Halpin, B. (2016)

Missingness and truncation in sequence data: A non-self-identical missing state

in G. Ritschard & M. Studer (eds), Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016, pp 443-444.

Missingness and truncation in sequence data: A non-self-identical missing state

Brendan Halpin

Abstract

Missingness in sequence data is a problem that has not received a great deal of attention. Since longitudinal data is more vulnerable to missingness than cross-sectional data, this is a lacuna. Two approaches are suggested in the literature: treating missingness as a separate state in its own right Gabadinho, Ritschard, Studer and Müller, 2009, and multiple imputation Halpin, 2015. The former is simple to implement but has shortcomings, not least in treating a missing–missing combination as a match; the latter is somewhat onerous. In this paper I propose a new concept: a non-self-identical missing state, where missing–missing combinations are treated as mismatches. I apply this to both normal missingness (gaps) and truncation of sequences (where entry or exit may happen at different times, yielding sequences of different length).

General missingness (gaps) and truncation are related but practically and conceptually different phenomena. If general missingness is treated as a special state, there is a risk that sequences will cluster according to how much missingness they contain.

Truncation is usually dealt with (if using the optimal matching distance measure) by deletion, but this may lead to sequences being disproportionately clustered according to length. In circumstances where sequences either begin or end on a specified event (e.g., begin with the first labour market participation spell, or end with retirement) the length of the sequence may be informative (e.g., late entry may be associated with higher education) and we may wish to differentiate between similarity represented in the length and similarity represented as matches between the observed portion. For instance, in both Halpin and Chan, 1998 and Bukodi, Goldthorpe, Halpin and Waller, 2016, class-career sequences are padded on the left by a “pre-entry” state to represent left-censoring, but (in the latter in particular) the con-

Brendan Halpin
University of Limerick, e-mail: brendan.halpin@ul.ie

cern is with how education affects later career: if we treat the pre-entry state as a non-self-identical missing, are the resulting distances more independent of sequence length than with OM-style deletion?

Using simulation and analysis of real data I show that a non-self-identical missing state can yield better results than the conventional approach for general missingness, but that locating a self-identical or non-self identical missing state in a maximally neutral location (as similar to all other states as is consistent with metricity) is also important. While multiple imputation functions better, it is more demanding in terms of setup and computation.

As regards truncation, OM's built-in capacity to compare sequences of different length is shown to work surprisingly well.

References

Bukodi, E., Goldthorpe, J. H., Halpin, B. & Waller, L. (2016). *Is education now class destiny? Class histories across three British birth cohorts*. Nuffield College, Oxford.

Gabadinho, A., Ritschard, G., Studer, M. & Müller, N. S. (2009). *Mining sequence data in R with the TraMineR package: a user's guide for version 1.2*. University of Geneva.

Halpin, B. (2015). *MICT: multiple imputation for categorical time-series* (Working Paper No. WP2015-02). Dept of Sociology, University of Limerick. Ireland. Retrieved from http://www.ul.ie/sociology/pubs/wp2015-02.pdf

Halpin, B. & Chan, T. W. (1998). Class careers as sequences: an optimal matching analysis of work-life histories. *European Sociological Review*, *14*(2).