Cross Validation of Multidimensional Representations by testing order hypotheses

Willem J. Heiser
and
Jacqueline Meulman

Department of Data Theory
University of Leiden
Middelstegracht 4
2312 TW Leiden
The Netherlands

Note: Slightly revised version of the paper "Cross Validation of Multidimensional representations" presented by the first author at the Workshop Table Ronde "Analyse des Donneés", january 1984, Toulouse, France.

1. Estimation subordinate to prediction

The original method of cross validation is a very simple short-cut to confirmation. The idea is to by-pass the necessity of collecting fresh data by splitting the current sample randomly into two parts: a learning sample and a testing sample. The learning sample is used for model fitting, eliminating outliers, finding suitable transformations, selection of predictors, etc. The testing sample then serves to assess the succes or failure of these efforts by evaluating their predictive power. Such a two-step procedure is still widely used in careful applications of Multiple Regression and Discriminant Analysis (but unfortunately it is mostly absent in applications of Principal Components Analysis, Multidimensional Scaling and Correspondence Analysis).

In the recent general statistical literature two refinements are frequently encountered. In the first place, rather than dividing the data into two parts, it is more commonly recommended to leave out one data point (or a small subset of points) at a time, to fit a model to the remaining points and repeatedly predict the excluded point(s). The average prediction error is then used for the assessment of the quality of the particular proposal to analyse the data. Secondly, apart from this so-called cross-validatory assessment, a procedure called cross-validatory choice has been introduced (Stone, 1974; Geisser, 1975). This involves the determination of the most predictive 'prescription' out of a well-defined class by actually minimizing the average prediction error. For example, 'shrinkers' (estimators of location in between a conventional one - such as the mean - and zero) and 'flatteners' (estimators of regression surfaces in between, e.g., the least squares one and a flat surface) are motivated by the anticipated reduction of fit of the model to fresh data.

Similar arguments were presented by Akaike (1974), who proposed to consider the average log-likelihood as a measure of fit and to choose between distinct model specifications $\pi \in \P$ by means of the rule

Cross Validation of Multidimensional Representations

1. Estimation subordinate to prediction

The original method of cross validation is a very simple short-cut to confirmation. The idea is to by-pass the necessity of collecting fresh data by splitting the current sample randomly into two parts: a learning sample and a testing sample. The learning sample is used for model fitting, eliminating outliers, finding suitable transformations, selection of predictors, etc. The testing sample then serves to assess the succes or failure of these efforts by evaluating their predictive power. Such a two-step procedure is still widely used in careful applications of Multiple Regression and Discriminant Analysis (but unfortunately it is mostly absent in applications of Principal Components Analysis, Multidimensional Scaling and Correspondence Analysis).

In the recent general statistical literature two refinements are frequently encountered. In the first place, rather than dividing the data into two parts, it is more commonly recommended to leave out one data point (or a small subset of points) at a time, to fit a model to the remaining points and repeatedly predict the excluded point(s). The average prediction error is then used for the assessment of the quality of the particular proposal to analyse the data. Secondly, apart from this so-called cross-validatory assessment, a procedure called cross-validatory choice has been introduced (Stone, 1974; Geisser, 1975). This involves the determination of the most predictive 'prescription' out of a well-defined class by actually minimizing the average prediction error. For example, 'shrinkers' (estimators of location in between a conventional one - such as the mean - and zero) and 'flatteners' (estimators of regression surfaces in between, e.g., the least squares one and a flat surface) are motivated by the anticipated reduction of fit of the model to fresh data.

Similar arguments were presented by Akaike (1974), who proposed to consider the average log-likelihood as a measure of fit and to choose between distinct model specifications $\pi \varepsilon \P$ by means of the rule

$$\min_{\pi} \{- 2 \ln L(\pi) + 2 n_{\pi}\}$$

where n_{π} is the number of independently adjustable parameters in model π and $\ln L(\pi)$ the associated maximum log-likelihood. The quantity over which the minimum is taken, called AIC, is derived from the ML estimate of the average prediction error.

Cross-validatory choice procedures have been used in curve-fitting to determine an optimal degree of complexity (Wahba and Wold, 1975) and in Principal Components Analysis for the determination of the number of components to retain (Wold, 1978; Eastment and Krzanowski, 1982). In these applications interest is apparently in global choices that are external to the optimization of a loss function. For techniques aiming at a multidimensional representation of the data (PCA, MDS, Correspondence Analysis), however, it will certainly be of interest to validate more specifically the actual characteristics of any one solution. After all, in most applications one not merely wants to make statements about dimensionality, but primarily about the relative positioning of points, clustering effects, or the salience of a priori specified contrasts among subsets of points. Therefore it may be not without value to return to the original idea of first freely finding out what seems to be the case and next testing one or more specific hypotheses directly on the set-aside data. The aim of this paper is to sketch a uniform framework for such tests and their use in cross validation.

2. Order hypotheses on proximity data

For one-dimensional solutions or other relatively simple situations (clustering, periodicity) it is often possible to derive from the model equality or proportionality conditions on the data. If the Spearman two-factor model holds, for example, the rows (columns) of the correlation matrix should all be proportional. Such patterns may readily be tested by standard methods. Another possibility, useful in the case of binary response data, is to employ the preliminary findings for the allocation of potential row profiles into two sets - the 'admissible' ones and the 'inadmissible' ones. Then tests of quasi-independence are possible within the framework

of log-linear analysis of contingency tables (Goodman, 1975; Davison, 1980).

In more complicated situations much flexibility is provided by framing hypotheses in terms of an ordering on equivalence classes of proximities. (Although this approach remains feasible for the study of proximity relations among t-tuples, attention will be restricted to the case t=2). Ordering of proximities is more useful for the present purposes than ordering along dimensions. For instance, while there are an embarrassing number (2^{n-1}) of ways to order n objects organized in a hierarchical tree so that its branches do not cross, there is only one 'préordonnance ultramétrique' (Lerman, 1970) - i.e., only one particular partitioning of the set of objects pairs into n ordered subsets from close to distant. Likewise most solutions of multidimensional procedures proper, while being subject to well-known indeterminacies such as reflections and rotations, do give unequivocal distance information.

Thus the idea is to translate precisely those characteristics of the representation that are both invariant and interpretable into predictions of (simple functions of) observables. The choice of an order hypothesis rather than some parametric family of prediction functions serves to maintain comparability between metric and nonmetric methods. A model that has been constructed with a metric method should a fortiori predict the ordinal information right - provided, of course, that there has been no under- or overfitting. Finally note that it no longer matters very much how many degrees of freedom or number of parameters the previously fitted models have; they all become equivalent to an a priori idea with about the same degree of freedom, as far as they enable the specification of the same number of equivalence classes. (This is not precisely true, as will become apparent below).

3. Isotonic regression

The ordered equivalence classes that make up the prediction are most easily displayed in a design matrix like Table 1. The classes

== INSERT TABLE 1 ABOUT HERE ==

are numbered from 1 to 9 (generally, with index k=1,...,K) and all pairs i,j with an equal value of k form a class of equal proximity. In figure 1, the multidimensional representation is

== INSERT FIGURE 1 ABOUT HERE ==

shown from which the classes were derived. (This example will be discussed in section 5). In order to set the problem in the context of the available statistical theory (the main reference is Barlow, Bartholomew, Bremner and Brunk, 1972), it is convenient to bring it into the format of a one-way classification. This is done in Table 2, which lists the proximity classes again, from

== INSERT TABLE 2 ABOUT HERE ==

smallest to largest, recording the corresponding subscripts of the cells of Table 1. The fresh data can be grouped in this fashion and then interpreted as independent identically distributed observations y_{ik} $(i=1,\ldots,n_k)$ of the random variable y, dissimilarity, with means μ_k and variances σ_k^2. The objective is to test the null hypothesis

$$H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_K$$

against the alternative hypothesis

$$H_1: \quad \mu_1 \leq \mu_2 \leq \ldots \leq \mu_K$$

with not all μs equal. Frequently a test of the homogeneity of the variances σ_k^2 would be in order too, but here it will simply be assumed that they are equal ($\sigma_k^2 = \sigma^2$ for all k). Under H_0, it is well-known that the maximum likelihood estimates of the means are given by $\hat{\mu}_k = \bar{y} = \hat{\mu}$ for all k, where

$$\bar{y} = \sum_k^K n_k \bar{y}_k \; / \sum_k^K n_k$$

and $\bar{y}_k$ are the column means. On the other hand, under H_1 it can be

shown that the maximum likelihood estimates are $\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_K$, where $\tilde{\mu}_k$ is the isotonic regression of $\bar{y}_k$, with weights n_k, with respect to the given partial ordering.

Actually, the isotonic regression is the minimizer of

$$\sum_k^K n_k (f_k - \bar{y}_k)^2$$

over all f_k for which $k \leq l$ implies $f_k \leq f_l$. There are several algorithms for the efficient minimization of this function (cf. Barlow et al., 1972). Although it is true that finding the isotonic regression is also one of the crucial steps in a nonmetric scaling procedure, it should be noted that what is done here is regressing the data on the 'model values', whereas in nonmetric scaling the model values are regressed on the data. This reversal is a natural consequence of the present purpose of testing rather than mapping the data.

Assuming normality, the likelihood function is (N = total number of observations):

$$(2\pi\sigma^2)^{-N/2} \quad \exp\{-\frac{1}{2\sigma^2} \sum_k^K \sum_i^{n_k} (y_{ik} - \mu_k)^2\}$$

and a likelihood ratio test can be devised that rejects H_0 for large values of the statistic

$$\bar{E}_K^2 = \sum_k^K n_k (\tilde{\mu}_k - \hat{\mu})^2 / \sum_k^K \sum_i^{n_k} (y_{ik} - \hat{\mu})^2$$

or, in words, the ratio of the between groups variance under monotonicity and the total variance. The null hypothesis distribution of $\bar{E}_K^2$ turns out to be (cf. Barlow et al., 1972, theorem 3.2) a weighted sum of Beta densities with parameters $\frac{1}{2}(l-1)$ and $\frac{1}{2}(N-l)$. Here l is the number of levels in the isotonic regression function and the weights $P(l,K)$ are probabilities of getting l levels when regressing K means (under H_0).

For the case of an equal number of observations in each group, $P(l,K)$ is easily computed; but for unequal n_k their computation is more problematical. The appearance of l, a number that is only

known after the computations, also explains the earlier hesitance to state that hypotheses in terms of an equal number of classes have exactly the same degree of freedom. In fact, the exact number of free parameters is ℓ. The AIC statistic will also be useful in comparing various hypotheses; note that it simply becomes (up to an additive constant):

$$AIC = N \ln \sum_{k}^{K} \sum_{i}^{n_k} (y_{ik} - \tilde{\mu}_k)^2 + 2\ell$$

or, essentially, the pooled within groups variance under monotonicity, corrected for the number of levels in the regression function.

4. Ramsay data

Ramsay (1968) reported an experiment on color perception which will serve as the first example. In the experiment, color patches were selected from the Munsell color system in such a way that their structural relationships resemble a capital Q. There were 21 colors, and 20 subjects gave ratio judgments of dissimilarity with respect to a standard pair.

To illustrate the ideas of the previous sections, the subjects were randomly split into two groups of 10. The square symmetric dissimilarity matrices of the first group were employed for in total 6 analyses, and the two-dimensional solutions for 4 of these are

== INSERT FIGURE 2 ABOUT HERE ==

displayed in figure 2. These 4 are all more or less 'blind' approaches: the data are averaged across subjects and then mapped into two-space by the standard programs ANACOR, SMACOF2(classical) SMACOF2(numer.) and SMACOF2(ordinal). For the Correspondence Analysis, the data were first transformed with a negative exponential before averaging.

It is clear that the solutions, matched in location and orientation, are all very similar; but especially Q and N (the ordinal and the numerical options of SMACOF2) are very close. A (ANACOR) and C (classical MDS à la Torgerson) differ most; they optimize the same type of loss function, but with a different weighting scheme

(cf. Heiser and Meulman, 1983). The results of the cross validation on the second group are reported in Table 3 in terms of $\bar{E}_K^2$ and AIC.

== INSERT TABLE 3 ABOUT HERE ==

The values of $\bar{E}_K^2$ are all very significant. The ordinal and numerical solutions are best (ex aequo) by both measures. It is interesting to note that this example sustains the idea that it is not so much ordinality which makes MDS programs like SMACOF2 good procedures for modeling proximities, but their Least Squares loss function defined directly on the distances. Also note that only if the hypotheses are very similar (cf. SMACOF2(ordinal) and SMACOF2(numerical)), the correction implicit in AIC is decisive.

The two other analyses on the learning sample were constrained MDS runs, one restricting the points to form a (not necessarily equally spaced) Q, the other using the dominant wavelength of the color patches as a directional constraint. The latter is psychologically unwise, and indeed such a solution turns out to be inferior, though still significantly better than a random hypothesis (cf. Table 4.). Both the Q-restricted SMACOF2 solution and the

== INSERT TABLE 4 ABOUT HERE ==

completely a priori prediction of the Munsell system do remarkably well, almost as good as the unconstrained SMACOF2 solutions.

The type of scatter obtained and the isotonic regression function are illustrated in figures 3 and 4, for the best representation

== INSERT FIGURES 4 & 5 ABOUT HERE ==

and the worst one, respectively. It might be possible to reduce the within scatter by rescaling the individual data, rather than giving them merely identical overall sums of squares, as was done here. This opens lots of room for refinements and will not be discussed in this paper.

5. Fisher data

The second example uses data from Fisher (1940) concerning the cross-classification of 5387 individuals with respect to eye color (points 1,...,4 in figure 1 and Table 1) and hair color (points 5,...,9). The following procedure was employed. A randomly chosen subset of 2693 individuals was used as the learning sample. Three analyses were performed, a Correspondence Analysis (see figure 1), a numerical, and an ordinal MDS (see figure 3). In all

== INSERT FIGURE 3 ABOUT HERE ==

cases, an attempt is made to approximate the χ^2-distances among the nine categories in a space as small as possible. ANACOR essentially gives three clusters, numerical SMACOF2 a much more scattered solution and the ordinal SMACOF2 a different type of clustering.

Next, from the remaining data 6 groups of 449 individuals were formed, and 6 tables of χ^2-distances were computed in order to have a reasonable number of replications in each postulated proximity group (K=9 in all cases). Results are given in Table 5, and it is clear that the ANACOR configuration gives a poor prediction of the χ^2 distances. This is also dramatically shown in the scatter

== INSERT TABLE 5 ABOUT HERE ==
== INSERT FIGURES 6,7,8 ABOUT HERE ==

plots of observations against hypothesized proximity classes.

6. Discussion

This study has followed a suggestion made by Heiser and Meulman (1983b) for reaching confirmatory evidence in geometric modelling without reliance on simultaneous parameter estimation. In a similar vein, cross-validation of covariance structures has recently been advocated by Cudeck and Browne (1983). To reiterate, the idea is to view models merely as reflections of or approximations to reality, and - as Cudeck and Browne expressed it: "When a model is only an

approximation to reality, however, and when a test of fit is carried out, statistical power theory virtually guarantees that the null hypothesis will be rejected if n is sufficiently large" (o.c., p. 149-150). Rather than trying to achieve optimality in a single sample of data, we should be concerned with identifying models which will perform optimally in future samples.

A whole range of classical hypothesis testing methods is available for studying the testing sample. Our choice of order hypotheses on grouped proximity values goes beyond an ordinary analysis of variance, in order to obtain more power. Yet it is still rather "omnibus", flexible and rough, and could be improved upon when experience with the model accumulates.

Critical values for the $\bar{E}_K^2$ test have been tabulated by Nelson (1977) for a limited number of values of K and an equal number of observations in each group. Generally, the computation of mixtures of Beta distributions is quite involved (for an algorithm, see Bremner, 1978) and simpler significance tests for (incomplete) predictions of order do exist, such as the "outcome counting" test of Green and Nimmo-Smith (1982), based on the rationale of Chassan's (1962) test. For a review, see Smith and Macdonald (1983), who also performed power simulations for $\bar{E}_K^2$ and some alternative tests (Macdonald and Smith, 1983).

There is no need in cross-validation to split the available sample in subsamples of equal size. From a statistical point of view, one would presumably prefer to minimize the size of the learning sample. It seems unfortunate that in many applications of Multidimensional Scaling and related methods it is the size of the testing sample that is reduced to zero.

7. References

Akaike, H. A new look at the statistical model identification. IEEE Transactions Aut.Contr., 1974, AC-19, 716-723.

Barlow, R.E. Bartholomew, D.J. Bremner, J.M. & Brunk, H.D. Statistical Inference under Order Restrictions. New York: Wiley, 1972.

Bremner, J.M. Algorithm AS 123. Mixtures of beta distributions. Applied Statistics, 1978, 27, 104-109.

Chassan, J.B. An extension of a test for order. Biometrics, 1962,

18, 245-247.
Cudeck, R. and Browne, M.W. Cross-validation of covariance structures. Multivariate Behavioral Research, 1983, 18, 147-167.
Davison, M.L. A psychological scaling model for testing order hypotheses. Brit.J.Math.Stat.Psych., 1980, 33, 123-141.
Eastment, H.T. and Krzanowski, W.J. Cross-validatory choice of the number of components from a principal component analysis. Technometrics, 1982, 24, 73-77.
Fisher, R.A. The precision of discriminant functions. Ann. Eugenics, 1940, 10, 422-429.
Geisser, S. The predictive sample reuse method with applications. J.Amer.Stat.Ass., 1975, 70, 320-328.
Goodman. L.A. A new model for scaling response patterns: An application of the quasi-independence concept. J.Amer.Stat.Ass., 1975, 70, 755-768.
Green, T.R.G. and Nimmo-Smith, I. 'Outcome-counting'- Significance tests from incomplete predictions of order. Brit.J.Psych., 1982, 73, 41-49.
Heiser, W.J. and Meulman, J. Analyzing rectangular tables by joint and constrained Multidimensional Scaling. J.Econometrics, 1983, 22, 139-167. (a)
Heiser, W.J. and Meulman, J. Constrained Multidimensional Scaling, including Confirmation. Appl.Psych.Meas., 1983, 7, 381-404. (b)
Lerman, J.C. Les bases de la classification automatique. Paris: Gauthiers Villars, 1970.
Macdonald, R.R. and Smith, P.T. Testing for differences between means with ordered hypotheses. Brit.J.Math.Stat.Psych., 1983, 36, 22-35.
Meulman, J. Correspondence Analysis and Stability. Internal report RR-84-01, Department of Data Theory, Leiden, 1984.
Nelson, L.S. Tables for testing ordered alternatives in an analysis of variance. Biometrika, 1977, 64, 335-338.
Ramsay, J.O. Economical method of analyzing perceived color differences. J.Opt.Soc.Amer., 1968, 58, 19-22.
Smith, P.T. and Macdonald, R.R. Methods for incorporating ordinal information into analysis of variance: Generalizations of one-tail tests. Brit.J.Math.Stat.Psych., 1983, 36, 1-21.
Stone, M. Cross-validatory choice and assessment of statistical predictions. J.Roy.Stat.Soc., 1974, 36 Ser.B, 111-148.
Wahba, G. and Wold, S. A completely automatic French curve: fitting spline functions by cross-validation. Comm.Statist., 1975, 4, 1-17.
Wold, S. Cross-validatory estimation of the number of components in Factor and Principal Component models. Technometrics, 1978, 20, 397-405.

Table 1 Order of distances from ANACOR solution for randomly halved Fisher data.

0	1	5	6	1	1	5	5	8
1	0	5	6	1	1	4	5	8
5	5	0	6	6	4	1	5	8
6	6	6	0	7	5	6	1	3
1	1	6	7	0	2	5	6	9
1	1	4	5	2	0	4	5	8
5	4	1	6	5	4	0	5	8
5	5	5	1	6	5	5	0	4
8	8	8	3	9	8	8	4	0

Table 2. Order hypothesis of Table 1 brought into the format of a one-way classification.

1	2	3	4	5	6	7	8	9
d(1,2)	d(5,6)	d(4,9)	d(2,7)	d(1,3)	d(1,4)	d(4,5)	d(1,9)	d(5,9)
d(1,5)			d(3,6)	d(1,7)	d(2,4)		d(2,9)	
d(1,6)			d(6,7)	d(1,8)	d(3,4)		d(3,9)	
d(2,5)			d(8,9)	d(2,3)	d(3,5)		d(6,9)	
d(2,6)				d(2,8)	d(4,7)		d(7,9)	
d(3,7)				d(3,8)	d(5,8)			
d(4,8)				d(4,6)				
				d(5,7)				
				d(6,8)				
				d(7,8)				

Table 3. Ramsay data: $\bar{E}^2_K$ and AIC for four different analyses.

ANACOR	.617	-4262.33
SMACOF2(classical)	.606	-4210.68
SMACOF2(numerical)	.641	-4404.64
SMACOF2(ordinal)	.641	-4400.94

Table 4. Ramsay data: $\bar{E}^2_K$ and AIC for constraints and hypotheses.

SMACOF2(Q-constr.)	.636	-4371.07
SMACOF2(Wavelength)	.471	-3572.32
MUNSELL	.633	-4343.25
RANDOM	.013	-2250.92

Table 5. Fisher data: $\bar{E}^2_K$ and AIC for three different analyses.

ANACOR	.588	-371.06
SMACOF2(numerical)	.893	-651.52
SMACOF2(ordinal)	.962	-880.77

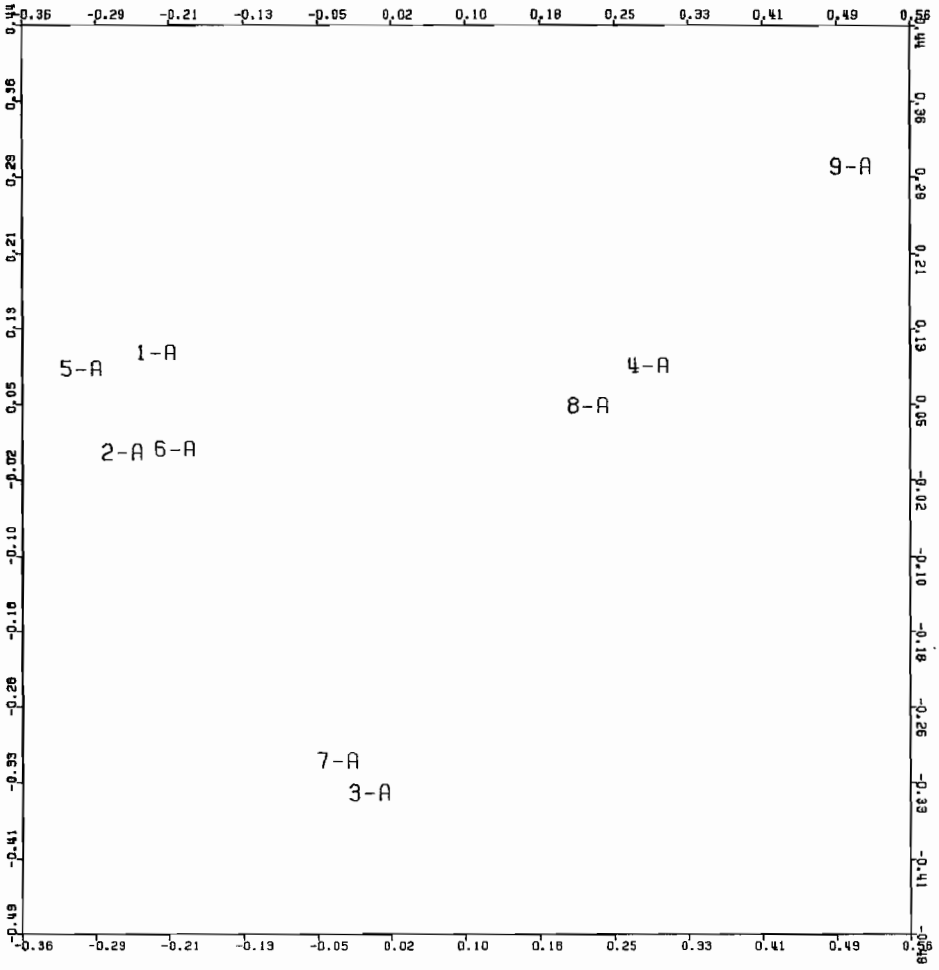

figure 1. ANACOR solution for randomly halved Fisher data.

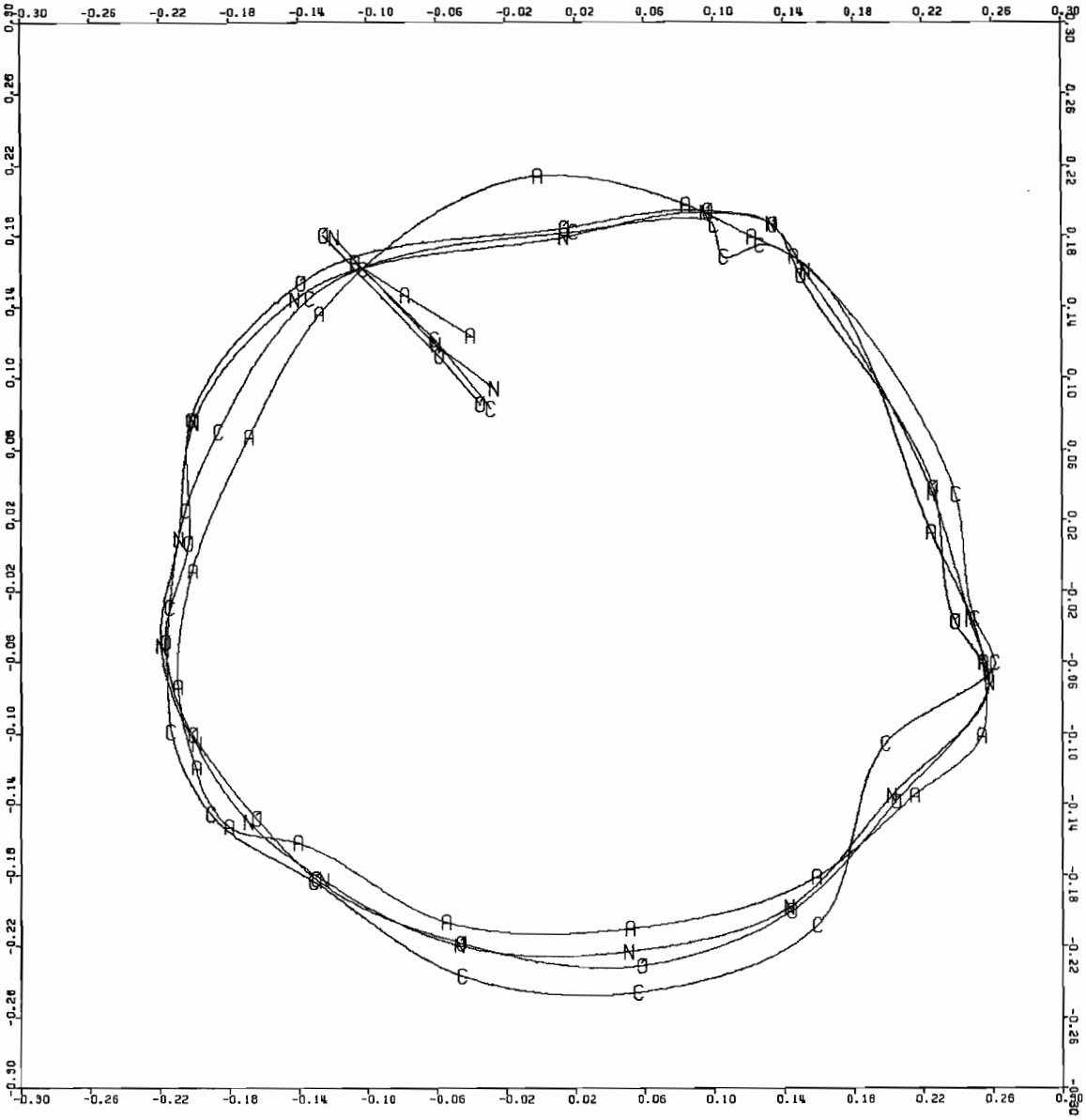

figure 2. Four solutions for Ramsay's data, matched.

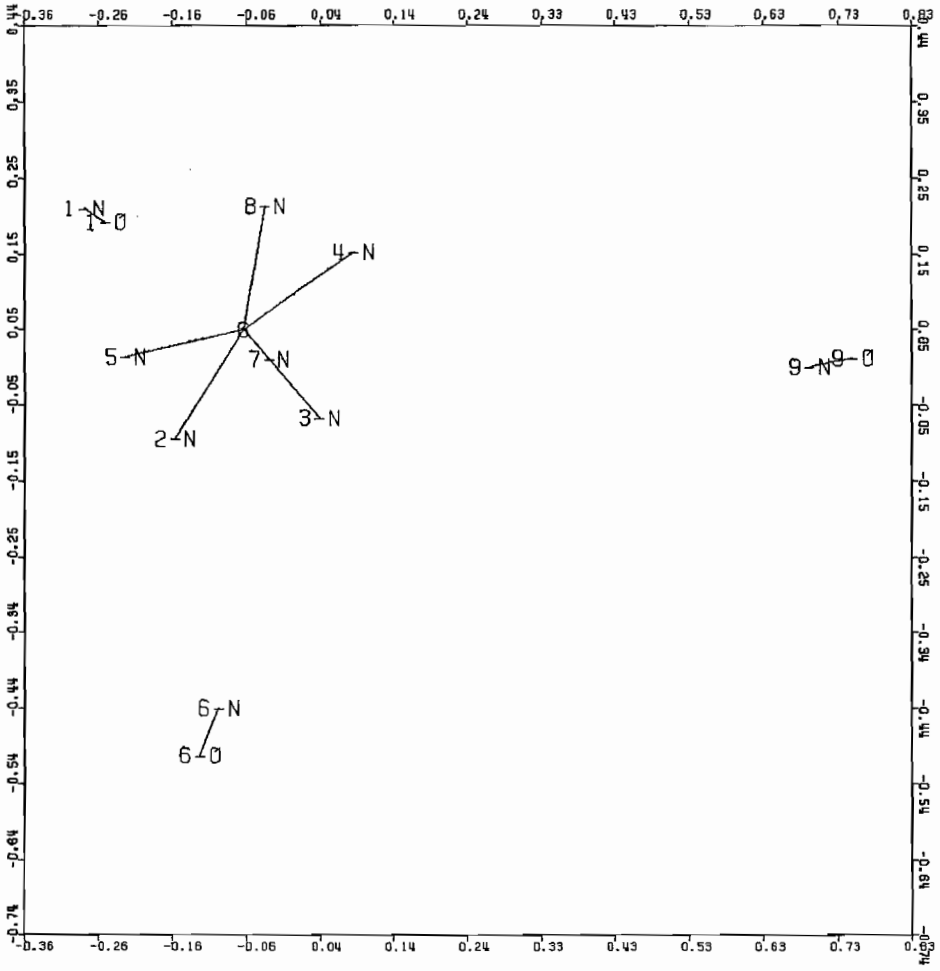

figure 3. Ordinal and numerical SMACOF2 solutions for the Fisher data.

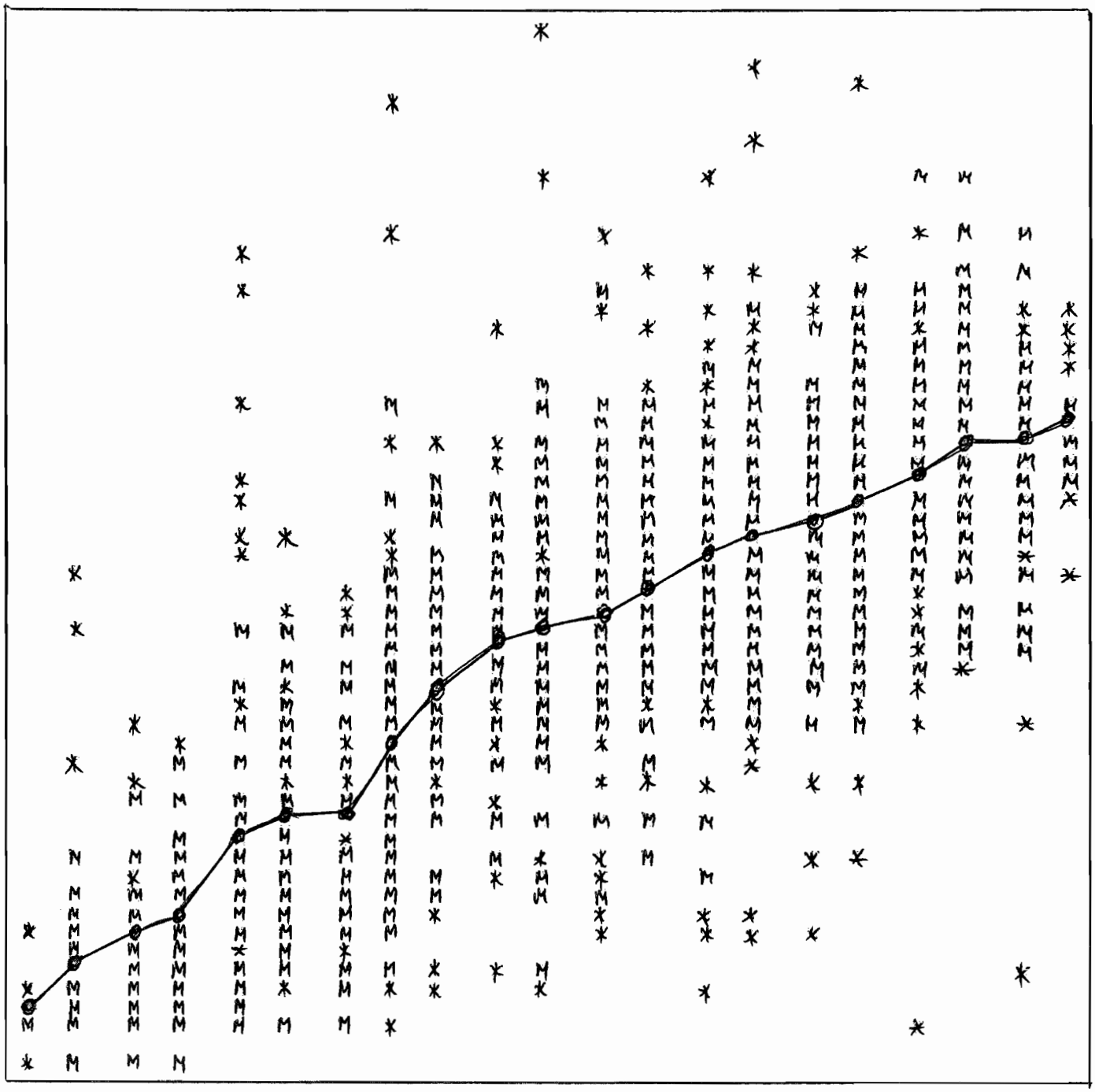

Figure 4. Scatter SMACOF2 numerical, Ramsay data.

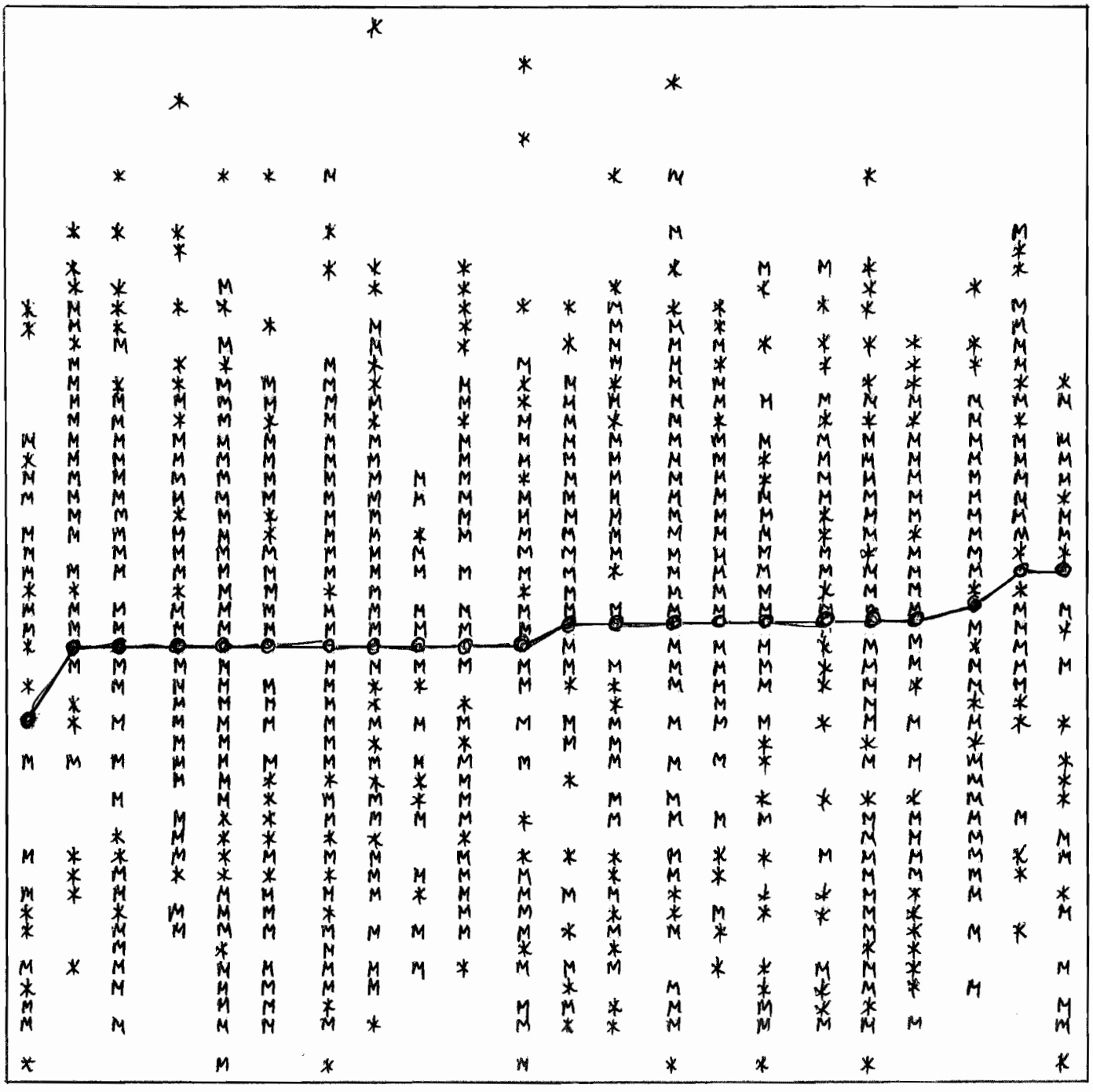

Figure 5. Scatter from random hypothesis, Ramsay data.

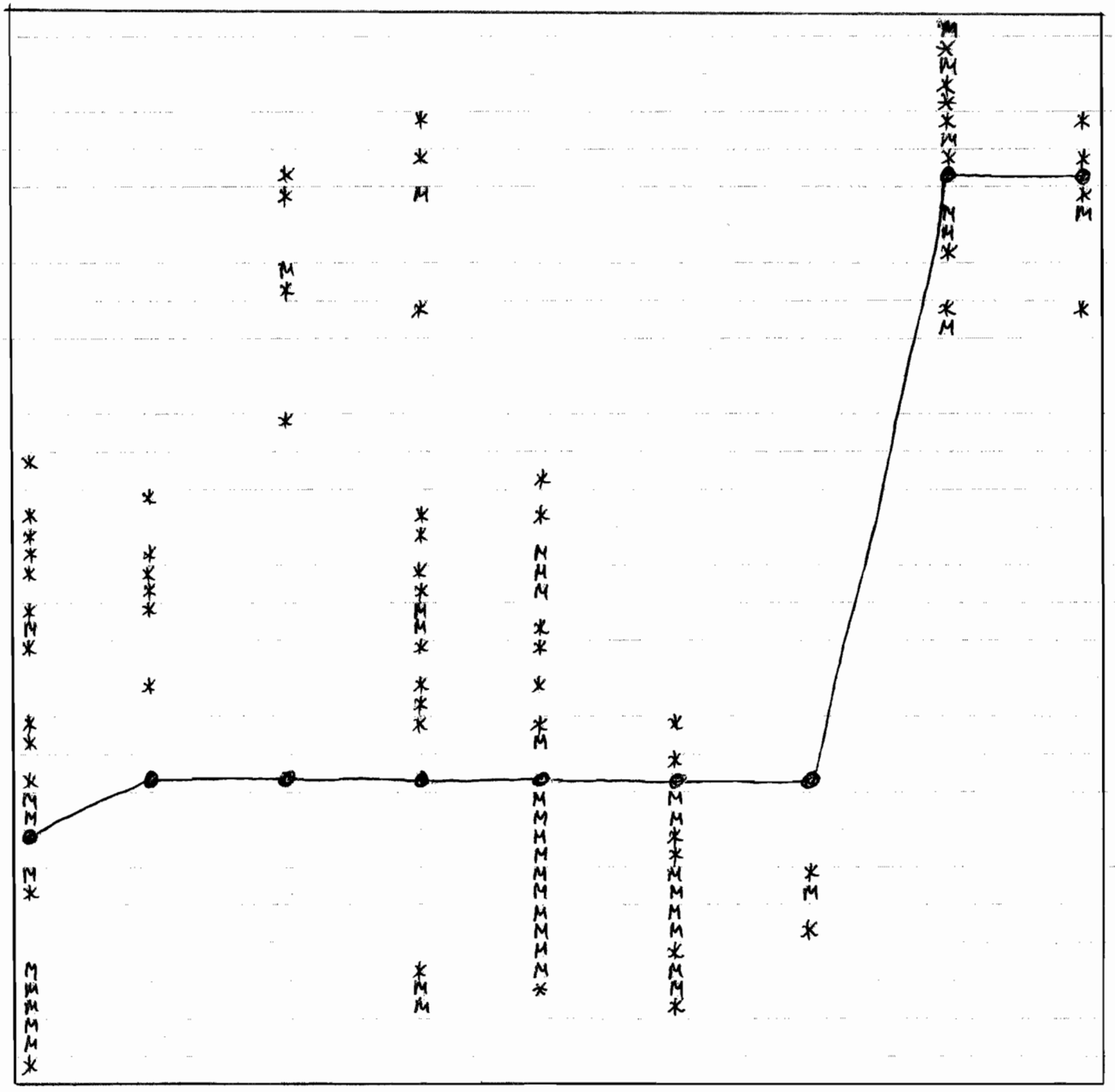

Figure 6. Scatter Fisher data, ANACOR.

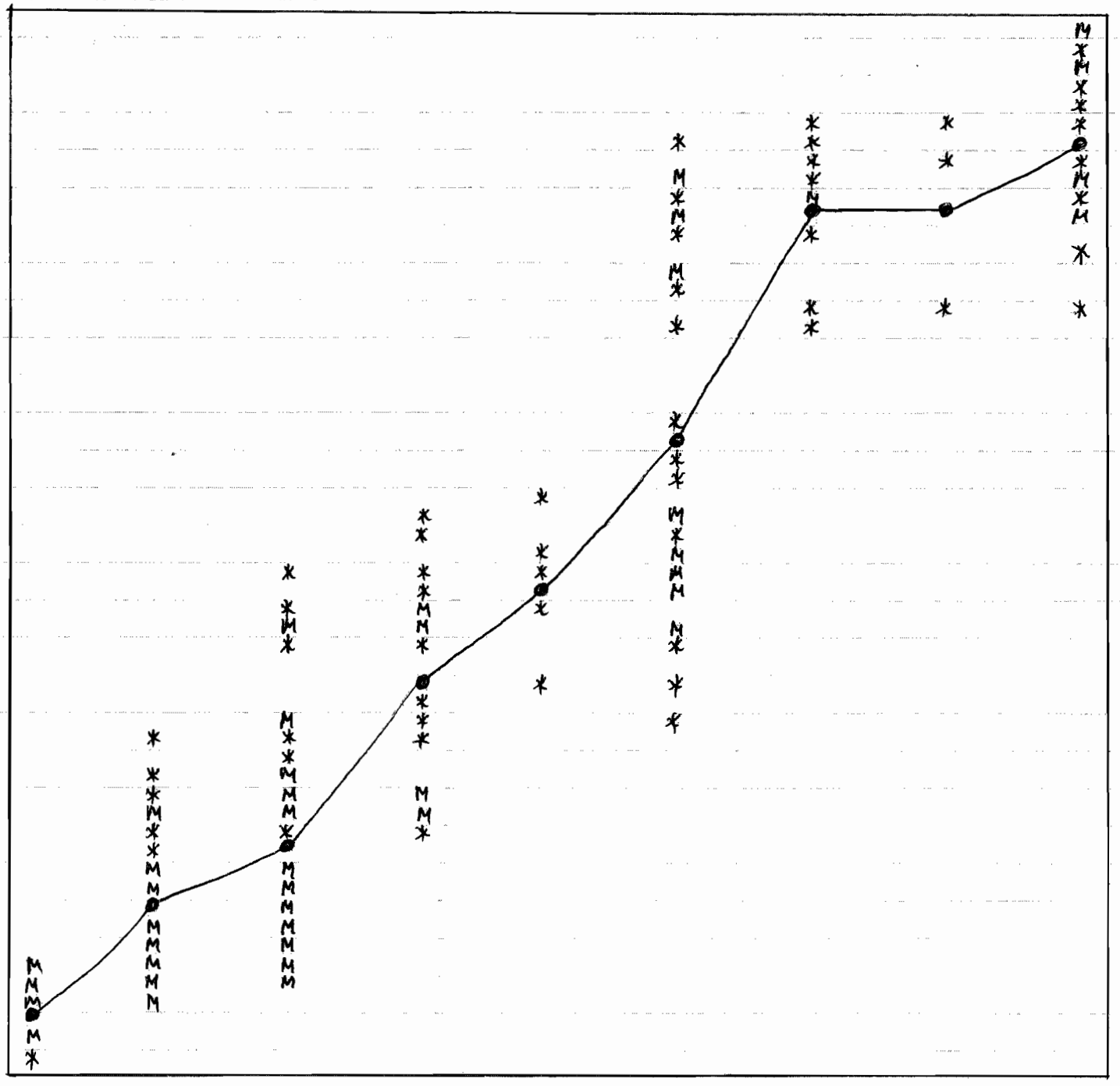

Figure 7. Scatter Fisher data, numerical SMACOF2.

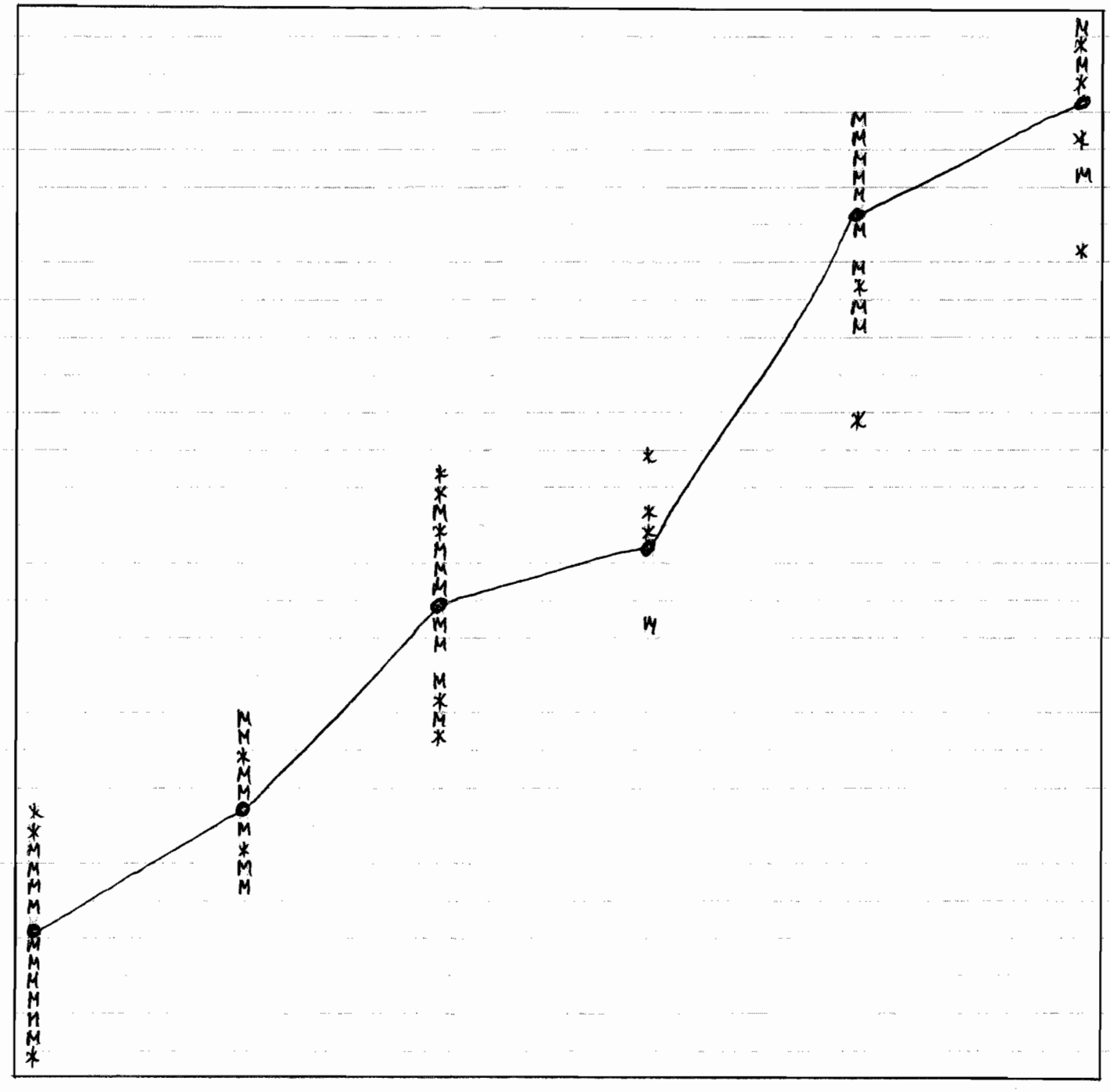

Figure 8. Scatter Fisher data, ordinal SMACOF2.