

# A distance metric for ordinal data based on misclassification

Dreas Nielsen  

Integral Consulting Inc., 508 Yale Ave. N. Suite 204, Seattle WA 98109, United States

Received 01 October 2023, Accepted 30 December 2023, Published 20 January 2024

---

**Abstract.** Distances between data sets are used for analyses such as classification and clustering analyses. Some existing distance metrics, such as the Manhattan (City Block or  $L_1$ ) distance, are suitable for use with categorical data, where the data subtype is numeric, or more specifically, integers. However, ordinality of categories imposes additional constraints on data distributions, and the ordering of categories should be considered in the calculation of distances. A new distance metric is presented here that is based on the number of misclassifications that must have occurred within one data set if it were in fact identical to another data set. This "misclassification distance" is equivalent to the number of reclassifications necessary to transform one data set into another. This metric takes account not only of the number of observations in corresponding ordinal categories but also of the number of categories across which observations must be moved to correct all misclassifications. Each stepwise movement of an observation across one or more categories that is required to equalize the distributions increases the distance metric, thus this method is referred to as a stepwise ordinal misclassification distance (SOMD). An algorithm is provided for the calculation of this metric.

**Keywords:** ordinal, distance, multinomial, categorical, misclassification


---

## 1 Introduction

Distance metrics are used for classification and clustering analyses, most commonly for the analysis of multivariate data sets. Distance metrics have been developed for use with continuous, categorical, and ranked numerical data, with bit strings, and with character strings. The choice of an appropriate distance metric for a particular data set is constrained to some extent by the data subtype. For example, Euclidean distance [9, 10] is appropriate for continuous variables (real numbers), whereas Manhattan distance [9] may be more appropriate for discrete variables (integers). Other distance metrics have been defined to quantify differences in relative magnitude (i.e., 'larger' or 'smaller') on a continuous scale [4, 11], for ranks [2, 12] and for binary comparisons [5].

This paper presents a new distance metric for ordinal data with an integer subtype, such as counts. Some existing distance metrics for ordinal data are based on the rankings of those

---

 Corresponding author. Email: [dnielsen@integral-corp.com](mailto:dnielsen@integral-corp.com)

categories for different objects. The Kendall tau distance is an example [1]. The distance metric described here is based on the number of misclassifications in one ordinal data set (of integers) relative to another. The number of misclassifications can also be interpreted as the number of corrections that are needed to transform one data set into the other. Misclassification by multiple ordinal steps is considered to be a larger error, and thus a larger distance, than misclassification by a single ordinal step.

This metric does not represent a distance in multidimensional space, as does, for example, the Manhattan distance, but represents a distance in terms of the number of errors that have been made or the number of corrections that would need to be made.

Other distance metrics that are conceptually similar, in that they are also based on quantifying the operations needed to transform one distribution into another, are the Levenshtein distance (or edit distance) [7] between text strings, and the Wasserstein distance (or transport distance) [8]. Neither of these are designed for, or applicable specifically to, categorical data.

Types of ordinal data to which this metric applies include Likert-scale data produced by questionnaires [6], pain-scale data used in medical assessments, and successional stages and life stages measured in ecological assessments [3].

## 2 Misclassification in nominal and ordinal categorical data

A simple misclassification distance metric for nominal categorical data—without consideration of ordinality—is simply the number of values in one data set that must be moved from one category to another to make the distributions of observations identical in both data sets. This misclassification metric applies only when both data sets have the same number of total observations. For such data sets, the misclassification metric is always one-half of the Manhattan distance metric.

The simple misclassification metric for nominal categorical data can be extended to ordinal categorical data by incorporating the number of steps between categories that an observation must be moved. That is, if an observation must be re-classified (moved) from one category to an adjacent category, that is considered to be a single misclassification error, and to lead to a misclassification distance of 1. If an observation must be moved from farther away (along the ordinal scale), then the distance is increased by 1 for each additional step that must be taken.

For some pairs of data sets, there is potentially more than one set of observation reclassifications that could be made to equalize the distributions of the two data sets. These may result in different distance measures. The stepwise ordinal misclassification distance (SOMD) should be taken as the minimum number of possible reclassification steps that might be taken.

The SOMD metric, in its simplest form, assumes that all misclassification errors between adjacent ordinal categories have the same weight. That is, if there are three ordinal categories, misclassification of an item from category 1 as category 2 has the same weight as a misclassification of an item from category 2 as category 3, and further that these weights are symmetrical (i.e., misclassification from 2 to 1 is equal to misclassification from 1 to 2). However, weights could be applied in an extended form of this metric if some types of misclassification error are considered to be more severe, or less likely, than others. For example, if the number of issues found during a software code review were classified into ordered categories of importance such as "undocumented", "inefficient", and "logic fault", a misclassification of an "inefficient" error as a "logic fault" error, or vice-versa, might be given a higher weight than a misclassification of an "undocumented" error as an "inefficient" error.

The difference between the simple misclassification distance and the stepwise misclassification can be illustrated with the data in Table 2.1. The simple misclassification distance between these two data sets (instances) is 4: four items must be moved from one category to another, in either data set, to make the distributions equivalent. The stepwise misclassification distance, in contrast, is 8, because each of those four observations must be moved by two steps, where each step represents a separate misclassification or error.

Table 2.1: Example ordinal data

Instance	Category 1	Category 2	Category 3
1	8	8	8
2	4	8	12

### 3 Comparison to categorical distance metrics

The SOMD metric will always be equal to or greater than the simple misclassification distance. When no observation is misclassified by more than a single category, these distance measures will be identical. However, as the previous example illustrates, when any observation must be moved by more than a single category to equalize the two distributions, the SOMD will be greater than the simple misclassification distance.

Although the simple misclassification distance has a fixed relationship to the Manhattan distance (i.e., it is always half, for data sets of equal size), the SOMD index has no fixed relationship to the Manhattan distance. The SOMD distance may be either greater or less than the Manhattan distance.

### 4 Algorithm for calculating the minimum stepwise ordinal misclassification distance metric

Algorithm 1 describes a process for calculating the SOMD metric.

This algorithm uses equal weights for each misclassification step within the ordered set of categories. If different weights are to be used for different steps, then an  $n \times n$  matrix should be used to specify the weights for different step sizes. The appropriate weight for values of  $i$  and  $j$  should be obtained from this matrix and used to update the distance measure on lines 8 and 17. If this matrix is not symmetric about the diagonal, then SOMD is no longer a metric because it is no longer necessarily symmetric (i.e.,  $dist_{SOMD}(A, B) \neq dist_{SOMD}(B, A)$ ) nor does it adhere to the triangle inequality.

---

**Algorithm 1** Stepwise ordinal misclassification distance (SOMD)

---

**Require:**  $n$  is the number of categories**Require:**  $A, B$  are vectors of counts in  $n$  categories**Require:**  $\sum_{i=1}^n A = \sum_{i=1}^n B$ **Ensure:**  $d = \text{SOMD}$  between  $A$  and  $B$ 

```

1:  $d \leftarrow 0$ 
2: for  $i \leftarrow 1$  to  $n$  do
3:   if  $i > 1$  and  $A[i] < B[i]$  then
4:     for  $j \leftarrow 1$  to  $i - 1$  do
5:       while  $A[j] > B[j]$  and  $A[i] < B[i]$  do
6:          $A[i] \leftarrow A[i] + 1$ 
7:          $A[j] \leftarrow A[j] - 1$ 
8:          $d \leftarrow d + (i - j)$ 
9:       end while
10:    end for
11:  end if
12:  while  $A[i] < B[i]$  do
13:    for  $j \leftarrow i + 1$  to  $n$  do
14:      while  $A[j] > B[j]$  and  $A[i] < B[i]$  do
15:         $A[i] \leftarrow A[i] + 1$ 
16:         $A[j] \leftarrow A[j] - 1$ 
17:         $d \leftarrow d + (j - i)$ 
18:      end while
19:    end for
20:  end while
21: end for
22: return  $d$ 

```

---

## 5 Conclusion

A new distance metric for ordinal integer data is presented. This metric is conceptually related to distance metrics such as the Manhattan distance, but incorporates information on the ordered nature of the data. This stepwise ordinal misclassification distance metric can be regarded as a measure of the *process* by which one data set may be transformed into another, not simply of the result of such a transformation.

A constraint of this method is that it is applicable only to equally sized groups. With this constraint, this method is applicable to matched-pairs data sets such as are produced by before/after studies. Other experimental or observational studies with equally sized groups can also make use of this method.

The provided algorithm is straightforward and allows this metric to be incorporated into data analyses such as classification and clustering methods. Modification of the algorithm to incorporate symmetrical or asymmetrical weighting of misclassification errors is also described, for other potential uses of this method.

## Declarations

### Funding

This work was supported by Integral Consulting Inc.

### Conflict of interest

The author has no conflicts of interest to declare.

## References

- [1] V. CICIRELLO, *Kendall tau sequence distance: extending Kendall tau from ranks to sequences*, EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, **7**(23) (2020), 1–20. [DOI](#)
- [2] W. D. COOK, *Distance-based and ad hoc consensus models in ordinal preference ranking*, European Journal of Operational Research, **172** (2006), 369–385. [URL](#)
- [3] D. FERNÁNDEZ AND S. PLEDGER, *Categorising count data into ordinal responses with application to ecological communities*, Journal of Agricultural, Biological, and Environmental Statistics, Springer Science and Business Media LLC, **21** (2015), 348–362.
- [4] K. JAJUGA, M. WALESIAK, AND A. BAK, *On the general distance measure*, Exploratory Data Analysis in Empirical Research, Springer Berlin Heidelberg, 2003. pp.104–109. [DOI](#)
- [5] M. KLEINDESSNER, AND U. VON LUXBURG, *Lens depth function and k-relative neighborhood graph: versatile tools for ordinal data analysis*, Journal of Machine Learning Research, **18**(58) (2017), 1–52. [URL](#)
- [6] R. LIKERT, *A technique for the measurement of attitudes*, Arch. Psychol. **22**(140) (1932), 1–55.
- [7] G. NAVARRO, *A guided tour to approximate string matching*, ACM Computing Surveys **33**(1) 31–88, (2001). [URL](#)

- [8] V.M. PARARETOS AND Y. ZEMEL, *Statistical aspects of Wasserstein distances*, *Annual Review of Statistics and Its Application*, **6**(1) (2019) 405–431. [DOI](#)
- [9] A.S. SHIRKHORSHIDI, S. AGHABOZORGI AND T.Y. WAH, *A Comparison study on similarity and dissimilarity measures in clustering continuous data* *PLoS ONE*, **10**(12): e0144059. (2015). [DOI](#)
- [10] J. TABEK, *Geometry: the language of space and form*, *History of Mathematics Series*, Infobase Publishing, 2014.
- [11] M. WALESIAK, *Distance measure for ordinal data*, *Argumenta Oeconomica*, **2** (1999), 167–173.
- [12] A. ZABORSKI, *Distance measures in aggregating preference data*, *Folia Oeconomica*, **3**(302) (2013), 183–190. [URL](#)