Novel Non-Metric MDS Algorithm with Confidence Level Test [1]

Y-h. Taguchi[2]
Department of Physics,
Chuo University, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

Y. Oono
Loomis Laboratory of Physics,
University of Illinois at Urbana-Champaign, 1110 W Green, Urbana, IL 61801, USA.

Abstract

A novel algorithm for non-metric multidimensional scaling (NMDS) method is proposed that is closely related to a means to evaluate the statistical confidence level of the resultant configurations. Such a statistical feature was not naturally associated with NMDS or MDS before. Our algorithm is so efficient that the relations among 1000 items can easily be analyzed with an inexpensive personal computer. The algorithm is illustrated with its application to, e.g., DNA sequence data.

Key Words: non-metric multidimensional scaling, confidence level test

[1] Submitted to *Annals of the Institute of Statistical Mathematics*

[2] e-mail:tag@granular.com

1 Introduction

Increasing importance of bioinformatics certainly requires efficient methods to handle very large complicated data sets to extract patterns and rules. To this end in mind in this paper we report a new efficient algorithm for non-metric multidimensional scaling (NMDS). Multidimensional scaling (MDS) (see, e.g., Borg and Groenen (1997)) is a major branch of multivariate analysis chiefly used in social sciences, psychology, and occasionally in ecology to visualize hidden relations among objects of interest (henceforth we call them operational taxonomic units (OTU)). Its essence is to find, in a certain metric space, a configuration (constellation) of the points corresponding to OTUs that is compatible with the given dissimilarity relations among them. The resultant configuration often visually exhibits the relations/structures hidden in the original data.

The metric MDS attempts to construct a configuration of OTUs in which the dissimilarity measure $\delta(i,j)$ for each OTU pair (i,j) is proportional to the distance $d(i,j)$ in the constructed configuration. The NMDS (Shepard and Kruskal 1964) attempts to construct a configuration of OTUs whose distances $d(i,j)$ preserve the rank ordering of the dissimilarities in the original data. That is, for any two pairs of OTUs, (i,j) and (k,l), the constructed configuration satisfies $d(i,j) > d(k,l)$ if $\delta(i,j) > \delta(k,l)$. NMDS can be applied to the rank data of dissimilarities (that is, qualitative data), and can often recover the metric MDS results when there are sufficiently many OTUs. Therefore, NMDS is one of the most versatile multivariate analysis methods. Usually, NMDS can deal with much more robust data than MDS can handle. Thus, in this paper, we construct a "pure" NMDS, maximally eliminating metric elements (even the so-called $\hat{d}$) in contrast to the conventional NMDS.

Our novel algorithm is closely related to the optimization of the measure of statistical confidence level of the resultant configuration. Besides, the algorithm is very efficient so that it even enables us to analyze $N = 10^4$ OTUs (this is <u>not</u> the number of binary relations that is a much larger number $\sim N^2/2$) with an inexpensive desktop machine. A 1000 OTU example is illustrated in this paper. Since our algorithm is of order $N^2 \ln N$ for updating the configuration (that is, essentially the number of the relations we handle, so greater efficiency is hardly possible), it is clear that much larger number of OTUs can be practically analyzed with NMDS.

The paper is organized as follows. In Section 2, the conventional NMDS methods are reviewed to show the presence of additional and unnecessary constraints. Section 3 discusses what "pure NMDS" should be, and our new NMDS algorithm is proposed in

Section 4. Its capability of handling a large sized data is demonstrated in Section 5. A method to measure the statistical confidence level of the resultant configurations is proposed in Section 6. The dimensionality of the embedding space is discussed in Section 7 based on this statistical confidence level. Finally, we exhibit an application of NMDS to a molecular biological example in Section 8. Section 9 is a summary.

2 Conventional NMDS

Suppose an (increasing) ordering of the pair dissimilarities of N points (OTUs) $\{1, 2, \cdots, N\}$ is given. That is, for any two pairs of points (i, j) and (k, l) such that $i, j, k, l \in \{1, 2, \cdots, N\}$ we assume that we can tell which dissimilarity of the two pairs is not smaller than the other. A typical NMDS problem may be summarized as follows: In a given metric space $\mathcal{R}$ find a configuration of N points the (increasing) order of whose pair distances is as close as possible to that of the given pair dissimilarities

All the conventional NMDS methods assume a certain intermediate pair distance (e.g., $\hat{d}$ below) that is compatible with the actual ordering of the pair dissimilarities and is simultaneously as compatible as possible with the pair distances d in the estimated configuration of the OTUs. There are many varieties of the conventional approach.

Kruskal's approach may be summarized as follows. We start with a set of pair distance values d that are computed from a trial N-point configuration in $\mathcal{R}$. The greatest convex minorant (that is, the lower boundary of the convex hull) of the plot of d against the ranking of the distances in the original data defines $\hat{d}$. d is then revised to minimize the stress: a suitably normalized discrepancy between d and $\hat{d}$. After this, the whole procedure is iterated until the stress becomes tolerably small. There are several implementations of this idea, but we do not discuss this approach any further (See, for example, Cox and Cox (1994) and Borg and Groenen (1997)).

Guttmann (Cox and Cox 1994) constructed d^*, which is the rearrangement of d according to the actual ordering of the given dissimilarities δ. Then, the stress $S_G = \|d - d^*\|$ is minimized for d. He showed that this procedure and Kruskal's procedure are equivalent. In the actual calculation, d^* is not directly considered as a functional of d, but is treated as given just as $\hat{d}$ in each updating process. Then, the procedure is iterative as in the methods in the preceding paragraph.

As has been seen above, the essence of the conventional NMDS is to choose $\hat{d}$ as close to d as possible under the condition that it is monotone with respect to the actual ranking of the dissimilarities in the original data. Depending on the interpretation of "as close as," different methods have been proposed as summarized on p43 of Green *et al.* (1970). KYST is the program based on the least square fit. Spline function fitting, etc., are conceivable and actually used. On p52 - 58 of Green *et al.* (1970), the letter 'R' is presented with about 30 representative points, and NMDS is performed with various algorithms. For example, the results of MDSCAL and KYST are different (KYST does not affinely skew 'R' as MDSCAL does). That is, the choice of $\hat{d}$ affects the outcome.

Thus it is clear that the conventional NMDS introduces extra constraints. If we wish to be faithful to the basic idea of NMDS according to its creator (Shepherd), it is an inevitable conclusion that $\hat{d}$ is required only by technical reasons for implementation. The input data is the ranking of δ. Then, what we must compare with it is the ranking of d. If the data set is purely without metric, but only with rank ordering, then the actual size of d should not matter (should not affect the outcome[3]). We must note that all the procedures to determine $\hat{d}$ (or d^*) so far proposed are more or less affected by the actual values (or their ratios) of d.

3 Pure NMDS

To implement NMDS in a way faithful to the idea of its creator, we must not introduce any extra information, bias, structure, etc., into the given information: the ranking of dissimilarities between pairs. We do not have any information about the actual pair distances. We wish to look for configurations of N points in $\mathcal{R}$ whose pair distances have the ranking maximally compatible with the given dissimilarity ordering.

Logically, there can be two approaches:
(A) We compare the distances d obtained from the configuration in $\mathcal{R}$ and the given ordering of the pair dissimilarities δ through their ranks only.
(B) We compare the distances d obtained from the configuration in $\mathcal{R}$ and all the possible distances that are compatible with the pair dissimilarity ranking.

The method (B) was recently proposed and implemented by Trosset (1998) .

[3] if it does, it implies that the result of the NMDS is not unique.

The approach (B) may be interpreted as an unbiased encoding of the rank ordering into the actual distances. In order not to bias the information due to encoding schemes , (B) takes into account of the totality of possible encoding schemes consistent with the rank ordering in the original data. Thus, logically (A) and (B) are equivalent, but since encoding into distances is an extra step, (B) is conceptually and practically less direct than (A).

4 New NMDS Algorithm

In the following, we propose a new algorithm for NMDS that realizes approach (A).

The basic idea of the algorithm is as follows: in a metric space $\mathcal{R}$ (in this paper, for simplicity, we assume this to be a D-dimensional Euclidean space $\mathcal{R} = \boldsymbol{R}^D$) N points representing the OTUs are placed as an initial configuration. For this initial trial configuration we compute the pair distances $d(i,j)$, and then rank them according to their magnitudes. Comparing this ranking and that according to the dissimilarity $\delta(i,j)$, we compute the 'force' that moves the points in $\mathcal{R}$ to reduce the discrepancies between these two rankings. The 'force' along the line connecting OTUs i and j is chosen to decrease (resp., increase) $d(i,j)$, if the rank of $d(i,j)$ is larger (resp., smaller) than that of $\delta(i,j)$. After moving the points according to the 'forces', the new 'forces' are computed again, and the whole adjusting process of the OTUs in $\mathcal{R}$ is iterated until the positions of OTUs converge sufficiently.

In this 'overdamped dynamics' the point configuration is driven by the potential energy

$$\Delta \equiv \sum (T_n - n)^2, \tag{4.1}$$

where T_n is the actual rank of the distance between the pair in $\mathcal{R}$ whose dissimilarity has the true rank n in the original data. $\Delta = 0$ is the ideal case. This Δ may be regarded as a counterpart of the stress in the conventional NMDS. As we will see in the next section, we can use quantities related to Δ to evaluate the confidence level of the resultant configuration statistically. Now, we can set up a null hypothesis to reject at a given confidence level. Furthermore, we can even discuss the plausibility of the substructures of the resultant configuration.

Thus an important feature of our NMDS algorithm is that the optimization process is directly connected to a process that improves the confidence level of the resultant configuration.

In order to describe our idea unambiguously, we describe it in a preliminary algorithm as follows:

1. The rank ordering of dissimilarities $\delta_{ij}, (i, j = 1, ..., N)$ for N OTUs are given.
2. Put N points randomly in a metric space $\mathcal{R}$ (here, we use D-dimensional Euclidean space for simplicity) as an initial configuration. Let their position vectors be $\boldsymbol{r}_i$.
3. Scale the position vectors in $\mathcal{R}$ as $\sqrt{\sum_i |\boldsymbol{r}_i|^2} = N$.
4. Compute $d_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ for all OTU pairs i and j in $\mathcal{R}$.
5. Set the mismatch counters $C_{ij} = 0$ for all pairs (i, j).
6. For all i, j, k, l, compare δ_{ij} with δ_{kl}, and d_{ij} with d_{kl}. If $d_{ij} > d_{kl}$ and $\delta_{ij} < \delta_{kl}$ (resp., $d_{ij} < d_{kl}$ and $\delta_{ij} > \delta_{kl}$), add -1 (resp., 1) to counter C_{ij} and 1 (resp., -1) to C_{kl}. Otherwise, the mismatch counters are intact.
7. For OTU i, if C_{ij} is positive (resp., negative), the point corresponding to it in $\mathcal{R}$ is moved toward (resp., away from) the one corresponding to OTU j by the amount $s|C_{ij}|$, where s is a small positive constant. This is actually performed at once for each i through calculating the following displacement vector for i:

   $$s \sum_j C_{ij} \frac{\boldsymbol{r}_i - \boldsymbol{r}_j}{|\boldsymbol{r}_i - \boldsymbol{r}_j|},$$

   where $\boldsymbol{r}_i$ is the current position of OTU i in $\mathcal{R}$.
8. Return to 3, and continue until the "potential energy" becomes sufficiently small.
9. To check convergence, we use the statistical confidence level based on a quantity related to Δ in (4.1) (see the next section).

In the above algorithm, we can deal with asymmetric data as well, i.e., $\delta_{ij} \neq \delta_{ji}$ if we compare δ_{ij} with d_{ij} while δ_{ji} with $d_{ji}(= d_{ij})$. Needless to say, if the mismatch between δ_{ij} and δ_{ji} is large, then representing the pair by a pair of points in a metric space is questionable. Therefore, we will not discuss this problem any further in this paper.

One may notice that this preliminary algorithm is time consuming; the total computing time grows as N^4 for one updating step for the configuration, where N is the number of OTUs. However, we can considerably reduce this with a modification that streamlines the force calculation step. This leads to the "pure NMDS Algorithm" we propose.

Algorithm "Pure NMDS"

1. Dissimilarities δ_{ij} $(i, j = 1, \cdots, N)$ for N OTUs are given. Order them as follows:
$$\cdots \leq \delta_{ij} \leq \delta_{kj} \leq \cdots. \tag{4.2}$$
2. Put N points randomly in $\mathcal{R}$ as an initial configuration.
3. Scale the position vectors in $\mathcal{R}$ as $\sqrt{\sum_i |\boldsymbol{r}_i|^2} = N$.
4. Compute d_{ij} for all OTU pairs (i, j) in $\mathcal{R}$, and then order them as
$$\cdots \leq d_{ij} \leq d_{kj} \leq \cdots. \tag{4.3}$$
5. Suppose δ_{ij} is the mth largest in the ordering in (4.2) and d_{ij} is the nth largest in the ordering (4.3). Assign $C_{ij} = n - m$. Remaining procedure is the same as preliminary Algorithm.

In the above algorithm (and throughout this paper) we fix s. In practice, we could choose an appropriate schedule to vary s as is often done in optimization processes. In this paper, for simplicity, we will not attempt such a fine tuning.

Consider an example with $N = 4$:
$$\delta_{12} < \delta_{34} < \delta_{23} < \delta_{13} < \delta_{14} < \delta_{24}.$$
Suppose d_{ij} orders as follows at a certain step of the calculation:
$$d_{12} < d_{23} < d_{34} < d_{13} < d_{14} < d_{24},$$
i.e., the ordering of d_{23} and d_{34} is reversed ($T_6 = 6, T_4 = 5, T_5 = 4, T_3 = 3, T_2 = 2, T_1 = 1$ in (4.1)). Needless to say, both algorithms give $C_{23} = 1, C_{34} = -1$, and $C_{ij} = 0$ for the remaining pairs i, j $(i < j)$. d_{ij} can be ordered with the aid of a sorting algorithm with the computation of order $N^2 \ln(N^2)$, so that the single updating process requires the computation of order $N^2 \ln N$.

Except for how to treat the ties (See, e.g., Lehmann 1975), both the preliminary algorithm mentioned first and the "pure NMDS" Algorithm given subsequently are the same. However, if breaking ties affect the results (i.e., the result lacks structural stability), then we cannot say any strong conclusion from the result, so that the tie problem is, except for the stability issues, unimportant.

5 Large Size Data

To demonstrate the efficiency of the pure NMDS Algorithm, i.e., its capability of handling a very large data set, we analyzed the data of 1,000 cities on the globe. All the computation was done with an inexpensive PC (PII 300MHz, 256MB SDRAM, 9GB HDD, Linux 2.0.35 with g77 0.5.21). Number of iterations were 300 steps and total CPU time was within 15 minutes. Data was taken from http://www.globalserve.net/ nac/city.html that contained more than 2000 cities all over the world, but more than the three fourths were in the US. We picked up all the cities outside the US (ca., 500), and then selected some from the US — first ca., 10 cities in the alphabetical order for each state.

δ_{ij} should be the arc length between two cities i and j. Actually, we used $-\cos\theta_{ij}$ where θ_{ij} is the angle between radial vectors of the two cities. This can easily be calculated by scalar product of the vectors. For NMDS, we need only the rank order of δ_{ij}. Since the arc length is proportional to θ_{ij} and $\cos\theta_{ij}$ is a monotone decreasing function of θ_{ij} for $0 < \theta_{ij} < \pi$, we can use $-\cos\theta_{ij}$ as δ_{ij} instead of the arc length. This is one important merit of NMDS over metric MDS.

We attempted to imbed these cities into a $3D$ Euclidean space so that $d_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ are compatible with the original rank order of δ_{ij}. $s = 1 \times 10^{-4}$ and we reached a converged solution after 300 iterations. Although we did not get a $\Delta = 0$ solution, Δ became very small. In Fig, 1, we compare the original and the imbedded configurations of the 1000 cities. It is difficult to show spherical configurations here, so we draw $2D$ projections from the same direction for both plots. In spite of such a non-uniform distribution of cities on the globe, recovering of the configuration is excellent. Moreover, in the $3D$ Euclidean space, the 1000 cities locate correctly on a sphere.

This example clearly demonstrates the capability of our method to handle very large point sets. We have never seen any example of NMDS or MDS that handled such a large

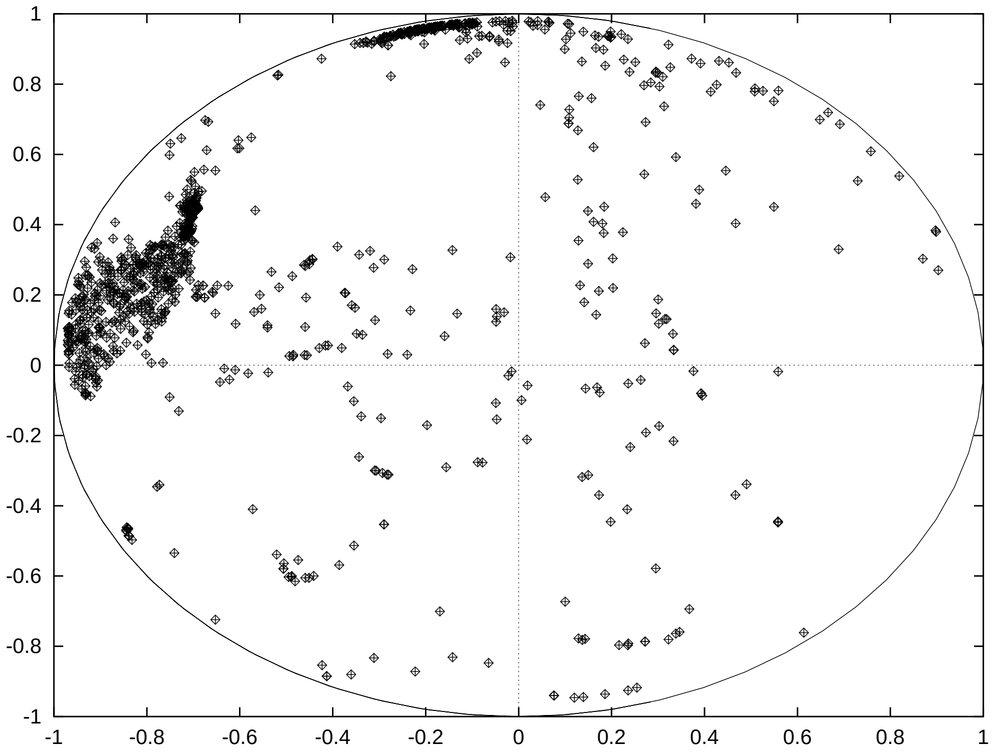


Figure 1: Positions of 1,000 cities (◇) and their reconstructed positions (+) by the pure NMDS Algorithm. The reconstructed map is scaled, rotated, and inverted to compare with the actual map.

data set. (For example, a currently available commercial software, PC-MDS[4], limits the number of OTU to be less than 100. For NMDS in Statistica[5] upper limit of number of OTU is 90.)

6 Measure of Confidence Level

Our algorithm is not free from the problem of local minima as all of the previously proposed algorithms. Despite the demonstration in Fig. 1, generally speaking, NMDS cannot give

[4] http://marketing.byu.edu/htmlpages/pcmds/pcmds.htm

[5] http://www.statsoftinc.com/toc.html

the perfect solution with $C_{ij} = 0$ for all pairs (i, j). We usually have several local minimum solutions depending on initial trial configurations.

In the conventional NMDS (Shepard and Kruskal 1964), stress is used to judge the plausibility of the obtained configuration. Since we do not have such an extra device as $\hat{d}$, we cannot use stress to check the quality of the result. The statistical properties of the stress has never been studied to our knowledge. Our measure of discrepancy Δ defined by (4.1) is statistically not trivial (because d_{ij} are not independent, even if the positions of the OTUs are), and we do not know its statistical property. However, the following closely related quantity can be defined for each OTU j

$$\Delta(j) \equiv \sum \left[T_n(j) - n\right]^2 , \tag{6.1}$$

where $T_n(j)$ and n are rank order only within $N - 1$ (i, j) pairs for the given j. Thus, $\Delta(j)$ can be regarded as a statistical variable for the relative position of the jth OTU with the remaining OTUs. We can estimate the probability $P(d)$ of $\Delta(j) < d$ with the null hypothesis that the rank ordering of d_{ij} $(i \in \{1, 2, \cdots, N\} \setminus \{j\})$ is totally random with respect to the rank ordering of δ_{ij} $(i \in \{1, 2, \cdots, N\} \setminus \{j\})$. If N is sufficiently large, then $\Delta(j)$ obeys a normal distribution $\mathrm{N}((M^3 - M)/6, M^2(M + 1)^2(M - 1)/36)$ $(M \equiv N - 1)$. For smaller N there is a table for $P(d)$ (Lehmann 1975). Thus we can always test the null hypothesis with a given confidence level for jth OTU.

As can easily be noted, instead of Δ defined in (4.1), we could use $\sum_j \Delta(j)$ (or an appropriately weighted sum) or the corresponding confidence levels as the potential function for the dynamics. This algorithm will be studied separately elsewhere.

The above quantity is also related to Kendall's τ (Lehmann 1975),

$$\tau = 2P\{(d_{ij} - d_{k\ell})(\delta_{ij} - \delta_{k\ell}) > 0\} - 1, \tag{6.2}$$

where $P\{Q\}$ is the probability of the event Q. When there are no tie data, i.e., for all $(i, j), (k, \ell)$, $d_{ij} \neq d_{k\ell}$ and $\delta_{ij} \neq \delta_{k\ell}$, $P\{(d_{ij} - d_{k\ell})(\delta_{ij} - \delta_{k\ell}) > 0\}$ is equal to

$$\frac{\sum_n [(T_n - n) > 0]}{[\#\mathrm{of}(i, j), (k, \ell)\mathrm{pairs}] = [N(N - 1)/2][N(N - 1)/2 - 1]/2}, \tag{6.3}$$

if two rankings d and δ are statistically independent, where N is total number of OTUs and summation is taken only when $T_n - n > 0$. Kendall's τ is also used to check whether there is a significant monotonic relationship between the two variables, although we do not employ this in the present paper.

In practice, we proceed as follows. First, choose a few small confidence levels , e.g., 5%, 1%, and 0.5% and count the number of OTUs whose confidence level is worse than each value. Then as the most reliable solution choose the configuration with the smallest number of OTUs with inferior confidence levels. If an OTU j cannot reject the null hypothesis that the order of d_{ij} $(i \in \{1, 2, \cdots, N\}; i \neq j)$ is uncorrelated to δ_{ij} with the confidence level better than q%, let us call j a 'q% level OTU' (here, q = 5, 1 or 0.5; we choose the smallest possible for each OTU). Suppose that solution A has two 5% level OTUs, three 1% level OTUs, and five 0.5% level OTUs, and that the solution B has one, two, and three OTUs for each confidence level, respectively. In this case, we regard solution B is more reliable than solution A.

Of course, there can be undecidable cases, and also there may be different but appropriate criteria. For example, it is difficult to decide which is more reliable the result with two 0.5% level OTUs or the other with only one 1% level OTU. For these cases we may employ both solutions, may increase dimensionality, or may use only OTUs with smaller confidence levels (See below). In any case we must accumulate more experience in this respect.

One should notice that our criterion can be used for conventional NMDS with $\hat{d}$. We demonstrate this when we compare our algorithm with KYST in Appendix. The difference between ours and the conventional ones is that our NMDS uses an optimization procedure directly related to the improvement of the confidence level of the resultant configuration.

7 Dimensionality of Imbedding Space

In MDS, the dimensionality of the imbedding space must be decided. In the conventional NMDS, stress values are plotted against the imbedding dimensions and the minimum dimensionality for saturation of the stress value is chosen.

Here, we wish to propose a slightly different criterion. As seen later in actual examples, there is a strong tendency for only a few OTUs to cause most discrepancies. In other words, most of the OTU positions converge to more or less the same relative positions independent of the initial configurations and the locations of only a few OTUs fluctuate. This fluctuation diminishes as dimensionality increases. Therefore, our criterion for determining the dimensionality is as follows. First, obtain different solutions starting from various initial configurations. Then, select the solutions with relatively small $\Delta(j)$s as mentioned in the previous section. In such a solution, if the locations of OTUs do not fluctuate considerably despite the change of initial configurations, we regard the solution to be reliable enough. Even if some OTUs fluctuate significantly, if there are stable subconfigurations with small $\Delta(j)$s (computed with the OTUs in the subconfigurations only) independent of the initial random configurations, then we may regard these substructures meaningful.

To illustrate how to choose the imbedding dimension, we use an example of imbedding 20 randomly placed points in $5D$ Euclidean space into $D(< 5)$ dimensional spaces by NMDS. $s = 1 \times 10^{-4}$ and the number of iterations is 2000. For each imbedding dimension D, we computed solutions starting from five different initial configurations. For $D = 2$, there were some 5% level OTUs. Therefore, $D = 2$ is not large enough. On the other hand, when $D = 3$, only one solution out of five solutions had one 0.5% level OTU and the remaining four had no 0.5% or worse level OTUs. For $D = 4$, no solutions had OTUs worse than 0.5% level. Thus, in this example, we can conclude that $3D$ is a reasonable dimension to recover the 20 points in the $5D$ space with 0.5% confidence level.

Figure 2 gives the Shepard plot for $D = 2, 3, 4$. When $D = 2$, the plot is scattered, but for $D = 3$ or 4 the rank order of δ_{ij} is well reproduced by the rank order of d_{ij}. Thus, our criterion for the choice of the embedding dimension is consistent with the conventional Shepard plot criterion.

Incidentally, one may wonder why this $5D$ example can be reduced to $3D$. One reason is that 20 points are not enough to express 5 dimensions. For example, to construct a unit cube, we already need $2^5 = 32$ points. Of course, to make a 5-simplex we need only 6 points, so there are really $5D$ configurations made of fewer than 20 points. However, if 6 points are randomly placed in a $5D$ space, it is highly likely that they lie close to some lower dimensional subspace. Thus, the smallness of the number of OTU coupled with fluctuations (statistical effects) seems to allow 20 points in a $5D$ space to lie close to a certain $3D$ subspace. Actually, principal component analysis told us that cumulative percentage up to the 3rd principal component of this sample configuration was about 80%. Thus, our $5D$ random data example is essentially a $3D$ configuration.

To illustrate our criterion for meaningful configurations, we deal with Wish's country data ($N = 12$). Both Kruskal and Wish (1978) and Trosset (1998) analyzed this data. In our analysis, $s = 1 \times 10^{-4}$ and the number of iterations was 2000. We started with five different configurations in 2 and $3D$. The results had always several 5% level OTUs. Thus, neither $2D$ nor $3D$ imbedding provided us with reliable solutions. This conclusion is the same as that of Kruskal and Wish. We conclude that our criterion of reliability is consistent with the conventional ones for this example.

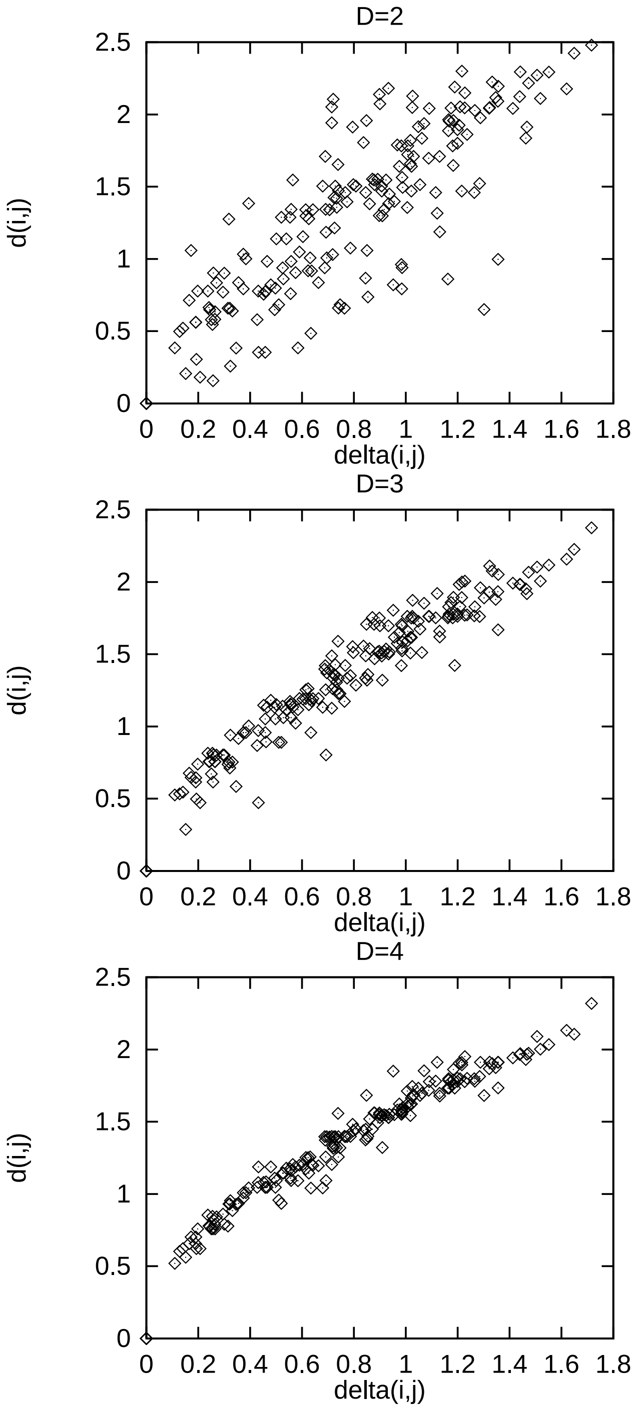


Figure 2: Shepard plots for $D = 2, 3, 4$ imbedding of the 20 points randomly distributed in $5D$. One sample configuration is shown for each case. $D = 3, 4$ looks more reasonable than $D = 2$.

Runs	>0.5%	>1%	>5%
1	11	9,12,13,24	10,27,30
2	—	11,25	10,24,26,27,28,29,30
3	28,29	9,27	10,24,30
4	—	10	24,27,28,29,30
5	24	28	27

Table 1: Results of NMDS for DNA sequence data ($D = 2$) for cichlid fish in both Lake Tanganyika and Lake Malawi. See figure caption of Fig. 4 for the whole list of fish names. Confidence levels of OTU are shown. Rows represent trial runs starting with five different initial configurations.

8 Molecular Biological Example

To illustrate the use of our algorithm we deal with DNA sequence data of cichlid fish in both Lake Tanganyika and Lake Malawi (Kocher *et al.* 1995). See figure caption of Fig. 4 for the whole list of fish names. For simplicity δ_{ij} is defined as the number of base (ATGC) mismatches, i.e.,

$$\delta_{ij} = L - \sum_k \delta[s_{ik}, s_{jk}],$$

where s_{ik} is the base (ATG or C) at kth position of the DNA sequence of OTU i and $\delta[s_{ik}, s_{jk}]$ takes 1 only when $s_{ik} = s_{jk}$ and L is the total number of bases. One might criticize that this dissimilarity is not a reasonable choice, because usually the number of uncommon bases is not proportional to the time since the speciation occurred. However, since what we need is just a rank order, any definition which does not alter the rank order gives the same result. Thus we believe that our results obtained by NMDS is robust and is not affected by a particular definition of distance between the sequences. In this example, $N = 31, s = 1 \times 10^{-5}(2D), 10^{-4}(3D, 4D)$, and the number of iterations is 2000. First, we try to imbed OTUs into $2D$ Euclidean space. Five trial runs had very different values of $\Delta(j)$ as shown in Table 1.

Clearly, all solutions have a few worse-than-1%-level OTUs. Thus, as a whole, $D = 2$ cannot be regarded as a good imbedding dimension. However, if we exclude $24, 27$, and 28th OTUs from solution 5, the solution can be regarded as a good $2D$ configuration with the confidence level better than 0.5%. Or, if we construct more solutions, there may be a better solution. This is just an illustration, so we do not go further in this paper.

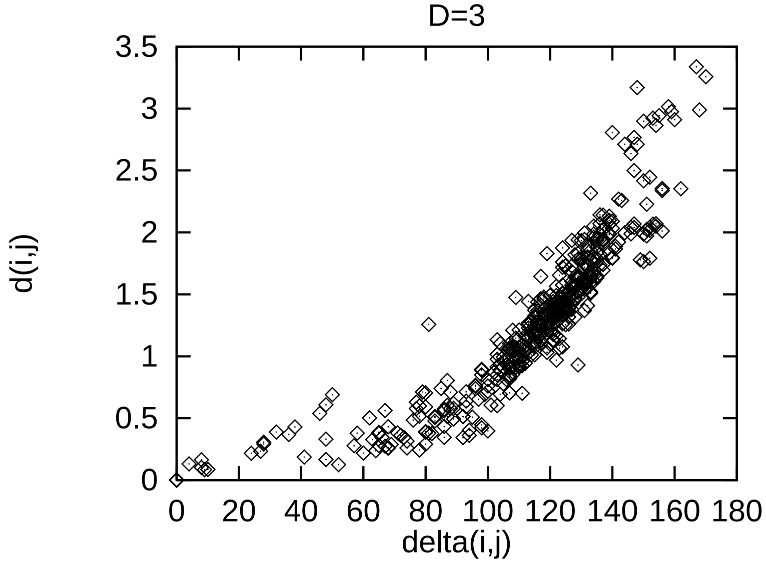
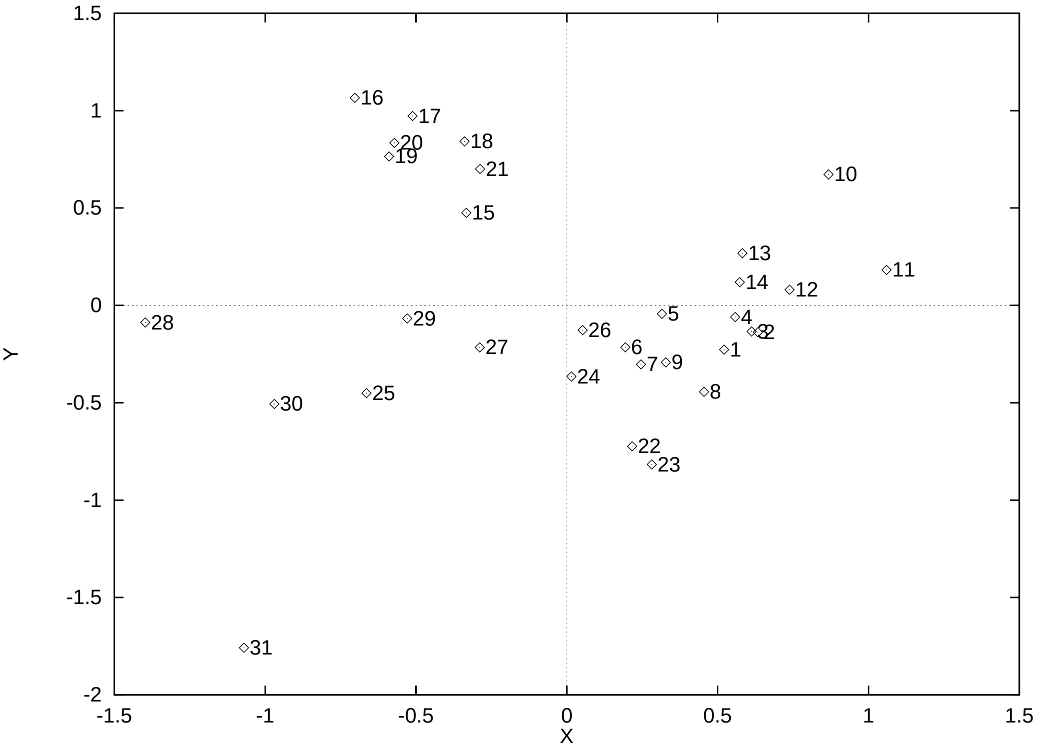


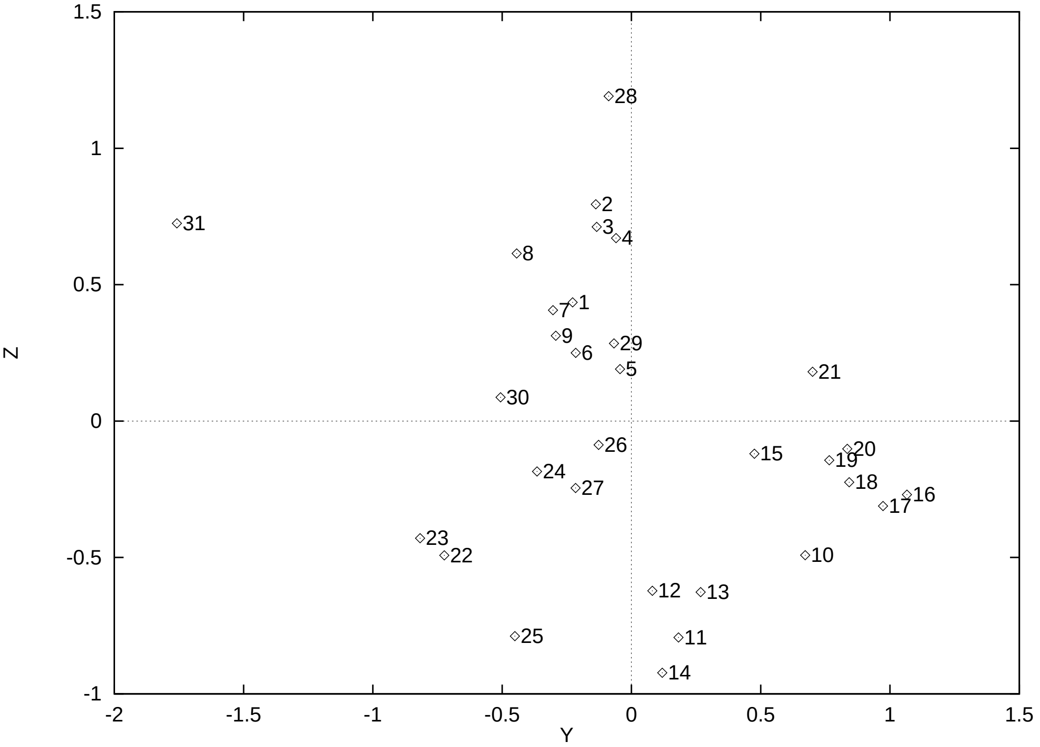
Figure 3: Shepard plot for $3D$ imbedding of DNA sequence data, where only 27th OTU has confidence level larger than 1 %.

For $3D$ Euclidean space as the imbedding space 27th OTU was with 1% confidence level and for three out of five runs 30th OTU was with 0.5% confidence level. Thus again, we cannot get any solution with confidence level 0.5%, but if we exclude 27th and/or 30th OTUs, we can get such a solution for the remaining OTUs.

Even for $D = 4$, we cannot get 0.5% confidence level solution for all OTUs. In two out of five runs, 24th OTU is with 1% confidence level, and in the remaining three runs 27th OTU is with 0.5% confidence level. For reference, we give the Shepard plot for $3D$ case (Fig. 3). Coincidence between d_{ij} and δ_{ij} is not bad as expected.

Our result is consistent with the phylogenetic analysis due to Hasegawa and K-ishino (1996), where some phylogenetic clades are reported. For example, $\{1, 2, \cdots, 9\}$, $\{10, 11, \cdots, 14\}$, and $\{15, 16, \cdots, 21\}$ are major clades. All of them can be seen as clusters in our results (Fig. 4; the former two clusters can be seen in the ZX-plane projection and the last one in the XY-plane).


Y
X
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
16
17
20
19
18
21
15
10
13
14
12
11
28
29
27
26
5
4
3
2
1
6
7
9
24
8
30
25
22
23
31


Z
Y
1.5
1
0.5
0
-0.5
-1
-2
-1.5
-1
-0.5
0
0.5
1
1.5
28
31
2
3
4
8
1
7
9
29
6
5
21
30
26
15
20
19
18
16
17
24
27
23
22
10
12
13
25
11
14

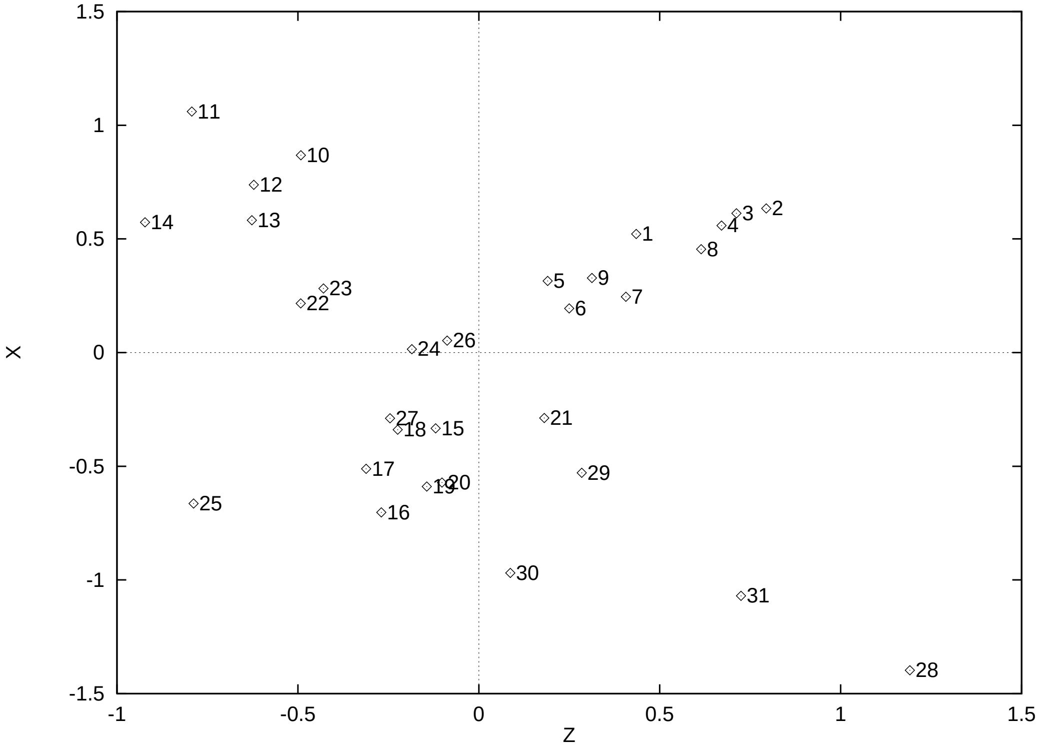


Figure 4: 2 dimensional projections of the obtained $3D$ configuration, i.e., projection to the XY-plane, YZ-plane, and ZX-plane (the directions of the axes are arbitrary), for cichlid DNA data. Major clusters correspond to the phylogenetic clades proposed by Hasegawa and Kishino (See text). Numbers denote (Names in parentheses other than Malawi represents tribes OTU belongs to. Malawi is the name of the lake OTU lives in.): 1: *Pseudotropheus zebra* (Malawi) 2: *Buccochromis lepturus* (Malawi) 3: *Champsochromis spilorhynchus* (Malawi) 4: *Lethrinops auritus* (Malawi) 5: *Rhamphochromis sp.* (Malawi) 6: *Lobochilotes labiatus* (Tropheini) 7: *Petrochromis orthognathus* (Tropheini) 8: *Gnathochromis pfefferi* (Limnochromini) 9: *Tropheus moorii* (Tropheini) 10: *Callochromis macrops* (Ectodini) 11: *Cardiopharynx schoutedeni* (Ectodini) 12: *Opthalmotilapia ventralis* (Ectodini) 13: *Xenotilapia flavipinnus* (Ectodini) 14: *Xenotilapia sima* (Ectodini) 15: *Chalinochromis popeleni* (Lamprologini) 16: *Julidochromis marlieri* (Lamprologini) 17: *Telmatochromis temporalis* (Lamprologini) 18: *Neolamprologus brichardi* (Lamprologini) 19: *Neolamprologus tetracanthus* (Lamprologini) 20: *Lamprologus callipterus* (Lamprologini) 21: *Lepidiolamprologus e-longatus* (Lamprologini) 22: *Perissodus microlepis 1* (Perissodini) 23: *Perissodus microlepis 2* (Perissodini) 24: *Cyphotilapia frontosa* (Tropheini) 25: *Tanganicodus irsacae* (Eretmodini) 26: *Limnochromis auritus* (Limnochromini) 27: *Paracyprichromis brieni* (Cyprichromini) 28: *Oreochromis niloticus* (Tilapiini) 29: *Tylochromis polylepis* (Tylochromini) 30: *Boulengerochromis microlepis* (Tilapiini) 31: *Bathybates sp.* (Bathybatini)

However, we do not claim that our results reproduce phylogenetic analysis fully. For example, the main purpose of Hasegawa and Kishino (1996) is to show that the clade $\{1, 2, \cdots, 5\}$ is monophyletic, which can never be seen in our results. In the phylogenetic analysis, even if genetic distance is short, OTUs cannot be regarded as neighbors when there are side branches between them. On the other hand, in our analysis, short genetic distance means neighboring OTUs. Thus, what we see can differ from phylogenetic relations. The biological meaning of this possible discrepancy must be explored.

As mentioned above, if there are no tie data in δ_{ij}, the preliminary algorithm and the pure NMDS algorithm give identical results. In the above example, there are tie data, but the number of such pairs is not large. Thus, the difference between two algorithms is expected to be small. In fact, two algorithms give almost the same solutions.

9 Summary and Concluding Remarks

We have proposed a novel algorithm for non-metric multidimensional scaling (NMDS) that is presumably the most faithful to the original idea of NMDS. This is why we call the proposed method the pure NMDS. The algorithm is closely connected to the statistical confidence level of the resultant spatial configuration of the OTUs. Thus, we can use statistical criterion to evaluate the plausibility of the OTU configuration and its subconfigurations in the imbedded result. The numerical efficiency of our algorithm allows us to handle a large number (even 10,000 with the aid of an inexpensive desktop computer within a day) of OTU. This feature is worthy of stressing from the practical (esp., bioinformatics) point of view.

Acknowledgements. YO was supported, in part, by the National Science Foundation Grant DMR-99-70690.

Appendix. Comparison with KYST.

One may wonder if our algorithm gives different results from the conventional methods. Here, we compare our results with those given by KYST, taken from http://www.netlib.org /mds/kyst2a.f. What we change are DIMMAX, DIMMIN, N(= number of OTUs), and the input data matrix.

- Map of 22 cities: We have chosen 22 cities in the US out of 1000 cities used in our paper. KYST correctly reproduces relative positions of 22 cities, and so does our new NMDS. Thus we can confirm that our KYST setup is correct.
- 20 random points in $5D$: DIMMAX=5 and DIMMIN=2. The obtained solution in $3D$ is the same as we obtained. STRESS obtained are 0 for $5D$, 0.045 for $4D$, 0.071 for $3D$, and 0.209 for $2D$. Thus significant decrease of STRESS occurs when D is increased from 2 to 3. This means, $D = 3$ is the plausible imbedding dimension. This conclusion agrees with ours.
- DNA Sequence: DIMMAX=5 and DIMMIN=2. We applied our statistical check for the obtained solution for $3D$ to find that 24th and 27th OTU had confidence level larger than 1%. This means that KYST and ours give the solutions with the same confidence level (to reject the null hypothesis).

  However, STRESS obtained are 0.050 for $5D$, 0.069 for $4D$, 0.104 for $3D$, and 0.173 for $2D$. Thus D dependence of STRESS is not useful to decide which D is the most plausible, while our criterion can. Therefore, our criterion is also useful to check if a conventional method like KYST gives plausible results or not.

From these results we may conclude that our novel algorithm gives the same results as KYST when it works. Our method is more advantageous than KYST, because it is conceptually direct, is numerically highly efficient, and is naturally connected with rank order statistical tests.

References

Borg, I. and Groenen, P.(1997) *Modern Multidimensional Scaling* (Springer, New York)

Cox, T. F. and Cox, M. A. A. (1994) *Multidimensional Scaling* (Chapman & Hall, London).

Green, P. E. Carmone, JR. F. J. and Smith, S. M. (1970) *Multidimensional Scaling : Concepts and Applications* (Allyn and Bacon, Massachusetts).

Hasegawa, M. and Kishino, Y. (1996) *Molecular Phylogeny* (in Japanese), Iwanami, Tokyo.

Kocher, T. D. Conroy, J. A. McKaye, K. R. Stauffer, J. R. and Lockwood, S. F. (1995) Mol. Phyl. Evol., **4** 420.

Kruskal, J. B. and Wish, M. (1978) *Multidimensional Scaling* (SAGE, Beverly Hills).

Lehmann, E. L. (1975) *Nonparametrics* Holden-Day, Inc., San Francisco.

Shepard, R. N. and Kruskal, J. B. (1964) Am. Psychol., **19**, 557.

Trosset, M. W. (1998) J. Classification, **15**, 15.