ROCKY MOUNTAIN
JOURNAL OF MATHEMATICS
Volume 10, Number 1, Winter 1980

A MIXED APPLICATION OF FUNCTIONAL AND NUMERICAL ANALYSIS TO BIOSTATISTICS

EMILIO GAGLIARDO

ABSTRACT. The results of investigations on viral diseases of insects, collected by the United States Department of Agriculture and coded into a computer-language list, posed several statistical-inference problems which are currently matters of investigation. One of these problems is here reported. Its solution did imply in a crucial step the analogue of a well known concept in classical analysis: the idea of the intermediate Hilbert space of N. Aronszajn.

1. *Data analysis.* More than 1000 (continuously growing) insect-virus reports are structured in the computer-based catalog programmed by M.E. Martignoni—P. Williams—D.E. Reineke—P.J. Iwai [4], [5], [6]. Among the statistical—inference problems posed by this large amount of data, the one considered as the most interesting by the microbiologists is:

What are the inferred probabilities of occurrence of viral diseases among the listed orders or families or genera? Is it possible to determine a 2- or 3-dimensional punctual representation of diseases, orders, and families, where the distances may easily point out extrapolated similarities, of interest in making judgments, predictions, decisions?

A rough data analysis shows that the insect viral diseases, which in the present paper will be coded as follows:

A = acute paralysis
B = chronic paralysis
C = CO_2 sensitivity
D = crystalline-array virosis
E = cytoplasmic polyhedrosis
F = densonucleosis
G = flacherie
H = gattine
I = granulosis
L = iridescent virosis
M = malaya disease
N = nucleopolyhedrosis
O = other nonoccluded-virus disease
P = other occluded-virus disease
Q = paralysis

Received by the editors on April 13, 1978.

Copyright © 1980 Rocky Mountain Mathematics Consortium

R = sacbrood
S = spheroidosis
T = watery disintegration
U = hairless-black syndrome

have non uniform similarities. It also appears that non uniform similarities, *from the point of view of viral diseases,* exist among the 9 orders of insects:

1 = acarina
2 = coleoptera
3 = diptera
4 = hemiptera
5 = hymenoptera
6 = lepidoptera
7 = neuroptera
8 = orthoptera
9 = trichoptera

On the other hand when the 1000 insect-virus reports [4], [5], [6] are considered individually, a trivial statistical analysis shows that most of these obvious similar'ties disappear in the too much detailed fragmentation. In other words, apparently many disease reports do not mention all diseases simultaneously present. It seems therefore reasonable to infer similarities between diseases (i.e. to infer "distances" of diseases) only after "grouping" the data in a suitable way: e.g. according to families or to orders. In the present paper we will describe a procedure which has been derived from functional analysis in order to determine "distances" relative, so to say, to an "intermediate" way of grouping the data. The resulting agreement with some other information about these diseases (see §4) seems to suggest the use of this new "intermediate" metric.

2. **Weighted correlation estimates.** Let us group the insect-virus occurrences in classes (e.g. families or orders: the way of grouping will be discussed in §3).

For each class k let n_k be the number of reported insects, and m_{ki} the number of those with disease i.

An over-all correlation between diseases i and j may be computed in terms of the weighted expressions:

$$a = \sum_k n_k \frac{m_{ki}}{n_k} \frac{m_{kj}}{n_k}, \qquad b = \sum_k n_k \frac{m_{ki}}{n_k} \frac{(n_k - m_{kj})}{n_k},$$

$$c = \sum_k n_k \frac{(n_k - m_{ki})}{n_k} \frac{m_{kj}}{n_k}, \quad d = \sum_k n_k \frac{(n_k - m_{ki})}{n_k} \frac{(n_k - m_{kj})}{n_k}$$

by means of the three estimates, well known in statistics:

$$Q^{ij} = \frac{ad - bc}{ad + bc}, \quad Y^{ij} = \frac{(ad)^{1/2} - (bc)^{1/2}}{(ad)^{1/2} + (bc)^{1/2}},$$
$$V^{ij} = \frac{ad - bc}{((a + b)(a + c)(b + d)(c + d))^{1/2}}.$$

As "distance" between the two diseases we may assume, respectively:

$$D_1^{ij} = 1 - (Q^{ij} + 1)/2, \; D_2^{ij} = 1 - (Y^{ij} + 1)/2,$$
$$D_3^{ij} = 1 - (V^{ij} + 1)/2.$$

Other experimented distances are:

$$D_4^{ij} = \frac{1}{2}\sum_k \left| \frac{m_{ki}}{\sum_h m_{hi}} - \frac{m_{kj}}{\sum_h m_{hj}} \right|, \quad D_5^{ij} = \frac{1}{2}\left(1 - \frac{\sum_k m_{ki} m_{kj}}{\sqrt{\sum_h m_{hi}^2 \sum_h m_{hj}^2}}\right).$$

All these "distances" obviously satisfy: $O \leqq D_s^{ij} \leqq 1$. In simulated experiments (where the insects are points of the 6-dimensional space with binary coordinates and diseases, orders, families are defined by Boolean functions or by linear thresholds) the "distance", (slightly) less sensitive to a change in the way of grouping, appeared to be D_4^{ij}.

3. **Two-dimensional display.** All distances D_s^{ij} between disease i and disease j considered in §2 and in particular the distances D_4^{ij} depend on the way of grouping the data: according to individual reports, according to families, or according to orders. In order to have an intuitive appreciation of these estimates it is useful to represent, for each choice of grouping, the 19 diseases as 19 points of the plane in such a way that the mutual actual distances d_{ij} are as close as possible to the distances D_4^{ij}. Of course a perfect agreement is usually possible only in a sufficiently high dimensional space (but also a 2-dimensional picture is helpful).

The minimum of $\sum_{ij} (d_{ij} - D_4^{ij})^2$, which in the n-dimensional space turns out to be a function of 19 variables (the coordinates of the 19 points which represent the 19 diseases), is easily determined by means of a fast method which minimizes a function *without evaluating derivatives* and whose store requirements are only proportional to the number of variables; see Gagliardo-Sacchi [3]; a short version is included in §6.

The 2-dimensional representations, when grouping the data according to families or to orders, are given respectively in Figure 1 and Figure 2. Their limited but positive agreement shows the existence of similarities (between diseases) which do not appear when individual reports are considered.

4. **Grouping by families or by orders.** Let E_{ij}, F_{ij} be the distances D_4^{ij} when grouping the data according to families or according to orders. In Figure 1 as well as in Figure 2 the continuous line shows the area of

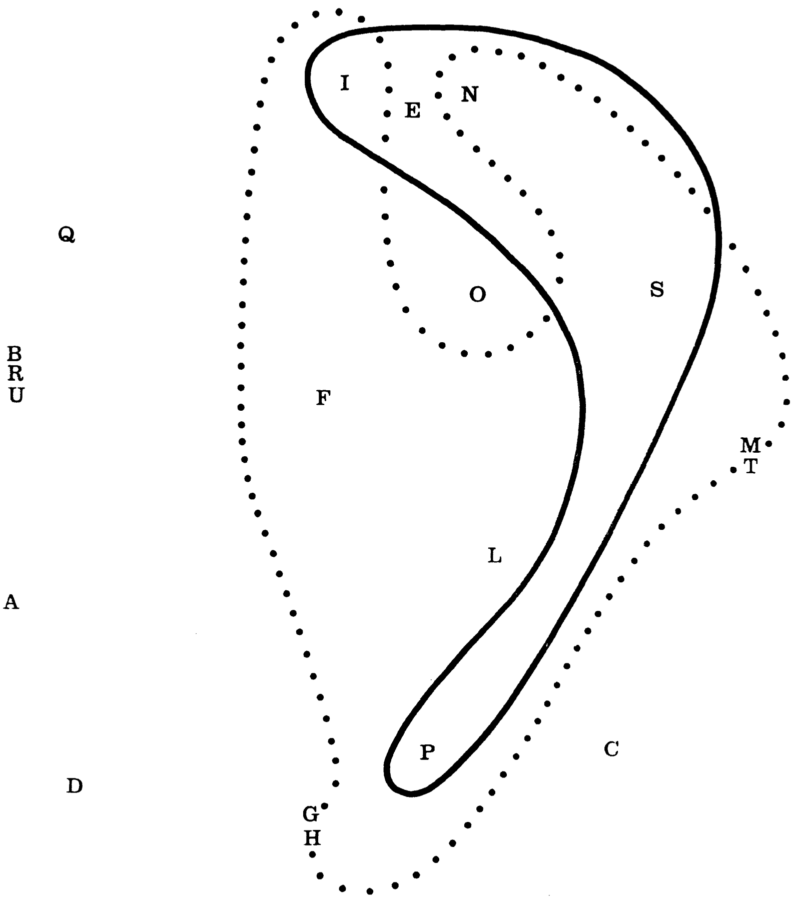


Figure 1

OCCLUDED diseases while the dotted line contains the DNA diseases (see [4] [5] [6]). These two notions were not coded in the input data in order to furnish a final test of reliability for these representations. The feeling that both metrics E_{ij}, F_{ij} contain useful information leads to the study of a metric G_{ij} which should correspond to an "intermediate" grouping between the grouping by families and the grouping by orders.

In functional analysis the first "intermediate" space C between Banach spaces A, B in the sense of the "intermediate Hilbert spaces" introduced

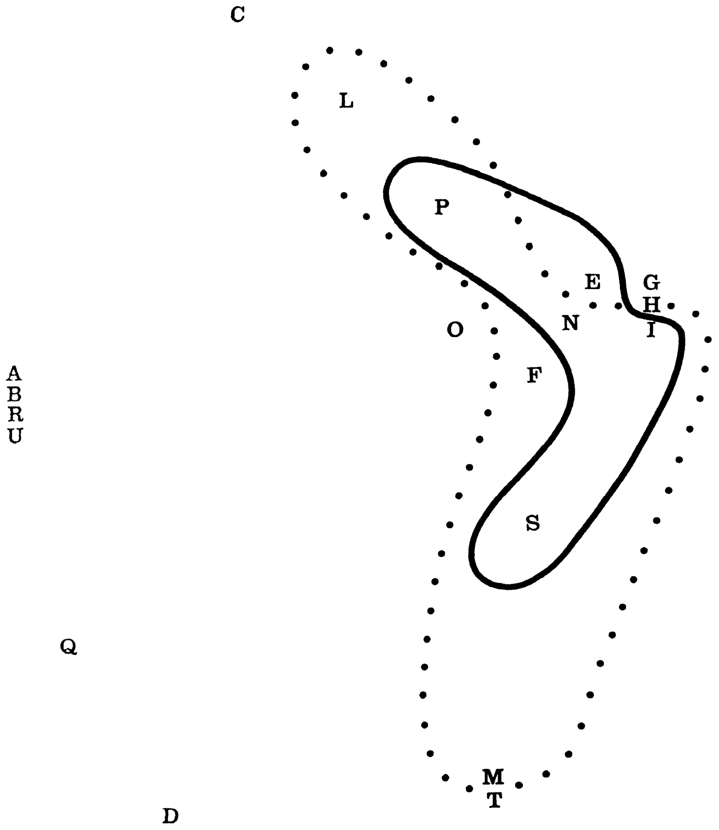


Figure 2

by N. Aronszajn, was constructed by:

$$\|u\|_C = \left(\int_{\partial D} y\, dx\right)^{1/2} \text{ where}$$

$$D = \left\{(x, y) \middle| u = v + w,\ \|v\|_A \leqq x,\ \|w\|_B \leqq y\right\}$$

and it was shown that if $A = L_1$, $B = L_\infty$ then C turns out to be L_2.

Many other functionals have been introduced in the development of that theory [**1**].

Following the general idea of these definitions we may now introduce the new distance between disease i and disease j:

$$G_{ij} = \min\left[\min_{k\neq i,\, \neq j} (E_{ik} + F_{kj}),\ \min_{k\neq i,\, \neq j} (F_{ik} + E_{kj})\right]$$

which leads to the 2-dimensional representation given in Figure 3. (Note that G_{ij} does not need to be between the numbers E_{ij}, F_{ij}). Both OCCLUDED as well as DNA diseases seem now to occupy "almost-convex" central areas. In this figure also the orders have been represented (in points determined by a trivial procedure of weighted means based on the frequencies of diseases in the orders).

5. **Inferences.** The matrix of distances G_{ij} and the resulting 2-dimensional display of diseases and orders (Figure 3) are intended to infer and show disease-disease and disease-order interrelations which could not

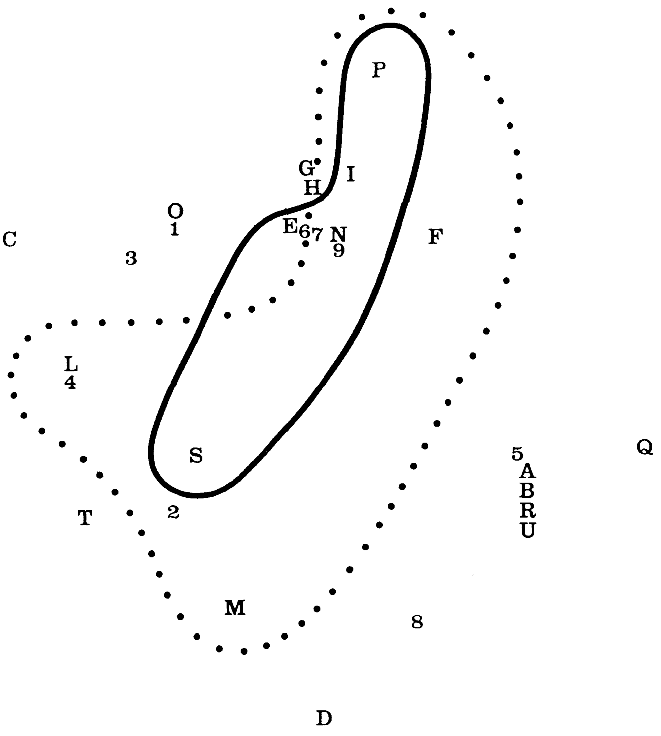
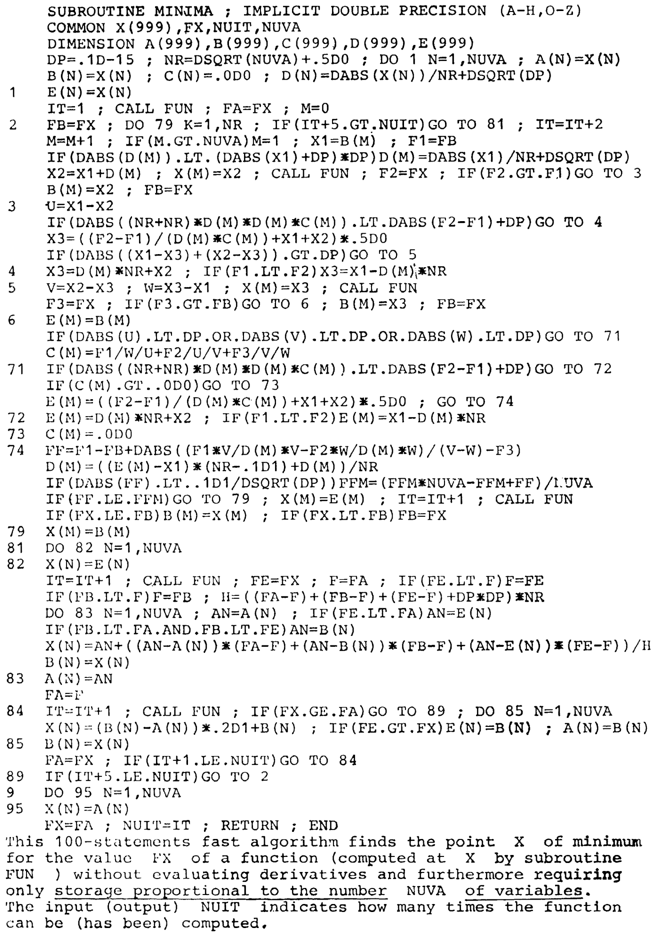


Figure 3

appear from the not always detected or not always reported frequencies of simultaneous occurrences. In fact, as already remarked, the distances E_{ij}, F_{ij}, and especially G_{ij} depend on extrapolated similarities based on analogies (like when inferences are made by means of procedures of artifical intelligence [2]).

As a final remark we may note that two of the three diseases H, Q, U of unknown type (DNA or RNA), and precisely Q, U, appear also in Figure 3 rather outside of the dotted line containing DNA diseases.



```
     SUBROUTINE MINIMA ; IMPLICIT DOUBLE PRECISION (A-H,O-Z)
     COMMON X(999),FX,NUIT,NUVA
     DIMENSION A(999),B(999),C(999),D(999),E(999)
     DP=.1D-15 ; NR=DSQRT(NUVA)+.5D0 ; DO 1 N=1,NUVA ; A(N)=X(N)
     B(N)=X(N) ; C(N)=.0D0 ; D(N)=DABS(X(N))/NR+DSQRT(DP)
1    E(N)=X(N)
     IT=1 ; CALL FUN ; FA=FX ; M=0
2    FB=FX ; DO 79 K=1,NR ; IF(IT+5.GT.NUIT)GO TO 81 ; IT=IT+2
     M=M+1 ; IF(M.GT.NUVA)M=1 ; X1=B(M) ; F1=FB
     IF(DABS(D(M)).LT.(DABS(X1)+DP)*DP)D(M)=DABS(X1)/NR+DSQRT(DP)
     X2=X1+D(M) ; X(M)=X2 ; CALL FUN ; F2=FX ; IF(F2.GT.F1)GO TO 3
     B(M)=X2 ; FB=FX
3    U=X1-X2
     IF(DABS((NR+NR)*D(M)*D(M)*C(M)).LT.DABS(F2-F1)+DP)GO TO 4
     X3=((F2-F1)/(D(M)*C(M))+X1+X2)*.5D0
     IF(DABS((X1-X3)+(X2-X3)).GT.DP)GO TO 5
4    X3=D(M)*NR+X2 ; IF(F1.LT.F2)X3=X1-D(M)*NR
5    V=X2-X3 ; W=X3-X1 ; X(M)=X3 ; CALL FUN
     F3=FX ; IF(F3.GT.FB)GO TO 6 ; B(M)=X3 ; FB=FX
6    E(M)=B(M)
     IF(DABS(U).LT.DP.OR.DABS(V).LT.DP.OR.DABS(W).LT.DP)GO TO 71
     C(M)=F1/W/U+F2/U/V+F3/V/W
71   IF(DABS((NR+NR)*D(M)*D(M)*C(M)).LT.DABS(F2-F1)+DP)GO TO 72
     IF(C(M).GT..0D0)GO TO 73
     E(M)=((F2-F1)/(D(M)*C(M))+X1+X2)*.5D0 ; GO TO 74
72   E(M)=D(M)*NR+X2 ; IF(F1.LT.F2)E(M)=X1-D(M)*NR
73   C(M)=.0D0
74   FF=F1-FB+DABS((F1*V/D(M)*V-F2*W/D(M)*W)/(V-W)-F3)
     D(M)=((E(M)-X1)*(NR-.1D1)+D(M))/NR
     IF(DABS(FF).LT..1D1/DSQRT(DP))FFM=(FFM*NUVA-FFM+FF)/NUVA
     IF(FF.LE.FFM)GO TO 79 ; X(M)=E(M) ; IT=IT+1 ; CALL FUN
     IF(FX.LE.FB)B(M)=X(M) ; IF(FX.LT.FB)FB=FX
79   X(M)=B(M)
81   DO 82 N=1,NUVA
82   X(N)=E(N)
     IT=IT+1 ; CALL FUN ; FE=FX ; F=FA ; IF(FE.LT.F)F=FE
     IF(FB.LT.F)F=FB ; H=((FA-F)+(FB-F)+(FE-F)+DP*DP)*NR
     DO 83 N=1,NUVA ; AN=A(N) ; IF(FE.LT.FA)AN=E(N)
     IF(FB.LT.FA.AND.FB.LT.FE)AN=B(N)
     X(N)=AN+((AN-A(N))*(FA-F)+(AN-B(N))*(FB-F)+(AN-E(N))*(FE-F))/H
     B(N)=X(N)
83   A(N)=AN
     FA=F
84   IT=IT+1 ; CALL FUN ; IF(FX.GE.FA)GO TO 89 ; DO 85 N=1,NUVA
     X(N)=(B(N)-A(N))*.2D1+B(N) ; IF(FE.GT.FX)E(N)=B(N) ; A(N)=B(N)
85   B(N)=X(N)
     FA=FX ; IF(IT+1.LE.NUIT)GO TO 84
89   IF(IT+5.LE.NUIT)GO TO 2
9    DO 95 N=1,NUVA
95   X(N)=A(N)
     FX=FA ; NUIT=IT ; RETURN ; END
```

This 100-statements fast algorithm finds the point X of minimum for the value FX of a function (computed at X by subroutine FUN) without evaluating derivatives and furthermore requiring only storage proportional to the number NUVA of variables. The input (output) NUIT indicates how many times the function can be (has been) computed.

A faster algorithm will appear in: Atti della Accademia Ligure 1979.

Figure 4

6. How to minimize the storage requirements for minimizing a function without evaluating derivatives. The shortest approach to describe the method mentioned in §3 (see [3]) for finding the point (X) of minimum of a function (FX, computed by subroutine FUN) *witthout evaluating derivatives nor difference quotients* but only computing (up to NUIT times) the function, and minimizing also the storage requirements (only proportional to the number NUVA of variables), is to list the short version in Figure 4 for NUVA $\leqq$ 999 variables, which minimizes also the number of its FORTRAN statements (83).

One of the ideas in the above program·is a kind of seemingly wrong extrapolation which along a "bended valley" keeps the minimizing path away from the "river", on which every straight-line local search would increase the function and consequently slow down the convergence.

ACKNOWLEDGMENTS. We are grateful to Microbiologist Dr. M. Martignoni of the United States Department of Agriculture for continually sending large amounts of structured data, with cooperative spirit, and for pointing out open questions.

We are indebted to Dr. C. Nadalet for selecting and programming all statistical experiments with the computer.

REFERENCES

1. N. Aronszajn and E. Gagliardo, *Interpolation spaces and interpolation methods*, Annali di Matematica P. e A. (1965), 51–117.

2. E. Gagliardo, *Applications of Artifical Intelligence to Astronomy, Molecular Biology, Nuclear Physics*, Bollettino della Unione Matematica Italiana (4) **9** (1974), 125–136.

3. E. Gagliardo and G. Sacchi, *An operator improving the performance of minimizing algorithms* (to appear).

4. M.E. Martignoni, P. Williams, and D.E. Reineke, *Computer-based Catalog of Viral Diseases of Insects*, Journal of invertebrate pathology (1) **22** (1973), 100–107.

5. M.E. Martignoni and P.J. Iwai, *A Catalog of viral diseases of insects and mites*, U.S. Dept. of Agriculture, general techn. report PNW-40.

6. ———, *A Catalog of viral diseases of insects and mites*, U.S. Dept of Agriculture, general techn, report PNW-40, 2nd ed, 1977

ISTITUTO MATEMATICO, UNIVERSITA, PAVIA, ITALY