Thesis Proposal: A blended distance to define "people-like-me"

Student: Anaïs Fopma
Student number: 6199356
Programme: Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Supervisors: Prof. dr. Stef van Buuren, Dr. Gerko Vink & Mingyang Cai
Word count: 663

1. Introduction

1.1. Growth curve modelling

Growth curve modelling is used to predict the future development of a child. It can provide answers to questions parents, physicians, and insurance companies may have concerning the future development of the child, the certainty of the future growth of the child, normal and healthy development, the certainty of prognosis, and effectiveness of treatment (Van Buuren, 2014). There are different approaches used in growth curve modelling. This study focuses on an approach termed curve matching.

1.2. Curve matching

Curve matching (Van Buuren 2014) is a nearest neighbour technique for individual prediction. The aim is to predict the growth of a target child by using the data of other, older children, of which we already have more data at a later age. To do this as accurately as possible, we want to use the data of a number of children (usually 5) that are most similar to the target child ("people like me"). In order to do this, we first need to match the children to the target child. The key question is: How do we obtain good matches? The current approach uses predictive mean matching. This means that we predict the values for all the donor children in the database and for the target child. The 5 donor children which have the closest predicted value are the best matches. Their growth curves are then plotted and point estimates can be calculated by averaging the measurements.

1.3 Historic similarity vs. future similarity

The current technique relies on using future similarity of the matches to the target child. Since different profiles may lead to the same predicted value, the values on separate predictors of some matches may be quite far from those of the target individual. Therefore, some users of curve matching question whether matches based on only future similarity are actually good matches, and whether historic similarity should be taken into account as well.

1.4. Research questions

The objective of this study is to answer this question by investigating the properties of a "blended distance" measure, which combines the future similarity and historic similarity. In order to find this out, the research questions to be answered are the following:

1. Does a higher blending factor result in increased similarity between target and matches on the observed predictors, as intended?
2. Is there a penalty from blending in terms of reduced predictability? In particular, is a blending penalty related to the dimensionality of X? What happens if y is unrelated to the first few principal components of X?
3. Can predictability ever exceed that of the unblended curve matching, e.g., if y is strongly related to the first principal component of X? Does pre-selection on similarity lead to increased standard errors in the regression coefficients of the analysis model, and what will be the effect the accuracy of predictions?
4. Is it possible to improve accuracy by boosting units that are more similar to the target unit (Efron & Hastie, 2016)?

2. Strategy

This study consists of simulation research and will report this according to the guidelines proposed by Morris, White, and Crowther (2019). Therefore, the planning of the study will follow the ADEMP structure, namely: Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures. An overview is provided below. In addition, empirical data collected by TNO will be used. R version 4.0.2 (2020-06-22) will be used to perform the analyses. Statistics in Medicine will be used as a reference journal for the format of the thesis.

2.1. Aims

See research questions.

2.2. Data-generating mechanisms

We want to see what the impact is of: a higher blending factor, if y is unrelated to the first few principal components of X, if y is strongly related to the first principal component of X, pre-selection on similarity, boosting units.

2.3. Estimands

Predicted growth of the child.

2.4. Methods

The models used by TNO?

2.5. Performance measures

- prediction accuracy
- standard errors in the regression coefficients of the analysis model
- Others?

3. References

1. Meigen C, Hermanussen M. Automatic analysis of longitudinal growth data on the website willi-will-wachsen. de. *Homo*. 2003;54(2):157-161.
2. Van Buuren S. Curve matching: a data-driven technique to improve individual prediction of childhood growth. *Annals of Nutrition and Metabolism*. 2014;65(2-3):227-233.
3. Ashworth A, Shrimpton R, Jamil K. Growth monitoring and promotion: review of evidence of impact. *Maternal & child nutrition*. 2008;4:86-117.
4. de Wilde JA, van Dommelen P, van Buuren S, Middelkoop BJ. Height of South Asian children in the Netherlands aged 0–20 years: secular trends and comparisons with current Asian Indian, Dutch and WHO references. *Annals of human biology*. 2015;42(1):38-44.
5. Straatmann VS, Pearce A, Hope S, et al. How well can poor child health and development be predicted by data collected in early childhood? *J Epidemiol Community Health*. 2018;72(12):1132-1140.
6. Berkey CS, Kent RL. Longitudinal principal components and non-linear regression models of early childhood growth. *Annals of human biology*. 1983;10(6):523-536.
7. Hauspie RC, Cameron N, Molinari L. *Methods in Human Growth Research*. Vol 39. Cambridge University Press; 2004.
8. Van Dommelen P, Van Buuren S. Methods to obtain referral criteria in growth monitoring. *Statistical methods in medical research*. 2014;23(4):369-389.
9. Hu Y, He X, Tao J, Shi N. Modeling and prediction of children's growth data via functional principal component analysis. *Science in China Series A: Mathematics*. 2009;52(6):1342-1350.
10. Zimmerman DL, Núñez-Antón V, Gregoire TG, et al. Parametric modelling of growth curve data: An overview. *Test*. 2001;10(1):1-73.
11. Ng K, Sun J, Hu J, Wang F. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits on Translational Science Proceedings*. 2015;2015:132.
12. Schönbeck Y, Van Dommelen P, HiraSing RA, Van Buuren S. Trend in height of Turkish and Moroccan children living in the Netherlands. *PLoS One*. 2015;10(5):e0124686.
13. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*. 2019;38(11):2074-2102. doi:10.1002/sim.8086
14. de Onis M, Onyango AW. WHO child growth standards. *The Lancet*. 2008;371(9608):204.
15. De Onis M, Onyango A, Borghi E, Siyam A, Blössner M, Lutter C. Worldwide implementation of the WHO child growth standards. *Public health nutrition*. 2012;15(9):1603-1610.
16. De Onis M, Wijnhoven TM, Onyango AW. Worldwide practices in child growth monitoring. *The Journal of pediatrics*. 2004;144(4):461-465.