RSiena: processing of missing and structural values

Ruth Ripley (with modifications by Tom Snijders)

November 4, 2014

1. Introduction

This document describes the processing of missing and structurally fixed values. It does not (yet!) include definitions of all the various missing data attributes used in RSiena.

Nothing yet about ML.

2. Outcome: see below for details

networks

1. Missing values are replaced by the most recent previous non-missing value, if any. Or zero.

behavior variables

1. Missing values are replaced by the most recent previous non-missing value, if any, or the earliest later one, if any, or by the mode for the wave.
2. Centering is done by subtracting the mean each time the value is used in a relevant effect.
3. No structurals yet.

constant and changing covariates

1. In siena07 and other functions relying on the C++ code (eg `sienaBayes`), missing values are replaced by 0, which is the mean of the centred data. Missing values are never altered in R.
2. Values are centered i.e. the ((global) mean is subtracted).

constant and changing dyadic covariates

1. Missing values are replaced by the (global) mean when preparing the data for input to C++.
2. Values are centered by *the relevant effect* subtracting the mean.

3. siena01 route

4. Object creation

4.1 networks: sienaDependent

1. For networks, all values must be missing or 0, 1, 10, 11

4.2 Data object

networks

1. Missing and structural values are excluded in calculation of distance between the waves.

behavior variables

1. No mean is calculated.
2. *moreThan2* treats missing values as a separate category, unlike for covariates.
3. *poszvar* always true if there are missing values. Otherwise only if not all values are the same.
4. Missing values are excluded in calculation of distance between the waves.

constant and changing covariates

1. Missing values are ignored in calculation of mean i.e. mean is average of non-missing values.
2. *moreThan2* ignores missing values.
3. *poszvar* always true if there are missing values. Otherwise only if not all values are the same.
4. Values are centered but NA's are never removed within R.
5. Range and similarity calculations ignore any missing values, excluding them from denominators as well.

dyadic covariates

1. Missing values are ignored in calculation of mean.

5. Reports

6. Input to C++: initializeFRAN

networks

1. Missing values are replaced by the most recent previous value if any. Otherwise by 0. *No carry back.*
2. Which values are treated as missing can be altered by composition change option. This may mean that a carried forward value is replaced by 0 later.

3. A new distance is calculated in R after these changes. Possibly is not used, but if it is used, its exact definition is not immediately apparent! See item below for details of composition change options. I think it is a hangover from days before the distance was calculated with the statistics.
4. Three edge lists are created, one for data which is to be used, one for missing indicators and one for indicators of structurally fixed values.

behavior variables Missing values are replaced by a previous value or a later value or the mode for the wave. Done in *unpackBehavior*, called by *initializeFRAN*.

constant and changing covariates When the data is read in C++: it is then zero'd if missing, and the fact that it is missing is stored. (Uses R ISNAN macro in C++, a convenient flag.)

dyadic covariates When making edgelists in preparation for C++, in the unpack functions called by*initializeFRAN*, NA's are replaced by the mean.

Composition change options

1 Never treated as missing, except in the initial report. Before the actor joins the network, row and column are set to 0 After the actor has joined, if later inactive, the previous value is carried forward.

2 Before the actor joins the network, row and column are set to 0 and the values are not treated as missing values. If temporarily missing after joining, the previous value is carried forward and this is not treated as a missing value. After finally leaving, any non-missing values in the data will be used, but they will be treated as missing values. If the data contained NA's then values are carried forward or zeroed as usual.

3 When inactive the values are treated as missing. If any values are present they will be used. If NA's are present then values are carried forward or zeroed as usual.

7. Processing in siena07

networks Each network has 3 network objects: one for the values and two boolean networks, one for missing indicators and one for structural indicators. In general, forward simulation does not concern itself with missing values until calculating statistics. Structural values are used alongside any composition change data when defining the flags indicating whether an actor is active or not.

behavior variables Centering is achieved when required by the relevant effect subtracting the overall mean. (Which is calculated after set up in C++, not R. Might be good idea to alter this as too easy to get reports and values used out of sync.)

constant and changing covariates The missing flags are used when calculating distance 2 covariate effects: if all the values which are being averaged are missing the result is treated as missing.

dyadic covariates The relevant effect subtracts the mean from the value each time the latter is requested.

Statistics See `StatisticCalculator.cpp`

1. We create *predictor* versions of the dependent variables which are used in any effect of which the dependent variable is not the owner. For networks these have all values missing at either end of the period set to zero. For behavior variables we only set to zero the values which are missing at the start of the period.
2. In the simulated values data for any actors who are inactive at the end of the period are set to the values at the start of the period so as not to affect the statistics. This is done in a routine named *setLeaverBack* but in fact it affects all actors inactive at the time, not just leavers during the period.
3. Then each dependent variable is processed in turn. The three types of effects are processed differently, and differently for networks and behavior variables:

   networks

   evaluation

   (a) In a copy of the current state of the owner dependent variable we make the following changes:

   missing Values missing at either end of the period are set to 0.

   structural There are two cases to consider: initial calculation of targets, and calculation of simulated statistics after simulation. In the former case we have used the data for the end of the period as the current state so we copy values structurally fixed at the start of the period from the data to the state. In the other case we copy values structurally fixed at the end to our state. (The same two changes are made in each case, but only one has any effect.)

   (b) Since a network is not used at the same time as a predictor and an owner we temporarily overwrite the predictor state's copy of this network with our changed copy so we can use the predictor state for all our calculations.

   endowment Here we create a *LostTieNetwork* and also overwrite the predictor network.

   lostTieNetwork

   (a) A copy is made of the initial state of the network, from the data

   (b) In a copy of the current state of the owner dependent variable, any values that are structurally fixed are changed as for evaluation effects.

   (c) Any values which are 1 in our changed copy of the current state are set to zero in the lost tie network.

   (d) Any values missing at the end of the period are set to zero in the lost tie network. ?? why not start too?

   predictor network(a) Values missing at the start of the period are set to zero in a copy of the current state of the network. (Values missing at the end are not changed).

   (b) Since a network is not used at the same time as a predictor and an owner we temporarily overwrite the predictor state's copy of this network with our changed copy so we can use the predictor state for all our calculations.

creation Here we create a *GainedTieNetwork* and also overwrite the predictor network as for evaluation effects.

predictor network A changed copy of the current state is created as for evaluation effects.

GainedTieNetwork A copy is made of our altered predictor network, and any values missing or 1 at the start of the period are set to zero.

behavior (a) A copy of the current state of values (less overall mean) is made and any values missing at start or end of the period are set to zero.

(b) A vector of differences, start of period value less (unaltered) current state is created. Entries missing at either end of the period are set to zero.

(c) Evaluation effects are calculated using the altered current centred values.

(d) Endowment and creation effects are calculated using the altered current centred values and the differences.