Confusion-choice model for multidimensional scaling:
Modification of Nakatani's model

HIROSHI HOJO

Department of Educational Psychology, School of Education, Waseda University, Shinjuku-ku, Tokyo 160

A modification of Nakatani's model (Nakatani, 1972, Journal of Mathematical Psychology, 9, 104-127) is proposed, in which stimulus $S_i$ and response $R_j$ are represented as a common point in a multidimensional space. When $S_i$ is presented, the distances between $S_i$ and all the responses are perturbed by an identical error process. If the distance between $S_i$ and $R_j$ is below the response threshold $t_j$, $R_j$ is acceptable. The observer chooses a response from among the acceptable responses according to a choice probability distribution. The advantages over the original are a great reduction in computing time, incorporation of the INDSCAL model, and use of the maximum likelihood estimation procedure, but the fit of the proposed model is slightly worse than Nakatani's original model. An application to a letter recognition data revealed that relative importance of the differences in shape of the lower part of letters (C, E, F, G, P, and R) to that of the right or left part of them increased as display times were increased.

Key words: recognition process, confusion models, multidimensional scaling, maximum likelihood estimation, confusion among letters.

The confusion-choice model presented by Nakatani (1972) attempts to model the confusion process of stimuli in a pure recognition experiment. It is, at the same time, recognized as a multidimensional scaling model in that it represents stimuli as points in a multidimensional subjective space. This model is successful in predicting confusion matrices and in finding stimulus configurations, but it has practical limitations on its applicability. The aim of this paper is to put forward a modification of his model which is free from such limitations.

Outline of Nakatani’s Model

The essential assumptions of Nakatani’s model (N-model) are as follows:

N1. Each stimulus $S_i$ and its correct response $R_i$ are represented as a common point in a multidimensional subjective space.

N2. The distance between $S_i$ and $R_j$ is computed by

$$d_{ij} = \left(\sum_{a=1}^{A}(x_{ia}-x_{ja})^2\right)^{1/2}$$

($i, j = 1, ..., n$), (1)

where $x_{ia}$ is the coordinate of stimulus $i$ on dimension $a$, $A$ is the dimensionality of the space, and $n$ is the number of stimuli.

N3. When $S_i$ is presented, the distance $d_{ij}$ is perturbed by an error process defined as

$$\lambda_{ij} = d_{ij} + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, 1)$.

N4. Each response $R_j$ is associated with its threshold $t_j$. If $\lambda_{ij}$ is below $t_j$, $R_j$ is acceptable, otherwise it is unacceptable.

N5. The observer chooses a response from the acceptable responses according to a choice probability distribution.

N6. The response thresholds are a function of the stimulus configuration.

Free parameters of this model, $\{x_{ia}\}$ and $\{b_j\}$, are estimated by the method of least squares. Given a fixed configuration, the
thresholds are also numerically estimated by a specific procedure under assumption N6.

Modification of Nakatani’s Model

The modification of N-model will be referred to as H-model in this paper. The following four modifications are made to N-model:

1. We introduce the idea underlying the INDSCAL model (Carroll & Chang, 1970), replacing assumption N2 by H2.

H2. The distance between Si and Rj for subject (or condition) k, d_{ij,k}, is defined by

\[ d_{ij,k} = \left[ \sum_{a=1}^{n} w_{ka}(x_{ia} - x_{ja})^2 \right]^{1/2} \]

where \( w_{ka} \) is the weight attached to dimension \( a \) by subject \( k \) and \( n \) is the number of subjects.

2. According to N-model a set of \( n \) stimuli yields \( 2^n \) possible confusion states. Thus, as Nakatani points out, N-model can hardly be applied when \( n \) exceeds ten because of the tremendous amount of computing time and memory space needed. We make assumption H3 to overcome this difficulty.

H3. The distance \( d_{ij,k} \) is perturbed by the error process

\[ \lambda_{ij,k} = d_{ij,k} + \varepsilon_{ij,k}, \]

where

\[ \varepsilon_{ij,k} \sim \mathcal{N}(0, 1). \]

The point to be emphasized is this: H3 assumes that an identical error perturbation is exerted on all distances between \( S_i \) and \( R_j (j = 1, \ldots, n) \), yielding \( n+1 \) confusion states (instead of \( 2^n \) for N-model) as will be shown later.

3. The stimulus coordinates \( [x_{ia}] \) and the thresholds \( [t_j] \) are estimated as free parameters in H-model. As to the estimates of bias probabilities, we use the values calculated directly from raw data, that is,

\[ \hat{b}_j = \frac{\sum_{k=1}^{n} \sum_{i=1}^{n} f_{ijk}}{\sum_{k=1}^{n} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ijk}}, \]

where \( f_{ijk} \) is the frequency with which the \( k \)th subject responds with \( R_j \) when \( S_i \) is presented. In words, \( \hat{b}_j \) is the relative frequency with which response \( R_j \) occurred to the total frequency of responses, and these \( \hat{b}_j \) values generally differ from one another although the subject is told that each stimulus is equally often presented. Therefore, \( \hat{b}_j \) can be an index of response bias \( b_j \) which will formally be defined later.

4. We adopt a maximum likelihood estimation procedure in place of the least squares estimation since the former allows model comparisons through a statistic called AIC (Akaike, 1974). The AIC of model \( \pi \) is defined by

\[ \text{AIC}(\pi) = -2 \ln L + 2n_s, \]

where \( L \) is the maximum likelihood and \( n_s \) is the effective number of parameters in model \( \pi \).

From the above-mentioned assumptions, N1, H2, H3, N4, and N5, we can calculate the confusion probability, \( \epsilon_{ij,k} \), that the \( k \)th subject responds with \( R_j \) when \( S_i \) is presented.

Let \( p_{ijk} \) denote the probability that \( t_j \) is below \( \lambda_{ij,k} \). By the normality assumption given in (5) we obtain

\[ p_{ijk} = \Pr(t_j < \lambda_{ij,k}) = \int_{\lambda_{ij,k}-t_j}^{t_j} \frac{1}{\sqrt{2\pi}} \phi(z) dz, \]

where \( \phi(z) \) is the density function of the standard normal distribution.

A binary-valued mediating variable is used to indicate whether or not \( R_j \) is acceptable. Thus

\[ Y_j = \begin{cases} 1 & \text{if } \lambda_{ij,k} \leq t_j, \\ 0 & \text{if } \lambda_{ij,k} > t_j. \end{cases} \]

It follows from assumption H3 that \( p_{ijk} \leq p_{ijk} \) is equivalent to \( d_{ij,k} - t_j < d_{ik} - t_j \), which can be written \((d_{ij,k} + \varepsilon_{ik}) - t_j < (d_{ik} + \varepsilon_{ik}) - t_j\).
Confusion-choice model

or \( \lambda_{ijk} - t_j < \lambda_{ik} - t_i \). Therefore, if \( p_{ijk} < p_{iuk} \) and \( Y_j = 1 \) \((\lambda_{ikt} - t_i < 0)\), then \( Y_j = 1 \) \((\lambda_{ikt} - t_i < 0)\), while if \( p_{ijk} < p_{iuk} \) and \( Y_j = 0 \), then \( Y_j = 0 \). To put this more generally, if \( p_{ijk} < p_{irk} \) \(< \ldots < p_{ijk} \), \( Y_j = 1 \), and \( Y_{j+1} = 0 \), then \( Y_j = 1 \), \( \ldots \), \( Y_{j+1} = 1 \), and \( Y_{j+1} = 0 \).

Now, a confusion state is defined as the \( n \)-tuple

\[
\mathbf{s}_h' = (Y_{ih}, \ldots, Y_{jk}, \ldots, Y_{nk})
\]

\((h = 1, \ldots, n+1)\), \((10)\)

where \( Y_{jh} \) is the value of \( Y_j \) for confusion state \( h \). Furthermore, define \( Y_{jh} \) as

\[
Y_{jh} = \begin{cases} 0 & \text{for } j < h - 1, \\ 1 & \text{for } j = h - 1, \\ 0 & \text{for } j > h - 1 \end{cases}
\]

\((11)\)

Then it is obvious from the above consideration that \( \mathbf{s}_h \) can represent any confusion state possible \((n+1) \text{ states}) in \( H \)-model. It can also be shown that the probability that confusion state \( \mathbf{s}_h \) occurs when \( S_i \) is presented, \( \epsilon_{i,h} \), is given by

\[
\epsilon_{i,h} = \begin{cases} p_{ijk} & \text{for } h = 1, \\ p_{ijk} - p_{i,jh} & \text{for } h = 2, \ldots, n, \\ 1 - p_{ijk} & \text{for } h = n + 1, \end{cases}
\]

\((12)\)

where \( p_{ijk} < p_{i,jh} \) and \( \sum_{h=1}^{n+1} \epsilon_{i,h} = 1 \).

On the other hand, given a confusion state \( \mathbf{s}_h \), the probability of responding with \( R_j \), \( g_{j,h} \), is given by assumption N5, which is formally expressed as

\[
g_{j,h} = \begin{cases} b_j & \text{for } h = 1, \\ b_j Y_{jh} / \sum_{i=1}^{n} b_i Y_{ih} & \text{for } h = 2, \ldots, n+1, \end{cases}
\]

\((13)\)

where \( b_j \) is defined as \( b_j = \Pr(R_j | \mathbf{s}_h) \)

\((\sum_{j=1}^{n} b_j = 1)\).

Consequently, we can obtain the confusion probability by summing the product \( \epsilon_{i,h} g_{j,h} \) over all confusion states. That is,

\[
\epsilon_{i,h} = \sum_{h=1}^{n+1} \epsilon_{i,h} g_{j,h}.
\]

\((14)\)

The effective number of parameters in \( H \)-model is \((n \text{ points}) \times (A \text{ coordinates per point}) \times (A \text{ coordinates for arbitrary origin}) \times (A(A-1)/2 \text{ coordinates for arbitrary rotation only for } N=1 + (A(N-1) \text{ weights}) + (n \text{ thresholds}) = nA - A(A-1)/2 \text{ only for } N=1 + A(N-1) + n). \)

To determine the estimates of these parameters, the log likelihood stated as

\[
\ln L = - \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ijk} \ln \epsilon_{i,jk}
\]

\((15)\)

is maximized over the stimulus coordinates and thresholds. The numerical method we employ in the present analysis is a quasi-Newton method, a BASIC program of which was taken from Tone (1981).

Other Versions of \( H - \) and \( N - \) Model

Two versions of \( H \)-model and one of \( N \)-model, called \( HV1 \)-, \( HV2 \)-, and \( NV1 \)-model, respectively, will also be considered in model comparisons in the next section. Each of them differs from its original in the following respect:

- **HV1-model** assumes a common threshold to all responses \((n_{HV1} = n_h - (n-1) \text{ thresholds})\).
- **HV2-model** allows for individual differences in stimulus configurations, thresholds, and bias probabilities \((n_{HV2} = N(nA - A - A(A-1)/2 + n))\).
- **NV1-model** deletes assumption N6 and estimates the thresholds as free parameters, using the values given by \((6)\) as the estimates of bias probabilities \((n_{NV1} = n_h)\).

In model comparisons this \( NV1 \)-model, instead of \( N \)-model, will be considered for its simplicity of estimation procedure.

In addition, we will consider the null model (Takane, 1981) which is based on no further assumptions than \( \epsilon_{i,j} = \{c_{i,j}\} \)

where \( c_{i,j} = f_{i,j} / \sum_{i=1}^{n} f_{i,ik} \) \((n_{null} = N(n-1))\).

Example of Application

In order to collect the relevant data, a small recognition experiment was conducted using six letters (C, E, F, G, P, and
Table 1

Summary of applications of three models to the letter recognition data

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>Models</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>HV1</td>
<td>HV2</td>
</tr>
<tr>
<td>1</td>
<td>11,973.7</td>
<td>12,018.9</td>
<td>11,998.8</td>
</tr>
<tr>
<td>n_r</td>
<td>14</td>
<td>9</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>11,798.9</td>
<td>11,851.3</td>
<td>11,823.1</td>
</tr>
<tr>
<td>n_r</td>
<td>22</td>
<td>17</td>
<td>60</td>
</tr>
</tbody>
</table>

R) as stimuli. Each letter was randomly presented on the video display under a display time which was randomly chosen for each trial from among eight different ones. A single subject (the author) observed each letter about 200 times under each exposure time. As the experiment was controlled by a microcomputer (TRS-80, Tandy Radio Shack) using a BASIC program, we cannot expect rigorous control of presentation time. Such being the case, eight display times are assigned the number 1, 2, ..., 8, respectively, at an ordinal scale T in the present analysis. We actually analyzed, however, only four confusion matrices for T=2, 4, 6, 8 to save time for parameter estimation.

Firstly, we took the confusion matrix for T=6 as an example of case N=1, and compared the goodness of fit of the models. The AIC of H-, HV1-, NV1-, and the null model were 2,862.60, 2,865.67, 2,860.57 (each in two-dimensional solution), and 2,878.62, respectively, indicating that NV1-model is the best fitting one. Notice, however, that the difference in AIC between H- and NV1-model is small.

Secondly, the results for N=4 (T=2, 4, 6, 8) are summarized in Table 1, where H-, HV1-, and HV2-model are compared in two different dimensionalities (1 & 2). (We were not able to obtain the AIC of the extended version of NV1-model for N>1 since estimating, by our BASIC program, parameters used in this model needed too much time, but judging from the above result for N=1, we can safely expect that this version would give the minimum AIC value.) The table shows that H-model in two-dimensional solution is the most appropriate of the three (AIC (null)=11,869.3, n_null=120).

Figure 1 represents the common stimulus space in (a) and condition (T) space in (b) from a two-dimensional H-model analysis of the letter recognition data. The radius of the circle which is drawn around each stimulus in (a) is equal to the threshold t_j, and the figure attached to it represents its bias probability b_j. 

![Figure 1](image-url)
space and the condition space (the configuration of weights) from the two-dimensional H-model analysis. Dimension 1 is interpreted as "curve on the right part vs. curve on the left part" of letters. Dimension 2 contrasts "closed bottom" with "open bottom". Figure 1 reveals that in this recognition task relative importance of the differences in shape of the lower part of letters to that of the right or left part of them increased as display times were increased. The ratios of the weights of dimension 2 to those of dimension 1 were 0.695, 0.696, 1.036, and 1.274 for $T=2, 4, 6, 8$, respectively.

**Discussion**

1. The major advantage of H-model over N- or NV1-model is that H-model takes much less computation time and memory space required. For example, when $N=1$ and the dimensionality is two, an evaluation of the log likelihood in (15) by the BASIC program (TRS-80) needs 0.3, 0.6, 1.1, 1.9, 2.9, and 4.3 (min) for $n=3, 4, 5, 6, 7, 8$, respectively, in case of H-model, but 0.4, 1.0, 2.7, 7.0, 18.2, and 44.7 for NV1-model. It should be noted that the difference between the two grows very rapidly with $n$.

2. H-model was also applied to Munsell color recognition data ($n=9$, Shepard, 1958), with the result that the null model fitted to the data better (AIC=19 499.5) than H-model in two-dimensional solution (AIC=19 540.4). One possible reason for this is that in application of the models we assumed that a single subject repeated all observations in spite of the fact that each color was presented to 36 subjects about 22 times. Another possible one is that the assumption about dimensionality is wrong. (There is some evidence (Takane, 1982) that the color space for these stimuli is more than two-dimensional.) Further data should be analyzed to see if the above result suggests a serious drawback in H-model.

**References**


Tone, K. 1981 BASIC. (Microcomputer programs series 1.) Baifukan. (In Japanese)

(Received Feb. 17, 1982; accepted Jan. 22, 1983)