Improving Maximum Likelihood Difference Scaling method to measure inter content scale
Andreas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, Patrick Le Callet

To cite this version:
Andreas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, Patrick Le Callet. Improving Maximum Likelihood Difference Scaling method to measure inter content scale. ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing, May 2022, Singapour, Singapore. hal-03632982

HAL Id: hal-03632982
https://hal.science/hal-03632982
Submitted on 13 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
The goal of most subjective studies is to place a set of stimuli on a perceptual scale. This is mostly done directly by rating, e.g. using single or double stimulus methodologies, or indirectly by ranking or pairwise comparison. All these methods estimate the perceptual magnitudes of the stimuli on a scale. However, procedures such as Maximum Likelihood Difference Scaling (MLDS) have shown that considering perceptual distances can bring benefits in terms of discriminatory power, observers’ cognitive load, and the number of trials required. One of the disadvantages of the MLDS method is that the perceptual scales obtained for stimuli created from different source content are generally not comparable. In this paper, we propose an extension of the MLDS method that ensures inter-content comparability of the results and shows its usefulness especially in the presence of observer errors.

Index Terms— Difference scaling, supra-threshold estimation, human perception, subjective experiment

1. INTRODUCTION

In general, it is important to perform subjective tests to validate systems (e.g. coding algorithms), benchmark objectives metrics, or create datasets to train machine learning or deep learning models. However, collecting these annotations can be time-consuming and expensive to perform. Due to the subjectivity of the data, we don’t always have an agreement in the judgment of people and accurate estimation is important to reduce noise and uncertainty on collected data. In that case, to boost these subjective tests is essential to allocate annotation resources on the right stimuli.

Multiple methodologies exist to rate stimuli, with direct estimation like Absolute Category Rating (ACR) or Double Stimuli Impairment Scale (DSIS), or indirectly, two-Alternative Forced Choice (2AFC) or pairwise comparison (PC). PC is a more reliable method because observers only need to provide their preference on each pair, comparisons are more sensitive, which is important to improve the discriminability.

PC experiments are generating matrices of values indicating how many times a stimulus has been preferred over another. These Pair Comparison Matrix (PCM) can be translated to a continuous scale, using models (e.g. Thurstone[1], Bradley and Terry [2], Tversky [3]). Due to the pairwise manner of presenting stimuli, the PCM size, and the number of possible comparisons is growing quadratically with the number of stimuli, introducing efficiency in a subjective protocol.

A lot of previous works have focused on active sampling solutions [4, 5, 6, 7, 8, 9] and more recently [10, 11, 12, 13, 14, 15, 16] to select the most informative pairs, and minimize experimental effort while maintaining accurate estimations.

In this work, we want to present an extension to the MLDS methodology [17, 18]. MLDS can be used to separately estimate scales of different stimuli sets with intra-set supra-threshold perceptual differences, see in section 2.1. But it isn’t possible to estimate a global scale where the stimuli of different sets can be related, section 2.2. These sets of stimuli from different contents are compared using different sampling strategies, introduced in section 3. With simulations, objectives are to demonstrate that our propositions can achieve good discriminability over content with robustness against bad annotator behavior and with efficiency.

Fig. 1: Example of intra-content trial presented to an observer to rate where he perceives a greater difference in the quadruplet.

2. MAXIMUM LIKELIHOOD DIFFERENCE SCALING

In this section, we explain the principle of MLDS[17, 18] and its extension to estimate the scale of cross-content perception.
The content \( C_i \) consists of a reference stimulus \( S_i \) and \( n-1 \) modified versions of this reference stimuli: \( S_2, ..., S_n \). This set of stimuli \( C_i = \{S_i, S_2, ..., S_n\} \) is pre-ordered along a physical continuum, with the assumption that larger alteration introduce higher perceptual difference:

\[
S_1 < S_2 < ... < S_n
\]  

(1)

The observer is presented during a trial with a quadruplet of stimuli \( (S_i, S_j, S_k, S_l) \), see figure 1 for an example, and asked to estimate where the perceptual distance is greater: inside the pair \( (S_i, S_j) \) or \( (S_k, S_l) \). A trial outcome is either 0 or 1, corresponding to the judgment of

\[
|S_i - S_j| - |S_k - S_l| > 0.
\]  

(2)

During solving, MLDS estimates scalar values \( (\phi_1^i, ..., \phi_n^i) \) to predict the observer judgment in each quadruplet.

2.1. Intra-content difference scaling

This system of equations can be reduced to a set of linear equations with the previous assumption and solved with a General Linear Model (GLM), using a link function \( \eta = \text{logit}(\pi(x)) \) or \( \text{probit}(\pi(x)) \) where \( \pi(x) = P(Y = 1|X_1 = x_1, ..., X_n = x_n) \) and,

\[
\pi(x) = F(\phi_1 X_1 + \phi_2 X_2 + ... + \phi_n X_n)
\]  

(3)

where \( F \) is the inverse of the link function \( \eta \). An example for 5 quadruplets of a content with 7 stimuli (1-7):

\[
\begin{pmatrix}
1 & 2 & 4 & 5 & 6 \\
1 & 2 & 3 & 7 \\
1 & 5 & 6 & 7 \\
1 & 2 & 4 & 6 \\
3 & 5 & 6 & 7
\end{pmatrix}
\]  

(4)

yield the following matrix,

\[
X = \begin{pmatrix}
0 & 1 & 0 & -1 & -1 & 1 & 0 \\
1 & -1 & -1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & -1 & -1 & 1 \\
1 & -1 & 0 & -1 & 0 & 1 & 0 \\
0 & 1 & 0 & -1 & -1 & 1
\end{pmatrix}
\]  

(5)

Once the \( (\phi_1^i, ..., \phi_n^i) \) are estimated, they can be plotted as a perceptual curve, for example one of the line in fig. 2.

2.2. Inter-contents difference scaling

In the previous section, we introduced the MLDS method and how it estimates perceptual differences inside a set of stimuli from a content. However, this method can’t directly relate and scale the difference of perception of a group of contents.

To estimate a scaling factor for each perceptual curve, one perceptual curve per content, we added inter-content comparisons. Where a quadruplet is composed of a pair of stimuli from a content \( C_i \) and a pair from a content \( C_j \): \( (S_a^i, S_b^i, S_c^j, S_d^j) \). The observer is asked, similarly as in intra-content comparison, to judge where he perceives a large perceptual distance:

\[
|S_a^i - S_b^i| - |S_c^j - S_d^j| > 0
\]  

(6)

With the addition of this set of inter-content trials, we can construct a larger matrix \( X \), solved with GLM, and estimate the difference of scales between a group of perceptual curves, figure 2.

\[
X = \begin{pmatrix}
X_1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & -1 & 1 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & 0
\end{pmatrix}
\]  

(7)

\( X \) is composed of submatrices \( (X_1, X_2) \) of intra-content comparisons for a content \( C_1 \) and \( C_2 \) and a third matrix of inter-content comparisons (e.g. last 2 rows in the example above).

Fig. 2: Example of 5 contents with 7 stimuli each and their perceptual curves, estimated using GLM and different scaling. On the right, the ground truth data to recover, and on the left estimations produced with simulated annotations of 15 participants, CIs are obtained via bootstrapping [19].

3. EXPERIMENTAL SIMULATIONS

In this section, we will experiment with the modification that we added to the MLDS methodology and check how we can select relevant quadruplets to minimize the experimental effort and be robust to spammer behavior. All comparisons are performed with simulated annotations.

3.1. Quadruplet selection

3.1.1. Intra-content quadruplet selection

The quadruplets for intra-content comparisons are generated following the recommendation of MLDS paper [17, 18],
where there are $\binom{L}{4}$ quadruplets for a L levels difference scaling experiment. In our case, we selected $L = 7$ levels to be estimated (a reference stimuli + 6 distorted versions) yielding 35 quadruplets to evaluate for each content.

3.1.2. Inter-content quadruplet selection

In the case of inter-content comparisons, we can create for a first content $M = \frac{L(L-1)}{2}$ pairs to be compared with the $M$ pairs of a second content, which is not tractable when the number of contents is large: if $N$ is the number of content, there are $\frac{N(N-1)}{2} \times M^2$ possible quadruplets. This will be the baseline fulldesign that we will improve from in the next section.

Next, we describe some sampling strategies to reduce the fraction of comparisons to perform and how we validate them in simulations.

3.2. Sampling strategies

3.2.1. Similar alteration levels comparisons

To reduce the amount of inter-content quadruplets to evaluate, we decided to compare only the same alteration levels pair in inter-content where a quadruplet is formed as follows $(S_i^a, S_i^b, S_j^a, S_j^b)$, with $a$ and $b$ 2 different alteration levels. On top of that to reduce the number of pairs available in comparison we constrained in a case named "consecutive design" to have $b = a+1$ and in a second case "reference design" $a = 1$ where $S_i^1$ and $S_j^1$ are the references.

With these 2 cases, and the associated design constraints, we reduce the number of comparisons to $L - 1$ quadruplets: where the $L - 1$ pairs, created from $L$ levels of alterations, of a content A are compared once to the $L - 1$ pairs of a content B. Then when comparing $N$ contents together we get $(L - 1) \times \frac{N(N-1)}{2} \times M^2$ quadruplets.

3.2.2. $k$-connection contents comparisons

The second part we can improve on is to reduce comparisons to only in sets of $k$ contents, instead of the $N-1$ other available contents: $(L - 1) \times (N \times k)$ comparisons with $k$ the number of compared contents.

The $k$-connection is performed over all contents $A = A_1, ..., A_n$ by connecting a content $A_i$ to the set of contents $(A_{i+1}, A_{i+2}, ..., A_{i+k})$ with a modulo operation $(i + k)\%n$, to loop back on the first contents if needed for the last $k - 1$ contents. We experimented with different values of $k$ and report the most interesting one in the result section.

3.2.3. AFAD: Adaptive Far Apart Discard design

When performing selecting comparisons, it is better to focus on close enough stimuli to avoid comparing perceptual distances that have a large difference, see eq.8: if we present these pairs to an observer, at best it will result in the same judgment as to all other previous observers, or in the worse case (e.g. spammer or error behavior) be an inverse judgment that will mislead the solving algorithm.

\[(S_i - S_j) >> (S_k - S_l)\] (8)

To avoid these cases, we will discard far apart quadruplets using the estimation of $\phi^i_j$ values. After each set of annotations from an observer, our difference scaling implementation fits and estimates the difference in the stimuli, then discards the bottom 20% of quadruplets with the largest difference scale. This strategy is reported as "AFAD" in the results section.

4. SIMULATIONS RESULTS

Fig. 3: RMSE results for 25 contents with 7 alteration levels: X axis log scale RMSE (left: number of annotations, right: number of annotations - log scale) with a 5% probability of inverting an annotation vote.

Fig. 4: Pearson correlation results for 25 contents with 7 alteration levels: X axis log scale Pearson correlation (left: number of annotators, right: number of annotations - log scale) with a 5% probability of inverting an annotation vote.
scores were randomly selected using the following principle. To construct the scores of each alteration in a content, we uniformly sample on \([0, \frac{5}{6}]\) 6 values and add a first value of 0 corresponding to the reference. Then we compute the cumulative sum over this array of 7 elements to yield the final 7 subjective scores to retrieve through simulation for a content, starting at 0 and ranging until a maximum of 5. An example of designed contents can be checked in figure 2, right part.

The following assumptions were made for simulation: 1) each stimulus has a single score, described above; 2) in each observation, the observed judgment of a quadruplet follows a Gaussian random variable, \(X, N(|ma - mb| - |mc - md|, 1)\), with \(m_i\) 4 scalars corresponding to the scores of the stimuli in the quadruplet and a fixed standard deviation of 1 to model stochasticity of the participant perception; 3) each observer has a 5% probability to produce a mistake, i.e., inverting his judgment, to test the robustness of the solution against outlier behavior; and 4) each trial is independent. Raw trial judgments are obtained by taking the sign of the sampled value from \(X\) and mapping it to the values 0 or 1.

The simulations use our python implementation of the original difference scaling algorithm, with GLM to convert raw trials judgment to estimated scores. RMSE and Pearson correlation are reported between the estimated scores and the designed ones. Each of the simulations was run at least 100 times to report 95% confidence intervals.

In figures 3 and 4, we present the simulation results for the different sampling strategies that we investigated. It shows that for a fixed number of trials (e.g. \(10^4\)), the original \textit{full design} performs worse than some designs and especially \textit{AFAD}. The proposed design converges faster to smaller estimation errors and a higher correlation score. To estimate the number of annotations saved by one design over another, we linearly fitted the data:

\[ \log(\text{number Annotations}) = A \times \log(\text{RMSE}) + B \quad (9) \]

**Fig. 5: RMSE results for 25 contents with 7 alteration levels:** X axis log scale RMSE (left: number of annotations, right: number of annotations - log scale) with a 20% probability of inverting an annotation vote.

**Fig. 6: Pearson correlation results for 25 contents with 7 alteration levels:** X axis log scale Pearson correlation (left: number of annotators, right: number of annotations - log scale) with a 20% probability of inverting an annotation vote.

We use coefficients \(A\) and \(B\) to predict the number of annotations needed for RMSE scores of \(R = [0.01, ..., 0.1]\). Finally, we compute the percentage of annotators saved for each RMSE score in \(R\) and average them. In Table 1, we read that AFAD design needs 39.7% less annotations compared to full design to achieve the same RMSE value on average.

<table>
<thead>
<tr>
<th>Design</th>
<th>% Annotations Saved</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-connect design</td>
<td>-0.68%</td>
</tr>
<tr>
<td>reference design</td>
<td>-26.67%</td>
</tr>
<tr>
<td>consecutive design</td>
<td>228%</td>
</tr>
<tr>
<td>AFAD</td>
<td>-39.7%</td>
</tr>
</tbody>
</table>

**Table 1:** Distance between curves in percent.

4.1. Observer Error Resiliency

When we increase the ratio of annotation error from 5% to 20% in fig. 5-6, the proposed solution \textit{AFAD} still estimates better than the other methods. We can see as well that the estimation error is higher than the one report in fig. 3-4, especially with a small number of annotations.

5. CONCLUSION AND FUTURE WORK

As we emphasized previously, MLDS methodology has a limitation when we need to compare perceptual scales of different source stimuli. The extension that we introduce and the \textit{active sampling} strategies that we suggested, is providing a solution to perform a subjective experiment in this paradigm of supra-threshold estimation. The strategy presents robustness to bad annotator behavior and can effectively reduce the number of comparisons to perform.

This work could be applied to collect robustly and effectively, in a large-scale crowdsourcing experiment, human judgments on distorted video patches, via video coding algorithms.
6. REFERENCES


