Tertiary Approach Considered Harmful

Jan de Leeuw

November 21, 2025

TBD

Table of contents

Note: This is a working manuscript which may be expanded/updated multiple times. All suggestions for improvement are welcome. All Rmd, tex, html, pdf, R, and C files are in the public domain. Attribution will be appreciated, but is not required. The files can be found at https://github.com/deleeuw/tertiary

1 Introduction

In monotone regression the data is

- the *target*, a numerical vector y of length n,
- a partial order $\preceq$ on $\mathcal{N} := \{1, 2, \cdots, n\}$

We define

- $i \approx j$ if both $i \preceq j$ and $j \preceq i$,
- $i \prec j$ if $i \preceq j$ but not $j \preceq i$.

In this paper we will only consider the case of a totalk order, which means that for all (i, j) either $i \preceq j$ or $j \preceq i$ (or both). If $i \approx j$ we say the pair (i, j) is a *tie*. Since being tied is an equivalence relation it partitions $\mathcal{N}$ into equivalence classes, called *tie-blocks*.

In the least squares version of monotone regression we minimize the weighted least squares loss function

$$\sigma(x) := \sum_{i=1}^{n} w_i (x_i - y_i)^2 \tag{1}$$

over all x for which $x_i \leq x_j$ if $i \prec j$. This definition of the monotone regression problem does not say what to do with ties. If there are no ties then the x_i must satisfy a total order: if $1 \prec \cdots \prec n$ we require $x_1 \leq \cdots \leq x_n$.

If there are ties in the data then the users of a non-metric scaling programm typically has to choose from various options. The two most prominent ones, proposed by Kruskal (1964a) and Kruskal (1964b), are

- Primary Approach: $x_i \leq x_j$ if $i \prec j$,
- Secondary Approach: $x_i \leq x_j$ if $i \preceq j$.

In the primary approach there are no constraints within tie-blocks, in the secondary approach it follows that we require $x_i = x_j$ if $i \approx j$. Also see Guttman (1968) for an extensive discussion of these options.

Kruskal (1964a) showed that primary approach monotone regression can be solved by redefining the constraints.

- Primary Approach Redux: $x_i \leq x_j$ if $i \prec j$ and if $(i \approx j) \wedge (y_i < y_j)$.

Thus y is used to order the indices within tie-blocks, and there is no pair (i, j) with $i \approx j$ and $y_i = y_j$ the resulting constraints on x define a total order. Kruskal does not give an explicit rule to deal with $(i \approx j) \wedge (y_i = y_j)$, but seems to suggest to complete the total order by requiring either $x_i \leq x_j$ or $x_j \leq x_i$. Which one of the two choices we make does not matter for the outcome of the monotone regression.

We need some additional notation for solving the monotone regression problem with the secondary approach to ties. Suppose there are $m < n$ tie-blocks. Compute the tie-block weighted averages $\overline{y}_k$. Also compute the tie-block weights $\overline{w}_k$ as the sum of the w_i in the tie-block.Then minimize

$$\sigma(\overline{x}) := \sum_{k=1}^{m} \overline{w}_k (\overline{x}_k - \overline{y}_k)^2$$

with the constraints

- Secondary Approach Redux: $\overline{x}_k \leq \overline{x}_\ell$ if $k < \ell$.

See De Leeuw (1977) for explicit proofs of the optimality of the two redux versions.

- Tertiary Approach: $\overline{x}_k \leq \overline{x}_\ell$ if $I_k \prec I_\ell$.

Suppose there are only $m < n$ different values in y. Define the tie-block averages $\overline{y}_k$ as the weighted average of the y_i in tie-block k.

the tertiary approach is in De Leeuw (1977). All three approaches are implemented as options in the smacof package (De Leeuw and Mair (2009), Mair, Groenen, and De Leeuw (2022)).

2 The Tertiary Approach

If there are only a few ties and there is a good fit the difference between the three approaches will be small. But in general there are several problems with the tertiary approach, and users of the smacof program should think twice before using it.

It is shown in De Leeuw (1977) that the solution of is given by

$$y_i = \overline{y}_j + (x_i - \overline{x}_j)$$

In the first place the monotone regression solution with the tertiary approach to ties may produce a vector y which is

3 References

The first is to ignore ties.

$$y_i = \overline{y}_j + (x_i - \overline{x}_j)$$

Example: two classes, equal number of elements. If the two group means are out of order $\overline{x}_1 > \overline{x}_2$ we have

$$\overline{y}_1 = \overline{y}_2 = \frac{1}{2}(\overline{x}_1 + \overline{x}_2)$$

Thus for $i \in I_1$

$$y_i = \frac{1}{2}(\overline{x}_1 + \overline{x}_2) + (x_i - \overline{x}_1) = x_i - \frac{1}{2}(\overline{x}_1 - \overline{x}_2)$$

Thbus $y_i < 0$ if $x_i < \frac{1}{2}(\overline{x}_1 - \overline{x}_2)$

For $i \in I_2$

$$y_i = \frac{1}{2}(\overline{x}_1 + \overline{x}_2) + (x_i - \overline{x}_2) = x_i + \frac{1}{2}(\overline{x}_1 - \overline{x}_2) \geq 0.$$

$x = (1, 9, 1, 3)$ Then $\overline{x}_1 = 5$ and overline $\overline{x}_2 = 2$. Thus $\overline{x}_1 - \overline{x}_2 = 3$ and y is $(-.5, 7.5, 2.5, 4.5)$

De Leeuw, J. 1977. "Correctness of Kruskal's Algorithms for Monotone Regression with Ties." *Psychometrika* 42: 141–44.

De Leeuw, J., and P. Mair. 2009. "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software* 31 (3): 1–30. https://www.jstatsoft.org/article/view/v031i03.

Guttman, L. 1968. "A General Nonmetric Technique for Fitting the Smallest Coordinate Space for a Configuration of Points." *Psychometrika* 33: 469–506.

Kruskal, J. B. 1964a. "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis." *Psychometrika* 29: 1–27.

———. 1964b. "Nonmetric Multidimensional Scaling: a Numerical Method." *Psychometrika* 29: 115–29.

Mair, P., P. J. F. Groenen, and J. De Leeuw. 2022. "More on Multidimensional Scaling in R: smacof Version 2." *Journal of Statistical Software* 102 (10): 1–47. https://www.jstatsoft.org/article/view/v102i10.