$ $\quad \underline{f}^T_{12L}(\underline{X}) = (10000)$ $\quad \underline{f}^T_{13L}(\underline{X}) = (00100)$

$\underline{f}^T_{21L}(\underline{X}) = (00001)$ $\quad \underline{f}^T_{22L}(\underline{X}) = (01001)$ $\quad \underline{f}^T_{23L}(\underline{X}) = (01000)$

$\underline{f}^T_{31L}(\underline{X}) = (00010)$ $\quad \underline{f}^T_{32L}(\underline{X}) = (00010)$ $\quad \underline{f}^T_{33L}(\underline{X}) = (00000)$.

2. LA VARIABILE CASUALE $N_{BW}(2A)$.

Sia E la variabile casuale $\begin{pmatrix} B & W \\ p & q \end{pmatrix}$ osservabile su ogni unita' territoriale a_i e sia $\underline{X}=(X_i)_{i\in N_n}$ una successione di n eventi indipendenti osservabile sulla suddivisione territoriale finita.

Al variare di $\underline{X}$ resta definita la v.c. $N_{BW}(L)$ delle diversita' contigue BW, dipendente dal particolare operatore di associazione L, le cui determinazioni η altro non sono che frequenze $f_{BWL}(\underline{X})$.

In particolare, per l'operatore V, la cui matrice delle contiguita' $\mathbb{V}$ ha generico elemento $v_{ij} = \begin{cases} 1 & \text{per } i \neq j \\ 0 & \text{per } i=j \end{cases}$, la v.c. $N_{BW}(V)$ ha la seguente distribuzione (2.1):

ν	0	n-1	...	$k_1 k_2$	...
$P(\nu)$	p^n+q^n	$n(pq^{n-1}+p^{n-1}q)$	...	$\frac{n!}{k_1!k_2!}\left(p^{k_1}q^{k_2}+p^{k_2}q^{k_1}\right)$	...

...	[n/2][(n+1)/2]
...	.....

$$\begin{cases} k_i \geq 0 \ (i \in N_2) \\ k_1+k_2=n \end{cases} .$$

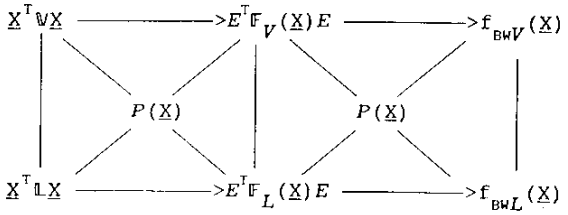
Per attribuire all'evento [n/2][(n+1)/2] la sua probabilita', si distinguono due casi a_0) e a_1) rispettivamente riferiti ad n pari e n dispari:

a_0) $P(n^2/4) = \binom{n}{n/2}(pq)^{n/2}$,

a_1) $P([n/2][(n+1)/2]) =$

$$= \binom{n}{[n/2]}\left(p^{[n/2]}\, q^{[(n+1)/2]} + p^{[(n+1)/2]}\, q^{[n/2]}\right).$$

Ove si consideri un generico operatore simmetrico L, la corrispondente variabile casuale $N_{BW}(L)$, le cui determinazioni $\nu(L)$ sono le $f_{BWL}(\underline{X})$, resta definita attraverso la commutativita' dei diagrammi ottenuti al variare delle n-ple osservabili $\underline{X}$:



Si noti tuttavia che la cardinalita' 2^n dell'insieme delle

n-ple $\underline{X}$ rende eccessivamente oneroso il calcolo per i valori di n riscontrabili nei casi concreti.

Se introduciamo nell'insieme degli operatori di associazione $\{L\}$ la relazione di equivalenza ~ cosi' definita:

$$L_1 \sim L_2 \Longleftrightarrow \sum_{i=1}^{n} L_{1i} = \sum_{i=1}^{n} L_{2i} = 2A,$$

ed indichiamo con 2A la classe degli operatori equivalenti $[L]$ allora ad ogni n-pla ordinata $\underline{X}$ corrisponde una $\left(\binom{n}{2}\right)$-pla non ordinata $\left(f_{BW_L}(\underline{X})\right)_{L\in 2A}$.

Quindi, indicata con $N_{BW}(2A)$ la variabile casuale le cui determinazioni $\eta(2A)$ sono le $f_{BW}(L,\underline{X})$ al variare di L nella classe 2A e di $\underline{X}$, la probabilita' di $\eta(2A)$ e' somma di probabilita' composte di eventi indipendenti di cui l'uno fa riferimento ad una distribuzione ipergeometrica e l'altro alla distribuzione binomiale (2.1):

$$(2.2) \qquad P(\eta(2A)) = \sum_{\nu\in N_{BW}(V)} \frac{\binom{\nu}{\eta(2A)}\binom{\binom{n}{2}-\nu}{A-\eta(2A)}}{\binom{\binom{n}{2}}{A}} P(\nu)$$

con le condizioni $\begin{cases} 1\le A\le \binom{n}{2} \\ \nu-\eta(2A)\ge 0 \\ \binom{n}{2}-A-(\nu-\eta(2A))\ge 0 \end{cases}$

In particolare per $A=\binom{n}{2}$ si ha $N_{BW}\left\{2\binom{n}{2}\right\} = N_{BW}(V)$ mentre per A=1 la v.c. $N_{BW}(2)$ ha distribuzione:

η	0	1
$P(\eta)$	p^2+q^2	2pq

.

Facendo riferimento all'es. per p =.5 la v.c. $N_{BW}(8)$ ha la seguente distribuzione:

η	0	1	2	3	4
$P(\eta)$	.0878	.1905	.4018	.2738	.0461

.

3.LA VARIABILE CASUALE $N_{BWvWB}(\underline{L}")$.

Se $V"$ e' l'operatore limitazione di V, la cui matrice W" ha elementi $v"_{ij}=\begin{cases} 1 \text{ per } i<j \\ 0 \text{ per } i\geq j \end{cases}$, allora la v.c. $N_{BWvWB}(V")$, che ha per determinazioni il numero complessivo delle diversita' tipo BW o WB, e' esattamente la v.c. $N_{BW}(V)$ (2.1) e, piu' in generale, per la simmetria e' anche $N_{BWvWB}(L") = N_{BW}(L)$ come anche $N_{BW}(2A) = N_{BWvWB}(A)$, essendo A la classe degli operatori $L"$ tali che $\sum_{i=1}^{n} L"_i=A$.

Ora gli $\left(\binom{n}{2}\right)_A$ operatori equivalenti di classe $\dot{A} = [L"]$ possono essere ripartiti in sottoclassi $[\underline{L}"] = [(L"_i)_{i\in N_n}]$ attraverso la nuova relazione d'equivalenza:

$$L"_1 \approx L"_2 \iff (L"_{1i})_{i\in N_n}=(L"_{2i})_{i\in N_n}.$$

In particolare gli insieme quozienti $\{L"\}/-$ e $\{L"\}/\approx$ hanno rispettivamente la classe $A=\binom{n}{2}$ e la classe $[\underline{L}"]=[(n-i)_{i\in N_n}]$ costituita dal solo operatore $V"$.

Al fine di definire la v.c $N_{BWvWB}([\underline{L}"])$, che ha per determinazioni il numero complessivo delle diversita' di tipo BW o

WB al variare dell'operatore nella classe [$\underline{L}$"], introduciamo la v.c. $\underline{N}_{BWvWB}(\underline{V}")$, che ha come determinazioni delle successioni di n valori il cui elemento i-esimo e' il numero delle diversita'(o BW o WB) esistenti tra la u.t. a_i e le successive u.t. a_j con $j\in(N_n-N_1)$.

E' possibile rappresentare il processo di generazione delle diverse determinazioni di $\underline{N}_{BWvWB}(\underline{V}")$ con il seguente schema (3.1), in cui distinguiamo:

a) due tipi di frecce ———> e - - ->, le quali esprimono rispettivamente il realizzarsi dell'evento B e dell'evento W;

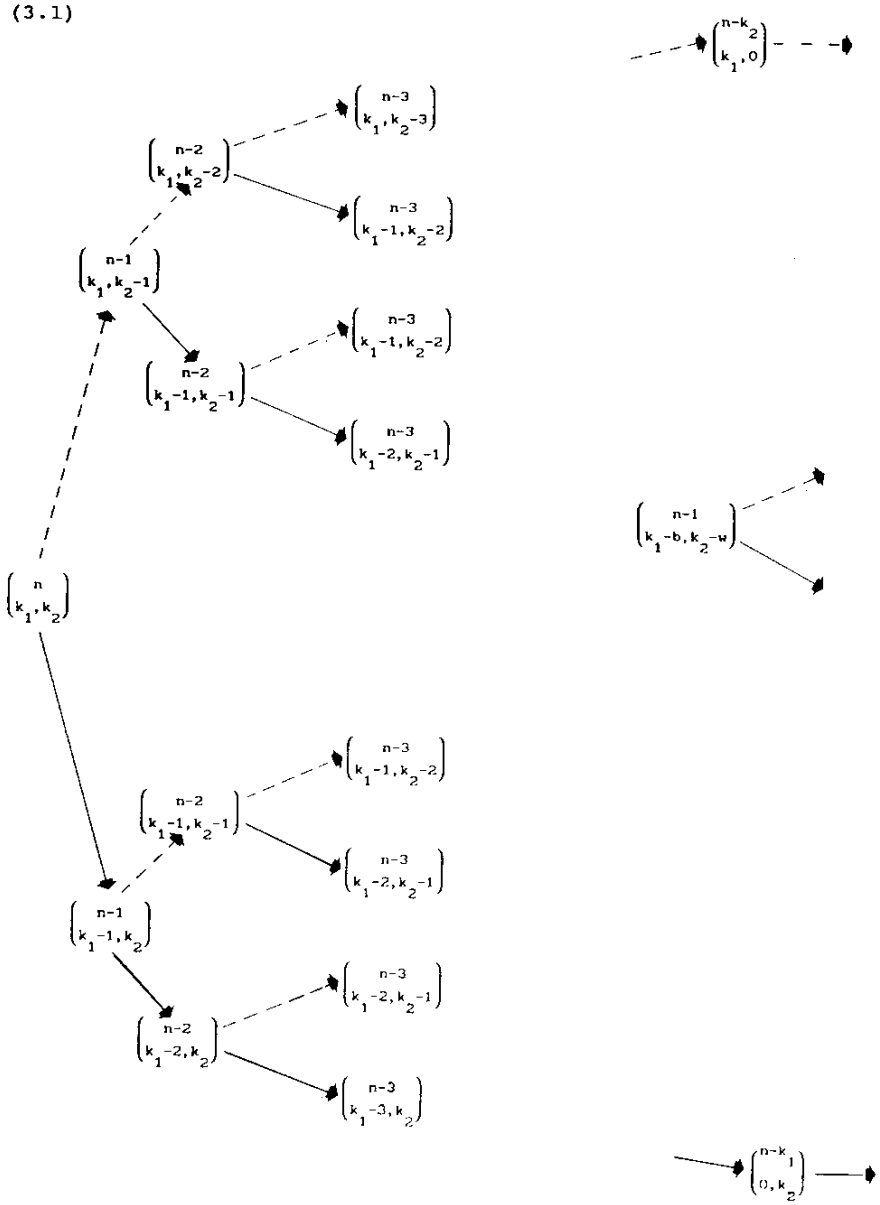
b) un vertice iniziale $\binom{n}{k_1,k_2}$ al variare di k_1 e k_2 con $k_1+k_2=n$;

c) vertici individuati dalle terne (i,b,w) con $1\le i\le n$ in cui i e' l'ordine dell'unita' territoriale di riferimento mentre b e w sono rispettivamente il numero degli eventi B e W che si sono realizzati dalla prima u.t. alla i-esima u.t. per cui $b+w=i$ con le limitazioni $0\le b\le k_1$ e $0\le w\le k_2$;

e dove la biforcazione dall'elemento generico $\binom{n-1}{k_1-b,k_2-w}$ e' eseguibile se il suo valore e' maggiore di 1.

Infatti, $\binom{n}{k_1,k_2}$ per ogni k_1, k_2 rappresenta il numero delle osservazioni $\underline{X}$ per le quali si e' realizzato k_1 volte l'evento B e k_2 volte l'evento W e se questo valore e' maggiore di 1, significa che per la prima u.t. puo' realizzarsi sia l'evento B che l'evento W e si analizzano i due casi si vede che $\binom{n}{k_1,k_2}$ si scinde in $\binom{n-1}{k_1,k_2-1}$ e $\binom{n-1}{k_1-1,k_2}$ la cui somma e' $\binom{n}{k_1,k_2}$ e, se uno dei due valori trovati e' maggiore di 1, significa che nella seconda u.t. e' di nuovo possibile la realizzazione sia dell'evento B che dell'evento W, cosi' proseguendo.

(3.1)


$\binom{n}{k_1,k_2}$
$\binom{n-1}{k_1,k_2-1}$
$\binom{n-2}{k_1,k_2-2}$
$\binom{n-3}{k_1,k_2-3}$
$\binom{n-3}{k_1-1,k_2-2}$
$\binom{n-2}{k_1-1,k_2-1}$
$\binom{n-3}{k_1-1,k_2-2}$
$\binom{n-3}{k_1-2,k_2-1}$
$\binom{n-1}{k_1-1,k_2}$
$\binom{n-2}{k_1-1,k_2-1}$
$\binom{n-3}{k_1-1,k_2-2}$
$\binom{n-3}{k_1-2,k_2-1}$
$\binom{n-2}{k_1-2,k_2}$
$\binom{n-3}{k_1-2,k_2-1}$
$\binom{n-3}{k_1-3,k_2}$
$\binom{n-k_2}{k_1,0}$
$\binom{n-1}{k_1-b,k_2-w}$
$\binom{n-k_1}{0,k_2}$

Inoltre l'elemento $\binom{n-1}{k_1-b,k_2-w}$, con $1 \le i \le n$, esprime in sintesi e le componenti i-esime delle frequenze $\underline{\nu} = \underline{f}_{BWvWB|V''}(\underline{X})$ attraverso i valori (k_1-b) o (k_2-w), secondo che si realizzi il w-esimo evento W o il b-esimo evento B sulla i-esima u.t., ed i modi di osservare tali valori nella quantita' data dallo stesso binomiale e le probabilita' $p^{k_1}q^{k_2}$ per ciascuna delle $\binom{n}{k_1,k_2}$ successioni differenti $\underline{X}$ di k_1 eventi B e k_2 eventi W.

Facendo riferimento allo schema (3.1) ed all'esempio, in cui e' n=5, la v.c. $\underline{N}_{BWvWB}(V'')$ e':

$\underline{\nu}$	0 0 0 0 0	41111 03111 00211 00011 00000	3333222222 3111222222 0211211112 0011011110 0000000000	2222223333 2222221113 2111121120 0111101100 0000000000	11114 11130 11200 11000 00000	0 0 0 0 0
$P(\underline{\nu})$	p^5	pq^4	p^2q^3	p^3q^2	p^4q	q^5

che ridotta a forma normale diventa:

ν	0	4			
$\underline{\nu}$	0 0 0 0 0	4 0 0 0 0	1 3 0 0 0	1 1 2 0 0	1 1 1 1 0
$P(\underline{\nu})$	p^5+q^5	p^4q+pq^4	p^4q+pq^4	p^4q+pq^4	$2p^4q+2pq^4$

6				
3 3 0 0 0	3 1 2 0 0	3 1 1 1 0	2 2 2 0 0	2 2 1 1 0
$p^3q^2+p^2q^3$	$p^3q^2+p^2q^3$	$2p^3q^2+2p^2q^3$	$2p^3q^2+2p^2q^3$	$4p^3q^2+4p^2q^3$

Se ora facciamo riferimento ad una specifica u.t., ad es. la i-esima, attraverso lo schema (3.1) possiamo determinare la v.c. $N_{BWvWB}(V_i'')$ (3.2), numero delle diversita' BW o WB riferite alla stessa u.t., le cui determinazioni $\nu_i \in [0,n-i]$ e le probabilita' $P(\nu_i)$ si ottengono per riduzione, secondo il principio delle probabilita' totali, dalle $P(\underline{\nu})$.

(3.2)

i \ ν_i	0	1	2	3	4
1	p^5+q^5	$4(p^4q+pq^4)$	$6p^2q^2$	$4p^2q^2$	p^4q+pq^4
2	p^4+q^4	$3(p^3q+pq^3)$	$6p^2q^2$	p^3q+pq^3	
3	p^3+q^3	$2pq$	pq		
4	p^2+q^2	$2pq$			
5	1				

Pertanto la v.c. $N_{BWvWB}([L_i''])$ ha come determinazioni i valori $\eta([L_i'']) \in [0,L_i'']$ con probabilita' (3.3)

$$P(\eta([L_i'])) = \sum_{\nu_i=\eta([L_i''])}^{n-i} \frac{\binom{\nu_i}{\eta([L_i''])}\binom{n-i-\nu_i}{L_i''-\eta([L_i''])}}{\binom{n-i}{L_i''}} P(\nu_i)$$

sotto le condizioni $\begin{cases} 0 \le L_i'' \le n-i \\ (n-i)-L_i''-(\nu_i-\eta([L_i'])) \ge 0 \end{cases}$.

Piu' in generale, per la v.c. $\underline{N}_{BWvWB}([\underline{L}''])$ le determinazioni $\underline{\eta}([\underline{L}''])$ sono costituite dalle successioni di n elementi, con n-esimo elemento 0, tali che queste abbiano almeno una successione $\underline{\nu}$ tra le determinazioni della v.c. $\underline{N}_{BWvWB}(\underline{V}'')$ per cui $\eta_i([\underline{L}'']) \le \nu_i$ ($i \in N_n$); allora per le $\underline{\eta}([\underline{L}''])$, sotto l'ipotesi di indipendenza, le

probabilita' sono espresse da (3.4):

$$P(\underline{\eta}([\underline{L}'']))=\sum_{\underline{\nu}\in \underline{N}_{BWvWB}(V'')}\left\{\prod_{i=1}^{n-1}\frac{\binom{\nu_i}{\eta_i([\underline{L}''])}\binom{n-i-\nu_i}{L_i''-\eta_i([\underline{L}''])}}{\binom{n-i}{L_i''}}\right\}P(\underline{\nu})$$

con le condizioni $\begin{cases}\nu_i-\eta_i([\underline{L}''])\geq 0\\ 0\leq L_i''\leq n-i\\ (n-i)-L_i''-(\nu_i-\eta_i([\underline{L}'']))\geq 0\end{cases}$

Infine, si determina la v.c. $N_{BWvWB}([\underline{L}''])$ per mezzo della v.c. $\underline{N}_{BWvWB}([\underline{L}''])$ ponendo $\eta([\underline{L}''])=\sum_{i=1}^{n}\eta_i([\underline{L}''])$ con $\eta_i([\underline{L}])$ componente i-esima di $\underline{\eta}([\underline{L}])$.

Per l'es. avendosi $(\underline{L}'')^T=(1,2,1,0,0)$ per p=.5 si hanno:

a) Le v.c. $N_{BWvWB}([L_i''])$

i \ $\eta([L_i''])$	0	1	2
1	.5	.5	
2	.25	.5	.25
3	.5	.5	
4	1		
5	1		

Si osservi, attraverso l'esempio, che $N_{BWvWB}([L_i''])=N_{BWvWB}(V_j'')$ per $j=n-L_i''$.

$\underline{L}''=\begin{vmatrix}00001\\00011\\00010\\00000\\00000\end{vmatrix}$, $(\underline{L}'')^T=(1,2,1,0,0)$ ed A=4, per p=.5 si hanno le distribuzioni:

η / P	0	1	2	3	4
$P(\eta(A))$	.0878	.1905	.4018	.2738	.0461
$P(\eta([\underline{L}'']))$	.0833	.2083	.3750	.2917	.0417

rispettivamente con media e varianza:

$E[N_{BWvWB}(A)]=2$ $\quad$ $E[N_{BWvWB}([\underline{L}''])]=2$

$Var[N_{BWvWB}(A)]=1$ $\quad$ $Var[N_{BWvWB}([\underline{L}''])]=1$

valori coincidenti con quelli ricavabili utilizzando le formule proposte da P.A.P. Moran (1948) per il calcolo dei momenti della statistica N_{BW}.

Ringrazio il Sig. F. Critani per la collaborazione prestata nella stesura dei programmi mirati alla realizzazione di questa nota e per i suggerimenti propositivi ai fini di una generalizzazione dei risultati sin qui ottenuti.

BIBLIOGRAFIA

Badaloni M.-Vinci E.,1988-"Contributi all'analisi dell'autocorrelazione spaziale", Metron vol.XLVI, n.1-4, pp.119-140.

Cliff A.D.-Hagget P.-Ord J.K.-Basset K.A.-Davies R.B.-"Elements of spatial structure"- Cambridge University Press, London, 1975.

Cliff A.D.-Ord J.K.- "Spatial autocorrelation" - Pion Limited, London, 1973.

Cliff A.D.-Ord J.K.- "Spatial process" - Pion Limited, London, 1981.

Coli M., 1982- "Sulla previsione dei dati territoriali: la

b) La v.c. $N_{BWvWB}([\underline{L}''])$

η	0	1			2			
$\underline{\eta}$	0 0 0 0 0	1 0 0 0 0	0 1 0 0 0	0 0 1 0 0	1 1 0 0 0	1 0 1 0 0	0 1 1 0 0	0 2 0 0 0
$P(\underline{\eta})$	.0833	.0833	.0833	.0417	.0833	.0417	.1667	.0833

3			4
1 1 1 0 0	1 2 0 0 0	0 2 1 0 0	1 2 1 0 0
.1667	.0833	.0417	.0417

c) La v.c. $N_{BWvWB}([\underline{L}''])$

η	0	1	2	3	4
$P(\eta)$	.0833	.2083	.3750	.2917	.0417

4. CONCLUSIONI

La ricerca presentata in questa nota evidenzia come attraverso l'acquisizione di informazioni quali la simmetria, la numerosita' dei legami e la distribuzione degli stessi per unita' territoriale, si possano considerare distribuzioni di probabilita' che "tendono" alla distribuzione $N_{BW}(L)$ del numero delle diversita' BW per lo specifico operatore di contiguita' L.

Infatti riferendosi all'esempio proposto per il quale:

funzione di autocorrelazione spazio-temporale per l'identificazione e la stima di modelli della classe STARMA" - Atti della XXXI Riunione Scientifica S.I.S., Torino.

Dacey, M.F., 1965- "A review on measures of contiguity for two and k-color maps", Technical Report N.2, Spatial Diffusion Study, Department of Geography Northwestern University, Evanston, Illinois.

Krishna Iyer, P.V.A., 1949- "The first and second moments of some probability distributions arising from points on a lattice, and their applications", Biometrika 36,135-141.

Moran P.A.P., 1948- "The interpretation of statistical maps" - J.R.S.S., Series B,10,243-251.

Rényi A.- "Calcul des probabilités" - Dunod, Paris, 1966.

Visini G.-Sclocco T.,1981-"Sull'indice di autocorrelazione spaziale C di Geary",Atti del V Convegno A.M.A.S.E.S., pp.277-298, Perugia 22-24 ottobre 1981.

Visini G.-Sclocco T.,1982-"Sui primi due momenti della statistica "join count" di Moran",Atti del VI Convegno A.M.A.S.E.S., pp.621-642, Marina di Campo (Isola d'Elba) 1-3 ottobre 1982.

Visini G.,1985-"Sull'analisi di carte tematiche bicromatiche attraverso l'autocorrelazione spazio-temporale",Atti Giornate di Lavoro A.I.R.O. 1985,pp.211-229, Venezia 30 settembre-2 ottobre 1985, Ed. Tecnoprint Bologna.

Visini G.,1989-"Considerazioni sul test z per gli indici di Moran",Pescara, accettata per la pubblicazione sugli Atti del I Convegno di Matematica Applicata all'Economia e alla Ingegneria,26-28 gennaio 1989,Pescara.

Visini G.,1990-"Considerazioni sulla statistica join count e l'informazione di Shannon",Convegno Giornate di Lavoro A.I.R.O. 1990, Sorrento 3-5 ottobre 1990, pp.853-865, Ed.Fast Print, Napoli.

Visini G.,1990-"Considerazioni sull'indice di correlazione spaziale",Giornate di Studio "Analisi dei dati spaziali e classificazione dei dati", Pescara 11-12 ottobre 1990, pp.477-488, Ed. Solfanelli M., Chieti.

I NUMERI DI FIBONACCI E LE EQUAZIONI DIFFERENZIALI

Bruno BARIGELLI, Luca OLIVIERI e Clara VIOLA
Ist. Mat. Stat. - Università di Ancona

Consideriamo i numeri individuati dalla relazione di ricorrenza

$$a_{n+3} = a_{n+2} + a_{n+1} + a_n \qquad (n\varepsilon N,\ n\geq 1)$$

con le condizioni iniziali $a_1=0$, $a_2=1$, $a_3=1$, che possiamo chiamare numeri di Fibonacci generalizzati o del terzo ordine, introduciamo i numeri di Fibonacci del terzo ordine di indice negativo e diamo una espressione che permette di calcolare a_n in funzione dell'indice n. Di seguito, otteniamo l'equazione differenziale lineare del primo ordine che ha per soluzione una funzione sviluppabile in serie di potenze con coefficienti proprio i numeri di Fibonacci del terzo ordine.

1- INTRODUZIONE

L'uso della matematica è ormai sempre più diffuso nei campi più diversi: dalla Biologia alla Statistica alla Ricerca Operativa, all'Economia.

In generale però molti problemi applicativi conducono a situazioni di tipo discreto, in particolare a relazioni di ricorrenza e quindi ad equazioni alle differenze finite.

E' noto che, per diversi motivi, in matematica sono

più agevoli da trattare problemi di tipo continuo e quindi è opportuno riuscire a trovare collegamenti stretti tra i due campi. Questo lavoro tratta appunto uno di questi temi.

Gli elementi della successione

$$a_n = a_{n-1} + a_{n-2} \qquad (n\geq 3) \qquad (1)$$

con le condizioni iniziali $a_1=1$, $a_2=1$, sono i numeri di Fibonacci individuati da una relazione di ricorrenza del secondo ordine ed è noto (vedi [1]) che la serie di potenze $\sum_0^\infty a_n x^n$ converge, per $|x|<(\sqrt{5}-1)/2$ alla funzione $x/(1-x-x^2)$.

In [4] viene illustrato un procedimento che permette di ottenere, a partire dalla relazione (1), l'equazione differenziale lineare del 1° ordine

$$-\frac{dy}{dx} + y\,\frac{1-x^2}{x(1-x-x^2)} = 0$$

che ha per soluzione una funzione che, se si impone la condizione $y(0)= 0$, è sviluppabile in serie di potenze in un opportuno intorno dell'origine con coefficienti proprio i numeri di Fibonacci (1).

In questo lavoro vengono presi in considerazione i numeri individuati dalla relazione

$$a_{n+3} = a_{n+2} + a_{n+1} + a_n \qquad (n\geq 1) \qquad (2)$$

con le condizioni iniziali

$$a_1=0,\ a_2=1,\ a_3=1 \qquad (3)$$

che, per analogia con i precedenti, verranno chiamati numeri di Fibonacci del terzo ordine. Dopo aver dimostrato alcune proprietà aritmetiche dei numeri (2) vengono introdotti i numeri di Fibonacci del terzo ordine con indice negativo dimostrando, anche per questi numeri, alcune proprietà aritmetiche; viene ripresa per a_n l'espressione in funzione di n: $a_n= c_1 t_1^{\,n} + \varepsilon_n$ ottenuta in [2] e viene fornito un programma per il calcolo di un valore approssimato di t_1 sia per eccesso che per difetto. Inoltre si dimostra che la serie di potenze

$$\sum_{1}^{\infty n} a_n x^n$$

i cui coefficienti a_n sono i numeri (2) converge in un intervallo con centro l'origine degli assi, alla funzione

$$S(x) = \frac{x^2}{1-x-x^2-x^3} \qquad (4)$$

Di seguito, utilizzando un procedimento indicato in [4], si ottiene l'equazione differenziale lineare del

1° ordine

$$-\frac{dy}{dx} + y\,\frac{2-x-x^3}{x-x^2-x^3-x^4} = 0$$

che ha per soluzione una funzione, che, se si impone la condizione $y(0) = 0$, è sviluppabile in serie di potenze in un opportuno intorno dell'origine con coefficienti proprio i numeri che verificano la (2).

2- NUMERI DI FIBONACCI DEL TERZO ORDINE.

2.1- PROPRIETA' ARITMETICHE.

2.1.1 Somma dei primi n numeri: Se S_n indica la somma dei primi n numeri allora

$$S_n = \tfrac{1}{2}(a_{n+3} - a_{n+1} - 1). \qquad (5)$$

Dim: Infatti, scritta la (2) con i valori dell'indice 1, 2,...,n, e sommando membro a membro, si ottiene

$$2S_n = a_{n+3} - a_{n+1} - a_3 + a_1$$

da cui, tenendo presenti le condizioni iniziali (3), si ottiene la (5).

NOTA BENE: è possibile esprimere S_n in funzione di a_n e precedenti, anzichè dei numeri seguenti a_n, e si ottiene $S_n = \tfrac{1}{2}(3a_n + 2a_{n-1} + a_{n-2} - 1)$.

2.1.2 Somma dei primi n numeri di indice dispari: Se S_d^n indica la somma dei primi n numeri di indice di-

spari, allora

$$S_d^n = \tfrac{1}{2}(a_{2n+2} - a_{2n+1} - 1). \qquad (6)$$

Dim: Infatti, scritta la (2) nella forma

$$a_{n+2} + a_n = a_{n+3} - a_{n+1},$$

ponendo successivamente l'indice uguale ad $1,3,\ldots,2n-1$ e sommamdo membro a membro, si ha $2S_d^n = a_{2n+2} - a_{2n+1} - 1$ da cui, ricordando le (3), segue la (6).

Corollario 1: due elementi consecutivi della (2), il primo dei quali di indice dispari, hanno parità diversa.

Corollario 2: due elementi consecutivi della (2), il primo dei quali di indice pari, hanno la stessa parità.

Corollario 3: la successione dei numeri di Fibonacci (2) alterna, a coppie, numeri pari e dispari e quindi numeri che si trovano a distanza di due posizioni sono di parità diversa.

NOTA BENE: è possibile esprimere S_d^n in funzione di a_{2n-1} e precedenti, invece dei termini seguenti a_{2n-1} e si ottiene $S_d^n = a_{2n-1} + \tfrac{1}{2}(a_{2n-2} + a_{2n-3} - 1)$.

2.1.3 Somma dei primi n numeri di indice pari: Se S_p^n indica la somma dei primi n numeri di indice pari, allora

$$S_p^n = \tfrac{1}{2}(a_{2n+3} - a_{2n+2}) \qquad (7)$$

Dim: Infatti, sottraendo dalla (5) scritta per i primi 2n numeri la (6), si ottiene la (7).

NOTA BENE: è possibile esprimere S_p^n in funzione di a_{2n} e precedenti anzichè dei termini seguenti a_{2n} e si ottiene $S_p^n = a_{2n} + \frac{1}{2}(a_{2n-1} + a_{2n-2})$.

2.2- NUMERI DI FIBONACCI CON INDICE NEGATIVO

Ponendo nella (2) -n al posto di n, si ottiene

$$a_{-n} = a_{-n+3} - a_{-n+2} - a_{-n+1} \qquad (8)$$

che è la relazione che caratterizza i numeri di Fibonacci del terzo ordine di indice negativo.

Ponendo nella (8) successivamente n = 0,1,2,..., si ha

$a_0 = a_3 - a_2 - a_1 = 0$

$a_{-1} = a_2 - a_1 - a_0 = 1$

$a_{-2} = a_1 - a_0 - a_{-1} = -1$

.....

Considerando come primo elemento della successione a_{-1}, i primi venti numeri sono

1, -1, 0, 2, -3, 1, 4, -8, 5, 7, -20, 18, 9, -47, 56, 0, -103, 159, -56, -206.

Anche per i numeri di questa successione si possono stabilire alcune proprietà aritmetiche.

2.2.1 Somma dei primi n numeri: Se S_{-n} indica la somma dei primi n numeri, allora

$$S_{-n} = \frac{1}{2}(a_{-n} - a_{-n+2} + 1). \qquad (9)$$

Dim:Infatti, sostituendo successivamente ad n nella (8) i numeri 1,2,...n, e sommando membro a membro, si ottiene

$$2S_{-n} = a_2 - a_0 - a_{-n+2} + a_{-n}$$

da cui, tenendo presenti le condizioni iniziali (3), la (9).

2.2.2 Somma dei primi n numeri di indice dispari: Se $S'_d{}^n$ indica la somma dei primi n numeri di indice dispari, allora

$$S'_d{}^n = \tfrac{1}{2}(1 - a_{-2n} - a_{-2n-1}) \qquad (10)$$

Dim: Infatti, scritta la (8) nella forma

$$a_{-n} + a_{-n+2} = a_{-n+3} - a_{-n+1}$$

sostituendo successivamente ad n i numeri 1, 3,...,2n++1, sommando membro a membro, si ha

$$2S'_d{}^n = a_2 - a_{-2n} - a_1 - a_{-2n-1}$$

e tenendo presenti le condizioni iniziali (3) si ottiene la (10).

Corollario: due numeri consecutivi, di cui quello con indice maggiore pari, sono di diversa parità.

Corollario: due numeri consecutivi, di cui quello con indice maggiore dispari, hanno la stessa parità.

Corollario: la successione (8) alterna, a coppie, numeri pari e numeri dispari e quindi numeri che si trovano a distanza di due posizioni sono di diversa parità.

NOTA BENE: è possibile esprimere $S'_d{}^n$ in funzione di a_{-2n+1} e precedenti anzichè dei numeri seguenti a_{-2n+1} e si ottiene

$$S'_d{}^n = \tfrac{1}{2}(1 + a_{-2n+1} - a_{-2n+2}). \qquad (10')$$

2.2.3 Somma dei primi n numeri di indice pari: Se $S'_p{}^n$ indica la somma dei primi n numeri di indice pari, allora

$$S'_p{}^n = \tfrac{1}{2}(a_{-2n} - a_{-2n+1}) \qquad (11)$$

Dim: Infatti, sottraendo dalla (9) scritta per i primi 2n+1 numeri la (10'), si ottiene la (11).

3- CALCOLO DI a_n IN FUNZIONE DELL'INDICE

Per ottenere a_n in funzione di n dalla (2), (vedi [2]), come è noto dalla teoria delle equazioni alle differenze finite (vedi ad esempio [5]) occorre risolvere l'equazione caratteristica ad essa associata

$$t^3 - t^2 - t - 1 = 0$$

Dette t_1, t_2, t_3 le sue soluzioni, si avrà

$$a_n = c_1 t_1{}^n + c_2 t_2{}^n + c_3 t_3{}^n$$

con c_1, c_2, c_3 costanti da determinare in base alle condizioni iniziali (3). Per la soluzione reale t_1 si ha: $1<t_1<2$ e per ottenere un valore approssimato abbiamo utilizzato due metodi: il metodo delle corde ed

il metodo delle tangenti.

Abbiamo elaborato un programma in linguaggio BASIC per ognuno dei due metodi, utilizzando variabili e costanti in doppia precisione. Con il metodo delle corde abbiamo ottenuto il valore maggiormente affidabile t_1 = 1,839286 alla settima iterazione ed il valore migliore, ma che non riteniamo molto affidabile, t_1 = 1,83928671 all'ottava iterazione. Con il metodo delle tangenti, alla terza iterazione è stato ottenuto il valore migliore di t_1 = 1,83928690 ed anche il valore più affidabile t_1 = 1,839286. Gli errori di calcolo dovuti all'approssimazione del computer non ci hanno consentito di ottenere valori maggiormente affidabili.
Le altre due radici dell'equazione caratteristica sono complesse e coniugate; posto $t_2=\beta+i\Gamma$, $t_3=\beta-i\Gamma$, si ottiene, con l'approssimazione di 10^{-5},
$-0,41965 < \beta < -0,41964$; $0,60628 < \Gamma < 0,60629$.
Posto $c_2 = h+ik$, $c_3 = h-ik$, imponendo le condizioni iniziali (3), si ottiene $0,18280 < c_1 < 0,18280$,
$-0,09140 < h < -0,09140$; $0,20646 < k < 0,20647$.

Se si ricorre alla forma trigonometrica dei numeri complessi ed alla formula di De Moivre, la (13) si può

scrivere in forma diversa: posto $\sigma=\sqrt{\beta^2+\Gamma^2}$ e $\Theta=\text{arctg}\Gamma/\beta$, si ha

$$a_n =$$

$$= c_1 t_1^n+(h+ik)\sigma^n(\cos n\Theta+i\,\text{sen}n\Theta)+(h-ik)\sigma^n(\cos n\Theta-i\,\text{sen}n\Theta)$$

con $|\varepsilon_n| = \sigma^n|h\cos n\Theta-k\,\text{sen}n\Theta|<\frac{1}{2}$, cioè

$$a_n = c_1 t_1^n + \varepsilon_n$$

4- STUDIO DELLA SERIE $\sum_1^\infty{}_n\, a_n x^n$.

Consideriamo ora la serie di potenze

$$\sum_1^\infty{}_n\, a_n x^n \qquad (12)$$

i cui coefficienti sono i numeri individuati dalla (2) con a_1, a_2, a_3 numeri naturali prefissati.Tale serie converge per x=0 e converge a zero. Per il calcolo del raggio di convergenza, si ha

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{c_1 t_1^{n+1} + \varepsilon_{n+1}}{c_1 t_1^n + \varepsilon_n} = t_1$$

ed allora la serie (12) converge per $|x|<1/t_1 = x_1$ e non converge per $|x|=x_1$, in quanto per $x=x_1$ ed $x=-x_1$ si ottengono le due serie

$$\sum_1^\infty{}_n\, c_1 \quad \text{e} \quad \sum_1^\infty{}_n\, (-1)^n c_1$$

non convergenti. Allora per $|x|<x_1$ si può porre

$$a_1x + a_2x^2 +\ldots+ a_nx^n +\ldots = S(x). \qquad (13)$$

Per ottenere esplicitamente S(x) moltiplichiamo la (13) successivamente per x, x^2, x^3. Si ha

$$a_1x^2 + a_2x^3 +\ldots+ a_nx^{n+1} +\ldots= xS(x) \qquad (14)$$

$$a_1x^3 + a_2x^4 +\ldots+ a_nx^{n+2} +\ldots = x^2S(x) \qquad (15)$$

$$a_1x^4 + a_2x^5 +\ldots+ a_nx^{n+3} +\ldots = x^3S(x) \qquad (16)$$

Ora, sottraendo le (14), (15), (16) dalla (13), ricordando la (2), si ha, sempre per $|x|<x_1$

$$S(x) = \frac{a_1x + (a_2-a_1)x^2 + (a_3-a_2-a_1)x^3}{1-x-x^2-x^3}.$$

Imponendo le condizioni (3) si ottiene

$$S(x) = \frac{x^2}{1-x-x^2-x^3} \qquad (17)$$

da cui, in particolare,

$$S(\tfrac{1}{2}) = 2 = \sum_1^{\infty n} (\tfrac{1}{2})^n.a_n.$$

Imponendo come condizioni

$$a_1=0,\ a_2 = a_3= a \quad (a\varepsilon N) \qquad (18)$$

anzichè le (3), si ottiene per S(x) l'espressione

$$S(x) = \frac{ax^2}{1-x-x^2-x^3} \qquad (19)$$

La funzione (19) risulta somma anche di altre serie di funzioni su tutto l'asse reale privato del punto $x=x_1$. Infatti, poichè

$$\frac{x^2}{1-x-x^2-x^3} = \frac{x^2}{x_1(x^2+kx+t_1)} \cdot \frac{1}{1-x/x_1}$$

con $k \approx 1{,}54369$, allora per $|x|<x_1$, risulta

$$S(x) = \frac{ax^2}{x^2+kx+t_1} \cdot \sum_{0n}^{\infty} t_1^{n+1} \cdot x^n .$$

Ancora, poichè

$$S(x) = \frac{ax}{x^2+kx+t_1} \cdot \frac{1}{x_1/x-1},$$

allora, per $|x|>x_1$, si ha

$$S(x) = \frac{-ax}{x^2+kx+t_1} \cdot \sum_{0n}^{\infty} (x_1/x)^n .$$

5- LA TRASFORMATA DI MC LAURIN DI UNA SUCCESSIONE.

Siano f(n) e g(n) due successioni reali definite per $n \varepsilon N$, e nello stesso intervallo ICR risulti

$$F(x) = \sum_{0n}^{\infty} f(n)x^n \quad , \quad G(x) = \sum_{0n}^{\infty} g(n)x^n .$$

Allora la serie prodotto secondo Cauchy delle due serie converge, nello stesso intervallo, alla funzione F(x). G(x) ed il coefficiente di x^n nella serie prodotto è

dato da

$$f(n)*g(n) = \sum_{0}^{n}{}_{r} f(r)g(n-r).$$

Le funzioni F(x) e G(x) vengono chiamate anche le <u>trasformate di Mc Laurin</u> delle successioni f(n) e g(n) e vengono indicate, rispettivamente con M{f(n)} e M{g(n)}.

Se si indica con f(n) = a(n) l'elemento generico della (2), con $n^{(r)}$ il prodotto n(n-1)...(n-r+1), si dimostra che, cfr [4],

$$M\{(n+r)^{(r)}\} = \frac{d^r F}{dx^r}. \qquad (20)$$

Per altre proprietà delle trasformate di Mc Laurin cfr.[4] e la bibliografia ivi contenuta.

Stabiliamo ora il seguente

<u>Teorema</u>: Se F(x) = M{f(n)} e G(x) = M{g(n)}, si ha, per s=0,1,2,3

$$M\{f(n)*g(n+s)\} = \begin{cases} F(x)G(x) & s=0 \\ F(x)\dfrac{G(x) - \sum_{0}^{s-1}{}_{i} g(i)x^i}{x^s} & s=1,2,3 \end{cases}$$

Dim: per s = 0,1,2 cfr.[4]. Per s=3 basta tener presente che

$$\lim_{n\to\infty} \sum_{0}^{n} {}_r\, g(r)x^r = G(x),$$

che

$$\lim_{n\to\infty} \sum_{0}^{n} {}_r\, g(r+3)x^r = \frac{G(x)-g(2)x^2-g(1)x-g(0)}{x^3}$$

e ricordare il prodotto secondo Cauchy di due serie.

6- L'EQUAZIONE DIFFERENZIALE COLLEGATA ALLA RELAZIONE DI RICORRENZA.

Determiniamo ora, partendo dalla relazione di ricorrenza (2) scritta come a(n+3) - a(n+2) - a(n+1) -a(n) = = 0, l'equazione differenziale che con apposite condizioni relative al punto x_0=0, ammette una ed una sola soluzione soddisfacente le condizioni iniziali imposte. Tale soluzione dovrà essere definita in un opportuno intorno U di x_0=0 e risultare ivi sviluppabile in una serie convergente di potenze intere di x. Come è noto, cfr [4], Teorema I, l'equazione differenziale cercata dovrà risultare lineare, cioè del tipo

$$c_n(x)\frac{d^n y}{dx^n} + \ldots + c_0(x)y = b(x),$$

inoltre dovrà ammettere come soluzione, per x=0, la funzione (17) con le condizioni (3).

Secondo quanto esposto in [4], basta introdurre quattro serie di potenze convergenti in U i cui coefficienti non tutti nulli A_{rs} (s=0,1,2,3) soddisfino, per ogni n e per ogni r, le relazioni di ricorrenza

$$\sum_{0}^{3}{}_{s} \sum_{0}^{n}{}_{r} A_{sr} \frac{(s+n-r)!}{(n-r)!} a(s+n-r) = \sum_{0}^{3}{}_{s} c_s(n)a(n+s). \qquad (21)$$

Da notare che la (21) non determina univocamente i coefficienti in questione a partire dalla conoscenza dei $c_k(n)$ ed a(n).

Ponendo quindi

$$A_{01} = A(i);\ A_{11} = B(i);\ A_{21} = D(i);\ A_{31} = E(i);$$

si ottiene

$$\sum_{1}^{n}{}_{r} \left[A(r)a(n-r)+B(r)\frac{(n-r+1)!}{(n-r)!}a(n-r+1)+D(r)\frac{(n-r+2)!}{(n-r)!}\cdot \right.$$

$$\left. .a(n-r+2)+E(r)\frac{(n-r+3)!}{(n-r)!}a(n-r+3) \right] = \sum_{0}^{3}{}_{s} c_s(n)a(n+s),$$

cioè

$$\sum_{1}^{n}{}_{r} [A(r)a(n-r) + B(r)(n-r+1)a(n-r+1) + D(r)(n-r+2).$$
$$.(n-r+1)a(n-r+2) + E(r)(n-r+3)(n-r+2)(n-r+1).$$
$$.a(n-r+3)] = 0. \qquad (22)$$

Applicando ora la trasformazione di Mc Laurin al primo membro della (22), indicando con $\alpha(x)$, $\beta(x)$, $\mu(x)$, $\sigma(x)$ le trasformate di $A(r)$, $B(r)$, $D(r)$, $E(r)$, con y la trasformata di Mc Laurin della successione (2) con le condizioni (3), tenendo presente il teorema precedente e la (22), si ottiene l'equazione differenziale

$$\alpha(x)y + \beta(x)\frac{dy}{dx} + \mu(x)\frac{d^2y}{dx^2} + \sigma(x)\frac{d^3y}{dx^3} = 0. \qquad (23)$$

Dal momento che sono possibili diverse scelte di α,β,μ,σ, possiamo porre $\beta(x) = -1$, $\mu = \sigma = 0$, ottenendo per $\alpha(x)$ l'espressione

$$\alpha(x) = \frac{2-x+x^3}{x-x^2-x^3-x^4}.$$

Allora la (23) diventa

$$-\frac{dy}{dx} + \frac{2-x+x^3}{x-x^2-x^3-x^4}\,y = 0. \qquad (24)$$

Da notare che alla (24) medesima si sarebbe arrivati anche se invece delle condizioni (3) avessimo posto le condizioni (18).

Si può verificare facilmente che, se $y(x)$ è soluzione della (30) con le condizioni

$$y(0)=0, \quad y(x)= \sum_{1}^{\infty}{}_{n}\ a_n x^n$$

in un opportuno intorno dell'origine, allora dalla relazione

$$-\sum_{2}^{\infty}{}_{n}\ n a_n x^{n-1}(x-x^2-x^3-x^4) + \sum_{1}^{\infty}{}_{n}\ a_n x^n(2-x+x^3) = 0$$

si ottiene, annullando il coefficiente di x^n, che i numeri a_n verificano la relazione (2) e le condizioni (3).

Questo lavoro è il risultato della collaborazione dei tre Autori. Va precisato, se si intendono individuare i singoli contributi, che la stesura dei paragrafi 2 e 4 è dovuta essenzialmente a C. Viola; quella del paragrafo 3 a L. Olivieri; quella dei paragrafi 5 e 6 a B. Barigelli.

B I B L I O G R A F I A

[1] B.BARIGELLI, E.VICHI e C.VIOLA, "Sulle serie di funzioni che hanno per coefficienti i numeri di Fibonacci", **Quad. Ist. Matematica e Statistica, Fac. Ec. e Commercio.- Un. Ancona**, n.20 (1988).

[2] B.BARIGELLI, E.VICHI e C.VIOLA, "Numeri di Fibonacci di ordine superiore", **Quad. Ist. Matematica e Statistica, Fac. Ec. e Commercio.- Un. Ancona**, n.33 (1990).

[3] B.BARIGELLI e C.VIOLA, "I numeri di Fibonacci generalizzati e le equazioni differenziali", **Quad. Ist. Matematica e Statistica, Fac. Ec. e Commercio.- Un. Ancona**, n.35 (1990).

[4] E.PESSA e B.RIZZI, "Relazioni di ricorrenza ed equazioni differenziali", **Periodico di Matematiche**, (1987), 3-22.

[5] C.VIOLA, "Equazioni alle differenze finite", **CLUA Ed., Ancona**, (1981).

[6] N.VOROBIEV, "Caracteres de divisibilitè. Suite de Fibonacci", **Ed. MIR, Mosca**", (1973).

Bruno BARIGELLI Luca OLIVIERI Clara VIOLA

Ist. Mat. e Stat.
Facoltà Economia e Comm.
-Università Ancona-
Via Pizzecolli, 37 - 60100 - ANCONA -

Barigelli Bruno: Associato di Matematica Generale nell'Università di Ancona, si occupa di problemi connessi con le relazioni di ricorrenza, in particolare i numeri di Fibonaci e di problemi relativi alla Programmazione matematica non lineare. E' autore di numerose pubblicazioni a carattere scientifico ed a carattere didattico. E' socio della AMASES, dell'AIRO, dell'UMI.

Olivieri Luca: borsista presso l'Ist. di Matematica e Statistica dell'Univ. di Ancona, si interessa di problemi di Calcolo numerico e di Ricerca Operativa. Autore di una pubblicazione a carattere scientifico e di una a carattere didattico. Ha ottenuto il premio istituito dalla SIP per la Tesi di Laurea nel 1989 sul "Controllo di gestione".

Viola Clara: Stabilizzata di Matematica nell'Università di Ancona, si occupa di problemi connessi con le relazioni di ricorrenza, in particolare i numeri di Fibonacci e di problemi di Ricerca Operativa. E' autore di numerose pubblicazioni a carattere scientifico ed a carattere didattico. E' socio della AMASES.

UNA FORMULAZIONE VARIAZIONALE COMPLETAMENTARE DEL PROBLEMA DI VERIFICA DI RETI DI DISTRIBUZIONE DI FLUIDO INCOMPRESSIBILE

Francesco Russo Spena* Andrea Vacca**

Sommario

Nel presente lavoro si analizza il classico problema di verifica di reti di distribuzione di fluido newtoniano incompressibile in regime di moto stazionario. Viene presentato un principio variazionale complementare che consente di acquisire la dimostrazione delle proprietà di esistenza ed unicità della soluzione del problema di verifica.

Summary

The paper deals with the classical problem arising in the analysis of fluid distribution systems in steady state of flow. A variational complementary principle is stated allowing the proof about the existence and uniqueness properties of solution.

1. Introduzione

È ben noto [1] che il problema di verifica di reti di distribuzione di fluidi newtoniani incompressibili in regime di moto stazionario è strettamente analogo a quello di verifica di telai elastici soggetti all'azione di sole coppie nodali.

Dal punto di vista analitico, infatti, entrambi i problemi sono caratterizzati da due insiemi di variabili duali o coniugate mutuamente collegati da trasformazioni lineari: uno dei due insiemi di variabili soddisfa la condizione di presentare somma algebrica nulla nei nodi della rete; l'altro gode della proprietà di avere somma algebrica nulla lungo ciascun percorso chiuso o maglia della rete.

Basandosi su questa proprietà di aggiunzione algebrica delle variabili duali, la formulazione generale del problema di verifica può essere conseguita con approccio variazionale, mediante

*Professore Associato, Facoltà di Ingegneria, Università di Napoli "Federico II".

**Allievo "Dottorato di Ricerca in Ingegneria Idraulica", Università di Napoli "Federico II".

il quale la soluzione viene ottenuta in base alle condizioni di stazionarietà di una coppia di funzionali reciproci nelle variabili duali che, dal punto di vista fisico, sono connessi alla dissipazione di potenza per attrito espressa in termini di portate volumetriche o di perdite di carico.

Seguendo questa impostazione e, grazie alla finita numerabilità degli spazi vettoriali implicati, si è ricondotti a due indirizzi computazionali alternativi: il primo utilizza metodi della Programmazione Matematica [4, 5], specializzati per l'analisi di problemi di minimizzazione condizionata; il secondo si basa sulla soluzione diretta dei tre sistemi di equazioni algebriche che presiedono al problema e che esprimono rispettivamente:

i) le condizioni di continuità delle portate per ciascun nodo della rete;

ii) le equazioni tra le perdite di carico agli estremi di ciascun ramo della rete ed i carichi idraulici nei nodi collegati dal medesimo ramo;

iii) le equazioni del moto che esprimono il legame tra perdita di carico lungo il generico ramo della rete e portata volumetrica defluente lungo il medesimo.

La formulazione del problema può essere ulteriormente generalizzata allo scopo di schematizzare reti di distribuzione comprendenti serbatoi di alimentazione o di compenso, di dispositivi di regolazione, di intercettazione, di regolazione inseriti nella rete in modo affatto generico [6].

In assenza di dispositivi idraulici e con riferimento ad espressioni monomie per i legami tra portata e perdita di carico nelle condotte, le proprietà di esistenza e di unicità della soluzione del problema di verifica delle reti idrauliche sono state dimostrate da Andrea Russo Spena [2] in un contesto variazionale.

Per quanto risulta agli scriventi non sembra che le medesime proprietà siano state dimostrate con riferimento a reti provviste di dispositivi del tipo sopra richiamato e con riferimento ad espressioni non monomie per il legame portata–perdita di carico.

Nel presente lavoro viene presentata la formulazione del problema di verifica ambientandola in spazi vettoriali astratti normati secondo il prodotto scalare e vengono applicati noti risultati su funzionali convessi definiti in insieme convesso, e sui cosiddetti super–potenziali di dissipazione [11].

Nel paragrafo 2 il problema viene presentato nella sua classica forma algebrica.

Nel paragrafo 3 viene acquisita la formulazione variazionale astratta e, dopo averne dimostrato l'equivalenza con il problema algebrico si fornisce la prova di esistenza ed unicità della soluzione.

2. Formulazione del problema

Ai fini di una schematizzazione di validità generale, una rete di distribuzione di fluido newtoniano incompressibile è da considerarsi costituita di $\tilde{\ell}$ rami (tratti di condotta di diametro e scabrezza costanti); $\tilde{n}$ punti nodali (punti nei quali convergono due o più condotte ovvero punti di una medesima condotta in cui si ha un brusco cambiamento della scabrezza

o della portata); t dispositivi di intercettazione, di strozzamento parziale o totale del flusso, di apparecchi di misura o di erogazione.

Il modello matematico corrispondente dovrà pertanto far riferimento alle seguenti equazioni:

i) equazioni che esprimono la continuità tra portate volumetriche $\mathbf{q}$ defluenti nei rami della rete (condotte o dispositivi) e portate $\mathbf{p}$ esterne erogate o immesse nei nodi medesimi[1]

$$\mathbf{Tq} + \mathbf{p} = 0 \tag{2.1}$$

in cui $\mathbf{T}$ denota la matrice topologica $(\tilde{n}, \tilde{\ell})$ di rango $\tilde{n} - 1$ [7], il cui elemento generico t_{ij} è rispettivamente: uguale $a + 1$ se il j–mo ramo ha inizio nel nodo i–mo; uguale a -1 se il j–mo ramo ha termine nel nodo i–mo; uguale a 0 se il lato j–mo non afferisce al nodo i–mo.

ii) equazioni che esprimono la continuità geometrica tra perdite di carico $\mathbf{x}$ nei rami della rete e quote piezometriche nodali $\mathbf{y}$

$$\mathbf{x} = \mathbf{T}^t\mathbf{y} \tag{2.2}$$

iii_a) equazioni che esprimono il legame tra portate $\mathbf{q}$ defluenti lungo i rami della rete e perdite di carico associate $\mathbf{x}$ rappresentate nella seguente forma generale

$$q_j = \psi_j(x_j) \qquad \forall j \in \{1, \ldots, \tilde{\ell} + t\} \tag{2.3}$$

Alla relazione funzionale genericamente espressa dalla (2.3) occorrerà assegnare una specifica espressione secondo che essa si riferisca ad un ramo di condotta o ad un dispositivo. Per quanto attiene alla caratterizzazione dei dispositivi si farà riferimento alla seguente rappresentazione analitica

$$q_j = r_j + s_j x_j |x_j|^{\tau_j} \qquad \forall j \in \{1, \ldots, t\} \tag{2.4}$$

I parametri r_j, s_j e l'esponente τ_j sono ottenuti con algoritmi di interpolazione applicati al campo di prestazioni sperimentali esibite dal dispositivo stesso.
Con riferimento invece ai tratti di condotta cilindrica e relativamente a fluido newtoniano incompressibile in regime di moto stazionario, alla relazione (2.3) si attribuisce la forma analitica espressa dalla equazione di Darcy–Weisbach generalizzata

$$x = \lambda \frac{8q|q|}{g\pi^2 d^5} L \tag{2.5}$$

in cui, secondo consuetudine, si sono indicati con g l'accelerazione di gravità e con d, L e λ rispettivamente il diametro, la lunghezza e l'indice di resistenza della tubazione.

[1] Qui e nel seguito indicheremo con carattere maiuscolo grassetto matrici bidimensionali; con carattere minuscolo grassetto matrici–colonna o vettori numerici.

Con riferimento a condizioni di deflusso in tubi commerciali scabri, all'indice di resistenza λ si attribuisce l'espressione di Colebrook & White

$$\frac{1}{\sqrt{\lambda}} = -2\log_{10}\left(\frac{2.51}{Re\sqrt{\lambda}} + \frac{\epsilon}{3.71d}\right) \tag{2.6}$$

essendo ϵ/d la scabrezza relativa equivalente in sabbia e Re il numero di Reynolds. Per gli scopi applicativi la relazione implicita trascendente (2.6) può essere convenientemente approssimata dalla seguente forma esplicita proposta da Citrini [8]

$$\lambda = \lambda_\infty\left(1 + \frac{8}{Re\epsilon/d}\right) \tag{2.7}$$

in cui λ_∞ indica il valore di λ relativo a condizioni di turbolenza pienamente sviluppata ($Re \to +\infty$) dato da

$$\lambda_\infty = \frac{1}{4}\left(\log_{10} 3.71\,\frac{d}{\epsilon}\right)^{-2} \tag{2.8}$$

Tenuto conto della (2.7), la (2.5) può essere scritta

$$x = aq|q| + bq \tag{2.9}$$

in cui si è posto

$$a = \lambda_\infty\,\frac{8L}{g\pi^2 d^5} \qquad b = 16\lambda_\infty\,\frac{L\nu}{g\pi\epsilon d^3}$$

essendo ν la viscosità cinematica.
Pertanto l'equazione costitutiva di forma generale (2.3) può essere espressa nella seguente forma

$$q_j = \mathrm{sgn}(x_j)\left[(b_j^2 + 4a_j|x_j|)^{1/2} - b_j\right]\frac{1}{2a_j} \qquad \forall j \in \{1,\ldots,\tilde{\ell}\} \tag{2.10}$$

in cui $\mathrm{sgn}(\cdot)$ denota la funzione signum.
In situazioni di moto puramente turbolento ($Re \to +\infty$), l'equazione (2.10) degenera nell' espressione monomia

$$q_j = \mathrm{sgn}(x_j)c_j|x_j|^{1/2} \qquad \forall j \in \{1,\ldots,\tilde{\ell}\}$$

essendo $c_j = a_j^{-1/2}$ l'ammettenza idraulica dipendente dalla lunghezza, dal diametro e dalla scabrezza relativa della tubazione.
In regime di transizione possono altresì essere utilmente impiegate antiche formule empiriche [9] riferite a tubazioni commerciali prodotte e rivestite con varia tecnologia. Tutte sono riconducibili ad espressioni monomie del tipo

$$x_j = \mathrm{sgn}(q_j)w_j L_j d_j^{-\gamma_j}|q_j|^{\alpha_j} \qquad \forall j \in \{1,\ldots,\tilde{\ell}\} \tag{2.11}$$

ove $\gamma_j \in [4.71, 4.98]$, $\alpha_j \in [1.79, 1.85]$ e w_j rappresenta un opportuno coefficiente numerico positivo.

Quando si tenga conto della (2.11), alla (2.3) può essere data la seguente altra espressione

$$q_j = c_j |x_j|^{\beta_j} x_j \quad \forall j \in \{1, \dots, \tilde{\ell}\} \tag{2.12}$$

essendo

$$\beta_j = \frac{1}{\alpha_j} - 1$$

iii$_b$) equazioni che traducono la presenza di dispositivi di regolazione (rubinetti, idranti, gruppi di erogazione) presenti nei nodi di erogazione della rete mediante un legame tra portate nodali spillate $\mathbf{p}$ e quote piezometriche nodali $\mathbf{y}$

$$p_i = q_i^* + k_i (y_i - z_i)^{\gamma_i} \qquad \forall i \in \{1, \dots, \tilde{n}\} \tag{2.13}$$

in cui z_i sta ad indicare la quota piezometrica nota dell'ambiente nel quale la portata p_i affluisce. In particolare se la portata p_i sbocca direttamente nell'atmosfera, z_i denota la quota geometrica del nodo i. La espressione esplicita di k_i e dell'esponente γ_i che figurano nella (2.13), dipendono ovviamente dalla geometria del dispositivo di erogazione, dal grado di apertura del dispositivo medesimo e dal numero di Reynolds. Il termine q_i^* denota una portata di valore assegnato e costante, erogata dal nodo i indipendentemente dal dislivello $y_i - z_i$.
Le equazioni (2.3) possono essere simultaneamente rappresentate nella seguente notazione vettoriale

$$\mathbf{q} = \Psi(\mathbf{x}) \tag{2.14}$$

essendo $\Psi(x) = \{\psi_j(x_j)\}, \quad \forall j \in \{1, \dots, \tilde{\ell}\}$.
Le equazioni (2.13) possono a loro volta essere scritte nella notazione compatta

$$\mathbf{p} = \mathbf{q}^* + \Phi(\mathbf{y}) \tag{2.15}$$

dove

$$\Phi(\mathbf{y}) = \{k_i (y_i - z_i)^{\gamma_i}\} \qquad \forall i \in \{1, \dots, \tilde{n}\}$$

Alla stregua della notazione introdotta e sostituendo la (2.2) nella (2.14), otteniamo

$$\mathbf{q} = \Psi(\mathbf{T}^t \mathbf{y}) \tag{2.16}$$

Inoltre dalle (2.1) e (2.15) siamo condotti alla seguente equazione

$$\Psi(\mathbf{y}) + \mathbf{T}\Psi(\mathbf{T}^t \mathbf{y}) = -\mathbf{q}^* \tag{2.17}$$

Infine al sistema di equazioni (2.17) può essere attribuita l'espressione

$$\Xi(\mathbf{y}) = \Gamma(\mathbf{y}) + \mathbf{q}^* = 0 \tag{2.18}$$

dove si è posto

$$\Gamma(\mathbf{y}) = \Phi(\mathbf{y}) + \mathbf{T}\Psi(\mathbf{T}^t \mathbf{y}) \tag{2.19}$$

Per completare il modello matematico della rete occorre ancora tener conto della presenza dei dispositivi il cui comportamento idraulico è definito dal sistema di equazioni non lineari (2.4).
Di queste equazioni che si aggiungono a quelle non lineari espresse dalle (2.18), si può tener conto mediante due distinte strategie.
Nel primo approccio le relazioni (2.4) sono considerate alla stregua di vincoli non lineari tra prefissate coppie di incognite nodali (gradi di libertà) in conformità con una strategia computazionale largamente impiegata nell'analisi strutturale [10].
Nel secondo approccio si tiene conto della presenza dei dispositivi considerando le relazioni (2.4) come equazioni costitutive di un ulteriore insieme di rami fittizi nella rete da assemblare con lo stesso procedimento adoperato nella deduzione dell'equazione vettoriale (2.18)
La scelta tra questi due approcci è essenzialmente influenzata da considerazioni basate sull'efficienza computazionale.
D'altra parte poiché ciascuna delle strategie descritte poco sopra presenta specifici pregi e difetti, entrambe possono considerarsi equivalenti dal punto di vista computazionale.
In quel che segue faremo riferimento al secondo approccio. L'ordine dell'operatore di continuità $\mathbf{T}$ dovrà conformemente essere variato per rappresentare la rete topologica fittizia per la quale indicheremo con n ed ℓ rispettivamente il numero di nodi e di rami.

3. Formulazione variazionale astratta

Come già detto nell'Introduzione per dimostrare l'esistenza e l'unicità della soluzione del sistema di equazioni (2.1)–(2.3) si darà al problema una formulazione variazionale equivalente.

Questo approccio si mostra doppiamente utile: da un lato consente la dimostrazione dell'esistenza e dell'unicità della soluzione del problema posto, dall'altro può fornire indicazioni circa la costruzione di algoritmi di risoluzione dotati di proprietà di convergenza e stabilità.

Il ragionamento che sarà sviluppato è essenzialmente basato su recenti acquisizioni nel campo della meccanica computazionale non lineare ed in particolare sulla teoria dei cosiddetti superpotenziali dissipativi [11]. In effetti nelle questioni di meccanica dei solidi si pone il problema inverso del calcolo delle variazioni laddove, ad esempio, si analizzi, con formulazione differenziale, il contatto unilaterale con attrito alla Coulomb tra due solidi deformabili.

Tale problema consiste essenzialmente nell'associare alle equazioni differenziali che lo definiscono un funzionale le cui condizioni di stazionarietà esprimono le relazioni originali come equazioni di Euler–Lagrange.

Nel caso in esame la proprietà di aggiunzione algebrica dell'operatore $\mathbf{T}$, che interviene nelle equazioni (2.1), (2.2), assicura il verificarsi della condizione necessaria per l'esistenza di soluzione del problema inverso.

È, dunque, necessario individuare un potenziale normale di dissipazione il cui gradiente esprima l'equazione costitutiva non lineare (2.3).

Per concretezza ci riferiremo nel seguito alla seguente espressione generale

$$a_j x_j |x_j|^{\beta_j} + b_j x_j + c_j \tag{3.1}$$

con a_j e b_j entrambi non negativi e $\beta_j > -1$.

Il potenziale normale di dissipazione per ciascun ramo della rete può essere facilmente dedotto mediante integrazione e, a meno di una costante additiva inessenziale, assume la forma

$$f_j(x_j) = \frac{1}{\beta_j + 2} a_j |x_j|^{\beta_j + 2} + \frac{1}{2} b_j x_j^2 + c_j x_j \qquad \forall j \in \{1, \dots, \ell\} \tag{3.2}$$

In base all'ipotesi fatta sul coefficiente b_j per provare la stretta convessità della (3.2) basterà mostrare la stretta convessità del termine

$$\xi(x_j) = \frac{1}{\beta_j + 2} a_j |x_j|^{\beta_j + 2}$$

che è evidentemente di classe $C^1(\mathbb{R})$, essendo $\beta_j > -1$. In tal modo sarà sufficiente provare che:

$$\forall z_1, z_2 \in \mathbb{R} \ : \ z_1 \neq z_2$$
$$\xi(z_2) - \xi(z_1) - \nabla_z \xi\big|_{z_1}(z_2 - z_1) > 0$$

ovvero, essendo $a_j > 0$, che:

$$\frac{1}{\beta_j + 2} \left[|z_2|^{\beta_j + 2} - |z_1|^{\beta_j + 2} \right] - z_1 |z_1|^{\beta_j} (z_2 - z_1) > 0$$

Dopo alcuni passaggi si perviene alla seguente limitazione inferiore

$$\frac{1}{\beta_j + 2} \left[|z_2|^{\beta_j + 2} + (\beta_j + 1) |z_1|^{\beta_j + 2} \right] - z_1 z_2 |z_1|^{\beta_j} > 0 \tag{3.3}$$

La disuguaglianza (3.3) è ovviamente soddisfatta quando $|z_1|\ |z_2| \neq z_1 z_2$.

Rimane quindi da esaminare solo il caso $z_2 = \vartheta z_1$ con $\vartheta \in \mathbb{R}^+ - \{1\}$ per il quale la disuguaglianza (3.3) diviene

$$\frac{1}{\beta_j + 2} \vartheta^{\beta_j + 2} + \frac{1}{\beta_j + 2} (\beta_j + 1) - \vartheta > 0$$

ed è immediato verificare che essa è soddisfatta per i valori di β_j ipotizzati, $(\alpha_j > 0)$.

Il sistema di equazioni (2.1)-(2.3) è pertanto equivalente al seguente problema di minimizzazione condizionata strettamente convesso

$$\min \left[\mathcal{F}(\mathbf{y}, \mathbf{q}) = tr\{\operatorname{diag} \mathbf{f}(\mathbf{T}^t \mathbf{y})\} - \mathbf{y}^t \mathbf{T} \mathbf{q} \right] \tag{3.4a}$$

soggetto a

$$\mathbf{T}\mathbf{q} + \mathbf{q}^* = 0 \tag{3.4b}$$

dove $\mathbf{f}(\mathbf{T}^t\mathbf{y})$ denota la matrice colonna di funzioni f_j, $j \in \{1, \dots, \ell\}$, e $tr\{\}$ sta ad indicare l'operatore traccia di una matrice quadrata.

Il carattere di stretta convessità del funzionale

$$tr\{\text{diag } \mathbf{f}(\mathbf{T}^t\mathbf{y})\}$$

è preservato $\forall \mathbf{y} \in \mathbb{R}^{n-1}$

Infatti il funzionale $\Lambda(x) = tr\{\text{diag } \mathbf{f}(\mathbf{x})\}$ è strettamente convesso essendo somma di termini strettamente convessi e quindi si ha:

$$\begin{aligned} &\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^l \ : \ \mathbf{x}_1 \neq \mathbf{x}_2 \\ &\Lambda(\mathbf{x}_2) - \Lambda(\mathbf{x}_1) - (\mathbf{x}_2 - \mathbf{x}_1)^t \nabla_x \Lambda\big|_{x=x_1} > 0 \end{aligned} \tag{3.5}$$

Posto

$$\mathcal{M}(\mathbf{y}) = \Lambda\big[\mathbf{x}(\mathbf{y})\big] \quad \text{con} \quad \mathbf{x}(\mathbf{y}) = \mathbf{T}^t\mathbf{y}$$

basterà provare che:

$$\begin{aligned} &\forall \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^{n-1} \ : \ \mathbf{y}_1 \neq \mathbf{y}_2 \\ &\mathcal{M}(\mathbf{y}_2) - \mathcal{M}(\mathbf{y}_1) - (\mathbf{y}_2 - \mathbf{y}_1)^t \nabla_y \mathcal{M}\big|_{y=y_1} > 0 \end{aligned}$$

essendo

$$\nabla_y \mathcal{M}(\mathbf{y}) = \nabla_y \Lambda\big[\mathbf{x}(\mathbf{y})\big] = \nabla_y \mathbf{x}^t(\mathbf{y}) \nabla_x \Lambda(\mathbf{x}) = \mathbf{T} \nabla_x \Lambda(\mathbf{x})$$

Dopo semplici passaggi si ottiene

$$\mathcal{M}(\mathbf{y}_2) - \mathcal{M}(\mathbf{y}_1) - (\mathbf{y}_2 - \mathbf{y}_1)^t \nabla_y \mathcal{M}\big|_{y_1} = \Lambda(\mathbf{x}_2) - (\mathbf{x}_2 - \mathbf{x}_1)^t \nabla_x \Lambda\big|_{x_1}$$

ed in virtù della (3.5), la dimostrazione è acquisita.

La sostituzione dei vincoli di uguaglianza (3.4b) nella forma non lineare (3.4a) consente di ricondursi al seguente problema di minimizzazione incondizionata:

$$\min\big[\mathcal{H}(\mathbf{y}) = tr\{\text{diag } \mathbf{f}(\mathbf{T}^t\mathbf{y})\} + \mathbf{y}^t\mathbf{q}^*\big] \tag{3.6}$$

per cui essendo $\mathcal{H}(\mathbf{y})$ in $\mathbb{R}^{n-1}$ strettamente convesso ed inferiormente limitato la soluzione ottima esiste ed è unica.

Mostriamo ora l'equivalenza tra il problema di minimizzazione incondizionata (3.6) ed il sistema di equazioni algebriche (2.1), (2.14) e la trasformazione lineare iniettiva $\mathbf{x} = \mathbf{T}^t\mathbf{y}$ [13], formalmente espresso da

$$\mathbf{T}\Psi(\mathbf{T}^t\mathbf{y}) + \mathbf{q}^* = 0 \tag{3.7}$$

Le seguenti proposizioni sono infatti equivalenti

a) $\bar{\mathbf{y}} \in \mathbb{R}^{n-1}$ è un punto di ottimo per il problema (3.6).

b) $\bar{\mathbf{y}} \in \mathbb{R}^{n-1}$ è soluzione della (3.7).

Proviamo che a) $\Rightarrow$ b).

Sia $\bar{\mathbf{y}} \in \mathbb{R}^{n-1}$ un punto di ottimo per la (3.6) allora risulta

$$\nabla_y \mathcal{H}(\mathbf{y})\big|_{y=\bar{y}} = 0$$

da cui si ottiene

$$\nabla_y \mathcal{H}(\mathbf{y}) = \nabla_y \left[\Lambda(\mathbf{x}(\mathbf{y})) + \mathbf{y}^t \mathbf{q}^*\right] = \nabla_y \left[\Lambda(\mathbf{x}(\mathbf{y}))\right] + \mathbf{q}^* = (\nabla_y \mathbf{x}^t)(\nabla_x \Lambda(\mathbf{x})) + \mathbf{q}^* =$$

$$= \mathbf{T}\nabla_x \Lambda(\mathbf{x}) + \mathbf{q}^* = \mathbf{T}\Psi(\mathbf{x}) + \mathbf{q}^* = \mathbf{T}\Psi(\mathbf{T}^t\mathbf{y}) + \mathbf{q}^*$$

L'ultima uguaglianza riferita a $\mathbf{y} = \bar{\mathbf{y}}$ prova l'asserto.

La dimostrazione che b) $\Rightarrow$ a) può essere ottenuta non appena si osservi che non esiste $\hat{\mathbf{y}} \in \mathbb{R}^{n-1}$ tale che le relazioni

$$\mathbf{T}\Psi(\mathbf{T}^t\hat{\mathbf{y}}) + \mathbf{q}^* = 0$$
$$\nabla_x \mathcal{H}(\mathbf{y})\big|_{\mathbf{y}=\hat{\mathbf{y}}} \neq 0$$

sussistano entrambe.

I risultati ottenuti si riferiscono a reti costituite da rami caratterizzati da espressioni del tipo (3.1) con a_j e b_j entrambi positivi e $\beta_j > -1$.

Per equazioni costitutive espresse dalla (2.10), si ottiene una diversa espressione del funzionale su $\mathbb{R}^{n-1}$ ma ancora strettamente convessa (cfr. Appendice), così che restano ancora provate le proprietà di esistenza e di unicità della soluzione.

Per quanto attiene alla presenza nella rete di dispositivi di regolazione il cui funzionamento idraulico può essere espresso da relazioni del tipo (2.15) o (2.13), con k_i e γ_j indipendenti dal numero di Reynolds, la linea di ragionamento rimane sostanzialmente invariata poiché il problema di minimizzazione incondizionata viene ad essere sostituito da un problema di minimizzazione condizionata su un insieme convesso. Infatti il funzionale $\mathcal{H}(\mathbf{y})$ si modifica in base all'aggiunta del termine strettamente convesso $tr\{\text{diag } \mathbf{h}(\mathbf{y})\}$, dove:

$$h_i(y_i) = k_i \frac{1}{\gamma_i + 1} (y_i - z_i)^{\gamma_i + 1} \qquad i \in \{1, \ldots, n-1\}$$

In tal modo il problema diviene:

$$\min\left[\mathcal{H}(\mathbf{y}) = tr\{\text{diag } \left[\mathbf{f}(\mathbf{T}^t\mathbf{y}) + \mathbf{h}(\mathbf{y})\right]\} + \mathbf{y}^t\mathbf{q}^*\right]$$
$$y_i \geq z_i \qquad i \in \{1, \ldots, n-1\}$$

che ammette una sola soluzione grazie della stretta convessità dei vincoli [14].

4. Conclusioni

La formulazione variazionale presentata nel Par. 3 per il problema di verifica di reti idrauliche comprendenti dispositivi idraulici e rami per i quali le relazioni tra portate volumetriche e perdite di carico sono espresse da relazioni non monomie, generalizza ben note proprietà provate in [2] e successivamente in [4], per reti magliate.

La linea di ragionamento seguita in questo lavoro si basa essenzialmente sulla teoria dei potenziali normali di dissipazione introdotti nelle questioni di meccanica non lineare dei solidi.

In base alla stretta convessità del potenziale normale di dissipazione corrispondente alle relazioni costitutive tra le variabili coniugate e alla proprietà di aggiunzione algebrica dell'operatore di continuità $\mathbf{T}$ che interviene nelle (2.1), (2.2), si acquisisce agevolmente la dimostrazione di esistenza e unicità della soluzione del problema non lineare.

L'equivalenza tra il problema algebrico (2.1)-(2.3) e il problema di minimizzazione incondizionata (3.6) può altresì fornire utili indicazioni per la scelta di algoritmi risolutivi dotati di proprietà di convergenza e stabilità.

Appendice

In questa appendice si definisce il potenziale normale di dissipazione associato alla relazione costitutiva(2.10) e se ne prova la stretta convessità.

Si ha infatti:

$$f(x) = \begin{cases} \dfrac{1}{2a} \displaystyle\int \left[(b^2 + 4ax)^{1/2} - b\right] dx & \forall x \geq 0 \\ -\dfrac{1}{2a} \displaystyle\int \left[(b^2 + 4a(-x))^{1/2} - b\right] dx & \forall x \leq 0 \end{cases}$$

e, a meno di una inessenziale costante additiva, risulta

$$f(x) = \begin{cases} \dfrac{1}{12a^2}(b^2 + 4ax)^{3/2} - \dfrac{b}{2a}\, x & \forall x \geq 0 \\ \dfrac{1}{12a^2}\left[b^2 + 4a(-x)\right]^{3/2} - \dfrac{b}{2a}\,(-x) & \forall x \leq 0 \end{cases}$$

che è ovviamente di classe $C^2(\mathbb{R}^{n-1})$.

La stretta convessità è allora conseguita non appena si osserva che:

$$\frac{d^2 f}{dx^2} = \begin{cases} \dfrac{1}{(b^2 + 4ax)^{1/2}} > 0 & \forall x \geq 0 \\ \dfrac{1}{\left[b^2 + 4a(-x)\right]^{1/2}} > 0 & \forall x \leq 0 \end{cases}$$

Bibliografia

[1] Fenves S.J.: *Network Topological Formulation of Structural Analysis*, Jour. ASCE, St. Div., 8, 1963.

[2] Russo Spena A.: *Su alcune proprietà caratteristiche delle reti*, Rend. Acc. Sc. Fis. Mat., Napoli, Serie 4, vol. XVII, 1950.

[3] Carpentier G., Cohen G., Hamam Y.: *Water Network Equilibrium Variational Formulation and Comparison of Numerical Algorithms*, in: Computer Application in Water Supply, Vol. I, J. Wiley & S., N.Y., 1987.

[4] Contro R., Franzetti S.: *A New Objective Function for Analyzing Hydraulic Pipe Networks in the Presence of Different State of Flow*, Meccanica, 18, 1983.

[5] Contro R., Franzetti S.: *Verifica delle reti idrauliche in pressione mediante programmazione non lineare*, Atti XVII Conv. Idr. e Costr. Idr., Palermo 1980.

[6] Russo Spena A.: *Il problema di verifica di sistemi di distribuzione idrica ad uso potabile o irriguo*, Giorn. Genio Civile, Fasc. 10, 11, 12, 1979.

[7] Narshing D.: *Graph Theory with Applications to Engineering and Computer Science*, Prentice–Hall, N.J., 1974.

[8] Citrini D.: *Una formula semplice per il moto uniforme delle correnti fluide nella zona di Colebrook*, Energia Elettrica, 1962.

[9] Nebbia G., Ippolito G, Russo Spena A., Viparelli M.: *Idraulica*, Liguori, Napoli, 1968

[10] Narayanaswamy O.S.: *Processing Nonlinear Multipoint Constraints in the Finite ment Method*, Int. Jour. Num. Meth. Eng., Vol. 21, 1985.

[11] Panagiotopoulos P.D.: *Variational Inequalities in Mechanics and Applications*, Birkäuser, Basel, 1985.

[12] Kantorovic L.V., Akilov G.P.: *Analisi funzionale*, Editori Riuniti, Roma, 1980.

[13] Ben–Israel A., Greville T.N.E.: *Generalized Inverses: Theory and Applications*, J. Wiley & S., N.Y., 1974.

[14] Wismer D.A., Chattergy R.: *Introduction to Nonlinear Optimization*, North Holland, New York, 1979.

Indirizzo degli Autori:

Dr. Francesco Russo Spena
Professore Associato di Fondamenti degli Equilibri non–lineari
Dipartimento di Scienza delle Costruzioni
Facoltà di Ingegneria
P.le V. Tecchio, 80125 Napoli
Ph. 081/7682111

Dr. Andrea Vacca
Allievo "Dottorato di Ricerca in Ingegneria Idraulica"
Facoltà di Ingegneria
Via Claudio 21, 80125 Napoli
Ph. 081/7683463