Predictive Ratio Matching Imputation

of Nested Compositional Data with Semicontinuous Variables

Gerko Vink Jeroen Pannekoek Stef van Buuren

JSM 2014

Universiteit Utrecht

TNO innovation for life

COMPOSITIONAL DATA

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

$$x_0 = x_1 + x_2 + ... + x_D,$$

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

SOME MORE BACKGROUND

All the information about compositional data is encapsulated in the ratios between the components[1]. Consequently, the proportions of the different parts of x obey

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

where

$$x_1 \geq 0, x_2 \geq 0, ..., x_D \geq 0,$$

such that the sample space is defined as simplex S^D

$$S^D = \{(x_1, x_2, \ldots, x_D) : x_j \geq 0; j = 1, 2, \ldots, D; \sum_j^D x_j = x_0\}.$$

[1] Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall.

MISSINGS IN A SINGLE COMPOSITION

$$x_0 = x_1 + x_2 + x_3$$

$$x_1 + x_2 = x_0 - x_3$$

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

$$x_1^* = \pi_{(21)}^*(x_0 - x_3),$$

and its complement

$$x_2^* = (1 - \pi_{(21)}^*)(x_0 - x_3),$$

where π_{21}^* is the imputed ratio for pair 21 and comes from the distribution

$$\Pr(\pi_{21}^* | \pi_{21}, x_0, x_3)$$

of donors with both x_2 and x_1 observed.

MULTIVARIATE MISSINGS IN A SINGLE COMPOSITION

$$x_0 = x_1 + x_2 + x_3 + x_4$$

$$x_1 + x_2 + x_3 = x_0 - x_4$$

We can solve this by finding starting values for x_1, x_2 and x_3 and iteratively updating the bivariate ratios from probable donor records.

Any starting value will be sufficient as long as the compositional structure remains intact.

All $\binom{j}{2}$ unique pairs, where j is the number of variables can be imputed by means of this approach and variables outside of the currently imputed pair, as well as the total of the composition, can serve as covariates in the prediction of π^*.

MULTIVARIATE MISSINGS IN A SINGLE COMPOSITION

$$x_0 = x_1 + x_2 + x_3 + x_4$$

The $\binom{j}{2}$ unique pairs in the above problem are

$$\begin{array}{cc} x_1 & x_2 \\ x_1 & x_3 \\ x_1 & x_4 \\ x_2 & x_3 \\ x_2 & x_4 \\ x_3 & x_4 \end{array}$$

MULTIVARIATE MISSINGS IN A SINGLE COMPOSITION

Some pairs do not need updating, as x_4 is observed and we would not want to alter observed values.

x_1 x_2
x_1 x_3
x_1 x_4
x_2 x_3
x_2 x_4
x_3 x_4

Missings in pairs where one of the variables is observed will be imputed in another pair where both variables are missing (if only one variable is missing, the equation could have been solved deductively)

Only having to consider jointly missing pairs makes PRM much faster.

PRM APPROACH

We require that starting values have been filled in and that any deductive imputation has been applied.

Carry out the following steps for all $\binom{j}{2}$ unique pairs (simplified version).

1. Calculate π^{obs} (if not defined: $\pi^{obs} = 0.5$).
2. Impute π^{mis}
3. Redistribute amounts

Repeat the above algorithm until convergence is reached. For multiple imputation do this $m \geq 2$ times

Imputations could be obtained by e.g. PMM with π^{obs} conditional on the remaining variables and the total.

x_0	x_1	x_2	x_3
32	10	15	7
18	0	9	9
22	6	3	—
14	0	—	—
8	22	—	4
30	—	—	—

BENEFITS OF PRM

By addressing the ratios between pairs, the compositional problem can be solved outside the simplex space

No need for post-hoc fixes when components are 0

The flexibility of imputation by chained equations - (m)ice

NESTED COMPOSITIONS: UNOBSERVED NESTED SUM

Suppose that we have a nested composition, where x_4 is a combination of x_5 and x_6, such that

$$x_4 = x_5 + x_6.$$

For the cases where x_4 is missing, the problem can be simplified to

$$x_1 = x_2 + x_3 + x_5 + x_6,$$

where x_4 can be deductively calculated after x_5 and x_6 are imputed.

This reduces the problem to a single composition, which can easily be solved by the proposed PRM algorithm.

NESTED COMPOSITIONS: OBSERVED NESTED SUM

For the cases where x_4 is observed, the problem can be divided into the independent imputation problems

$$x_1 = x_2 + x_3 + x_4 \quad \text{and} \quad x_4 = x_5 + x_6.$$

Solving these separate compositions is also straightforward with the proposed PRM algorithm.

In both cases donors are drawn from within the compositional level of the missing values.

MISSINGS IN MULTIPLE NESTED COMPOSITIONS

$$
\begin{array}{lcccccccc}
x_0 & = & x_1 & + & x_2 & + & x_3 & + & x_4 \\
 & & = & & & & & & = \\
 & & x_9 & & & & & & x_5 \\
 & & + & & & & & & + \\
 & & x_{10} & & & & & & x_6 \quad = \quad x_7 \quad + \quad x_8
\end{array}
$$

A solution for this data where x_1, x_4 and x_6 are known is simply the summation of a sumscores respective parts, such that

$$
\begin{aligned}
x_0 &= x_1 + x_2 + x_3 + x_4 \\
x_1 &= x_9 + x_{10} \\
x_4 &= x_5 + x_6 \\
x_6 &= x_7 + x_8
\end{aligned}
$$

MISSINGS IN MULTIPLE NESTED COMPOSITIONS

For unknown x_6, all components from $x_6 = x_7 + x_8$ are moved to the higher level, such that

$$\begin{aligned} x_0 &= x_1 + x_2 + x_3 + x_4 \\ x_1 &= x_9 + x_{10} \\ x_4 &= x_5 + x_7 + x_8. \end{aligned}$$

For unknown x_4 and x_6 it holds that

$$\begin{aligned} x_0 &= x_1 + x_2 + x_3 + x_5 + x_7 + x_8 \\ x_1 &= x_9 + x_{10}, \end{aligned}$$

and for unobserved x_1, x_4 and x_6 there remains one composition to be imputed, namely

$$x_0 = x_9 + x_{10} + x_2 + x_3 + x_5 + x_7 + x_8.$$

DIVIDE AND CONQUER APPROACH

Any set of nested compositions can be imputed by means of the following scheme (highly simplified!).

Start with the lowest level composition and carry out for all (nested) compositions

1. For cases where the sumscore of a nested composition is missing
   - 1.1 Promote the missingness problem to the next higher level composition and save it for later.
2. For cases where the sumscore of a nested composition is observed
   - 2.1 Impute the missing parts in the composition by means of PRM.
   - 2.2 Calculate unobserved nested totals (if any) in the current composition based on the imputed parts.

Repeat the above until convergence is reached. For multiple imputation do this $m \geq 2$ times

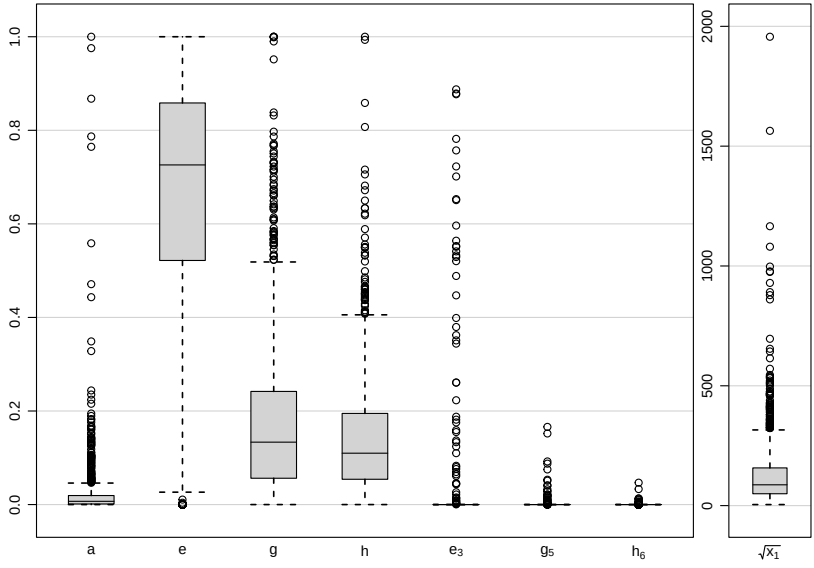
EVALUATING PRM

Simulation study: Dutch Wholesaler Data (DWD) for 2007. The DWD dataset contains edited information on 1067 wholesalers for a set of cost statistics (a, e, g and h) that sum op to a set total x_0, leading to composition

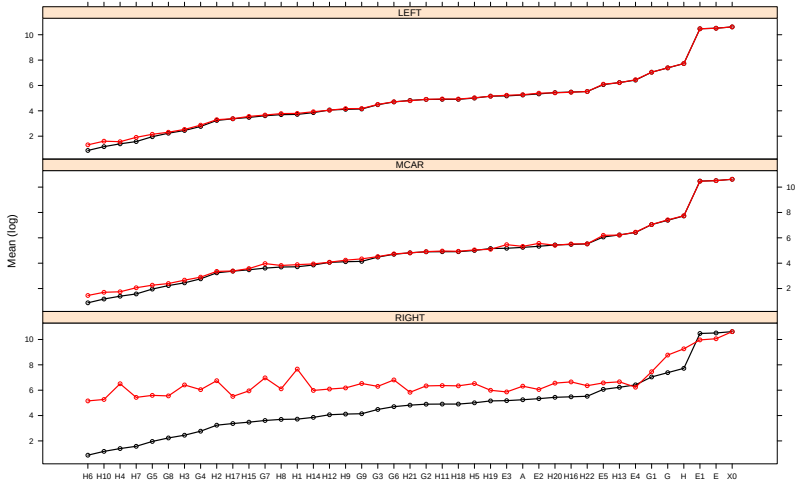
$$x_0 = a + e + g + h,$$

x_0	the total operating costs		
a	company depreciation	single measure	
e	buying costs	5 parts	
g	personnel costs	9 parts	
h	other costs	21 parts	$h_1 = h_2 + h_3 + h_4$

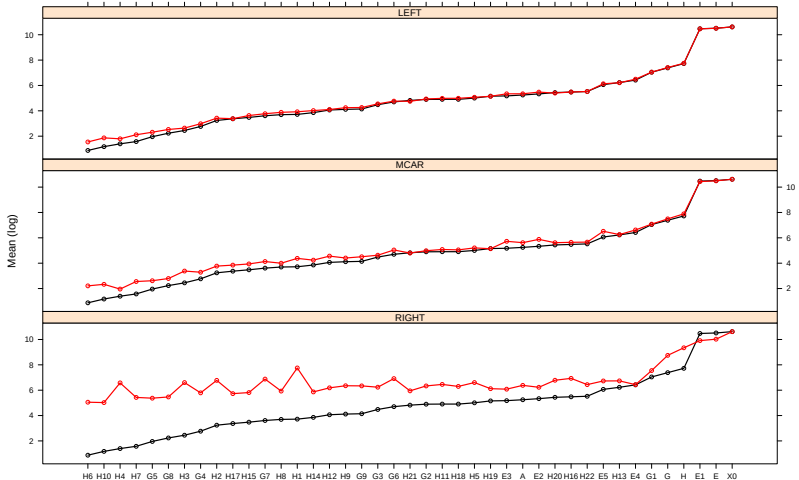
3 missingness mechanisms: left-tailed MAR, right-MAR and MCAR

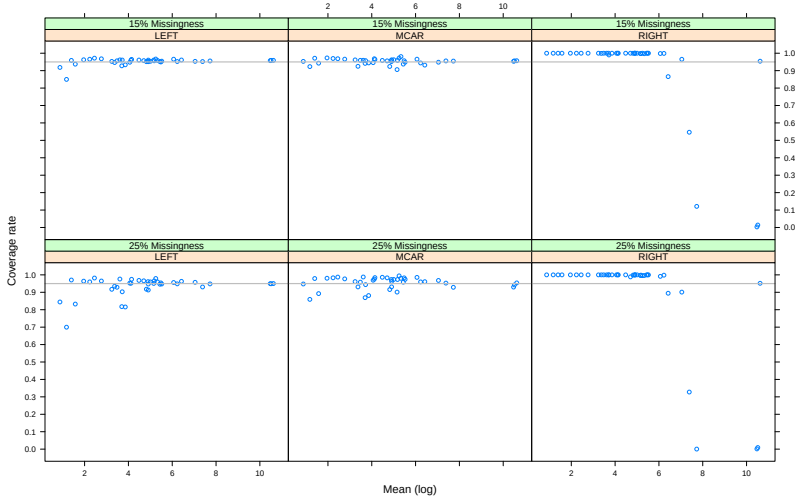

1.0
0.8
0.6
0.4
0.2
0.0
2000
1500
1000
500
0
a
e
g
h
e3
g5
h6
√x1

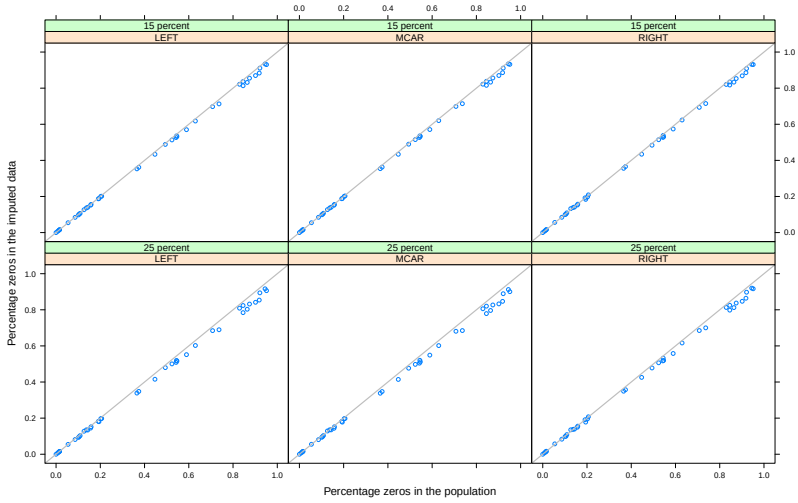
15 % Missingness (27.75 % bivariate MCAR)

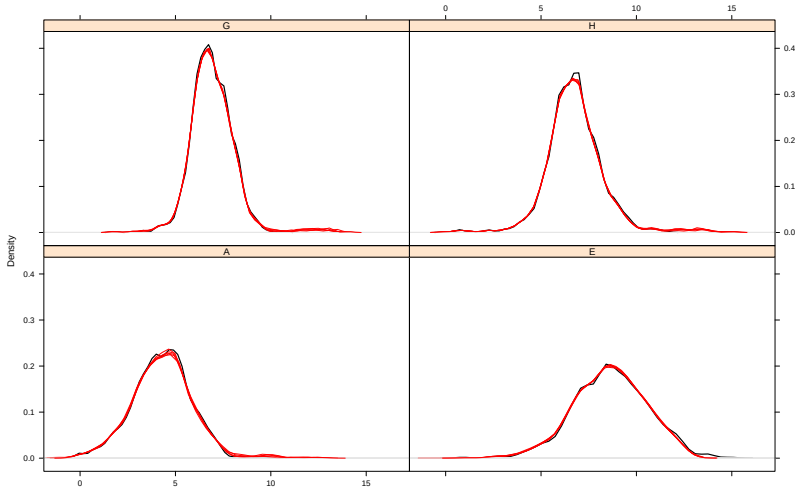


25 % Missingness (43.75 % bivariate MCAR)




15% Missingness
LEFT
15% Missingness
MCAR
15% Missingness
RIGHT
25% Missingness
LEFT
25% Missingness
MCAR
25% Missingness
RIGHT
Coverage rate
Mean (log)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
2
4
6
8
10


15 percent
LEFT
15 percent
MCAR
15 percent
RIGHT
25 percent
LEFT
25 percent
MCAR
25 percent
RIGHT
Percentage zeros in the imputed data
Percentage zeros in the population
0.0
0.2
0.4
0.6
0.8
1.0


G
H
A
E
Density
0
5
10
15
0.0
0.1
0.2
0.3
0.4

TO SUM UP / TO DO

PRM emerges as a very effective imputation approach for intricately nested compositional data

- Skewed semicontinuous (or zero-inflated) data
- No need for transformations of the data
- No need for post-hoc fixes to handle zeros
- Flexibility of mice-based approaches
- Works also when all components are missing

However, the top-level compositional total needs to be observed (or imputed beforehand)

A comparison with imputation approaches for single compositions needs to be made