EFFICIENT CONVERSION OF NON-METRIC INFORMATION INTO METRIC INFORMATION

By: Robert P. Abelson and John W. Tukey
Yale University and Princeton University

The title of this paper may prove misleading. "Conversion of non-metric information into metric information" may sound like getting something for nothing. In fact, we are concerned with getting a convenient modest something for an inconvenient modest something. Further, our methods are thus far limited to a particular class of situations. We do not have anything like a universal recipe for converting the qualitative into the quantitative, though a good deal can be done within the confines of the situations with which we are here concerned.

Consider n points to which we wish to assign numerical values, $X_1, X_2, \ldots X_n$. Suppose we have insufficient information to provide a natural or "correct" assignment of numerical values; our knowledge is limited to a set of constraints on the values, i.e. a set of inequalities on the X's. (For example, $X_1 \leq X_2 \leq \ldots \leq X_n$).

Consider first the purposes that an assignment of numerical values can serve in such a situation. It has become somewhat fashionable, particularly in certain areas of the behavioral sciences, to frown upon "arbitrary" numerical assignment (scaling) procedures.Some take the position that in the absence of a compelling rationale for numerical assignment, no numerical assignment whatever should be attempted. (Stevens, 1951). Thus if a set of points are known only up to a rank order, one is limited to the declaration of an "ordinal scale". Further manipulations using the scale are limited, so the dictum goes, to techniques appropriate to ordinal scales -- in particular, to those non-parametric statistical techniques designed for the analysis of rankings. The net effect of this dictum is to restrict the flexibility of statistical analysis severely and unnecessarily.

Reliance in such circumstances upon non-parametric procedures seems to us to be unwise, not because such procedures always lack power (90% power is no cause for disdain), but because they are poorly adapted to the variety of uses one requires for good insight into bodies of data. Often when adaptation to new uses is attempted, it is only at considerable sacrifice of power (as in the situations discussed here). Furthermore, the typical state of knowledge short of metric information is not rank-order information; ordinarily, one possesses something more than rank-order information. For example, one may know that X_1, X_2, and X_3 are ordered and in addition that X_2 is closer to X_3 than it is to X_1. Non-parametric techniques which take full advantage of such types of situation are generally unavailable. We would like to probe more deeply here, to gain some idea of what lies between rank-order scales and metric scales.

Consider now the kind of problem for which a numerical assignment procedure is useful. Suppose that the n points represent levels of an independent variable and that we wish to carry out the regression of a dependent variable (about which we have metric information) upon this independent variable (about which we have only non-metric information). To be even more specific, the independent and dependent variables might be imbedded in an analysis of variance design where we were interested in forming a single degree of freedom contrast among the levels of the independent variable. The appropriate coefficients to use in forming such a contrast would be a direct outcome of an assignment of numerical values.

To sum up thus far: we seek a procedure for assigning numerical values to a set of n entities, given a set of inequalities which the assigned values must obey. The problem is of interest because a) it sheds light upon the nature of knowledge intermediate between rank-order knowledge and metric knowledge, and b) the solution makes powerful regression techniques, particularly the formation of contrasts, applicable to many situations when the entities represent levels or versions of an independent variable.

The criterion for good numerical assignment.

The sequence of n numerical values to be assigned must obey certain inequalities. Likewise, the "ideal" values, the values one would assign if one had full scale knowledge, must obey the same inequalities. That is, both the sequence we choose and the sequence we ought to have chosen lie in the convex set of sequences permitted by the inequalities. Denote the chosen sequence by $[X_1, X_2, \ldots X_n]$ and the ideal sequence by $[Y_1, Y_2 \ldots Y_n]$. A convenient and reasonable criterion of the success of our choice is the square of the formal product-moment correlation between [X] and [Y]:

$$r^2 = \frac{\left[\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2 \quad \sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

To avoid confusion, one must note that this correlation coefficient is purely "formal" and is not to be thought of in terms of a bivariate distribution from which points are sampled. This r^2 plays a key role in the specific application discussed earlier. In testing the significance of a contrast, the power of the test increases directly with r^2. The same r^2 is almost ubiquitous in other aspects of regression analysis.

There is an obvious difficulty with r^2 as a criterion: one does not know the ideal sequence, [Y]. The sequence [X] is of our choice; but, in our ignorance, [Y] might be any sequence within a certain range of possibilities. A further choice must be made in order to provide a usable criterion. On the one hand, one might make some kind of distributional assumption about the possible [Y]'s, and average r^2 over this a priori distribution. It is difficult to do this in any reasonable and meaningful way. (Indeed the resulting mathematical problem is rather difficult to attack.) On the other hand, one might make the conservative, fixed assumption that the Y-sequence may well be such as to minimize r^2 for the chosen X-sequence. This minimum r^2, for [Y]

satisfying the inequalities and [X] fixed, is the criterion we have chosen to assess any fixed [X]. The mathematical problem then becomes a maximin problem: how should one choose [X] such that the minimum r^2 is maximized?

In other words, we play a "game against Nature" in which we fear the worst. For any choice of numerical assignments for [X], assume that Nature chooses a set of "true" values [Y] which obey the inequalities but yield r^2 min., the lowest possible squared product-moment coefficient with [X]. We play the game by choosing [X] such that r^2 min. is maximized. (We refer to this choice as the "maximin sequence", denoted by [C]). This results in a guarantee that r^2 cannot be less than a certain value, denoted as r^2 maximin, so long as Nature obeys the inequalities. The strategy amounts to optimizing the conservative guarantee, rather than maximizing some kind of average value.

Mathematical properties of the problem.

Since the criterion is correlational, the units and origins of the sequences [X] and [Y] are immaterial to the maximin problem. In what follows, only the relative spacing of the n numerical values is of consequence, while the units and origins are chosen for convenience.

With a given set of inequalities there is associated a special set of sequences which we call "corners", such that any admissible sequence [Y] can be generated as a positive linear combination of the corner sequences. (The name corner arises from the geometrical conception of the permissible sequences as a convex set of vectors). As an example, consider the rank order case: $Y_1 \leq Y_2 \leq Y_3 \leq Y_4$. Fixing $Y_1 = 0$, a simple set of corners is the triplet:

(0, 0, 0, 1)
(0, 0, 1, 1)
(0, 1, 1, 1)

Any [Y] satisfying rank order (with $Y_1 = 0$) can be expressed as a positive linear combination of these corners.

The corners provide the key to the maximin solution, via the following two theorems proved by Tukey. Proofs are given in our more extended article. (Abelson and Tukey, 1958).

Theorem I. For any fixed [X], minimum r^2 is reached for [Y] equal to one of the corner sequences. In other words, whatever choice we make for [X], Nature plays her most damaging game at one of the corners. Consider the rank order case again, and suppose we "play" the equal-interval sequence -3,-1,1,3. Nature achieves r^2 min. = .600 by playing 0,0,0,1 or 0,1,1,1. No worse than this can be done to us when we play -3,-1,1,3. (However, we have a better play in the maximin sense).

Theorem II. (Oversimplified) The maximum r^2min is achieved by the sequence which correlates equally with all corner sequences. In the rank order case with n = 4, we need simply find the sequence (C_1, C_2, C_3, C_4) which correlates equally with (0,0,0,1), (0,0,1,1), and (0,1,1,1). This is a matter of simultaneous linear equations in the unknown C's which are readily solved.

Theorem 2 as it has been given here is not correct for all sets of inequalities. In particular, the theorem fails when the sequence which correlates equally with all corners does not itself satisfy the appropriate inequalities. Further complications arise when there are more than (n-1) corners in a given case. The fuller paper goes into the subtle details involved. The correct but more involved theorem will simply be stated here in passing: For any system of inequalities with its associated corner sequences, there exists one and only one sequence which a) is a positive linear combination of a set of the corner sequences such that b) it correlates equally highly with all these corners and c) more highly with corners not in the set, if any.

Results for the rank order case

First we present in some detail the results for the rank order case. Then, more briefly, the results for other cases. Throughout we use for the values of the maximin sequence the convenient normalization

$$\sum_{i=1}^{n} C_i = 0 \qquad \sum_{i=1}^{n} C_i^2 = 1/r^2 \text{ maximin}$$

The maximin sequence in the rank order case for n=4 is: -.866, -.134, .134, .866. For n=8, the values are: -.935, -.289, -.144, -.045, .045, .144, .289, .935. For indefinitely large n, the limiting values for the extreme points are: -1.000, -.414, -.318, -.268.....,.268, .318, .414, 1.000.

The solution is markedly non-linear. The values at the two ends have very large relative separation from the next values inwards. This comes about because the solution is guarding against the possibility that Nature will play 0, 0, ... 0,0,0,1 or 0,1,1,1, ... 1,1. A linear sequence can fail rather badly against these possibilities, especially for large n. However, in practice one is often unwilling to acknowledge sequences as pathological as a,a,... a,a,a,b as reasonable possibilities for the "true" sequence. If so, then one may attempt to rule out such unusual sequences from Nature's repertoire. This means reformulating the inequalities so that these pathological corners do not occur. This is possible in a number of ways, all of which require that something more stringent than mere rank order be assumed. When this is done, one finds that the end values of the maximin sequence are not forced to lie so far from the body of the sequence as in the rank order case, and a linear sequence does not fare as poorly (in the maximin sense) as a basis for numerical assignment. Of this, more later.

The following brief display gives r^2 maximin in the rank order case for various values of n; by way of comparison, the values of r^2 min. against a linear assignment are shown.

n	5	10	20	50
r^2 maximin	.596	.478	.406	.339
r^2 min.(linear)	.500	.273	.143	.059

100	200	500	1000	∞
.303	.274	.244	.225	$2/2+\log.(m-1)$
.030	.015	.006	.003	$3/(m+1)$

Asymptotically, r^2 maximin approaches zero, but very slowly, whilst r^2 min. for a linear sequence approaches zero rather rapidly.

The form of the maximin sequence may be roughly approximated with a simple pattern of integers by the following device: write down a linear sequence with mean zero, quadruple the extreme values and double the next-to-end values. At n=8, for example, this quick approximation to the maximin sequence would be (-28,-10,-3,1,1,3, 10,28). For n less than 50, the r^2 min. for this approximation is at least 90% as high as r^2 maximin. This scheme, which we dub as the "linear-2-4" sequence, is easily remembered.

If Nature is really playing a near-linear sequence, then of course we would be better off by playing a linear sequence than by guarding against wild behavior of Nature by playing the maximin solution or its surrogate, the linear-2-4. If we would like to achieve higher r^2 in case Nature's behavior is near-linear without risking too great a drop in r^2 below the maximin value in case Nature's behavior is wild, a good hedge for small n is to choose a "linear-2" sequence; that is, a linear sequence with the end values doubled. In passing, it might be mentioned that "rankits", like linear coefficients, fare poorly when Nature is behaving wildly.

Other orderly cases

Definitions

I. Symmetric Rank order

$X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_{n-1} \leq X_n$

$D_1 = D_{n-1}$, $D_2 = D_{n-2}$, $D_3 = D_{n-3}$ ...

where $D_1 = (X_2-X_1)$, $D_2 = (X_3-X_2)$ ...

$D_{n-1} = (X_n-X_{n-1})$

II. Symmetric, Extremes Bunched

$X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_{n-1} \leq X_n$

$(D_1 = D_{n-1}) \leq (D_2 = D_{n-2}) \leq (D_3 = D_{n-3}) \cdots$

III. Non-symmetric, Extremes Bunched

$X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_{n-1} \leq X_n$

$D_1 \leq D_2 \leq D_3 \ldots; D_{n-1} \leq D_{n-2} \leq D_{n-3} \cdots$

IV. Symmetric, Extremes Spread

$X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_{n-1} \leq X_n$

$(D_1 = D_{n-1}) \geq (D_2 = D_{n-2}) \geq (D_3 = D_{n-3}) \ldots$

V. Non-symmetric, Extremes Spread

$X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_{n-1} \leq X_n$

$D_1 \geq D_2 \geq D_3 \geq \ldots; D_{n-1} \geq D_{n-2} \geq D_{n-3} \cdots$

VI. "Diminishing Returns"

$X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_{n-1} \leq X_n$

$D_1 \geq D_2 \geq D_3 \geq \ldots \geq D_{n-1}$

(or mirror image)

Maximin r^2

n	I	II	III	IV	V	VI
3	1.000	1.000	.750	1.000	.750	.933
4	.853	.947	.909	.974	.667	.887
5	.853	.974	.778	.947	.625	.834
6	.786	.940	.874	.922	.599	.827
7	.786	.964	.787	.901	.578	.806
8	.744	.936	.856	.882	.561	.789
9	.744	.958	.791	.865	.548	.774
10	.714	.935	.845	.850	.536	.761
11	.714	.953	.793	.837	.526	.750
12	.692	.935	.838	.826	.517	.740
13	.692	.950	.794	.815	.507	.731
14	.674	.935	.832	.805	.501	.724
15	.674	.949	.795	.796	.492	.716
16	.659	.934	.827	.788	.487	.710
17	.659	.948	.795	.780	.479	.704
18	.646	.934	.825	.773	.475	.699
19	.646	.947	.796	.767	.467	.694
20	.636	.934	.823	.762	.464	.689

Maximin weights exemplified: n = 8

C_1	-.707	-.477	-.548	-.707	-.935	-.935
C_2	-.293	-.395	-.418	-.219	-.110	-.160
C_3	-.225	-.312	-.289	-.131	-.055	-.062
C_4	-.189	-.230	-.159	-.044	0	.036
C_5	.189	.230	.159	.044	0	.134
C_6	.225	.312	.289	.131	.055	.231
C_7	.293	.395	.418	.219	.110	.329
C_8	.707	.477	.548	.707	.935	.427

Further cases for small n

In the literature on psychological scaling, the case in which the first differences of a ranked sequence are ranked is called an "ordered metric scale" (Coombs,1950). The cases treated above are a limited coverage of this variety of scale. Next we consider all possible ordered metric scales with n=3,4,or 5. Also, we consider for n=4 all possible "higher-ordered metric scales" (Siegel, 1956). These are cases in which all differences of a ranked sequence are ranked. In addition, the rank order case for n=3 and 4 is considered with a numerical constraint on the relative size of the biggest or the smallest interval.

Ordered metric scales with n=3

There is only one case here. We have $X_1 < X_2 < X_3$ and $X_2-X_1 > X_3-X_2$. (The other possibility is simply a mirror image of this one). The maximin sequence can be approximated with the simple integer sequence (-7,2,5) with r^2 min. = .923.

n=4

Maximin coefficients and maximin r^2, for all cases of n=4 involving only simple inequalities among differences.

1. The three differences (X_2-X_1), (X_3-X_2) and (X_4-X_3) are represented by the digits 1,2, and 3.
2. When the inequalities specify that a particular difference is greater than another, the larger difference is written first (e.g. the system of inequalities $(X_3-X_2) \geq (X_2-X_1) \geq (X_4-X_3) \geq 0$ is written 213.)
3. When the relative size of two differences is not specified, they are enclosed in parentheses. (e.g. the system of inequalities $(X_2-X_1) \geq (X_3-X_2) \geq 0$; $(X_2-X_1) \geq (X_4-X_3) \geq 0$ is written 1 (23).

System	C_1	C_2	C_3	C_4	r^2
(13)2	-87	00	00	87	.667
(12)3	-87	-13	50	50	.789
1(23)	-87	05	27	55	.887
132	-87	05	27	55	.887
123	-87	04	29	54	.887
2(13)	-66	-34	34	66	.909
213	-66	-34	42	58	.941

n=5

System	C_1	C_2	C_3	C_4	C_5	r^2
(124)3	-89	-20	00	20	89	.595
(123)4	-89	-20	00	55	55	.694
(14)(23),(14)23	-89	00	00	00	89	.625
(13)(24),(13)24, (13)42	-89	-10	-10	55	55	.704
(12)(34),(12)34	-89	-20	33	36	40	.801
(23)(14),(23)14	-55	-55	00	55	55	.833
1(234)	-89	00	08	25	57	.840
1(34)2,1342, 14(23),1432	-89	01	04	29	55	.843
1(24)3, 1423	-89	-04	12	29	52	.853
1(23)4,12(34),(12)43, 1324,1234,1243	-89	-05	13	32	50	.854
2(134)	-69	-40	16	35	58	.892
2(14)3	-69	-40	26	26	58	.901
24(13)	-61	-49	16	36	58	.914
23(14)	-56	-54	14	40	56	.918
2(34)1,2341,2431	-56	-54	16	36	57	.921
2(13)4	-69	-40	18	44	47	.923
2413	-61	-49	21	30	58	.927
21(34),2134,2143	-69	-40	21	38	50	.931
2314	-63	-47	19	37	53	.935

Higher-ordered metric scales for n=4

These cases can be specified as elaborations of ordered metric scales. The notation is as in the previous displays, with the addition of constraints upon the sums of differences. For example (in the abbreviated notation), $1+2 \geq 3$ signifies $D_1+D_2 \geq D_3$.

System	C_1	C_2	C_3	C_4	r^2
123; $1 \geq 2+3$	-87	07	36	43	.933
123; $1 \leq 2+3$	-73	-19	35	58	.972
132; $1 \geq 2+3$	-87	16	16	55	.908
132; $1 \leq 2+3$	-72	-10	13	69	.974
213; $2 \geq 1+3$	-66	-34	50	50	.952
213; $2 \leq 1+3$	-73	-19	35	58	.972

A further case with n=3

Rank order is known, and the ratio of the larger D to the smaller D does not exceed K. (We do not know which D is larger, however).

K	r^2
9	.824
4	.893
2	.964
1.5	.987

The maximin sequence for any K can be represented most simply by (-1,0,1).

n=4

Rank order is known, and the largest D (whichever it is) does not exceed the fraction p of the range.

p	C_1	C_2	C_3	C_4	r^2
.40	-67	-24	24	67	.981
.50	-71	-24	24	71	.893
.60	-71	-26	26	71	.865
.70	-75	-22	22	75	.816
.80	-79	-16	16	79	.765
.90	-83	-14	14	83	.711

Rank order is known, and the smallest D, whichever it is, is not less than the fraction q of the range.

q	C_1	C_2	C_3	C_4	r^2
.05	-82	-15	15	82	.723
.10	-78	-16	16	78	.796
.15	-74	-18	18	74	.865
.20	-71	-19	19	71	.924
.25	-69	-20	20	69	.971
.30	-68	-22	22	68	.988

Discussion

Results for a large number of cases have been presented. Many of the maximin r^2's are seen to be in the .80's or .90's. This is quite good for most analytic purposes. Thus a little non-metric information will go a long way when it is converted to metric information. A comparison of cases makes it clear that maximin r^2 is most readily boosted above the rank order value when the inequalities put bounds upon the external intervals of the sequence. "Extremes bunched" is a more favorable case than "extremes spread"; for n=4 and 5, r^2 is higher when an internal interval is known to be biggest than when an external interval is known to be biggest. Restrictions on the fraction of the total range allotted the biggest interval result in powerful increases in r^2; this comes about because huge external intervals are thereby prohibited. One way to summarize this class of results is to say that scales with big gaps in the middle are more "robust" than scales with big gaps in the tails.

Certain other general conclusions are apparent from the results: symmetry is a fairly powerful condition; higher-ordered metric scales can be very close indeed to numerical scales; and so on.

Nevertheless, many of you are no doubt wondering about the proof of this pudding. How often can maximin sequences actually be put to good use?

The answer is not cut-and-dried. Consider the rank order case. Here the values of maximin r^2 are only fair; moreover, the maximin sequence has an unfamiliar flavor. The end values are moved far out to guard against wild plays of Nature. Are we seriously recommending that for a rank order case with, say n=6, the contrast (-20,-6,-1,1,6,20) be used to capture the single degree of freedom associated with the rank order?

The investigator might reject the appropriateness of this contrast. He might say, "It is too bizarre. Give me straight-forward linear weights, or perhaps rankits. I do not foresee that Nature will play tricks on me". Our reply would be, "If you say your non-metric information is rank order and nothing more, then you implicitly acknowledge the possibility of a "true" sequence of the form (a,a,a,a,a,b). A conservative man would protect himself against such a possibility. If you say that this possibility is inconceivable, then you really have more non-metric information than mere rank order. If you could define this extra knowledge precisely, it would lead to another maximin sequence, one that might strike you as intuitively more reasonable .

Here lies the heart of the situation. Quite commonly, when we say we only know rank order, we actually know more than this, but don't know how to express what else it is that we know. Typically, our excess "knowledge" is to the effect that the scale is no worse than mildly curvilinear, that Nature behaves smoothly in some sense. This

is a more vague and general conception than any of the highly specific cases considered in this paper. The maximin method needs extension to this general case, the problem being to specify the inequalities and corners in some reasonable way. The same problem, seen from a different standpoint, has been apperceived by Mosteller (1958). The problem is clear; the solution is not.

The murkiness of the general "rank-order-plus-smoothness" case should not obscure the fact that in a good many practical situations the maximin approach can straightforwardly be used to good effect. Perhaps the leading candidate for a clear-cut case is the ordered metric scale for n=3. Excellent use can be made of the contrast (-7,2,5) in the situation where X_2 is known to lie between X_1 and X_3 but nearer to X_3 than to X_1. One instance of the use of this contrast is already in the literature. (Sarnoff, 1960). It is hoped that other instances of practical application will make their appearance in the near future.

REFERENCES

Abelson, R.P. & Tukey, J.W. (1958) Efficient utilization of non-numerical information in quantitative analysis: I. General theory, the case of simple rank order, and some other systems of simple irregularities. Princeton University (dittoed).

Coombs, C.H. (1950) Psychological scaling without a unit of measurement. Psychol. Rev., 57, 145-158.

Mosteller, F. (1958) The mystery of the missing corpus. Psychometrika, 23, 279-289.

Sarnoff, I. (1960) Reaction formation and cynicism. J. Personal., 28, No. 1

Siegel, S. (1956) A method for obtaining an ordered metric scale. Psychometrika, 21, 207-216.

Stevens, S.S. (1951) Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.) Handbook of experimental psychology. New York: Wiley.