Predictive Ratio Matching for Compositional Data

Gerko Vink

Department of Methodology and Statistics, Utrecht University
Division of Methodology and Quality, Statistic Netherlands

Compositional Data

Let us consider x_0 as a combination of x_1 through x_D, such that

$$x_0 = x_1 + x_2 + ... + x_D \tag{1}$$

where the integers $1, 2, ..., D$ denote the parts and the subscripted letters $x_1, ..., x_D$ denote the components.

Compositional Data

All the information about compositional data is encapsulated in the ratios between the components (Aitchison, 1986). Consequently, the proportions of the different parts of x obey

$$\frac{x_1}{x_0} + \frac{x_2}{x_0} + ... + \frac{x_D}{x_0} = 1 \qquad (2)$$

which is equivalent to Equation (**??**), where

$$x_1 \geq 0, x_2 \geq 0, ..., x_D \geq 0 \qquad (3)$$

We define the sample space of a D-part composition as the simplex S^D

$$S^D = \{(x_1, x_2, \ldots, x_D) : x_j \geq 0; j = 1, 2, \ldots, D; \sum_j^D x_j = c\} \quad (4)$$

Let us assume that we have the following 3-part compositional data with missing values

x_1	x_2	x_3	x_0
10	15	7	32
0	18	0	18
6	3	–	22
0	–	–	14
–	16	–	–
22	–	4	8
–	–	–	30
5	10	15	–

For some of the missing values, it is possible to deductively impute the true value. For example, the third row yields $x_3 = 22 - (6 + 3) = 13$ and the bottom row yields $x_0 = 5 + 10 + 15 = 30$.

MISSINGS IN A SINGLE COMPOSITION

$$x_0 = x_1 + x_2 + x_3 \tag{5}$$

$$x_1 + x_2 = x_0 - x_3 \tag{6}$$

We can solve this by imputing by imputing the ratio $\pi = x_1/x_2$ from a probable donor record d, yielding

$$x_1^* = \frac{\hat{\pi}_d^{(12)}}{\hat{\pi}_d^{(12)} + 1}(x_0 - x_3), \tag{7}$$

and its complement

$$x_2^* = \frac{1}{1 + \hat{\pi}_d^{(12)}}(x_0 - x_3) \tag{8}$$

MISSINGS IN A NESTED COMPOSITION

$$x_0 = x_1 + x_2 + x_3 \tag{9}$$

$$x_3 = x_4 + x_5 \tag{10}$$

Let x_2, x_3 and x_4 be jointly missing for some, but not all, cases. For the cases where x_3 is missing, the problem can be simplified to

$$x_0 = x_1 + x_2 + x_4 + x_5, \tag{11}$$

where x_3 is simply the sum of x_4 and x_5 and does not need to be imputed, but can be deductively calculated afterwards. This reduces the problem to a single composition, yielding imputations

$$x_2^* = \frac{\hat{\pi}_d^{(24)}}{\hat{\pi}_d^{(24)} + 1}(x_0 - x_1 - x_5) \quad \text{and} \quad x_4^* = \frac{1}{1 + \hat{\pi}_d^{(24)}}(x_0 - x_1 - x_5). \tag{12}$$

The imputed value for x_3 can then be calculated as

$$x_3^* = x_4^* + x_5 \tag{13}$$

For the cases where x_3 is observed, the problem splits into the independent problems

$$x_0 = x_1 + (x_0 - x_1 - x_3) + x_3 \tag{14}$$

and

$$x_3 = \frac{\hat{\pi}_d^{(45)}}{\hat{\pi}_d^{(45)} + 1}(x_3) + \frac{1}{1 + \hat{\pi}_d^{(45)}}(x_3) \tag{15}$$

where donors are drawn from within the compositional level of the missing values.

MISSINGS IN MULTIPLE NESTED COMPOSITIONS

$$
\begin{array}{ccccccccc}
x_0 & = & x_1 & + & x_2 & + & x_3 & & \\
 & & = & & & & = & & \\
 & & x_8 & & & & x_4 & & \\
 & & + & & & & + & & \\
 & & x_9 & & & & x_5 & = & x_6 + x_7
\end{array}
\tag{16}
$$

A solution for this data where x_1, x_3 and x_5 are known is simply the summation of a sumscores respective parts, such that

$$
\begin{aligned}
x_0 &= x_1 + x_2 + x_3 \\
x_1 &= x_8 + x_9 \\
x_3 &= x_4 + x_5 \\
x_5 &= x_6 + x_7
\end{aligned}
$$

For unknown x_5, all components from $x_5 = x_6 + x_7$ are moved to the higher level, such that

$$\begin{aligned} x_0 &= x_1 + x_2 + x_3 \\ x_1 &= x_8 + x_9 \\ x_3 &= x_4 + x_6 + x_7 \end{aligned}$$

where the unobserved x_5 is calculated afterwards as $x_6 + x_7$. For unknown x_3 and x_5 it holds that

$$\begin{aligned} x_0 &= x_1 + x_2 + x_4 + x_6 + x_7 \\ x_1 &= x_8 + x_9 \end{aligned}$$

and for unobserved x_1, x_3 and x_5 there remains one composition to be imputed, namely

$$x_0 = x_8 + x_9 + x_2 + x_4 + x_6 + x_7$$

For any D-part composition, the number K of ratios to be considered is the number of unique pairs, without considering the order of the element of a pair, which equals

$$K = \frac{D(D-1)}{2} = \binom{D}{2}. \tag{17}$$

For any pair x_k, with $k = 1, ..., K$, we can compute the ratio

$$\pi_k = \frac{x_{k.1}}{x_{k.2}}, \tag{18}$$

coming from the distribution

$$\Pr(\pi_k | x_{k.1}, x_{k.2}) \tag{19}$$

with $k.1$ and $k.2$ denoting the first and second part of the composition in k, respectively.

1. Start with the lowest level l. If there are multiple compositions at level l, repeat steps 2-4 for each composition.
2. For all $x_{0,mis}^{(l)}$ move the corresponding components to level $l - 1$
3. For all $x_{0,obs}^{(l)}$ find starting values for the components, if needed
4. Calculate all $K^{(l)}$ relevant pairs k
   4.1 Impute joint-missing ratios for all k pairs and redistribute the corresponding amounts
   4.2 If applicable, calculate the sumscores of the previous level
5. Set $l = l - 1$ and repeat step 2-4.
6. repeat steps 1-5 until convergence is reached. For multiple imputation do this m 2 times, each time saving the completed dataset.

x_1	x_2	x_3	x_0
10	15	7	32
0	18	0	18
6	3	13	22
0	14	0	14
7.1	15.6	7.3	30
5	12	15	32

x_1	x_2	x_3	x_0
10	15	7	32
0	18	0	18
6	3	13	22
0	5.6	8.4	14
5.3	11.4	13.3	30
5	12	15	32