PhD Thesis

Data Splitting and Its Applications

Vahid Nassiri

Response to the reviewers

We thank the reviewers for their critical assessment of our work. In the following we address their concerns point by point.

Reviewer 1

Prof. Helena Geys

Reviewer Point P 1.1 — Page 56: $b_{ij}(i = 1, .., 15, j = 1, \ldots, m_j)$ -¿ Are those totals correct?

Reply: We thank the reviewer to bring up this point. Indeed, the way we have displayed the totals could be misleading. In fact, i goes from 1 to 15, as we have 5 personality factors, each for mother, father, and child. Then j runs for various items in the 44-item BFI questionnaire correspond to each of these aspects. The number of items correspond to each personality aspect are not equal. Now, instead of m_j we show these by m_i as they corresponds to factor-role i. This is changed in the revised version as follows:

> b_{ij}'s $(i = 1, \ldots, 15, j = 1, \ldots, m_i)$ are the random effects (latent variables) each representing one factor-role and m_i is the number of items corresponds to factor-role i.

Reviewer Point P 1.2 —

- Page 65, +2: main concern in
- Page 80: eq. (ourgenpl)
- Page 133: iterative multiple outputation
- Page 186: preicisions
- Page 186: item 2: The MO correction

Reply: These are fixed in the revised version of the manuscript.

Reviewer Point P 1.3 — Update. For $m > M_0$ (maybe a bit confusing to use the index m again here?)

Reply: We agree with the reviewer that it could be confusing here, and the sentence is corrected as 'For sub-sampling size $m > M_0$' in the revised manuscript. In fact, m is is not just the index, it is still the sub-sampling size which is increased by one unit in every iteration.

Reviewer Point P 1.4 — Page 137 bottom: a 5x3 matrix is given and then columns are ordered and presented on top of page 138. However, the link with the indices of the ordered columns and the original matrix is not clear.

Reply: The matrix matrix on page 138 show the index corresponds to each element in the sorted matrix on page 137. We agree with the reviewer that currently this connection is not sufficiently clear, so we have changed it in the revised manuscript as follows:

> Order the columns of the matrix in the previous step. For example, the following matrix shows a permutation which rearranges columns of the matrix in the previous step in an ascending order:

Reviewer Point P 1.5 — Page 174, last sentence of first paragraph may need rephrasing?

Reply:

We agree with the reviewer on this point. This is changed as 'Therefore, each subject is only used in the sub-samples that the responses presented there are measured for it.' in the revised manuscript.

Reviewer Point P 1.6 — Page 177, sub-section 6.1.3 is mentioned twice but it is really 2 different paragraphs within that subsection?

Reply: We thank the reviewer for pointing this out. This is now corrected as follows:

> In this section the proposed idea will be explored and examined via two simulations studies. Sub-section 6.1.3 considers 1- joint modelling of linear mixed models, and 2- a joint model of ordinal data in a generalized linear mixed models context.

Reviewer 2

Prof. Katrijn Van Deun

Reviewer Point P 2.1 —

- P3, first line: add 'that' → when we realize THAT increasing....
- Top p4: replace "are" with 'can be done' in "and analyzing these different sub-samples are independent of each other"
- 2nd paragraph p4: add 'the' in front of 'MapReduce Methodology'
- P6, add 'the' in 'size of THE dataset makes it eligible ...'

- Bottom p7: remove 's' in 'breakS down'
- P13 top: replace 'its' by 'his/her own approach'
- P13 top: insert 'the' in 'Typically, with THE data splitting ...'
- P13, third bullet: remover 'of' in 'all of analysis results'
- P13, bottom: add 'and' in 'we will review different splitting approaches, AND also their appropriate combination rules'
- P18 replace 'that' by 'where' in 'An example in our motivation datasets WHERE random horizontal splitting can be beneficial is the Divorce in Flanders'
- P18, insert 's' in 'deign' → design
- P20, top: insert 'is' in 'it IS worth ...'
- P26, top: insert 'the' in 'with THE missing data issue'
- P43, replace 'that' by 'for which' in 'the ones FOR WHICH using our proposed methodology'
- P45, replace 'Big Five Inventory (BFI) is questionnaire' by 'THE Big Five Inventory (BFI) is A questionnaire'
- P55: replace personalty by personality
- P59, bottom: replace corresponds by corresponding
- P65, second line: replace is by iN in 'is this thesis'
- P80, above Equation 4.12, please fix the reference to Eq. in '(see also Eq. (ourgenpl)).'
- P 243, first line: insert 'the' in 'during the drug development process'
- P 246, replace 'created' by 'create' in 'can be used to created'

Reply: These are fixed in the revised version of the manuscript.

Reviewer Point P 2.2 — P3: something seems to miss with respect to explaining the predictor in the bullet point "Fitting a linear model to the generated sample as the predictor and y = 1 + 3x + _, where _ _ N(0, 0.01)."

Reply: We thank the reviewer to bring this up, indeed, the current sentence is vague. We meant the generated sample in the fist example as the predictor, but now it is changed as follows in the revised manuscript:

> Fitting a linear model to a random sample from standard normal as the predictor x, and $y = 1 + 3x + \epsilon$, where $\epsilon \sim N(0, 0.01)$.

Reviewer Point P 2.3 — P4: please check the sentence containing "but for others, it would even speeds down with a sub-linear rate" -¿ the 's' in speedS is grammatically not correct and you probably mean 'slow down'

Reply: Indeed, that was a mistake which is now corrected as follows:

> As we may see, for some analyses, the computation time increases linearly as the sample size increases, but for others, it would even slow down with a sub-linear rate.

Reviewer Point P 2.4 — Chapter 3, section 3.1., first paragraph: I miss a reference to the BFI

Reply: We totally agree with the reviewer. The revised manuscript includes necessary references now:

> Divorce in Flanders (Mortelmans et al., 2011) dataset presents data from Big Five Inventory (BFI) questionnaire (John and Srivastava, 1999; Denissen et al., 2008) answered by Flemish families.

Reviewer 3

Prof. Karl Meerbergen

Reviewer Point P 3.1 — Table 2.1 contains too few nonzero numbers to draw conclusions about the complexity of the faster methods. It may be better to use more digits behind the comma.

Reply: We thank the reviewer for point out to this issue. In the revised manuscript, bench-marking is done again and now the computation times are displayed in microseconds, so the comparison becomes possible also for faster methods.

Reviewer Point P 3.2 — The font in Figure 3.1 is too small for my eyes.

Reply: The font size is made larger in the revised manuscript.

Reviewer Point P 3.3 — I did not find a definition of the symbol J. Therefore, I could not verify the result of Equation 4.1 on page 75.

Reply: J indicates a matrix of ones. The revised manuscript now includes this information: 'where I indicates an identity matrix and J indicates a matrix of ones.'

Reviewer 4

Prof. Philippe Lemey

Reviewer Point P 4.1 — The PhD candidate may consider including a list of abbreviations, but this is not absolutely necessary, so I leave this to his discretion. I will send the PhD candidate an annotated pdf version of thesis that marks a number of typos that can be corrected.

Reply: The list of abbreviations are added to the revised manuscript. We also thank the reviewer for all his comments in the annotated PDF file. We have made sure all of the corrections are applied.

Reviewer 5

Prof. Peter Hoet

Reviewer Point P 5.1 — The manuscript form is not (clearly) in line with the instructions provided by the doctoral school. The manuscript is much longer than the suggested 150 pages, do not want to be too strict on this but 402 is more than two time exceeding. The candidate can consider e.g. whether chapter 7 is strictly needed (?) and/or to place the appendix after the bibliography (in fact the appendix is not a part of the manuscript).

Reply: Chapter 7 is actually based on two of our publication, do despite its title, it is indeed necessary. But we totally agree with the reviewer and move the appendix after the bibliography.

Reviewer Point P 5.2 — The final manuscript should contain a scientific abstract in Dutch and English and a short description of the professional background of the PhD researcher.

Reply: The revised manuscript now contains a scientific abstract both in Dutch and English as well as a short description of professional background.