SOCPROG: PROGRAMS FOR ANALYZING SOCIAL STRUCTURE

Written by Hal Whitehead
Department of Biology
Dalhousie University
Halifax, Nova Scotia B3H 4J1
CANADA
hwhitehe@dal.ca

May 2017

SOCPROG2.8 (for MATLAB 9.1.0, release 2016b)

Contents

1 PREFACE

The set of programs that I call SOCPROG has had a punctuated evolution. I first produced an integrated set of programs for the analysis of social structure (SOCPROG 1.0) in 1997 using MATLAB4.2 and the MATLAB Statistics toolbox. In 1999 (SOCPROG1.3), I added population and movement modules. In 2004 SOCPROG was almost completely rewritten and restructured with a new set of programs (SOCPROG2.0) for the considerable improvements then available in MATLAB6.5. Later in 2004, SOCPROG2.1 was made available, with a compiled version so that users do not need to have MATLAB itself. In 2008, I made several improvements, in particular adding network analyses (SOCPROG2.3). SOCPROG2.4 (2009) used MATLAB7.7, fixed a few bugs and added better handling and analyses of interaction data. The latest version (SOCPROG2.8) uses MATLAB9.1 (it seems to work well on MATLAB8.4, but will have problems with MATLAB8.3 and earlier releases, as MATLAB substantially changed the way graphics is programmed in version 8.4), fixes a few bugs in SOCPROG2.5-2.7, and adds functionality.

Let me know if you find bugs or other problems and I will see what I can do.

With the compiled version, you do not need MATLAB to run SOCPROG. You will not have the same flexibility with the compiled version but most aspects of SOCPROG seem functional. See 2.5 for information on the compiled version.

A number of researchers have tried SOCPROG. The most common problems seem to have been:

- not having the MATLAB Statistics Toolbox installed or available;
- MATLAB is case-sensitive, so it is important to type in names of variables, etc., with the same combination of upper and lower-case letters in all parts of the analysis.

1.1 For more detail and background on analysis of animal social structure

See:
Whitehead, H. 2008. *Analyzing animal societies: quantitative methods for vertebrate social analysis*. University of Chicago Press. 320pp.

1.2 Please cite SOCPROG as:

Whitehead, H. 2009. SOCPROG programs: analyzing animal social structures. *Behavioral Ecology and Sociobiology* 63: 765-778.

2 INTRODUCTION

2.1 Purpose

The purpose of these programs is to help those with data on the associations of identified individual animals to analyze their data and develop models of social structure, population structure and movement. The programs are written in MATLAB.

2.2 Starting out: Master SOCPROG window

After starting MATLAB:

- Set the 'Current directory' on the MATLAB toolbar to the directory where you have stored the SOCPROG files.
- In the main command window type 'socprog' and then press the return key. A 'master' SOCPROG window should appear with a number of pushbutton options allowing you to access different parts of SOCPROG.

The compiled version starts differently. See 2.5.

2.2.1 *Master SOCPROG window*

The master SOCPROG window is arranged so that you usually work down it, entering data first, then setting sampling periods, associations and/or restrictions, and finally carrying out social, population or movement analyses. Elements of SOCPROG2 are shown as 'pushbuttons' on this window.

At the bottom, centre, is a pushbutton to allow you to change the default location where SOCPROG looks for, or saves, files. When you input data, the location you choose automatically updates this location.

2.3 Some potentially useful information about using MATLAB

- **Ctrl-C** aborts operation, although this may not happen immediately or ever (!)
- MATLAB is **case-sensitive** so get your capitals in the right place!
- You can alter graphs and other displays interactively (see MATLAB Manual or Help) or using the command line. For instance:

  You can add or change axis labels, or titles, on graphs by typing (in the command window) something like:

  xlabel('This is the x-axis')
  ylabel('This is the y-axis')
  title('My first graph')
- If you have a legend on a figure you can move it using a mouse (click and drag). You can also change the legend interactively on the figure or by typing the **legend** command in the command window. e.g.: **legend('Females','Males','Male-Female')**
- MATLAB and SOCPROG often treat *NaN* ('Not a Number') as a missing value

2.3.1 *MATLAB platforms*

The programs were written using the Windows operating system. They have not been tested on others (e.g., Mac, Linux, UNIX) but I have heard informally that most parts work reasonably well. The compiled version will only work on Windows operating systems (but may

have difficulties on Windows systems of much earlier or later vintage than the date the programs were compiled).

2.4 Test data sets

Included with SOCPROG2 are several test data sets in Excel format:

- 'simlina.xls': primary data file in *linear* mode with alphanumeric identification codes
- 'simlinn.xls': primary data file in *linear* mode with numeric identification codes
- 'simdyaa.xls': primary data file in *dyadic* mode with alphanumeric identification codes
- 'simdyan.xls': primary data file in *dyadic* mode with numeric identification codes
- 'simgrpa.xls': primary data file in *group* mode with alphanumeric identification codes
- 'simgrpn.xls': primary data file in *group* mode with numeric identification codes
- 'simsexa.xls': supplemental data file with alphanumeric identification codes
- 'simsexn.xls': supplemental data file with numeric identification codes
- 'simmovan.xls': primary data file of movements between study areas
- 'simmovco.xls': primary data file of movements in continuous 2-D space

Use simsexa.xls with simlina.xls, simdyaa.xls and simgrpa.xls, and simsexn.xls with simlinn.xls, simdyan.xls, simgrpn.xls, simmovco.xls.

2.5 Compiled SOCPROG

There is a compiled version of SOCPROG. The compiled version allows you to use SOCPROG as a stand-alone program without MATLAB. Click here for information on the compiled version.

3 INPUT OF DATA

The primary format of the observation data used by SOCPROG2 is the MS Excel worksheet (.xls or .xlsx). However, it is also possible to use data sets saved in SOCPROG1 format or use '.csv' (comma separated) ASCII files directly. So if you cannot deal with Excel, or Excel is causing problems, you can input your data as .csv files. You can save data input from Excel (or SOCPROG1 or .csv) in SOCPROG2 format and use and reuse them directly. This can save lots of time, as the restrictions, sampling periods and defined associations are also retained in the SOCPROG2 format, and the import is much faster.

You can also directly enter an association matrix from either Excel or an ASCII file (section 3.8), although in this case many fewer options for social analysis are available.

3.1 Data format in Excel

Some advice on coding data for social analysis is given by Whitehead (2008b, 71-79). You may wish to use the test data sets (2.4) as templates for setting up Excel worksheets for SOCPROG input.

3.1.1 *Primary data file*

The programs take data from Excel worksheets. The 'primary data file' contains lines, or records, each of which corresponds to an observation, either of an individual (*linear mode*), a dyad—i.e. two individuals—(*dyadic mode*) or a group of individuals (*group mode*) (Whitehead 2008b, 71-77). In *linear mode*, at each observation of an individual its presence is recorded with a variety of other information (e.g. date, time, position, behaviour, group identifier, quality of identification). In *dyadic mode*, each observation is of two individuals, usually interacting, whose identities are recorded together with a variety of other information (e.g. date, time, position, type of interaction). In *group mode*, as each group is observed its membership is recorded together with some ancillary information for the group (e.g. date, time, position, behaviour, duration of watch, group size). It is not necessary that all members of each real group are recorded, although a few analyses will not make sense if this is the case. In all three modes, the first line gives the names of the fields (pieces of information)–these names must not have spaces or punctuation in them. The last field notes the identity or identities of the individuals for that record. In *linear mode* there is only one name of an individual per record but in *group mode* there may be more than one, and in *dyadic mode* there are always two. In *group* and *dyadic modes* the names are separated by spaces. In *dyadic mode*, the order of the two individuals may be important (e.g. groomer and groomee). It is usual that one field (called 'Date') gives the date and time of the observation in any Excel date format (this also helps analyses). Interaction data are usually in *group* or *dyadic mode* (and it is strongly recommended that you do this) whereas association data are usually in *group* or *linear mode.*

Here are examples of primary data file worksheets in *linear mode, dyadic mode* and *group mode:*

Table 1. Linear mode primary data from Excel worksheet, with alphanumeric individual identities.

Date	Posn	Grp-code	ID
1/1/00 9:00	279.9	1	A1
1/1/00 9:00	279.7	1	I9
1/1/00 9:00	278.2	1	N14
1/1/00 9:00	280	1	O15
1/1/00 12:00	42.6	2	H8
1/1/00 12:00	40.3	2	K11
1/1/00 12:00	42	2	M13
1/1/00 12:00	41.1	2	T20
1/1/00 15:00	664	3	D4
1/1/00 15:00	663.6	3	G7
1/1/00 15:00	664	3	L12
1/1/00 15:00	664.8	3	Q17
1/1/00 15:00	663.6	3	S19
1/2/00 9:00	325	1	A1
1/2/00 9:00	325.9	1	I9

Table 2. Dyadic mode primary data from Excel worksheet, with numeric individual identities.

Date	Interaction	Ids
1/1/00 9:49	A	13 20
1/1/00 14:54	A	14 15
1/1/00 15:41	A	17 19
1/2/00 9:11	F	19 20
1/2/00 9:41	F	10 18
1/2/00 10:09	A	6 16
1/3/00 10:35	A	10 18
1/3/00 11:03	D	19 20
1/3/00 14:32	A	6 16
1/3/00 17:40	A	15 8
1/4/00 7:16	D	19 20
1/4/00 13:17	A	13 3
1/4/00 16:15	A	15 8
1/5/00 6:00	A	5 8
1/5/00 15:57	B	16 10
1/5/00 17:55	F	13 3
1/11/00 7:19	C	18 12
1/11/00 10:09	A	8 4
1/12/00 7:14	C	8 4
1/12/00 9:01	B	16 10

Table 3. Group mode primary data from Excel worksheet, with numeric individual identities.

Date	Place	Behaviour	Group
1/1/00 9:49	A	2	8 11 13 20
1/1/00 14:54	A	1	1 9 14 15
1/1/00 15:41	A	2	4 7 12 17 19
1/2/00 9:11	B	1	4 7 12 17 19 20
1/2/00 9:41	B	1	2 10 18
1/2/00 10:09	A	3	3 5 6 16
1/3/00 10:35	A	5	2 10 18
1/3/00 11:03	A	3	4 7 12 17 19 20
1/3/00 14:32	A	2	5 6 16
1/3/00 17:40	A	1	1 9 14 15 8
1/4/00 7:16	A	2	4 7 12 17 19 20
1/4/00 13:17	A	2	11 13 3
1/4/00 16:15	A	2	1 9 14 15 8
1/5/00 6:00	A	2	1 9 14 15 8
1/5/00 15:57	B	2	5 6 16 10
1/5/00 17:55	B	2	11 13 3
1/11/00 7:19	C	2	2 18 12
1/11/00 10:09	A	2	1 9 14 15 8 4
1/12/00 7:14	C	2	1 14 15 8 4
1/12/00 9:01	B	2	5 6 16 10

Identities may be alphanumeric (as in Table 1) or numeric (as in

Table 2 and Table 3). Numeric identities should be positive integers (not zero, fractions or negative). Field data can be alphanumeric (as in 'Place' in Table 3) or numeric (as in 'Behaviour' in Table 3).

3.1.2 *Supplemental data*

You can also input supplemental data about individuals, such as sex, age, or genetic data (e.g. haplotype) from another Excel worksheet (Whitehead 2008b, 77-79). The first field should be the names of the individuals, as in the data file, and the first row should give the names of the supplemental data fields (Table 4). Please don't use the same field names in both the data file and supplemental data (except maybe 'ID').

Table 4. Supplemental data from Excel worksheet, with numeric individual identities.

ID	Sex	Age
1	M	15.5
2	M	2.7
3	F	5.8
4	M	14.5
5	M	20.8
6	F	9.7
7	F	7.4
8	F	24.6
9	M	6.1
10	F	17.2
11	M	11.7
12	M	17.7
13	F	11.7
14	M	4.0
15	M	15.7
16	F	0.3

If data are missing (e.g. No behaviour type recorded for one group; or sex unknown for an individual), I suggest that you use blank cells for alphanumeric fields and **NaN** ('Not-a-number' to MATLAB) for numeric fields.

3.1.3 *Note: Alphanumeric and numeric data*

Fields which are a combination of numeric and alphanumeric data can cause SOCPROG problems. Avoid if possible. Sometimes an Excel field which looks to be numeric is partly alphanumeric. If you add zero to the field, put it in another field and convert it to 'Values' this can sometimes solve the problem.

3.2 How to input, examine and save Excel data

After checking that the Excel worksheets are in the correct format:

- Click on: ‘Input data’ in the master SOCPROG window.
- Click on ‘Excel or .csv data’ on the drop-down menu in the small ‘Data input’ window that appears (you can also choose whether to see a summary of the data, or the raw data, to check that the data have been input correctly), and press ‘Go’ .
- Choose the Excel file (or .csv file, see 3.5) with the primary data in the interactive box.
- If there is more than one worksheet in the Excel file, choose the one with the primary data from a list that appears in a new window.
- If data are in *group* or *dyadic mode*, then a dialog box appears asking whether you wish to convert the data to *linear mode*. This may be helpful in case where you have coded asymmetric association/interaction and is necessary if you are doing movement analyses. However, if you wish to examine interaction rates leave the data in *dyadic* or *group mode*. This option is discussed further in 3.3.1.
- A dialog box asks if there are supplemental data. If yes, you must choose the Excel file (and worksheet, if there is more than one in the file) or .csv file (see 3.5) where the supplemental data are stored.

  Note: If there are no individual names common to both the main and supplemental files, then a message to this effect appears in the command window, and SOCPROG proceeds as though no supplemental file was used.

3.3 Viewing and saving the data

A summary window then appears (as long as you did not clear the checkbox in the ‘Input data’ window) with the following information:

- primary data file name
- data mode: *linear*, *dyadic* or *group*
- number of lines, or records in the primary data file
- list of primary data file fields, with first and last (alphanumeric or date) or max and min (numeric) values
- number of individuals, with first and last (alphanumeric names) or min and max (numeric names)
- if supplemental data file used, list of supplemental data file fields, with first and last (alphanumeric) or max and min (numeric)

For each primary and supplemental field, and the individual names, there is a pushbutton at the right. Clicking on this gives more details in the command window: levels of fields (integer values for numeric; days for dates) and number of records or individuals with each level; information on each individual in main data file. In these summaries, a lack of numeric data is indicated by ‘NaN’, and a lack of alphanumeric data by a blank.

If the data look correct, save them as a SOCPROG2 (‘.mat’) file using the ‘Save’ button at the top right of the summary window.

I STRONGLY recommend that you use the information in the summary window, and appearing in the command window after using the pushbuttons, to check that the data were input correctly.

If in doubt, view the raw data, by checking the ‘View raw data’ checkbox on the ‘Input data’ window, or from the ‘Raw data’ pushbutton in the ‘Sampling period, restrictions,

associations' window. When viewing the data, formats may be different from the Excel sheet, and extra fields will be added (3.7).

3.3.1 *Converting group or dyadic mode data to linear mode data*

If inputting *group* or *dyadic mode* data from an Excel file, you can convert the data to *linear mode* by pressing 'Yes' in the dialog box that appears after you name the primary data file. (Do not do this if you wish to calculate interaction rates). The data will then be converted as though they were read from a file with just one individual per row, and two additional primary fields are added:

- *Line*: the line on which the individual appeared in the original Excel data file (i.e. the group), and
- *Order*: the order in which the individual was named (1, 2, 3, ...) on the line in the original Excel data file.

Thus the first two lines of the *group mode* data in Table 3 would appear as in Table 5.

Table 5. Result of converting group mode data to linear mode.

Date	Place	Behaviour	Line	Order	ID
1/1/00 9:49	A	2	1	1	8
1/1/00 9:49	A	2	1	2	11
1/1/00 9:49	A	2	1	3	13
1/1/00 9:49	A	2	1	4	20
1/1/00 14:54	A	1	2	1	1
1/1/00 14:54	A	1	2	2	9
1/1/00 14:54	A	1	2	3	14
1/1/00 14:54	A	1	2	4	15
...					

The conversion itself will take a little time with large data sets, and, with a *linear* mode conversion, producing associations and setting restrictions may also be a little more time consuming. However, the conversion has advantages:

- You can now do analyses of continuous movement (12.2).
- This is a way to code asymmetric associations/interactions (but generally not as easy as using the path for analyzing interactions 6.2). For instance for dyadic interactions, on each line the first ID is of the actor and the second the reactor. Then in defining association (4.3), the association may be defined by *Line* weighted as a function of *Order* (see 4.3.1.3).
- You can also use this option to code for a proximity measure of association. If individuals are arranged in one-dimensional groups (e.g. when roosting along a branch, or swimming in ranks), and are entered into the Excel file in this order, then association may be defined by the minimum difference of a combination of *Line* and *Order* (see 4.3.2).

The option of converting *group* or *dyadic mode* data to *linear mode* data is only available with Excel or .csv data sets.

3.3.2 *Problems entering Excel files*

MATLAB's routine for reading Excel files is not perfect. If you are having problems (error messages, data read in wrong), you can try the following:

- Check the header line, that there are names for all fields except that for the ID's
- Check all fields end in the same record
- Check no unintentionally blank cells
- Remove all blank lines at the end of the worksheet
- Convert dates and times to numbers in Excel (right click on date/time field, choose 'Format Cells…', and then 'Number' and 'OK')
- Check for fields which combine both numerical and alphanumerical data (3.1.3)
- Try saving as a different (perhaps earlier) version of Excel
- View the raw input data, by checking the 'View raw data' checkbox on the 'Input data' window, or from the 'Raw data' pushbutton in the 'Sampling period, restrictions, associations' window
- If problems continue, export Excel data to a .csv file and read it in (3.5)

3.4 How to input SOCPROG1 files

If you have data entered (from ASCII files) using older (1.*) versions of SOCPROG and saved as '.mat' files, these can be converted to SOCPROG 2. Click on 'SOCPROG1 file' in the drop-down menu in the 'Data input' window, and press 'Go'. Choose the '.mat' file with the SOCPROG 1 data in the interactive box. The summary window comes up, if you did not clear the checkbox in the 'Data input' window. You can, and should, check the data have been converted correctly, as described for Excel data in 3.2. Then save the data in SOCPROG 2('.mat') format using the Save button at the top right of the summary window. [This is a relic of the earliest versions of SOCPROG2 and I am not entirely sure that this option is still functional!]

3.5 How to input .csv files

SOCPROG can also use primary and supplemental data files input in comma-separated, .csv, format. You can produce such files from programs like Excel simply by saving a worksheet in .csv format. The data should be set up as for the Excel files (Table 1 to Table 4) with commas separating the fields (there should be no commas within the fields). Thus the primary *group mode* data of Table 3 becomes:

```
Date,Place,Behaviour,Group
1/1/00 9:49,A,2 ,   8   11   13   20
1/1/00 14:54,A,1 ,   1    9   14   15
1/1/00 15:41,A,2 ,   4    7   12   17   19
1/2/00 9:11,B,1 ,   4    7   12   17   19   20
1/2/00 9:41,B,1 ,   2   10   18
1/2/00 10:09,A,3 ,   3    5    6   16
1/3/00 10:35,A,5 ,   2   10   18
1/3/00 11:03,A,3 ,   4    7   12   17   19   20
```

```
1/3/00 14:32,A,2 ,   5    6   16
1/3/00 17:40,A,1 ,   1    9   14   15    8
1/4/00 7:16,A,2 ,   4    7   12   17   19   20
1/4/00 13:17,A,2 ,  11   13    3
1/4/00 16:15,A,2 ,   1    9   14   15    8
1/5/00 6:00,A,2 ,   1    9   14   15    8
1/5/00 15:57,B,2 ,   5    6   16   10
1/5/00 17:55,B,2 ,  11   13    3
1/11/00 7:19,C,2 ,   2   18   12
1/11/00 10:09,A,2 ,   1    9   14   15    8    4
1/12/00 7:14,C,2 ,   1   14   15    8    4
1/12/00 9:01,B,2 ,   5    6   16   10
```

Note that the identities in group or dyadic mode are separated by spaces, not commas, and there is no comma at the end of each line.

To enter primary data in .csv format, proceed as with Excel data, except on the file selection window, choose the '(*.csv)' option (rather than the default '(*.xls)', or '(.xlsx)') under 'Files of type:' at the bottom of the window, and then select your .csv file (this may be a little different with non-Windows operating systems).

Supplementary data can also be entered as a .csv file in the same manner. You can mix an Excel primary data file with a .csv supplemental data file, or vice versa.

I am told that the French version of Excel separates items with a semi-colon (";") rather than comma (","). These will need to be changed to input the file into SOCPROG by globally replacing the semicolons with commas in the .csv file.

3.6 How to input SOCPROG2 files

If you have already input data either from Excel, .csv or SOCPROG 1 files and saved them (sections 3.2 and 3.4), you can use them again simply by clicking on: 'SOCPROG 2 file' on the drop-down menu in the 'Data input' window, pressing 'Go', and then choosing the '.mat' file which you previously saved in the interactive box. The summary window (see 3.2) comes up again (for the restricted data if restrictions have previously been set, 4.2), if you did not clear the checkbox in the 'Data input' window.

3.7 Additional fields

SOCPROG adds some additional primary data fields to those you have entered (as long as you do not already have fields with these names):

Record the record number in the Excel file, although this is AFTER sorting by date and time, if you have entered *Date*

Day Julian day from 1 Jan 1900, only added if *Date* is a field

Month calendar month, only added if *Date* is a field

Year only added if *Date* is a field

Hour time of day if *Date* is a field

Sample sample number; added when you have set the sampling period (4.1)

Line and *Order* are also added if *group mode* data were converted to *linear mode* (3.3.1) SOCPROG also adds some additional supplemental data fields to those you have entered (as long as you do not already have fields with these names):

ID, a string (alphanumeric) field with the individuals' names (converted from numeric if necessary)

Numrec, the number of records containing an individual in the unrestricted data set

Numsamp, the number of samples containing an individual in the data set; added when you have set the sampling period, (4.1)

These fields may be useful in restricting analyses to certain years, only those individuals seen more than a certain number of times, etc. (see 4.2), and for other purposes.

3.8 Inputting association matrix directly

Association measures (which can be association indices calculated outside SOCPROG, measures of kinship, range overlap, or ...) can be entered directly from Excel or an ASCII file using this option. In Excel, the association measure should look like that in Table 7, with the names of the individuals in the first row and column. In an ASCII (.txt) file, the association measure should look like that in Table 8, with the names of the individuals separated by spaces in the first row, and in each other row a name followed by the values of the association measure. Missing association measures can be represented by 'NaN'. Association values and names should be separated by spaces. For either option, you then enter the multi-line description, and a short name of the association measure (no spaces in the short name!), in a dialog box, and the 'analyzing multiple association measures' screen comes up (chapter 8). To make sure that the Excel or ASCII file has been correctly entered, I advise that you display it after it has been entered (8.2.5).

4 SETTING SAMPLING PERIOD, RESTRICTIONS, ASSOCIATIONS

After data input, this is the module that you will generally use next. You will need to set a sampling period for most analyses; association must be defined for any analysis of social structure which is not based upon interactions; and the ability to restrict the data is extremely useful.

There is a pushbutton in the upper middle of the master SOCPROG window which takes you to a second window allowing you to set the sampling period, make restrictions on the data and define associations. At the top left of the window are lists of the primary and supplemental data fields. Fields which are string (equivalent to alphanumeric) are marked by 's' in the lists.

In this window, and other windows, to enter names of primary or supplemental fields into editable boxes, select the primary or supplemental field from the popup list and press the 's' button beside the box.

At the top of this window are pushbuttons which allow you to close the window, save the data including sampling periods, associations and restrictions (if set) as a SOCPROG2 .mat file, view the data with ('Restricted') or without ('Raw data') restrictions (if set) in the command window, or view the summary ('Summary') of the data (with restrictions, if set). When viewing the data, formats may be different from the Excel sheet, and extra fields will be added (3.7).

4.1 Setting sampling period

The sampling period is the period of time within which associations are examined (or, for the population and movement analyses, within which identifications are binned). It must be an expression based on one (or more) of the primary data fields in the data matrix, but these fields/variables must be numeric not string (equivalent to alphanumeric, marked by 's' in the list on the left of the window). Integer parts of the expression you use define different sampling periods. Some examples are:

Year <- Annual sampling periods
Day <- Daily sampling periods
***Date*/10** <- 10-day sampling periods
***Date**8** <- 3-hour sampling periods
***Record*/100** <- Groups of 100 records
***Year**4+(*Month*>3+*Month*>6+*Month*>9)** <- Years and Seasons

There is an editable string on the top right of the sampling-restrictions-associations window in which you can enter an expression for the sampling period. The defaults are *Day* if either *Date* or *Day* is a primary data field, and otherwise the first primary data field. Press 'Set' or 'Reset' to set the sampling period to that given in the editable string. If association has been calculated (4.3), it is recalculated if you press 'Reset' for the sampling period.

4.2 Setting restrictions

You may want to carry out analyses on only certain portions of your data set. Restrictions are set at the bottom left of the sampling-restrictions-associations window. Restrictions can be made on records, individuals or combinations of these (4.2.1), using the two editable strings.

On the left, you can enter a string representing a MATLAB expression using primary data fields which restricts the records that are used in subsequent analyses. e.g. Enter:

***Year*<1992** <- 1992 and subsequent years are omitted
(*Month*>4)&(*Month*<8) <- May-July only
~((*Sample*>1)&(*Sample*<8)) <- just sample 1, and from 8 on
strcmp(*Behaviour*,'G') <- just when *Behaviour* is 'G'
***Record*>100** <- just records 101 and on

Similarly, on the right you can make restrictions using the supplemental fields (attributes of individuals) which restricts the individuals that are used in subsequent analyses. e.g. enter:

~strcmp(*Sex*,'F') <- no females
strncmp(*ID*,'Q',1) <- just individuals for whom the first character of the ID is 'Q'
(*Numrec*>20)'&strcmp(*Sex*,'M') <- males with at least 21 records

Useful relational operators for numeric fields are:

'<', '>', '<=' (≤), '>= ' (≥), '==' (equal to), '~=' (not equal to).

You cannot use these operators for string fields; instead use 'strcmp' (compare strings), 'strncmp' (compare first n elements of strings), etc.

The '&' (and), '|' (or), and '~' (not) logical operators are very useful for combining or modifying restrictions.

Number variables should be compared with numerical constants, or other number variables, and string variables (those with '(s)' after them) with string constants (enclosed in single quotes, e.g. 'F') or other string variables using 'strcmp' or 'strncmp'. String variables can

be converted into number variables using the operator 'str2double' and numbers into strings using 'num2str' (although you should not need to do these often). e.g.

str2double(*ID*)>100 <- just ID numbers greater than 100 when ID's can be interpreted as numbers

Variables can be combined in different ways. e.g.

(*Year*+*Month*/12)>90.5 <- after June 1990

Notes:

i) If you combine string and numerical expressions in the same restriction, you must put a single quote (representing transpose) after the numerical parts (e.g. **(*Numrec*>20)'&strcmp(*Sex*,'M'))**; otherwise you may get the following error message:

??? Error using ==> &
Array dimensions must match.

ii) If you restrict on *Numsamp* or *Numrec* this works only on the number of samples or records in the unrestricted data set. However, if you view the restricted data you will see the number of samples or records in the restricted data set.

4.2.1 *Complex restrictions*

If data are in *linear mode* (*group mode* data can be converted to *linear mode*, 3.3.1), then it is possible to make restrictions combining primary data fields (on records) and supplemental data fields (on individuals), using the lefthand editable string (on records). For instance, with data collected over several years on individuals of known age, if there is a supplemental field, *DOB*, giving the date of birth of each individual, and the primary field *Date* indicates when information were collected, then, to use only individuals before their 10th birthday, type:

***Date*<*DOB*+10*365**

Generally then, if data are in *linear mode*, you can use supplemental field names in the restrictions on primary data fields. Remember to compare numbers with numbers and strings with strings.

4.3 Defining associations

Association must be defined for many SOCPROG analyses of social structure, as well as the estimation of mortality using social structure (10). Association is defined on the upper right of the sampling-restrictions-associations window, beneath the sampling period options.

An association is a number for each pair of individuals in each sampling period. Normally the association is one or zero (observed associated: not observed associated) but associations can be quantitative (e.g. number of minutes associated; see below). Most associations are symmetric (i.e. if A and B are associated in a sampling period then so are B and A), although this is not necessarily the case, for instance with nearest-neighbour (13.2) measures. A part of the association definition and calculation is the self-association of each individual with itself during the sampling period. This is usually '1' if the individual was identified during the sampling period, '0' if it was not, but quantitative measures are also appropriate (number of minutes studied in sampling period) in some cases. So, for each sampling period there is an n x n matrix of associations, where n is the number of individuals, with self-associations along the diagonal. This

matrix is usually, but not always, 1:0, and usually, but not always, symmetric.

To calculate the associations, press the 'Set'/'Reset' button half way up on the right of the sampling-restrictions-associations window. If associations have been set and calculated, they are recalculated automatically if you change the sampling period or restrictions.

There are several possibilities for defining association type which fall into four main categories, that you select using the pop-up menu beneath the sampling period options:

4.3.1 *Group variable*

The data set is divided into groups, using a group variable. All individuals with the same value of this variable (rounded down to the nearest integer) in a particular sampling interval are considered in the same group during that sampling interval. Group variables must be expressions based on one (or more) of the primary data fields, but these fields/variables must be numeric, not string (equivalent to alphanumeric). They could be of the form:

Record <- each record is a group (the default in *group mode*)
Line <- each original line of Excel *group mode* data is a group (default when *group mode* converted to *linear mode*; 3.3.1)
Groupnumber <- numbered groups
Hour <- individuals sighted in each hour of the day are considered grouped
floor(*Hour4)** <- individuals sighted in each 15 minutes are considered grouped

The group variable is entered in the editable text box half way up on the right.

If you choose the group variable form of association then you have three choices for association type (selected using the pop-up menu beneath the editable text box for the group variable).

4.3.1.1 Grouped in sampling period

Individuals are considered associated (association=1) in a sampling period if they were found at least once in the same group during the sampling period, and not associated (association=0) if they were never seen in the same group during the period. This is the default and should probably generally be used.

4.3.1.2 Number of groups in sampling period

Individuals are considered associated by the number of times they were seen grouped in a sampling period (so the association is the number of groups they were both seen in). This may make sense if observations of groups were independent.

4.3.1.3 Weighted by...

This is the same as 'No. groups in sampling period' except that each group is weighted by a function of the primary data fields evaluated for the first individual in the dyad which you enter in a box which appears. This can be useful in a number of situations.

For instance, if you wish to weight association by the number of minutes that a group was watched, select this option, and choose, as the weighting variable, a variable which contains the watch duration for each group. In this case association in a sampling period is the number of minutes that two individuals were observed together in a sampling period. e.g.:

Weighting variable = ***Watchduration***

If the data are from asymmetric behavioural measures, then the weighting option can be used to make an asymmetric association measure. So if the data were in *group mode*, converted to *linear mode* (3.3.1) with the first individual on each line being the actor and the second being the receiver, then the *Order* variable can be used so that association of a dyad within a sampling period gives the number of times the first individual was the actor in an interaction between the pair (i.e. when *Order*=1 for the first individual).

Weighting variable = **(*Order*==1)**

4.3.2 *Minimum difference in attribute*

With this option you set an attribute, which is one (or a combination) of the primary data fields, and two individuals are associated in a sampling period if the difference between the values of this attribute is less than some minimum value for some pair of observations in the sampling period. For instance,

Attribute: ***Hour***

Minimum difference: **0.25**

defines association such that two individuals are associated if they were identified at least once within 15 minutes of each other, during the sampling period.

As another example, if data on the ordering of individuals in one-dimensional groups (such as roosts on branches) are available and have been converted from *group mode* data to *linear mode* data (3.3.1), then the following settings allow association to be defined as 'side-by-side' (assuming there are fewer than 100 individuals in any group):

Attribute: ***Line**100+*Order***

Minimum difference: **1**

4.3.3 *More complex associations*

This allows you to define association in a variety of ways by referring to a MATLAB script file (m-file) which defines associations between individuals in each sampling period. This file uses as input any of the primary data fields, and the scalar *nid* (number of individuals in a sampling period, set by the program) to produce a *nid* x *nid* matrix *qq* defining associations, with effort for each individual on the diagonal. It also uses the array *rl*{1:*nid*} (set by the program) which gives a list of records for each individual in the sampling period under consideration. In the m-file you can define a string, *assstr*, which the program will use to describe the association type.

Examples of such m-files (for a continuous measure of temporal association, a nearest-neighbour analysis, and a measure, and an association measure standardized by effort on focal individuals) are given in the Appendix (13).

If you use this option, you should be careful when choosing an association index (6.1).

4.3.4 *Distance-based associations*

Use this option when you want the relative locations of the identifications of the individuals to define association. So, individuals are considered associated during any sampling period if they were observed within a certain range (and maybe time) of one another, and not associated if there were no pairs of identifications of the two individuals this close to one another.

There are seven settings:

x-variable: Name of x-variable of location, usually something like ***x*** or ***Lat***

y-variable: Name of y-variable of location, usually something like ***y*** or ***Long***

Time: Name of time variable, often ***Date***

Range: Choose either 'Euclidean', which is suitable for x- and y-coordinates in the same units, such as metres, or 'Rhumb line' in which the x- and y-variables are assumed to be degrees and fractions of a degree of latitude and longitude respectively, and the curvature of the Earth is taken into account when calculating the distance between identifications.

Association if distance <: Give the maximum distance between two identifications for them to be considered associated.

Association if time <: Give the maximum time between two identifications for them to be considered associated. This time lag is in the same units as the time variable, days if ***Date*** is used. So enter ***3/(60*24)*** is you only consider associations for pairs of identifications within 3 minutes of one another.

Distance multiplier: This relates the maximum distance to the units of the x- and y-variables. So, if the x- and y- variables are in metres, putting ***1*** here means that the maximum range for association is the number of metres entered in 'Association if distance<'. If you use latitudes and longitudes for your x- and y- variables, then, for different distance measures, enter here:

1 for degrees of latitude
60 for nautical miles (minutes of latitude)
111.12 for kilometres (the default if you use latitudes and longitudes)
11112 for metres

Leave the y-variable blank if you only have one-dimensional data.

Leave the time variable blank if you do not care when the individuals were identified within a sampling period, just where they were identified (the default).

5 BASIC DATA

Click on this pushbutton to get some basic information on your data set. You must have set a sampling period first (4.1). Then a selection screen with checkboxes appears allowing you to obtain information on:

- Number of individuals, identifications, sampling periods, ..., given defined sampling periods and restrictions.
- Number of individuals identified in each sampling period.
- A 'discovery' curve plotting the cumulative number of individuals identified against the cumulative number of identifications (with only one identification of each individual counted during each sampling period). This indicates the rate at which individuals enter the data set, and so what proportion of the population have been identified.
- A 'discovery' curve plotting cumulative number of individuals identified against the sampling period. This indicates the rate at which individuals enter the data set, and so what proportion of the population have been identified.
- If association has been set (4.3), then you may also obtain an estimate of the 'social

differentiation', the coefficient of variation of the true association indices (the proportion of time dyads spend together), as well as an estimate of the correlation coefficient between the true association indices and the calculated association indices (6.1) (Whitehead 2008b). The former is a measure of how varied the social system is (social differentiations less than about 0.3 indicating rather homogeneous societies, greater than about 0.5 well differentiated societies, and greater than about 2.0 extremely differentiated societies), the latter an indicator of the power of the analysis to detect the true social system (1.0 indicates a perfect job, 0.0 a useless one). If you have set an association index (6.1), and it is one of the standard ones that estimates the proportion of time individuals spend together (e.g., 'simple ratio'), then two estimates of the social differentiation and correlation between true and estimated association indices appear, calculated using methods described by Whitehead (2008a). Of the two methods the likelihood one is theoretically the best (confirmed using simulation) and should be used when available. The simpler Poisson approximation (it is extremely approximate!) is all that is shown if association indices have not been defined. The likelihood method of estimating social differentiation can take some time with larger data sets (there is a trade-off between the precision of numerical integration, and so estimation precision, and computer time, which you can adjust by changing the 'Resolution of integration'). You can also choose whether you want to calculate bootstrap standard errors for these measures, and the number of bootstrap replicates (which randomly resample sampling periods with replacement). With many bootstrap replicates, this may take some time.

6 ANALYZING ASSOCIATION INDICES AND INTERACTION MEASURES

Most analyses of social structure use 'relationship measures' which indicate the strength of a relationship between two individuals (Whitehead 2008b). The two most commonly used types of relationship measures are association indices and interaction rates. High values indicate that the individuals associate or interact a lot, low values that they rarely do. Many association indices are simple estimates of the proportion of time that the pair of individuals are associating, and interaction rates are fundamentally self explanatory. There are several ways that association indices can be calculated, or interaction rates defined.

An important distinction is between relationship measures that are symmetric (i.e. the relationship between A and B is the same as that between B and A) and those that are asymmetric (i.e. the relationship between A and B may be different from that between B and A). Association indices are usually symmetric, whereas interaction rates (e.g. grooming rates) may not be.

To calculate and examine association indices you must have entered a data set (3), set a sampling period (4.1) and defined and calculated associations (4.3). You often will have defined restrictions (4.2). For interaction rates, a definition of association is not used or needed, and setting a sampling period is optional, but useful.

To begin these analyses, click on 'Analyze associations/interactions' in the master SOCPROG window.

You are asked whether you wish to analyze associations or interactions. Click on the one you want (because of the flexibility in SOCPROG, you can actually do many analyses using either route with identical results, but it is usually much easier to use the appropriate option). A small window comes up which, depending on your choice, allows you either to define an

association index (6.1) or an interaction rate (6.2).

Once these are done, you enter the analysis screen. This contains a series of options, allowing you to list the matrix of association indices or interaction rates (6.4), as well as their standard errors (6.5), save the matrix as a SOCPROG MATLAB (6.6), VNA (6.8), GraphML (6.9) or ASCII file (6.7), examine the distribution of association indices or interaction rates (6.10 and 6.11), calculate network statistics (6.12), use network measures to divide the population into communities (6.13), display the association indices or interaction rates matrix using network diagrams (**Error! Reference source not found.**), principal coordinates analysis (6.15), multidimensional scaling (6.16) or hierarchical cluster analysis (6.17). The latter also lets you define clusters. You can also carry out hypothesis tests of the association or interaction data (6.18 and 6.20). Not all these analyses are available in any particular case (for instance tests for preferred companionship use association not interaction data), and only the available options will be displayed. At the top left of the screen is the option to redefine the association index (6.1) or interaction rate (6.2).

6.1 Choosing an association index

The possible association indices are listed in Whitehead (2008b, 98), with the addition of the option of defining your own index (6.1.1). 'Both identified' only considers associations for a pair of individuals within sampling periods during which both individuals were identified (Christal and Whitehead 2001), and 'joint occurrences' gives the sum of associations over all sampling periods (useful for interaction data, although it is generally much easier to use calculate these using the interaction rate screen). If defining association by presence in the same group, then the 'Simple ratio' is probably the most appropriate index (see Ginsberg and Young 1992), while for number of groups or weighted groups, the 'Half weight' seems to be best. These are the defaults. Both vary from 0 to 1, and are estimates of the proportion of time that a pair of individuals are associating. For further guidance on choosing association indices see Whitehead (Whitehead 2008b, 97-104).

Association indices between an individual and itself (diagonal elements of association matrix) are generally set at 1.0.

6.1.1 *Custom association indices*

The custom index option (not available with compiled SOCPROG) allows you to calculate many different indices, especially when used in conjunction with the custom option for defining association (4.3.3). The custom index should be described in a MATLAB script-file called 'assin.m' in the SOCPROG directory. It should take as input a combination of some of the following matrix variables (or possibly others calculated in earlier parts of the program):

inmat - a square matrix giving the sum, over sampling periods, of the associations between each pair of individuals (as defined in 4.3). If these associations are presence/absence of an association, then *inmat* gives the number of sampling periods in which the two individuals were associated; if they represent the number of minutes that a pair was associated in the sampling period, then *inmat* gives the total time associated during the study.

na - a square matrix giving the sum, over sampling periods, of the self-associations (as defined in 4.3) between the first individual in a pair and itself (thus each column of *na* is

identical). If these associations are presence/absence of an association, then *na* gives the number of sampling periods in which the first individual was seen; if they represent the number of minutes that a pair was associated in the sampling period, then *na* gives the total time the first individual in a pair was seen during the study.

nb - as *na* but for the second individual in a pair. Thus *nb* is *na* transposed.

yab - a square matrix giving the number of sampling periods in which both individuals were identified, but not together (only makes sense if association within a sampling period is defined in a 1:0 manner).

The output of 'assin.m' is the association index *assocm*.

So, for instance, to get an asymmetric, 'proportion of time together' index, 'assin.m' might contain:

***assocm=inmat./na*;**

As a final example, this is a 'Simple ratio' index, but with *NaN* ('not a number') in positions where there were less than five sampling periods in which both members of the pair were identified. This allows you to restrict subsequent analyses to more data-rich dyads:

assocm=inmat./(na+nb-inmat-yab);assocm=assocm./(0.0+(inmat+yab)>4);

If you use 'assin.m' to define an association index, then you can describe the association index in a way that will be printed out in subsequent analyses by defining the string variable *aistr* which is done by including a line in 'assin.m' such as:

***aistr*='Association index = Simple Ratio (>4 Sampling periods)';**

6.1.2 *Correcting for gregariousness*

A checkbox at the bottom of the screen where you choose your association index allows you to correct it for gregariousness, their tendency to associate, as proposed by Godde et al. (2013). Basically, the association index between individuals A and B is divided by the sum of the association indices involving A and the sum of those involving B, and multiplied by the sum of all the association indices. This removes differences in gregariousness between individuals. So, if two individuals preferentially associate, given their gregariousness, the index is greater than 1.0, and if they avoid each other it is less than one.

6.1.3 *Generalized affiliation indices*

There are a number of factors that may contribute to the association between two individuals. These include differences in gregariousness, habitat usage, same/different gender, and what can be called affiliation—how much the animals "like" one another. Generalized affiliation indices try to remove other potential causes of association from an association index leaving just affiliation (Whitehead and James 2015). The network of generalized affiliation indices can then be displayed and analysed in a similar way to association indices. Technically, generalized

affiliation indices are the residuals from a generalized linear model with association indices as the dependent variable and the different potential predictors of association (such as a measure of joint gregariousness, gender similarity, habitat overlap) as the predictor variables.

To construct generalized affiliation indices, when you reach the small "Define association index" screen, choose an association index as described above (6.1), and click on "Generalized Affiliation Index" (checking "Standardize for gregariousness" has no effect when constructing generalized affiliation indices). A larger screen appears. You use features on this screen to define the predictor variables used to calculate the generalized affiliation index. The association index (independent variable) is listed at the top left, beneath which is a list of the current predictor variables (initially empty). The supplemental fields that you can use to construct predictor variables are listed in the top centre. You may now enter these predictor variables in different ways, you can manipulate them, perform MRQAP tests to evaluate their significance (after which you may remove unimportant predictor variables), and then calculate generalized affiliation indices.

6.1.3.1 Input predictors: SOCPROG association, set of SOCPROG associations, Excel or ASCII measures

You can input predictor measures in a variety of ways using the popup menu on the upper right of the 'Generalized affiliation indices' screen. These predictor measures are input together with the names of the individuals. These names must include all of those names indexing the association index, otherwise the predictor measure will be rejected, and a message appears in the main window.

The easiest input format for a predictor measure is that created by SOCPROG itself. Matrices of association indices or interaction rates are created and then saved using the 'Saving association matrices (MATLAB)' pushbutton (6.6) so you can use one kind of association index (typically with a very broad definition of association) as a predictor for another, or an association measure saved from the 'Analysis of multiple association measures' screen (8.2.7), or one saved from the 'Generalized affiliation indices screen itself' (6.1.3.6). On clicking the 'Input predictor: SOCPROG association' pushbutton, small windows appear in which you select the file with the predictor measure, and choose a short name for the measure (without spaces), assuming one has not been previously chosen. If the names of the individuals on the input predictor matrix include all those for your association matrix, the predictor measure is included, and its short name appears on the list of predictors at the top left.

You can also enter as predictors a set of association measures previously saved from the 'Analysis of multiple association measures screen' (8.2.8), or from the 'Generalized affiliation indices screen itself' (6.1.3.6). The short name of each predictor then appears on the list at the left, once again assuming the names of the individuals on the predictor include all those for the previously input association indices.

Predictor measures can also be entered from Excel and ASCII files. In Excel, the predictor measure should look like that in Table 7, with the names of the individuals in the first row and

column. ASCII predictor measures should look like that in Table 8, with the names of the individuals separated by spaces in the first row, and in each other row a name followed by the association indices. Association values and names should be separated by one or more spaces. Once the Excel or ASCII file has been read in, you then enter the multi-line description, and the short name of the measure (no spaces), in a dialog box. To make sure that Excel and ASCII predictors have been correctly entered, I advise that you display them after they have been entered (6.1.3.5). Once again the predictor is only included and listed if the names of the individuals on the predictor include all those for the previously input association indices.

6.1.3.2 Calculate gregariousness predictor

Pressing this pushbutton adds a gregariousness predictor. Following Whitehead and James (2015), the gregariousness predictor between individuals A and B is the log of the sum of the association indices involving A (except the AB index) multiplied by the sum of those involving B (except the BA index). If any individual has only one associate, then its gregariousness with this associate becomes -∞, invalidating MRQAP tests and generalized affiliation indices. In such cases the gregariousness predictor is not logged for any dyad, and a warning appears on the command screen. The gregariousness predictor is added to the list at the top left.

6.1.3.3 Predictor measure from supplemental data

You can make predictor measures from the supplemental fields. First highlight the supplemental fields that you wish to use, and only these fields. Then, a window comes up with these fields listed at the left, and several options on the right.

One or two push buttons appear at the top right: one button if only one supplemental field was highlighted (‘Same (1); Different (0)’), two if more than one field were highlighted (‘All fields same (1); Any different (0)’ and ‘Any fields same (1); All different (0)’). These options produce 1:0 predictor measures. In the case of just one supplemental field, dyads have a ‘1’ if they have the same value of the supplemental field, ‘0’ if they have different values. With more than one field and ‘All fields same (1); Any different (0)’, then the dyads have a ‘1’ in the predictor matrix if they have the same value in all fields, and ‘0’ if there are differences in any field. Conversely, with ‘Any fields same (1); All different (0)’, then the dyads have a ‘1’ in the predictor matrix if they have the same value in any field, and ‘0’ if there are differences in all fields. You then enter the multi-line description, and the name of the measure, in a dialog box. You can use this to generate a predictor based on gender: same sex, different sex.

Lower on the right hand side of the small window is a popup menu with several distance measures available. These distance measures work with continuous variables, so string (alphanumeric) fields are converted to numeric, by making the first string ‘1’, the second ‘2’, etc. The conversion codes are given in the command window. Select a distance measure and press ‘Go’. The distance measures are described and defined in the MATLAB Help documentation under ‘pdist’. If you select the ‘minkowski’ distance measure, an editable box appears in which the minkowski exponent can be changed–the default is ‘2’. If you select ‘custom’, a dialog box appears allowing you to select a MATLAB function m-file describing a custom distance measure function of the

form:

function d = distfun(XI,XJ)

This function produces distances between pairs of individuals represented by the rows of both XI and XJ. The columns are the selected supplemental fields, converted if they are string variables, and the function produces an association measure 'd'. For instance, if there are two supplemental fields selected, 'Sex' (a string variable) and 'Age', the following function m-file gives '0' if the sexes of the two individuals are different, and the difference in age if sexes are the same:

```
function d=disfun(XI,XJ)
d=abs(XI(:,2)-XJ(:,2)).*strcmp(XI(:,1),XJ(:,1));
```

After pressing 'Go', you then enter the multi-line description, and the name of the measure, in a dialog box. To make sure that the new predictor measure has been correctly entered, I advise that you display it after it has been entered (6.1.3.5), and check one or two of the calculations.

6.1.3.4 Transform predictor measure

This takes the predictor measure highlighted on the list at the left (the first highlighted if several are highlighted), and allows you to transform it almost any way, into a new predictor measure (you cannot transform the association index here; in this case you need to define a different index; see 6.1). The 'Transform measure ...' window allows you to define the type of transformation. Examples include:

sqrt(X)
1-X
acos(X) [inverse of cosine]

You can also put any value you like on the diagonal (although diagonal values are ignored when performing MRQAP tests and calculating generalized affiliation indices), and give the transformed predictor whatever short name you choose, but not including spaces. The original untransformed predictor is retained in the list (you probably want to remove it; see 6.1.3.5) and the transformed predictor added.

6.1.3.5 Remove, display and rename predictor measure(s); display association index

To remove predictor measures from the list at the left, click on them to highlight, and then press the 'Remove predictor measure(s)' pushbutton.

If one or more predictor measures are highlighted in the listbox at the left, they can be listed in the command window, by clicking on the 'Display predictor measure(s)' button. The numeric format used is that given in the editable string at the top right: '4.2' means four characters with two decimal places, etc. You can use the select, copy and paste features of Windows to put the output measure or measures into other documents. Similarly, you can display the association indices by

clicking ‘Display association index’.

You can change the short name of one or more predictor measures by highlighting them on the list at the left and then pushing the ‘Change measure name(s) button’ (no spaces in short names).

6.1.3.6 Save predictor(s) as SOCPROG association measure, set of SOCPROG association measures

Pushing the ‘Save as SOCPROG association measure(s)’ allows you to save any predictor measures highlighted in the list box at the left. Each highlighted predictor measure is saved as a separate MATLAB (.mat) file. In this option, the short name of the measure is also saved.

Pushing the ‘Save as SOCPROG set of association measures’ saves all the predictor measures on the list at the left as one MATLAB file. They can later be reentered as a set (8.1.2, 6.1.3.1).

6.1.3.7 MRQAP tests of predictor variables

MRQAP (Multiple Regression Quadratic Assignment Procedure) tests allow you to examine the value of a predictor variable as an explanation for the pattern of association indices. The test considers whether each of the predictor matrices makes a significant contribution towards explaining the matrix of association indices while controlling for the presence of the other predictors. It uses permutation tests (the ‘double-semi-partialing’ technique of Dekker et al. (2007)), while measuring the effective contribution of each predictor using partial correlation coefficients. The number of permutations is that given in the top right of the ‘Generalized affiliation indices’ screen’. You can perform MRQAP tests on the current list of predictor variables by pressing the ‘MRQAP tests’ pushbutton in which case you stay on the ‘Generalized affiliation indices’ screen, and may wish to remove one or more ineffective predictor measures before moving on. Alternatively you can press the ‘MRQAP tests and Generalized affiliation index’ in which case you move on to the ‘Analyze association indices’ screen after the MRQAP test results are given in the main window. If there is only one predictor measure in the list, SOCPROG does a Mantel test between it and the association indices instead of the MRQAP tests.

6.1.3.8 Generalized affiliation index calculation

If you press either the ‘Generalized affiliation index’ or ‘MRQAP tests and Generalized affiliation index’ pushbuttons, generalized affiliation indices are calculated as the residuals of a generalized linear model, with the non-diagonal elements of the association index matrix as the dependent variable and the corresponding elements of the predictor variable matrices as independent variables. At the top of the screen in the middle is a popup menu that allows you to choose one from up to two of the following models:

‘Normal distribution’. Assumes that the errors of the model are normally distributed (so reducing the generalized linear model to a general linear model).
‘Binomial distribution’. This is appropriate when the association measure is of the form:

$$\frac{\text{Number of sampling periods pair observed associated}}{\text{Number of sampling periods of effort}}$$

The popular simple ratio and half weight association indices are of this form (6.1). Using the binomial model has the advantage that dyads with little data available (small denominator in above ratio) have little impact on the model.

'Poisson distribution'. This is appropriate when the index is a straight count, such as the joint occurrences index.

After calculating the generalized affiliation indices, SOCPROG moves on to the 'Analyze association indices' screen. As generalized affiliation indices are residuals, they can be negative and have a mean value near zero. Thus some options available for normal association indices are not available (particularly principal coordinates analysis, and some permutation test statistics, such as CVs).

When using generalized affiliation indices, a popup menu appears towards the top left of the 'Analyze association indices' screen, allowing one to change the type of residuals used to calculate the generalized affiliation indices. 'Raw' residuals are the default. 'Pearson', 'Anscombe' and 'Deviance' residuals transform the residuals so that they should be roughly distributed according to the normal distribution with zero mean and standard deviation one. Thus particularly high or low affiliation indices can be identified (greater than about 2.0 or less than about -2.0 may indicate unusually high affiliation or strong avoidance respectively).

6.2 Defining interaction rates

You must have *group* or *dyadic mode* data to define interaction rates.

On the left of the 'Defining interaction rate' window is a list of the primary data fields. Click on any to have its levels displayed beneath. On the right are four editable strings where you can define the interaction rate. The upper one ('Interaction rate, numerator') must have an entry. It defines which records (lines of the primary Excel file) are counted as an interaction for a pair of individuals. Here are some options:

1 – all records containing that pair of individuals

strcmp(*Beh***,'A')** – just those records with an 'A' in the *Beh* field

strcmp(*Beh***,'A')| Strcmp(***Beh***,'B')** – just those records with 'A' or 'B' in *Beh* field

Strcmp(*Beh***,'A') &(***Level***>3)'** – just those records with an 'A' in the *Beh* field, and *Level* (could be a measure of the strength of the interaction) greater than 3.

Strcmp(*Beh***,'A') .****Numint***'** – the sum of the *Numint* field (number of interactions for each record) for just those records with an 'A' in the *Beh* field.

If you are using numeric fields and expressions (*Numint* in the previous example), they must be followed by an apostrophe, and joined to other expressions with "**.***".

If you just want overall counts of interactions (or sampling periods containing interactions), then leave the next three editable, 'denominator', strings empty. However, some individuals may have been observed more than others, and you may wish to correct for this, by dividing by some measure of the amount of time each individual was observed, or you may wish to calculate the relative rate of one kind of interaction over another. If you enter something in more than one of the 'denominator' strings, records are counted if any of the conditions are satisfied. If you are using quantitative values (*Numint* in the example above), just fill in one of the denominator fields. So here are some examples (Table 6).

Table 6. Examples of filling in the editable strings in the 'Define interaction rate' screen (*Numint* is a quantitative field giving the number of interactions in a bout, with each record of the input file corresponding to a bout).

numerator	denominator (both inds)	denominator (1st ind)	denominator (2nd ind)	Interpretation
1				Count of all records with two individuals interacting in any way
strcmp(*Beh*,'A')				Count of all records with two individuals interacting with *Beh* = 'A'
Numint'				Total number of interactions between each two individuals
strcmp(*Beh*,'A'). *Numint*'				Total number of interactions between each two individuals with *Beh* = 'A'
1	1	1	1	Number of records with two individuals interacting divided by number of records with at least one individual present (= 'simple ratio' association index)
strcmp(*Beh*,'A')	1			Proportion of interactions between two individuals that have *Beh* = 'A'
1		1		Proportion of interactions with first individual as interactant that are with second individual
1			1	Proportion of interactions with second individual as interactee that are with first individual
Numint'		*Numint*'		Proportion of total number of interactions with first individual as interactant that are with second individual
strcmp(Beh,'A')	strcmp(Beh,'I')	strcmp(Beh,'I')	strcmp(Beh,'I')	Rate of interaction of type 'Beh' = 'A', relative to proportion of records where null behavior 'I' is recorded for either individual (useful for focal animal type data)

At the bottom right of the 'Define interaction rate' screen are two further important options:

'Asymmetric' checkbox. If you check this, interactions are assumed to be asymmetric (makes sense for grooming, threats, …). In *dyadic mode*, the first individual entered for each record is the interactant who does something, the second the interactee who has something done to them (i.e. if the individuals are 'E' and 'A', interaction 'E A' means 'E', the interactant, interacts with 'A', the interactee, i.e. E→A). In *group mode* each individual is the interactant for interactions with individuals entered to its right (i.e. if the individuals are 'E F A', interactions E→F, E→A and F→A are entered). If unchecked (useful for interactions such as touching, kissing) the interactions are assumed symmetric (so: in *dyadic mode* if the individuals are 'E A', interactions EA and AE are entered; in *group mode* if the individuals are 'E F A', interactions EF, FE, EA, AE, FA and AF are entered).

'Total counts / Number of sampling periods' popup menu. The 'Number of sampling

periods' option is only available if sampling periods have been set previously (4.1). 'Total counts' adds up the number of records for which the conditions are true for the numerator, and (if set) the denominator. 'Number of sampling periods' adds up the number of sampling periods within which there is at least one record for which the conditions are true for the numerator, and (if set) the denominator. This is particularly useful for cases when interactions occur in sequences, so we wish to move from 'how many times did A groom B?' to 'In how many hours was A observed grooming B?' If you select 'Number of sampling periods', then numeric field values (e.g. *Numint* in the examples above) are ignored.

Interaction rates between an individual and itself (diagonal elements of matrix of interaction rates) are set at zero.

6.3 General options when analyzing association indices or interaction rates

There are several options when analyzing association indices, which are set at the top (centre and right) of the 'analyzing association indices' window. A popup list of supplemental fields is provided at the top right.

6.3.1 *Class variables*

Here you can define a class variable, separating the individuals into classes using the supplemental fields or functions of them. These give additional possibilities for some analyses. e.g.

Sex	<- Supplemental field *Sex*
10*(floor(*Age*/10))+5	<- *Age* in 10 time unit divisions: 5,15,25, ...

To enter names of supplemental fields into the formula for the class variable, choose the supplemental field from the popup list and press the 's' button beside the class formula. Some of the other options on this page will cluster the individuals, and you can use these clusters as a class variable (6.13, 6.17).

6.3.2 *Numeric format*

You can select the numeric format used when listing results to the command window or ASCII files on the top right of the 'analyzing association indices' window: '4.2' means four characters with two decimal places (e.g. '7.36'), etc.

6.3.3 *Labels for individuals*

Here you can change how the individuals are labelled in lists shown in the command window and plots. You should write a combination of supplemental field names and other characters. Here are some possibilities

ID	<-Just the name of each individual (the default)
***ID*(*Sex*)**	<- This gives ID with Sex in parentheses, e.g. Moe(M)

<- Nothing, leaves labels blank, may be useful for cluttered plots

To enter names of supplemental fields into the formula for the labels, choose the supplemental field from the popup list and press the 's' button beside the label formula.

6.4 List association matrix or interaction rates

This displays the matrix of association indices or interaction rates in the command window. If the association matrix or interaction rates are symmetric (usually the case for association indices), only the lower left triangle is shown. You can use the select, copy and paste features of Windows to put it into other documents. You can change the number of decimal places in the output using the editable format string (6.3.2).

6.5 List SE association matrix or interaction rates

This displays the estimated standard error of the association matrix in the command window. A dialog box comes up allowing you to use either an analytical approximation using a binomial assumption or the bootstrap method in which sampling periods (or individual interactions if 'Total counts' has been selected as the interaction rate, 6.2) are sampled with replacement to produce bootstrap replicates. The number of bootstrap replicates is in the 'No. permutations' editable string at the top right. Use at least 100. The bootstrap is probably theoretically the best method, but may take some time, or cause a crash (due to exceeding memory limitations) on large data sets, and both methods usually produce similar results.

The analytical approximations are $SE(a) = a\sqrt{((1-a)/w)}$ and $SE(i) = i/\sqrt{w}$, where a is the association index, i the interaction rate, and w the number of sampling periods including an association or interaction (or individual interactions if 'Total counts' has been selected as the interaction rate, 6.2) (Whitehead 2008b, 96). This option is not available if the index was corrected for gregariousness (6.1.2) or is a generalized affiliation index (6.1.3).

Both bootstrap and analytical methods assume that the probabilities that sampling periods include associations or interactions are independent (or the interactions themselves are independent if the 'Total counts' option (6.2) was chosen).

If the association indices or interaction rates are symmetric (usually the case for association indices), only the lower left triangle of the standard errors is shown. You can change the number of decimal places in the output using the editable format string (6.3.2).

After printing the estimated standard errors of the association matrix, a dialog box allows you to save these estimated standard errors in either a MATLAB or ASCII file. The formats are the same as for saving the original association matrix in these formats (see 6.6, 6.7).

6.6 Saving association matrices or interaction rates (MATLAB)

You can save the association matrix or interaction rates as a MATLAB file. This in fact saves three things: '*assocm*', the association matrix or matrix of interaction rates, '*rnam*' a list of the names of the individuals used, '*atit*' which is a string array giving the name of the original data file, the sampling period, restrictions, association type and association index, or how the

interaction rates were calculated. Saving the matrix as a MATLAB file allows it to be used in the 'analysis of multiple association measures' module (see 8.1.1), or as a predictor in the calculation of generalized affiliation indices (6.1.3.1).

6.7 Saving association matrices or interaction rates (ASCII)

You can also save the association matrix or interaction rates as an ASCII (text) file. This writes the association matrix to a .txt file in the same format as it is listed on the screen (6.4), except that the full (square) association matrix is always written, even if it is symmetric. You can change the number of decimal places in the output using the editable string (6.3.2).

6.8 Saving in network format (VNA file)

This writes the association matrix or interaction rates to a file in the VNA format for entry into network analysis programs such as UCINET (www.analytictech.com/ucinet_5_description.htm) and Netdraw (http://www.analytictech.com/netdraw.htm). IDs and supplemental data are included. Note that in this format no distinction is maintained between association or interaction rates that are 0.0 and those missing (i.e. no available data). These are coded as 0.0 (zero) and 'NaN' (missing) respectively in MATLAB and ASCII formats. Do not use numeric formats in which the total number of characters is more than two plus the number of decimal places. i.e. "6.4" is OK; "7.4" will cause problems.

When saving generalized affiliation indices (6.1.3) in VNA format, the original association indices are also saved, so you can compare their networks, as well as the gregariousness (sum of original association indices) of each individual.

6.9 Saving in network format (GraphML file)

This writes the association matrix or interaction rates to a file in the GraphML format for entry into network analysis programs such as yED and Gephi.

IDs and supplemental data are included. Note that the attribute type of the supplemental/class variables is also exported: 'string' vs 'number' ('double' in GraphML). This may be an issue. For instance in Gephi, modularity partitions can only be displayed if they strings, and those created by Community division by modularity (6.13) in SOCPROG are numbers. To change attribute types in GraphML files, open the file in a text editor, such as Notepad, and change the attribute type from 'double' to 'string'. e.g.

```
<key id="d5" for="node" attr.name="Cluster" attr.type="double">
```

becomes

```
<key id="d5" for="node" attr.name="Cluster" attr.type="string">
```

Note that in GraphML format, missing data ('NaN' in MATLAB) is not generally read correctly. [Thanks to David Lusseau for this code, and Mauricio Cantor for feedback].

6.10 Distribution of associations or interaction rates (list)

This displays in the command window, for each individual, the mean association index or

interaction rate with all other individuals (excluding the individual with itself), the sum of all associations (which for association indices is similar, but not identical, to the 'typical group size' (see Jarman 1974)) or overall (with all other individuals) interaction rate, and the maximum association or interaction rate (excluding individuals with themselves). Also given are the means, and standard deviations of these measures over all individuals and by class (each class with all individuals, between pairs of classes, and overall within- and between- classes), if you have entered a class variable (6.3.1). You can change the number of decimal places in the output using the editable string (6.3.2; the digit before the decimal point is ignored in this case).

Furthermore, if you have entered a class variable, a Mantel test is carried out on the null hypothesis that 'associations/interaction rates between and within classes are similar' (see Schnell *et al.* 1985). Results are expressed as the t-value (with infinite degrees of freedom), p-value (for 2-tailed test) for the analytical approximation, the permutation p-value (if more than zero permutations have been set on the top right of the 'analyzing association indices' window), and matrix correlation coefficient. If within class associations/interactions are higher, t is positive. A more general set of options for Mantel tests is available within the 'Analysis of multiple association measures' module (8.4.4).

6.11 Distribution of associations or interaction rates (plot)

If you select this option, a window opens allowing you to plot histograms of association indices or interaction rates. These can be of four sorts:

- Association indices or interaction rates (all non-diagonal elements)
- Mean association indices or interaction rates (by individual, ignoring diagonal elements)
- Maximum association indices or interaction rates (by individual, ignoring diagonal elements)
- Sum of association indices (by individual, including diagonal elements, which are usually 1.0 for association indices) or overall interaction rates (with all other individuals)

You can also decide whether you want bars or symbols (symbols may be better for printing on black and white printers), totals or proportions, and set the number of bins on the histogram. The default is 10, i.e. 0.1 intervals for association indices. If you have previously run permutation tests (6.18), you can produce histograms of the association matrix of the last random permutation, by checking the box that appears. It may be illustrative to overlay random distributions on the real data.

If you entered a class variable (6.3.1), you can select 'from' and 'to' classes, e.g. 'From M' 'To F' will give distributions of associations or interaction rates from males to females.

You can make a new plot, or add data to an existing plot (useful, for instance, if you want to plot M-M and F-F together, or overlay random and real). If you overlay plots, a legend appears. This can be moved using the cursor if it appears in an unfortunate place. You can also change the legend (2.3).

You cannot plot distributions of association indices (or interaction rates) and their sums on the same histogram. It also does not make much sense to plot proportions and totals on the same histogram, and you will get an error message if you try to use two different bin sizes on the same plot.

6.12 Network analysis measures (this part of SOCPROG was developed in conjunction with David Lusseau)

This button calculates a number of measures treating the matrix of association indices or interaction rates as a weighted network (Wey *et al.* 2008; for use of network analysis in studying animal societies, see Croft *et al.* 2008; Whitehead 2008b). This module only uses symmetric association indices or interaction measure. To obtain network measures of asymmetric data, export the data into a specialist program such as UCINET (6.8). The measures (defined in Whitehead 2008b, 174) are:

- *strength* This is simply the sum of association indices of any individual with all other individuals (Barrat *et al.* 2004). It is the same as the sum of associations (6.10) minus one. High strength indicates that an individual has strong associations with other individuals.
- *eigenvector centrality* This (given by the first eigenvector of the matrix of association indices Newman 2004) is a measure not only of how well an individual is associated to other individuals, but also how well they are associated. So to have high eigenvector centrality, an individual will have relatively strong associations to other individuals which in turn have relatively strong associations.
- *reach* The reach of an individual is a measure of indirect connectedness, and is a useful concept in a society that possesses behavioural contagion, so that the behaviour of A towards B may influence the behaviour of B towards C. I define the reach of A as the sum, over other individuals B, of the sum of the products of all pairs of association indices linking A and B through another individual C (defined in Whitehead 2008b, 174).
- *clustering coefficient* This is a measure of how well the associates of an individual are themselves associated. I use the matrix definition of clustering coefficient for weighted networks of Holme et al. (2007).
- *affinity* The affinity of an individual is a measure of the strength of its associates, weighted by the association index between them. So an individual with high affinity has relatively high associations with individuals who have high strength.

The measures are calculated for each individual. Presented, for each measure, are the average and standard deviation for the whole population. If classes of individual, such as sex, are defined (see 6.3.1), the output gives the mean and standard deviation values of the measures for members of each class. Also shown are the within-class and between-class values (just strength for between classes). At the bottom of the output are given are the correlations, over individuals, of strength with clustering coefficient, and strength with affinity.

When you press the ‘Network analysis statistics’ button, if the association index, interaction rate, or association measure is asymmetric (i.e. the value for A with B is not necessarily the same as that for B with A), a dialog box asks you whether you wish to quit the network analysis, or proceed using the average (mean of the value for A with B and the value for B with A).

Then, a small window comes up with several checkbox options:

- *Bootstrap SE’s* Pressing this gives [in square brackets] estimated standard errors for all measures calculated using the bootstrap method by which the sampling periods are sampled with replacement to produce bootstrap replicates. The number of bootstrap replicates is in the ‘No. permutations’ editable string at the top right of the ‘Analyzing

association indices' window. Use at least 100. This option is only available for association indices or interaction rates calculated by SOCPROG (not association measures).

- *Output for individuals* This checkbox gives output of the network analysis measures in the Matlab command window for each individual (the default if less than 20 individuals).
- *Permutation tests* [THIS OPTION IS SOMEWHAT EXPERIMENTAL AT PRESENT—DO NOT USE WITHOUT THINKING VERY CAREFULLY ABOUT WHAT YOU ARE DOING!] This uses the tests for preferred/avoided association module routines (6.18) to perform hypothesis tests on the network measures. The options 'No. permutations', 'Flips per permutation', and permutation method are set as described in section 6.18. The options 'Between classes' and '2-sided significance level for dyads' are inoperative in this implementation. If using this option, beneath most output network measure statistics (but not between- and within-class statistics) are given the mean value over all random permutations (the expected value, 'E'), and the proportion of values from the random permutations that are less than the real value ('P'). These allow the hypothesis that individuals are randomly connected in the network, given the data structure, to be tested. Please read section 6.18 carefully before using this option (see also Whitehead 2008b, 182). You should use at least 1,000 permutations for this option, and maybe check the output by rerunning. This option is not available for interaction rates or association measures.
- *ASCII (text file) output* This produces a tab-delimited ASCII file with the network measures for each individual that you can import into Excel or other programs.

You can also save the association matrix in VNA or GraphML formats for analysis using specialized network analysis programs (see 6.8 and 6.9). These are much more sophisticated and flexible than SOCPROG for general network analyses, but do not contain all the measures and analyses that I think are most useful for analyzing association matrices (such as the permutation tests described above).

6.13 Community division by modularity

This option allows you to explore the possibility that the population you are studying can be usefully divided into clusters, such that association indices are generally high among individuals in the same cluster, and generally low among individuals in different clusters. The measure used for assessing such a division is modularity (Newman 2004): the difference between the proportion of the total association within clusters and the expected proportion, given the summed associations of the different individuals (as given in 6.10). This is modularity-1 as given in section 6.17; modularity-2, with expected values found by permutation, does not work with this method. With even a moderate number of individuals in the population, there are a huge number of possible divisions into clusters. How do we find the best delineation, which could be defined as that with the highest modularity? Newman (2006) suggests an eigenvector-based method as being generally efficient, and this is implemented by SOCPROG.

After clicking on the 'Community division by modularity' button, the output, given in the MATLAB command window, is a list of individuals arranged by cluster, with the eigenvector corresponding to the final bifurcation involving that individual (values near zero indicate uncertainty in the assignment of the individual), and the cluster in which the individual was

placed. At the bottom is given the modularity of this arrangement. Modularities greater than about 0.3 are considered to indicate useful division of the population (Newman 2004). There is also the option to save the cluster identifications as a supplemental variable (see 6.17.1).

6.14 Network diagram

This makes a network diagram (or sociogram) of the matrix of association indices or interaction rates (Whitehead 2008b, 155-156). Nodes representing the individuals are arranged in two-dimensions and the thickness of links between pairs of nodes indicates the strength of their relationship (value of association index or interaction rate between individuals). There are several options, most of which are set on the small window which appears first ('Network diagram Options'):

- Arrangement of nodes. There are three options (note: you can change the arrangement as you wish later (see below):

  a) Circular arrangement in which strongly linked individuals are usually close to one another.

  b) Two-dimensional non-metric multidimensional scaling arrangement in which more strongly associated individuals are usually closer together than pairs less strongly associated (6.16). You can set the maximum number of iterations for the multidimensional scaling by changing the 'No. permutations' option at the top right of the 'analyzing association indices' window (1,000 should generally suffice; the command window will tell you if the iteration limit is exceeded). Note: the multidimensional scaling option for the arrangement of nodes does not always work.

  c) Principal coordinates arrangement in which the distance between nodes is inversely proportional to their association (6.15). This is the default.
- Node size proportional to gregariousness. If you check this, the margin of the rectangular box containing the name of each node is proportional to the gregariousness of the individual (the sum of its associations). So individuals with more and stronger associations will tend to have larger boxes than those with few weak associations. If the association matrix is asymmetric, the gregariousness is calculated for the individual as actor not recipient.
- Font size for node labels on network diagram.
- Minimum value of association index shown as a link. Increase this to clean up 'messy' network diagrams. Press 'return' after changing the minimum value to automatically change the default legend values.
- Maximum width of link. This is the width of the link between the two individuals with the greatest association index.
- Legend values. You can put in numbers for the values in the legend (showing how the width of the link relates to the association index), separated by spaces. Usually 2-4 values are suitable (the default is three values: the maximum association; the minimum shown; and half way between them). The default numeric format for the legend is set on the 'analyzing association indices/interaction rates' window (6.3.2), but you can change this by simply editing the values shown in the 'Legend values' editable box. Remove numbers if you do not want a legend.

- The colour of the links can be changed by clicking on the 'Colour of link' tile (initially blue). A screen comes up allowing you to choose a new colour.
- The background colour of the node boxes can be changed by clicking on the 'Node colour' tile (initially white). A screen comes up allowing you to choose a new colour. If you have set a class variable (6.3.1), there are separate colours for nodes of each class (e.g. males and females). You can change these colours by clicking on the respective tiles (only possible if <16 classes).
- Labelling of individuals can be changed on the editable box of the 'analyzing association indices' window (6.3.3).

If an asymmetric measure is being used, then an additional 'Directional links?' checkbox appears. If this is checked (the default for asymmetric matrices) then arrows are drawn between the points that represent the individuals, potentially one in each direction (if both associations between the two individuals are greater than the minimum value to be shown). If the box is unchecked, then only a single line between each pair of points is shown, representing the mean of the two association indices. If directional links are to be shown, you can change the separation between bidirectional links, the colour of the arrow heads and their size.

To make a network diagram using the chosen options, click on the 'Make' pushbutton. You can make changes to the diagram using the MATLAB graphics options (if you are using the uncompiled version of SOCPROG). In particular you can change the positions of the nodes, which is useful in many cases, especially if they are plotted on top of, or close to, one another. Click on the  symbol in the menu at the top of the diagram. Then move the cursor to the node you wish to move. Press the left mouse button and hold it down while you move the node. When you have moved all nodes you wish, press 'Redraw' on the Network diagram screen. The appropriate links will now be moved to match the new node locations.

With many individuals (>~50 individuals) network diagrams will take some time to draw.

For additional possibilities in drawing network diagrams, export to a specialist network drawing routine (6.8, 6.9).

6.15 Principal coordinates analysis (classic metric multidimensional scaling)

This makes a 'classic' metric scaling, using principal coordinates analysis (Whitehead 2008b, 156-158), of the data. It produces an arrangement of points, each representing an individual, so that the distance between them is proportional to one minus the squareroot of their association index or interaction rate (if the measure is asymmetric, the A-B association/interaction is the mean of the A to B and B to A associations/interactions)–so strongly associated or frequently interacting individuals are plotted together, weakly associated ones are plotted apart. The dimensionality is arranged so that the first n dimensions (principal coordinates) of the resulting analysis give as much information as possible in this number of dimensions. In the command window are listed each principal coordinate, the percentage of the total variance in the squarerooted association/interaction matrix explained, the cumulative variance explained by all dimensions up to and including this one, and the eigenvalue. Eigenvalues greater than one indicate coordinates explaining more information than the average; large negative eigenvalues indicate poor performance of the principal coordinates analysis. You can plot any two dimensions against one another, and change the fontsize of the labels for the individuals, in the small window that

appears first ('Principal Coordinates Analysis Options'). Normally only the first two are plotted, although if these just explain a few percent of the original variance, then the plot does not provide a good representation.

The background colour of the boxes with the individual names can be changed by clicking on the 'Node colour' tile (initially white). A screen comes up allowing you to choose a new colour. If you have set a class variable (6.3.1), there are separate colours for nodes of each class (e.g. males and females) (only if <16 classes). You can change these colours by clicking on the respective tiles.

You can change the positions of the individuals in the representation, which is useful if they are plotted on top of, or close to, one another (as long as you are using the uncompiled version of SOCPROG). Click on the symbol in the menu at the top of the diagram. Then move the cursor to the node you wish to move. Press the left mouse button and hold it down while you move the node.

The results (scores) of the principal coordinates analysis are available as a global variable called 'pc', which you can use for further analysis if you have MATLAB. So for instance to do a 3-D 'spike' plot of the first 3 principal coordinates, type the following in the command window:

```
>>global pc
>>figure
>>plot3([pc(:,1) pc(:,1)]',[pc(:,2) pc(:,2)]',...
  [pc(:,3) zeros(length(pc(:,3)),1)]','-k',pc(:,1),pc(:,2),pc(:,3),'ok')
>>set(gca,'DataAspectRatio',[1,1,1]);
```

6.16 Multidimensional scaling

This makes a (non-metric or metric) multidimensional scaling representation of the matrix of associations or interactions (Whitehead 2008b, 158-161). If the measure is asymmetric then the A-B association/interaction is the mean of the A to B and B to A associations/interactions. Like principal coordinates analysis, it produces an arrangement of points, each representing an individual, so that the distance between them is inversely proportional to their association index or interaction rate–so strongly associated, or frequently interacting, individuals are plotted together, weakly associated or interacting ones are plotted apart. However, unlike principal coordinates, in multidimensional scaling, the number of dimensions is set beforehand, and the analysis tries, iteratively, to find a suitable representation of the points the given number of dimensions. It measures its performance by 'stress'. Lower stress is good, and the program proceeds until stress cannot be lowered by moving any one point. Final stress values less than about 10% are generally considered to indicate a useful representation (Morgan *et al.* 1976). In metric multidimensional scaling, the aim is to have the distances between points proportional to the squareroot of one minus the association index or interaction rate, so the result is often similar to that from principal coordinates. In non-metric multidimensional scaling, the relationship only has to be monotonic, so that more highly associated or more frequently interacting pairs of individuals are represented by closer points. Because this is an easier condition, stress is usually lower in non-metric plots. However, unlike principal coordinates, metric scaling is iterative: it takes time to find a solution, and this may not be unique or optimal. It may depend on the start position. I strongly suggest you repeat multidimensional scaling analyses two or more times (using different random start

positions, see below) to see whether the final representation is optimal and reproducible. The small window that appears first ('Multidimensional scaling') allows you a number of options:

- Change the fontsize of the labels for the individuals.
- Set the maximum number of iterations (default 200); if you receive a message 'Iteration limit exceeded. Minimization of criterion did not converge' in the command window, increase the number of iterations.
- Choose the scaling criterion (metric vs nonmetric, etc.; type '**help mdscale**' in the command window to see descriptions).
- Starting configuration. You can either use the output of a principal coordinates analysis (the most efficient option) or a random configuration of the points (different each time), which allows you to check for the repeatability of the analysis.
- Choose the number of dimensions in which to scale (default =2).

All pairs of chosen dimensions are plotted against one another, and the final stress is output in the command window.

The background colour of the boxes with the individual names can be changed by clicking on the 'Node colour' tile (initially white). A screen comes up allowing you to choose a new colour. If you have set a class variable (6.3.1), there are separate colours for nodes of each class (e.g. males and females) (only if <16 classes). You can change these colours by clicking on the respective tiles.

You can change the positions of the individuals in the representation, which is useful if they are plotted on top of, or close to, one another (as long as you are using the uncompiled version of SOCPROG). Click on the ↖ symbol in the menu at the top of the diagram. Then move the cursor to the node you wish to move. Press the left mouse button and hold it down while you move the node.

As with principal coordinates analysis (6.15), a global variable 'pc' gives the scores of the multidimensional scaling which can be used for further analysis or displays.

6.17 Hierarchical cluster analysis

This makes a hierarchical agglomerative cluster analysis of the association or interaction data, and displays the results as a dendrogram or tree-diagram (Whitehead 2008b, 161-168). The individuals are arranged on one axis and their degree of association on the other. The tree indicates the association index or interaction rate between hierarchically formed clusters of individuals. There are several options, most of which are set on the small window which appears first ('Hierarchical Cluster Analysis Options'):

- Font size for text on display.
- Linkage method–how succeeding clusters are linked. The options are single, complete, average, or Ward's (look up 'linkage' in Matlab Help for more information). It is recommended that you stick with the default 'average' linkage, or use Ward's linkage. Average-linkage and Ward's (unlike 'complete' or 'single' linkage) are among the generally preferred types (Milligan and Cooper 1987), and their results are generally similar to those from other recommended types. You can identify the most appropriate linkage type for your data by examining the cophenetic correlation coefficient (see below).
- Orientation. This allows you to orientate the dendrogram tree up, down, to the left or right.

- Modularity. For any division of a population into clusters, modularity (following the definition of Newman 2004) is the difference between the proportion of the total of the association indices, or interaction rates, within clusters and the expected proportion. Expected proportions are calculated in two possible ways (see Whitehead 2008b, 224-226):

  modularity 1: these are the expected proportions given the summed associations (equivalent to gregariousness as given in 6.10) or interaction rates of the different individuals.

  modularity 2: here the expected proportions are from the permutation of associations within samples in the preferred/avoided association module routines (6.18, using the number of permutations, and flips per permutation as outlined in that section). This controls for gregariousness and the structure of the data. Modularities are usually lower by this method than *modularity 1*. Clusters formed tend to have mutual association preferences rather than just being identified together. This option is not available for interaction rates or when association measures are input directly.

  Following a method developed by D. Lusseau, modularity of one or both types is presented in a separate window for the clusters defined by different cutoffs of the association index. The maximum modularity, and value of the association index at this modularity, is given in the command window for the types chosen. Modularities greater than about 0.3 are usually considered to indicate useful divisions of the data (Newman 2004). You can save the clusters defined by this maximum modularity criterion (6.17.1).
- Knot diagram. This option implements a method devised by Wittemyer et al. (2005) in which the cumulative number of bifurcations (joins) in the dendrogram is plotted against some function of the association index or interaction rate. In one figure, SOCPROG plots cumulative bifurcations against the straight association index or interaction rate (upper panel) and its negative logarithm (lower panel). The latter can be considered an association distance, and so is congruent with Wittemyer et al.'s (2005) presentation, who look for inflexions in this diagram, which indicate changes in the rate of clustering. When the bifurcation rate drops suddenly, this is termed a knot, and is a potentially useful level at which to define clusters. More than one knot can be identified, indicating hierarchical tiers of social structure (Wittemyer *et al.* 2005). Click on the 'Get x' button in the lower panel to identify a knot: move the cross-hairs over the knot (just in the lower diagram), and click. The value of the association index at the knot is written on the diagram, and in the 'Colour threshhold' box of the 'Hierarchical Cluster Analysis Options' screen, allowing you to rerun the cluster analysis to colour, and potentially save (6.17.1), the clusters formed by the knot.
- Colour threshold. This allows you to have clusters existing at a certain association level to be coloured distinctively. If this is set at zero (default), there is no colouring. You can save the clusters defined by this criterion, (6.17.1).
- Labelling of individuals can be changed on the editable box of the 'analyzing association indices' window (6.3.3).

Printed in the command window is the cophenetic correlation coefficient, which ranges from 0 to 1, and indicates how well the dendrogram matches the matrix of association indices or

interaction rates. Cophenetic correlation coefficients greater than about 0.8 indicate a good match (Bridge 1993). If the cophenetic correlation coefficient is less than 0.8, then the dendrogram is not a good representation, and probably should not be displayed. You can use the cophenetic correlation coefficient values to decide which linkage method gives the best representation of the data (highest coefficient).

AN IMPORTANT NOTE: It is easy to get misleading results using hierarchical cluster analysis, and to over-interpret dendrograms. Even random data, with no preferred or avoided associations can produce an interesting-looking dendrogram (Whitehead 2008b, 161-162). If working with association indices you should run the test for preferred associations first (6.18), and, if these exist (significant p-value in the Bejder et al. test) then consider a hierarchical cluster analysis, and only use it if the cophenetic correlation coefficient is greater than about 0.8. But:

ANOTHER IMPORTANT NOTE: Even if there are preferred associations and a large cophenetic correlation coefficient, a hierarchical cluster analysis may not be an appropriate way of displaying them if the society is not hierarchically arranged (i.e social entities nested within larger social entities). In many cases, network diagrams (**Error! Reference source not found.**) or principal coordinates analyses (6.15), which do not assume a hierarchically-organized social structure, are more appropriate methods of displaying a set of associations, and envisaging a society.

6.17.1 *Saving clusters as a supplemental field*

You can save the clusters from the hierarchical cluster analysis, if you have set modularity or the 'colour threshold' as a value other than zero. Clusters are defined as those set at this value on the dendrogram or by maximum modularity (colours showing clusters containing more than one individual). To use this option click 'Yes' on the small dialog box asking you whether you want to 'Save Clusters as Supplemental Field?', then enter the name for the field with the cluster assignations. You cannot enter a name already used and the default is *Cluster*. A new numerical supplemental field is then added with cluster assignation numbers (1, 2, ...) under this name (the value for individuals not in the cluster analysis, because of restrictions, is *NaN*). It is made the default class variable (6.3.1), so that then you can, for instance, list the distribution of associations (6.10) getting the mean associations within and between clusters, or colour the nodes in a network diagram to distinguish the clusters (6.14). You can go back to the sampling-restrictions-associations window and press 'View' to see the cluster assignations if you like (3.3).

REITERATING AN IMPORTANT NOTE: You can always produce clusters using this technique even with random data, but they may not mean anything to the individuals. I suggest you only use this option if clusters are very clear in the data, and the cophenetic correlation coefficient is high.

6.18 Tests for preferred/avoided associations, and differences in gregariousness

SOCPROG implements the test for preferred or avoided associations introduced by Bejder et al. (1998), but adds some variations (see Whitehead 2008b, 122-130). The general null hypothesis of these tests is that individuals associate with the same probability with all other individuals (or among some set of them), given their availability. These tests are only available for associations, not interaction rates, nor measures imported directly (3.8). For a general

introduction to hypothesis testing in animal social networks I suggest that you read Croft et al. (2011).

The routines permute the data randomly in different ways to allow you to test the null hypotheses of no preferred/avoided associations. In general, for the null hypothesis to be rejected, the distribution of association indices from the real data should be different from the distribution of association indices from a number of permuted data sets. So if you do 10,000 permutations and, for instance, only 150 of them have higher SD's of the association indices than the SD of the real association indices, then you can reject the null hypothesis that the distribution of real association indices is no different than would be expected from random association at p=0.03 (2-tailed test), or the null hypothesis that the variation of real association indices is no greater than would be expected from random association at p=0.015 (1-tailed test). The routines give several possible test statistics (each calculated for both random and real data):

- mean of association indices
- median of association indices (only if less than 15 individuals—with more than 15 individuals, calculating medians would be very time consuming)
- SD of association indices
- CV of association indices (SD/mean)
- proportion of non-zero association indices
- mean of non-zero association indices
- SD of non-zero association indices
- CV of non-zero association indices (SD/mean)
- SD of typical group sizes (see 6.18.6)

The results of the tests are presented as a table giving, for each of these test statistics, the real value, the mean of the values for the random data sets, and the number of times the statistic from the random data was less than the real value. Then, when appropriate, these values are translated into the P-value of a one-sided test that the real data are more variable (indicating preferences or avoidances, or greater variation in gregariousness) than expected. Recommended test statistics (see below) are noted.

Usually, the significance levels of these measures are well correlated. However, they are testing different aspects of the data, so it worth thinking carefully about what you are doing. For instance, if some individuals preferentially associate with other individuals over several sampling periods, then the SD and CV of association indices should be higher in the real data set than the random data sets, but the means may be the same. Also, if some individuals avoid others, the proportion of non-zero association indices should be lower in the real data than in random data.

The routines use the procedure described by Bejder et. al. (1998) and Manly (1995) which sequentially inverts the intersection of 2 rows and 2 columns in a 1:0 data matrix, a 'flip'. Some enhancements to the Manly/Bejder et al. procedure have been introduced (see Whitehead 2008b, 122-130). Miklós and Podani (2004) noted a bias in the Manly/Bejder et al. routine which made the test generally conservative (less likely to reject null hypothesis). This has now been fixed by enumerating each 'trial flip' rather than just the successful flips.

As each new random matrix is only slightly different from the previous one, this means that the random data sets are not independent of each other or the real data. Therefore, as the starting point is the real matrix, the p-values are biased against extreme values, and the number of permutations carried out is not, by itself, a good indicator of the accuracy of the p-value of the

test. Thus, you must generally carry out more permutations with this procedure (perhaps 20,000) than is usual for most Monte Carlo methods (often 1,000). If you do too few permutations, your test will generally be conservative and your p-value inaccurate (Manly 1995). How do you know how many permutations to do? One way is to keep increasing the number of permutations until the p-value stabilizes (Bejder *et al.* 1998). So, for instance, start off with say 1,000 permutations. Then try with 2,000. If the p-values are similar, you can stop. If not, try 4,000, etc., until your p-value stays fairly constant. For large data sets, these permutations can take some time, so I suggest you initially run them with just a few permutations (e.g. 100), so that you know whether you will need an overnight run to do 20,000. We have not found a reliable way to predict how many permutations will be needed to stabilize p-values.

If you have degenerate data so that permutations which satisfy the constraints cannot be made (e.g. if the definition of group is the same as the sampling period so that all individuals identified in a sampling period are grouped), then you will get a message that the matrix is degenerate, and a test cannot be carried out. Generally, increasing the size of the sampling period (4.1) helps with such problems.

A variation introduced in SOCPROG2 is that it is possible to do a number of flips (inversions of the intersection of 2 rows and 2 columns in a matrix) for each calculation of the association matrix (a 'permutation'). Simulations suggest that about 1,000 trials per permutation is optimum in most cases. You set the number of permutations in the dialog box at the top of the 'Analyzing Associations Indices' window and the number of trials per permutation (flips) in the framed area in the centre-right of the window.

If you have set a class variable (6.3.1), you can check the 'between classes' checkbox and two drop-down menus will appear, allowing you to choose 'from' and 'to' sets of individuals on which to perform the tests. Thus, for instance, you can test whether females have preferred/avoided associates among males.

Running these tests produces a random data set which you can compare with the real data set in a couple of other analyses (e.g. 'Distribution of associations', 6.11; and 'Temporal analyses', 7.3.2).

To run the tests for preferred/avoided association, you should check, and may adjust, the 'number of permutations' (on the top right of the 'analyzing association indices' window), the 'trials per permutation', '2-sided significance level for dyads' (6.18.4), and, if you have set a class variable, you can select 'from' and 'to' sets of individuals (see above). Before pressing 'Test' and doing the permutation test, you can also select the permutation method. Three methods are available (plus a separate and different test for differences in gregariousness when groups are not defined, 6.18.6):

6.18.1 *Permute groups within samples (only possible when groups are defined)*

In this case the null hypothesis is that there are no preferred or avoided companions (individuals who preferentially group together or avoid one another) given the number of groups each individual was seen in during each sampling period. So individuals seen in lots of groups in the same sampling period are likely to group together at random, and this is accounted for. In this test, for each sampling period, the elements of the incidence matrix of groups by individuals are permuted, keeping row and column totals constant. This test accounts for situations in which not all individuals are present in each sampling interval (because of birth, death, migration, etc.), but

not for differences in gregariousness between individuals.

These tests use a modification of the Manly/Bejder et al. procedure in which, at each step, a sampling period is randomly chosen within which the data are to be flipped (Whitehead 2008b, 127-128). This option tests for both long-term (between sampling-period) and short-term (within sampling-period) preferred companions. Simulations, using half-weight and simple ratio association indices (6.18.5) suggest that long-term preferred companionships are indicated by a significantly high SD of the real association indices, whereas short-term preferred companionships are indicated by a significantly low mean of the real association indices.

6.18.2 *Permute all groups (only possible when groups are defined)*

This is similar to the permute groups within samples test, in that the null hypothesis is that there are no preferred or avoided companions given the total number of groups each individual was seen in during the study. So individuals seen in lots of groups are likely to group together at random, and this is accounted for. In this test, the elements of the entire incidence matrix of groups by individuals are permuted, keeping row and column totals constant. This test does not account for situations in which not all individuals are present in each sampling interval (because of birth, death, migration, etc.), and so may indicate preferred companionships (rejection of null hypothesis) in such situations when this is only because a pair of individuals were in the study area together. Also the method does not account for differences in gregariousness between individuals, and groups must be independent. Differences in gregariousness or non-independent groups can cause rejection of the null hypothesis. Thus this test is generally only recommended for short, small studies, when all individuals are likely to be in the study area in every sampling period, and groups are sampled independently, such as studies of the social structure of a small number of captive animals. Preferred/avoided companionships are indicated by a significantly high SD of the real association indices.

6.18.3 *Permute associations within samples*

In this test, the null hypothesis is that there are no preferred companions *between* sampling periods, given the number of associations each individual has in each sampling period (Whitehead 2008b, 129-130). Thus it looks for long-term companionships or avoidances (the length of term depending on the interval between sampling periods). In this test, the elements of the symmetric association matrix are permuted for each sampling period keeping row (and thus column) totals constant. The permuted mean association indices should almost equal the real means (because of the constraints). These tests use a modification of the Manly/Bejder et al. procedure in which a symmetric matrix is sequentially flipped by first choosing two individuals for the rows, and then two more individuals, different from the first pair, for the columns. This test can only detect long-term (between sampling period) preferred/avoided companionships. Long-term preferred/avoided companionships are indicated by a significantly high SD of the real association indices. As this test does account for differences in gregariousness, and animals moving in and out of the study area, it is generally the most robust, and preferred, of the three alternatives offered by SOCPROG. This test is not available if association is not defined as 1:0 (e.g. if you have an association type using the number of groups in a sampling period, 4.3.1.2, or some other kind of weighting, 4.3.1.3)—a message appears in the command window: ‘Tests for preferred/avoided association by

permuting associations only work on 1:0 associations'.

6.18.4 *Dyadic significance levels*

One advantage of the Manly/Bejder et al. procedure is that it allows you to find dyads that have significantly large or small associations. To do this, type a number such as '0.05' in the editable box titled '2-sided significance level for dyads'. This will list all dyads whose real association index is either greater than 97.5% or less than 2.5% of their random association indices. The actual and expected numbers of such significant dyads is also given. If the actual number is less than, or similar to, the expected number, you should not pay too much attention to these. You can apply a Bonferroni correction to find highly significant dyads, although this will only be feasible with fairly few individuals and a large number of permutations (as you will be multiplying significance levels by (n-2).(n-1)/2 where n is the number of individuals, assuming the 'from' and 'to' sets of individuals are the same). However (VERY IMPORTANT) you will often need MANY more permutations to get accurate dyadic significance levels than for the overall association matrix, so before you use these dyadic significance levels, run the test several times to make sure you have stable dyadic p-values (Whitehead *et al.* 2005). Also (IMPORTANT), do NOT use these dyadic p-values as measures of the strength of association between pairs of individuals–they are a function of the structure of the data as well as the strength of association. Use the association indices as measures of dyadic strength of association, and the p-values as indicators of their reliability.

6.18.5 *Suggested strategies for running tests for preferred/avoided associations*

An important consideration when using the 'permute groups within samples' (6.18.1) or 'permute associations within samples' (6.18.3) options is the length of the sampling period. If it is too short, so few groups or associations occur within a sampling period, then the permutations are restricted and the tests have little power. If it is too long, then individuals may move into, or out of, the study area within sampling periods which can produce significant test results even when no preferred/avoided companionships exist. (The lagged identification rate analysis, 12.1.2, may help you choose a suitable sampling interval.) If association is not defined using groups then only the third option, 'permute associations within samples' (6.18.3), is possible, and I suggest you use the SD or CV of the association indices as a test statistic, rejecting the null hypothesis of no preferred/avoided long-term (between sampling period) companionships if the SD or CV of all association indices is significantly high.

If association is defined using groups, there are more possibilities. The results of simulations (Whitehead *et al.* 2005) using both simple ratio and half-weight association indices suggest that:

- Short-term (within sampling period) preferred/avoided companionships are best distinguished by:
  - Using the 'permuting groups within samples' option and rejecting the null hypothesis of no preferred/avoided short-term companionships if the mean of all association indices is significantly low.
- Long-term (between sampling period) preferred/avoided companionships are best

distinguished by:

> Using the 'permuting groups within samples' or 'permuting associations within samples' options and rejecting the null hypothesis of no preferred/avoided long-term companionships if the SD or CV of all association indices is significantly high. If there are short-term associations, this will tend to lower the SD of the associations as well as the mean, so it may be better to use the CV as a test statistic for long-term associations (Whitehead *et al.* 2005).

Despite these enhancements of the two methods which permute groups, they make more assumptions than the permutation of associations method (6.18.3), and I now believe that the permutation of associations should be generally preferred.

6.18.6 *Tests for differences in sociality or gregariousness*

The Manly/Bejder et al. procedure can also be used to test for differences in sociality, or gregariousness, among individuals: are some individuals found in consistently larger, or smaller, groups than others (Whitehead *et al.* 2005)? If association is defined using groups, and either the 'permute groups within samples' (6.18.1) or 'permute all groups' (6.18.2) options are chosen, then the output includes a test statistic 'SD of typical group sizes' (typical group sizes are the group sizes experienced by individuals Jarman 1974). High values of this statistic (compared with those from the random data sets), and correspondingly significant p-values, suggest that there are some individuals that are found in consistently large or small groups. If the dyadic option is chosen (6.18.4), the individuals with significantly large or small typical group sizes are listed, together with their p-values. If 'to' and 'from' classes of individual are selected, then the test is whether the number of 'to' individuals grouped with each 'from' individual differs among the 'from' individuals, so, for instance, whether males differ in the number of females they associate with.

If association is not defined using groups, this test is not feasible. However SOCPROG provides another option, using a different kind of permutation test (outlined by Whitehead 2008b, 93-95). Choose the 'Permutation test for gregariousness' option. Then the test statistic is the SD or CV of the mean number of associates of each individual in each sampling period. If these are unexpectedly large then this indicates that there are individuals with consistently high and low numbers of associates. The distribution of expected values under the null hypothesis is found by randomly permuting the identities of the individuals that were identified during each sampling period. Unlike the Manly/Bejder et al. test, these permutations are independent and so fewer are needed and the 'trials per permutation' option is not used. 1,000 or so permutations should normally suffice. Dyadic significance levels are not available. However, tests can be made between classes, which examine hypotheses such as 'Do females differ in the number of males that they associate with?'

6.19 Measures of asymmetry

If you have an association index or interaction rate that is asymmetric (so that the value for A with B is not necessarily the same as for B with A), then several options appear at the bottom of the 'Analyzing association indices/interaction rates' window. With such an asymmetric association index or interaction rate, we can calculate, for each dyad, a measure of the asymmetry.

SOCPROG can produce five such measures (see Whitehead 2008b, 115-117):

- Beilharz and Cox's (1967) measure which is simply the difference of the interaction rates between A and B and between B and A divided by their sum, so it varies between -1 and 1, with a value of 0 indicating no asymmetry. If the interaction rates are counts, then SOCPROG can estimate standard errors using the analytical formula on p. 116 of Whitehead (2008b), or the bootstrap method (if the data were input using SOCPROG). These both assume interactions (or probability of an interaction in a sampling period, if using sampling periods) are independent. The analytical method gives an estimated standard error of zero if either count is zero, which is perhaps not ideal.
- Van Hooff and Wensing's (1987) 'directional consistency index', which is the absolute value of the Beilharz and Cox measure, and so always positive. If the interaction rates are counts, then SOCPROG can estimate standard errors using the analytical formula on p. 116 of Whitehead (2008b), or the bootstrap method (if the data were input using SOCPROG). These both assume interactions (or probability of an interaction in a sampling period, if using sampling periods) are independent. The analytical method gives an estimated standard error of zero if either count is zero, which is perhaps not ideal.
- De Vries et al.'s (2006) dyadic dominance index which is the interaction rate between A and B (plus 0.5), divided by the sum of the interaction rates between A and B and between B and A (plus 1.0). This only makes sense (and SOCPROG will only calculate it) if the interaction rate is a count. SOCPROG can estimate standard errors using the analytical formula on p. 116 of Whitehead (2008b), or the bootstrap method (if the data were input using SOCPROG). These both assume interactions (or the probability of an interaction in a sampling period, if using sampling periods) are independent.
- The likelihood-ratio G-test of the null hypothesis that there is no asymmetry in the probability of interaction between A and B and between B and A. This only makes sense (and SOCPROG will only calculate it) if the interaction rate is a count, and is only valid if the counted interactions (or sampling periods) are independent. SOCPROG gives the value of G and the p-value for each dyad.
- The chi-squared test of the null hypothesis that there is no asymmetry in the interaction rates between A and B and between B and A. This only makes sense (and SOCPROG will only calculate it) if the interaction rate is a count, and is only valid if the counted interactions (or sampling periods) are independent. SOCPROG gives the value of chi-squared and the p-value for each dyad. The results should be very similar to those from the G-test.

These measures, as well as se's for the first three (optional), and p-values for the last two, are available by pressing the 'measures of asymmetry' pushbutton,

6.20 Tests for reciprocity/unidirectionality

Another pushbutton conducts tests for reciprocity or unidirectionality. These tests, which were introduced by Hemelrijk (1990b), investigate the hypothesis that an asymmetric interaction measure is reciprocal; in other words that the rate of interaction between individuals A and B is correlated with that between B and A. If there is no correlation, the interaction measure is said to be unidirectional (Hemelrijk 1990b). Reciprocity/unidirectionality is examined in SOCPROG by

correlating an association matrix with its transpose (receivers become actors and vice versa). Two types of reciprocity are tested:

Absolute reciprocity Here individuals return interactions to other individuals based upon the absolute frequency of interactions that they receive, compared with the overall distribution of these interactions within the population (Hemelrijk 1990b);

Relative reciprocity Here individuals return interactions to another individual based upon the frequency of interactions that they receive from that other individual, relative to the rates at which they receive interactions from other members of the population (Hemelrijk 1990b).

Relative reciprocity implies absolute reciprocity, but not necessarily the converse.

SOCPROG carries out four tests for reciprocity/unidirectionality:

- Mantel Z-test of the association matrix with its inverse–this tests for absolute reciprocity, but can be strongly affected by large (or small) outlying values.
- Dietz's (1983) R-test of the association matrix with its inverse–this is the same as a Mantel test but values of the matrix are replaced by their ranks, so this is analogous to Spearman's rank correlation coefficient. This tests for absolute reciprocity, and is much less strongly affected by large (or small) outlying values than the Mantel test.
- Rr-test of the association matrix with its inverse (Hemelrijk 1990b). In this case, the values in each row of the matrix are replaced by their within-row ranks. This tests for relative reciprocity. Hemelrijk (1990b) found this to be less powerful than the Kr-test in detecting relative reciprocity.
- Kr-test of the association matrix with its inverse (Hemelrijk 1990b). In this case, the values in each row of the matrix are compared with all other values within the row (as in Kendall's τ non-parametric correlation coefficient). This tests for relative reciprocity. Hemelrijk (1990b) found this to more powerful than the Rr-test at detecting relative reciprocity.

SOCPROG does all these tests using the number of permutations set on the top right of the 'analyzing association indices/interaction rates' window, and gives one-sided p-values for tests of reciprocity (alternative hypothesis) against unidirectionality (null hypothesis). In the first three cases it also produces matrix correlation coefficients, the correlation between elements of the test matrix with those in its inverse.

If there are 'NaN's ('not-a-number') in the association matrix, they are ignored when performing the tests. This can be very useful when data are unavailable for some dyads, or should be ignored for other reasons (e.g. mother-offspring).

These analyses can be performed on parts of the association matrix, between classes of individuals, if a class variable has been set (6.3.1), so you can look for reciprocity only among females or between females and males (Hemelrijk 1990b). In the latter case, both classes of individuals are permuted separately. To test for reciprocity between or within classes, set a class variable (6.3.1). Then, when you click on 'Test for reciprocity/unidirectionality', a dialog box comes up asking whether you wish to test between classes, click 'Yes', then select the 'from' and 'to' classes in the small window that appears, and click 'Go'.

These tests only look within one association measure; to compare two or more association measures, see the analyzing multiple association measures module (8.4.4).

6.21 Analyze dominance hierarchy

This option analyzes the matrix of interaction rates from the perspective that the animals may form a dominance hierarchy indicated by the rates of interaction of a strongly asymmetric behavior (such as the winners of agonistic encounters, submissive behavior, or priority access to resources), so that within a dyad the more dominant individual is more frequently the actor than the reactor. The dominance measures available are summarized by Bayly et al. (2006) and Whitehead (2008b, 186-195). A number of the measures (indicated below with asterisks, '*'s) are only appropriate for count data, so SOCPROG does not present them if the interaction rates are not all integers.

- Landau's (1951) *h*, a measure of linearity. A perfectly linear hierarchy is one in which there are no inconsistencies, so that if A dominates B, and B dominates C, then A dominates C. *h* ranges from 0 in which each individual dominates exactly half the others, to 1 for a perfectly linear hierarchy.
- De Vries' (1995) *h*', a modification of *h* to deal better with unknown dominance relationships (*).
- Also presented by SOCPROG is a permutation test of the null hypothesis that dominance is random against the alternative that dominance is somewhat linear (de Vries 1995) (*). This uses the number of permutations set on the top right of the 'analyzing interaction rates' window.
- De Vries et al.'s (2006) measure of steepness, the probability that a more dominant individual wins an interaction (*). In a very steep hierarchy, (steepness close to 1.0), dominants always win, whereas in a shallow one (steepness near zero) outcomes of contests are not predictable. SOCPROG also gives the results of a randomization test, suggested by De Vries et al. (2006), of whether the steepness is significantly greater than would be expected if the proportion of interactions won by each member of a dyad was random (*).

SOCPROG gives up to three dominance indices for each individual. The difference in dominance index between individuals indicates how much one individual dominates another (Whitehead 2008b, 190-191).

- The proportion of interactions won. This is simply the total of the interaction rates where the individual was the actor divided by the total of the interaction rates where the individual was the actor or reactor. This is equal to 1.0 if the individual was always the actor, and 0.0 if it was always the reactor.
- David's (1987) score, which has been considered to be the best of the dominance indices (Gammell *et al.* 2003). Animals that usually dominate have high positive scores, and those that are usually dominated have large negative scores.
- De Vries et al.'s (2006) modification of David's score for count data (*).

SOCPROG also gives several dominance rankings, in which animals are ranked based upon their dominance over other individuals. For details of these methods, see the cited papers or Whitehead (2008b, 189-190). For each ranking method, the individuals are listed from most to least dominant (calculated using a modification of the method of de Vries 1998). If more than one ranking performed equally well, all the optimal ones are presented.

- The 'I' method of Slater (1961), which minimizes the number of inconsistencies (where a lower ranking individual has a higher interaction rate as actor than its more highly ranked dyadic partner).

- The ‘I&SI’ method of de Vries (1998), in which parts of the hierarchy that are unresolved by the ‘I’ method are decided by minimizing the sum of the rank differences between individuals whose ranks are inconsistent.
- Brown’s (1975, 86) method which minimizes the proportion of dyadic interaction rates in which a lower ranking individual is actor.
- Crow’s (1990) method which maximizes the sum, over dyads, of the difference between the two interaction rates multiplied by the difference in ranks.

7 TEMPORAL ANALYSES

The analyses in this module are based on computing and displaying lagged association rates (estimates of the probability that if two individuals are associating now, they will still be associated various time lags later) as described by Whitehead (2008b, 195-214). However there are a number of options and extensions.

To perform the temporal analyses you must have entered a data set (3), set a sampling period (4.1) and defined and calculated associations (4.3). You often will have defined restrictions (4.2). The temporal analyses do not work with interaction data (they probably could be made to do so, but this has not been achieved yet). Then click on ‘Temporal analyses’ in the master SOCPROG window.

The analyses are designed to work on association defined as ‘Grouped in sampling period’ and may not work on ‘Number of groups in sampling period’, or ‘weighted’ options (4.3.1.2, 4.3.1.3).

7.1 Types of association rate

A pop-up menu at the top lets you choose the type of association rate to be analysed:

7.1.1 *Lagged association rate*

This is an estimate of the probability that if two individuals are associating now, they will still be associated various time lags later (Whitehead 1995). You will almost always want to do this analysis. Other options on this menu help you interpret the lagged association rate.

7.1.2 *Null association rate*

This is a modification (a more valid modification) of the null association rate proposed by Whitehead (1995). This is the expected value of the lagged association rate if there is no preferred association (i.e. if the probability that A and B associate is independent of whether they have associated before), given the sighting histories of the individuals (who was seen in which sampling period) and the number of associations of each individual in each sampling period. It will generally be less than or equal to the lagged association rate. When the lagged association rate equals the null association rate, this indicates no preferred associations over these time lags.

7.1.3 *Intermediate association rate*

The intermediate association rate is similar to the lagged association rate but, between any two individuals, only associations, and potential associations (of the first individual), between the first and last recorded association of this dyad are considered (Whitehead 1995). In this case the time lag is the minimum of the time between the sampling period and the first or last sighting of the individuals together. It will approximate 1.0 if associations with long lags are between members of permanent groupings which do not disassociate between observed associations. If long-term reassociations often follow periods of separation then the intermediate association rate's relationship with time lag may be similar to that of the lagged association rate. You need a fair amount of data to compute meaningful intermediate association rates. Other drawbacks of the intermediate association rate are discussed in Whitehead (2008b, 199-200). It should be used rarely and carefully.

7.1.4 *Saved curve*

This lets you plot (or add to a plot) a saved curve, or curves. See 7.6 for saving curves. This takes the curve, and legend for it, from a saved MATLAB file (.mat). You can add it to already completed plots or make a new plot, so, for instance, comparing lagged association rates from different study areas, species or time periods.

7.2 Plotting and calculating

To plot a lagged, null or intermediate association rate, or replot a saved curve, press either 'New plot' or 'Add to plot' buttons. (If you selected 'Saved curve' a window will ask you for the name of the file where the curve was saved.) 'Add to plot' will add what you have selected to the last plot you made with this temporal analysis window, if there is one. You cannot add standardized (7.3.3) curves to unstandardized curves or vice versa, and you cannot add curves with different sampling periods.

If two or more curves are plotted on a figure, then a legend is added. You can edit the legend (click on the arrow above the figure, right click on the legend, click on 'Show property editor', click on 'More properties...', click on 'String...'). You can also move the legend by dragging it with the mouse.

If the data set is large, calculating association rates may take a little time (see 7.7.1).

7.3 Other options

You can set a number of other options for the analyses of lagged (and other) association rates:

7.3.1 *Name of analysis*

You can enter a name for the analysis which will appear in the command window following an analysis, and in the legend if there are several curves plotted with different names given.

7.3.2 *Random data*

If you have already run a permutation test (6.18) a check box appears allowing you to run analyses on the random data from the last run of the permutation tests. This may be useful when trying to decide whether observed patterns in lagged association rates are important features. If you ran the "Permute all groups" option (6.18.2), the data should have been truly scrambled and you will likely get a lagged association rate plot similar to that of the null association rate (7.1.2). In contrast if you used the "Permute associations within samples" (6.18.3) or "Permute groups within samples" (6.18.1) options, only the social side of the data, not its temporal content, were scrambled. Then if the random data lagged association rate plot looks similar to that for the real data, the implication is that the changing patterns of association are more due to demographic patterns (individuals being born, dying or leaving the study area), than changing affiliations. In contrast, if the random and real plots are very different, this suggests that affiliation patterns change over time.

7.3.3 *Standardized rates*

If you check the 'standardized' check box, any lagged association rates (or null or intermediate rates) are calculated as 'standardized' (i.e. considering the effort during the second sampling period) using the terminology of Whitehead (1995, Appendix B), or Whitehead (2008b, 197-199). Standardized rates should be used when not all true associates of an individual are recorded during a sampling period in which it is seen. They can be interpreted as follows:

Standardized lagged association rate: this is an estimate of the probability that if two individuals are associated at any time, then, after the specified lag, the second individual is a randomly chosen associate of the first. The intercept on the y-axis (with lag zero) is an estimate of the inverse of the mean (over individuals) typical group size (number of associated individuals including itself Jarman 1974) minus one (Whitehead 1995).

Standardized null association rate: this is an estimate of the probability that if two individuals are associated at any time, the second is a randomly chosen associate of the first after the specified lag, if associations are completely random over that time lag. The standardized null association rate is the inverse of the population size minus one, and so does not change with time lag.

Standardized intermediate association rate: this is an estimate of the probability that, if two individuals are associated at any time, the second is a randomly chosen associate of the first after the specified lag, given that the lag is within the span of time over which the individuals were associated. The standardized intermediate association rate will approximate the standardized lagged association rate of that lag if individuals frequently disassociate between their first and last associations, and remain constant (at approximately the standardized lagged association rate with very small lag) if there is no disassociation.

7.3.4 *Log x-axis*

If you check this, the x-axis (time lag) of the lagged association rate plot is shown on a

log-scale. This is appropriate in many cases when time differences between sampling periods range over an order of magnitude or more.

7.3.5 *Moving average*

Lagged association rates (and null and intermediate rates) are plotted continuously against time lag, using a moving average method. You can change the number of potential associations over which the lagged association rate and its associated lag is calculated. High values will give a smoother curve, and truncate the ends; plots using low values tend to have lots of random noise. What is high or low will depend on the situation, so experiment. If you use too high a value (greater than the number of associations), the program will use the number of associations, and you will just get one point on the plot (a star). You may wish to try again with a smaller number of associations for the moving average. You can use different moving averages for different lines (of the same or different data) on the same plot. Standardized (7.3.3) rates usually need a higher number of potential associates over which to make the moving average. The median time lag (in sampling periods) over which the moving averages were calculated will also be reported in the command window. This is useful in assessing the precision of points along the x-axis.

7.3.6 *Analyses by classes*

You may enter a class variable in the editable box using the supplemental fields or functions of them (see 6.3.1 for examples). Supplemental variables are listed in a popup menu; press ‘s’ to put the names of the selected supplemental field into the class-defining editable box. Then you can carry out the analyses between the classes (e.g. lagged association rates from males to females), using the ‘From’ and ‘To’ pop-up menus.

7.3.7 *Maximum lag*

You can set a maximum lag (in sampling intervals) to be considered at the bottom of the window on the left. This is useful if you only are interested in short time intervals but your data span long periods, or to shorten calculation time for very large data sets. ‘0’ (the default) indicates that no maximum lag is set.

7.4 Jackknifing

In order to obtain estimates of precision for your lagged association (and other) rates you can use the jackknife procedure in which the analysis is run several times omitting one or more sampling periods each time (Efron and Stein 1981). To make jackknife runs, check the jackknife box. You can alter the number of jackknife groupings by changing the ‘Jackknife grouping factor’ which groups sampling periods for jackknifing. The default is 1 (jackknifing on each sampling period), but if this produces too many jackknife groupings, or if the sampling periods are not independent, increase the factor. For instance, if the sampling period is 1 day, and you put the jackknife grouping factor equal to 30, then you are jackknifing (approximately) on months of data. After resetting the jackknife grouping factor and pressing ‘Enter’, the display of the number of jackknife groupings on the bottom right is updated.

Estimates of precision from the jackknife procedure are approximate for several reasons: it is an approximate and conservative (i.e. tends to underestimate precision) procedure (Efron and Stein 1981), and it assumes independence of jackknife groupings, which may not be strictly true. However, it will generally give a reasonable idea of the approximate precision of your plots and parameter estimates.

On the plots, usually five equally spaced jackknife error bars (± 1 estimated standard error) will be displayed. Sometimes, there will be insufficient data to produce jackknife estimates. If so a message appears in the command window, and the plot is drawn without error bars. Try a smaller grouping factor and/or moving average.

7.5 Fit model(s)

After you have run a lagged (or other) association rate analysis, you can fit mathematical models to the data by pressing the 'Fit model(s)' button. After pressing this, you see a window with a number of models of how the lagged (or other association rate) changes with time lag, and an informal 'explanation' of each model. You will get a different set of models depending on whether or not the association rates have been standardized. The model explanations include the following abbreviations:

- 'Pref. comps' = Preferred companions: some pairs of individuals have a preference for associating, which is constant over time;
- 'Casual acqs' = Casual acquaintances, who associate for some time, disassociate, and may reassociate;
- 'Rapid dis.' = Rapid disassociation; some associates disassociate very quickly, within one time period.

The model explanations should not be taken literally without some thought and perhaps additional analyses as different types of social system can produce similar shapes of lagged association rate which fit the same mathematical model (see Whitehead 2008b, 204-211).

The models are of the exponential form proposed and used by Whitehead (2008b, 204-211). In these models the time lag is represented by 'td' and the parameters of the models by 'a1', 'a2, 'a3', ... Check beside one or more of these models to fit them.

Alternatively, or additionally, you can specify your own, custom, model at the bottom. You can also change the informal explanation from 'Custom', and you must also check the box on the left. In specifying this model, use the standard MATLAB (and 'C') mathematical notation, '**td**' for the lag, and '**a1**', '**a2**', etc for the parameters. If you have three parameters use '**a1**', '**a2**', and '**a3**'. Do not skip any in this sequence (so do not use '**a1- a3*td**'). You can also use functions like '**sin**' and '**cos**' to give cyclical lagged association rates. e.g.

a1*cos(a2*td)+a3

On the right of the screen you can set the start values of each parameter (**a1**, **a2**, ...) for each model. This is useful for checking whether the fitting, which is iterative, has reached a global optimum. So change the start values of the parameters, and see if you get the same result (or at least a worse one). Also changing the start values of the parameters is often a good way to resolve problems of model fitting (indicated by a MATLAB error message in the command window or a resultant fitted curve that looks little like the data curve).

The models are fitted using maximum likelihood and binomial loss to the full data set.

This is a more legitimate procedure than that used by Whitehead (1995), but in practice the procedure used for fitting curves seems to make little difference.

The fitted model is plotted on the graph of lagged (and/or other) association rates, and in the MATLAB command window are given the model type, the informal explanation (DO NOT TAKE AS PRESCRIPTIVE!), the fitted parameters (with standard errors if the jackknife has been used, 7.4) and the summed log-likelihood. Also given are the results of a goodness-of-fit chi-squared test of the number of repeat associations in each time lag bin against the expected number given the model, and lumping time lags so that the expected number is at least six in each bin, as well as the variance inflation factor (chi-squared statistic divided by degrees of freedom).

7.5.1 *Model selection*

Because of a lack of independence of data points (associations), the summed log-likelihoods from different models cannot be used for formal likelihood ratio tests. However, I have done some simulations (Whitehead 2007) which suggest that the quasi Akaike Information Criterion (QAIC) provides some basis for selection among models of lagged association rates:

$$\text{QAIC} = -[2 \cdot \text{Summed log-likelihood} / \hat{c}] + 2 \cdot K$$

where $\hat{c}$ is the variance inflation factor for the most general of the models being compared, and K is the number of parameters being estimated plus one (Burnham and Anderson 2002). The model with the minimum QAIC is selected, and the difference between the QAIC of any other model and the selected one, ΔQAIC, gives an indication of how well the data support the less favoured model (Burnham and Anderson 2002):

ΔQAIC: 0-2	substantial support for model
ΔQAIC: 4-7	considerably less support
ΔQAIC: >10	essentially no support

Although the AIC criterion is also given for each model, my simulations suggest these should not be used for model selection.

If using the QAIC for model selection, you must run all the models being considered simultaneously (just one press of the 'Fit model(s)' button), so that the variance inflation factor can be taken from the most general model, and applied to all of them. DO NOT COMPARE QAIC's FOR MODELS RUN SEPARATELY.

7.6 Saving curves

You can save the currently plotted curves, and their legends, by clicking this button. A window appears asking you the name of the file (the extension '.mat' is automatically attached). You can then use these to add to plots made later (7.1.4), such as those from different data sets, or made with different restrictions (4.2).

7.7 Some technical computation issues

7.7.1 *Speed of calculation*

Subsequent calculations of lagged association, and other rates, use saved data from the earlier calculation, and are much faster as long as you don't change: the type of analysis (lagged, null or intermediate); the 'To' and 'From' class choices; the maximum lag; whether randomized data are used; or whether the rates are standardized or not. Thus you can do repeat calculations over different moving averages, or with different jackknife intervals, quickly even when you have a large data set which slows the initial analysis.

7.7.2 *Calculation of intermediate association rates*

The method of calculation of intermediate association rates was chosen for its performance on a large, sparse test data set (~1,500 individuals, ~250 samples). However, a second method of calculating intermediate association rates is included in the code. This second method produces the same results as the default method but is substantially slower with the test data set. However, for a data set which is much less sparse, and, say, contains a lot much more information about the individuals over fewer sampling periods, the second method might be faster (although I have not tried this). To change methods, edit line 161 (currently–line number may be changed) of the m-file *temporal.m*, which now reads

intype=1;%type of calculation for intermediate association rate

Change this to:

intype=2;%type of calculation for intermediate association rate

8 ANALYSIS OF MULTIPLE ASSOCIATION MEASURES

This section of SOCPROG is for analyzing situations when there are two or more measures of relationship between pairs of individuals, as well as for when a precalculated association matrix is available (which can be entered either from the master SOCPROG window, 3.8, or here in various formats, 8.1.1, 8.1.3, 8.1.4). These are often symmetric matrices of association indices (6.1), for instance the half-weight association indices based on how often individuals are found in the same group, or sharing a roost, but they may be asymmetric relationship measures, such as grooming rates, or non-social measures of similarity or dissimilarity, for instance age differences between individuals or range overlaps or estimates of genetic relatedness. So in this module, we consider issues such as: 'Do dyads have different patterns of association in different behavioural states (such as feeding or resting)'; 'Do individuals that are closely genetically related spend more time in the same group than individuals which are not so closely related?'; 'Are rates at which individuals groom one another inversely related to the proportion of time they spend in the same groups?'; 'How can we best summarize several interaction/association measures to produce an overall relationship measure?'.

Unlike other modules in SOCPROG, 'analyzing multiple association matrices' does not necessarily have to directly follow entering a raw data set (3). You enter association measures directly in the module, although these may come from previously calculated (6.1, 6.2) and saved (6.6) association indices or interaction rates. You can also enter association matrices in Excel or ASCII files directly on the master SOCPROG window (3.8).

To run this module click on 'Multiple measure analyses' in the master SOCPROG

window, or enter it automatically after inputting an Excel or ASCII association matrix file on the master SOCPROG screen (3.8). You can then enter association measures (arranged in square matrices) in several ways, and perform a variety of analyses. Each measure has four pieces of information:

- the names of the individuals indexing it in the appropriate order;
- the square matrix of the measure itself giving the association measure between each pair of individuals;
- a multi-line description (such as data file, restrictions, association index type, ...);
- a short 'name' of the measure (which should not include spaces).

8.1 Entering association measures and supplemental data

Association measures can be entered singly, or as groups, into these analyses. Entered measures are listed in the list box at the left on the 'analyses of multiple association measures' window. Each association measure has a short name which is shown in the list. Clicking on one or more (using **Ctrl** for selecting several measures) of these names enables them to be entered into subsequent analyses. Association measures can be entered to this list directly from the SOCPROG master window (3.8), or from the popup menu at the right of the screen. Supplemental data can also be entered using the popup menu. Here are the possibilities (from top to bottom on the popup menu):

8.1.1 *Input SOCPROG association measure*

The easiest format for an association measure is that created by SOCPROG itself. Matrices of association indices or interaction rates are created and then saved using the 'Saving association matrices (MATLAB)' pushbutton (6.6). On clicking the 'Input SOCPROG association measure' pushbutton on the 'Analysis of multiple association measures' windows, small windows appear in which you select the file with the association measure, and choose a short name for the measure (without spaces), assuming one has not been previously chosen (by previously saving a measure from this window, 8.2.7). The short name of the measure then appears on the list at the left, and the measure is available for analysis.

8.1.2 *Input SOCPROG set of association measures*

This pushbutton allows you to enter a set of association measures previously saved using the pushbutton lower on the list (8.2.8). The file name used is placed in the title of the list box of association measures at the left.

8.1.3 *Input EXCEL association measure*

Association measures can be entered from Excel using this option. In Excel, the association measure should look like that in Table 7, with the names of the individuals in the first row and column. You can enter 'NaN' for missing data in cells. You then enter the multi-line description, and the short name of the measure (no spaces), in a dialog box. To make sure that the

EXCEL measure has been correctly entered, I advise that you display it after it has been entered (8.2.5).

Table 7. Example of EXCEL file containing association measure.

	Andy	Bert	Charlie	Deb	Elen	Fran	George	Harry
Andy	1	0.2	0.4	0.7	0	0	0.6	0.2
Bert	0.2	1	0.4	0.1	0.2	0.1	0.3	0
Charlie	0.4	0.4	1	0.3	0.1	0.4	0.1	0.2
Deb	0.7	0.1	0.3	1	0.4	0.4	0.1	0.1
Elen	0	0.2	0.1	0.4	1	0	0.2	0.2
Fran	0	0.1	0.4	0.4	0	1	0.5	0.3
George	0.6	0.3	0.1	0.1	0.2	0.5	1	0.5
Harry	0.2	0	0.2	0.1	0.2	0.3	0.5	1

8.1.4 *Input ASCII association measure*

Association measures can be entered from an ASCII (.txt) file using this option. The association measure should look like that in Table 8, with the names of the individuals separated by spaces in the first row, and in each other row a name followed by the association indices. Association values and names should be separated by one or more spaces. You then enter the multi-line description, and the name of the measure, in a dialog box. To make sure that the ASCII measure has been correctly entered, I advise that you display it after it has been entered (8.2.5).

Table 8. Example of ASCII file containing association measure.

```
          Andy  Bert Charlie Deb  Elen  Fran George Harry
Andy       1.0   0.2   0.4   0.7   0.0   0.0   0.6   0.2
Bert       0.2   1.0   0.4   0.1   0.2   0.1   0.3   0.0
Charlie    0.4   0.4   1.0   0.3   0.1   0.4   0.1   0.2
Deb        0.7   0.1   0.3   1.0   0.4   0.4   0.1   0.1
Elen       0.0   0.2   0.1   0.4   1.0   0.0   0.2   0.2
Fran       0.0   0.1   0.4   0.4   0.0   1.0   0.5   0.3
George     0.6   0.3   0.1   0.1   0.2   0.5   1.0   0.5
Harry      0.2   0.0   0.2   0.1   0.2   0.3   0.5   1.0
```

8.1.5 *Input EXCEL supplemental data*

Supplemental data on the attributes of individuals can be entered from Excel files, just as in the initial entry of data, and then used to build new association measures (8.2.1) and for other purposes. See 3.1.2 for format. Supplemental fields are listed in the centre of the window; fields

which are string (equivalent to alphanumeric) are marked by ‘s’ in the list; and the name of the Excel file containing the supplemental data is placed in the title of the list box of supplemental fields.

8.2 Creating, manipulating, saving, removing, listing and renaming association measures

A series of pushbuttons on the right of the window allow association measures to be created from supplemental data, manipulated, removed from the analyses, saved, listed or renamed. Here are the possibilities (from top to bottom):

8.2.1 *Make association measure from supplemental data*

Having input supplemental data from an Excel file (8.1.5), association-type measures can be produced in a range of ways. This ability can be very useful in producing measures showing similarity of gender, genetic relatedness, and so on. To make such measures, first highlight the supplemental fields that you wish to use, and only these fields. Then, a window comes up with these fields listed at the left, and several options on the right.

8.2.1.1 Same/different

One or two push buttons appear at the top right: one button if only one supplemental field was highlighted (‘Same (1); Different (0)’), two if more than one field were highlighted (‘All fields same (1); Any different (0)’ and ‘Any fields same (1); All different (0)’). These options produce 1:0 association measures. In the case of just one supplemental field, dyads have a ‘1’ if they have the same value of the supplemental field, ‘0’ if they have different values. With more than one field and ‘All fields same (1); Any different (0)’, then the dyads have a ‘1’ in the association matrix if they have the same value in all fields, and ‘0’ if there are differences in any field. Conversely, with ‘Any fields same (1); All different (0)’, then the dyads have a ‘1’ in the association matrix if they have the same value in any field, and ‘0’ if there are differences in all fields. You then enter the multi-line description, and the name of the measure, in a dialog box.

8.2.1.2 Distance measures

Lower on the right hand side of the small window is a popup menu with several distance measures available. These distance measures work with continuous variables, so string (alphanumeric) fields are converted to numeric, by making the first string ‘1’, the second ‘2’, etc. The conversion codes are given in the command window. Select a distance measure and press ‘Go’. The distance measures are described and defined in the MATLAB Help documentation under ‘pdist’. If you select the ‘minkowski’ distance measure, an editable box appears in which the minkowski exponent can be changed–the default is ‘2’. If you select ‘custom’, a dialog box appears allowing you to select a MATLAB function m-file describing a custom distance measure function of the form:

function d = distfun(XI,XJ)

This function produces distances between pairs of individuals represented by the rows of both XI and XJ. The columns are the selected supplemental fields, converted if they are string variables, and the function produces an association measure ‘d’. For instance, if there are two supplemental fields selected, ‘Sex’ (a string variable) and ‘Age’, the following function m-file gives ‘0’ if the

sexes of the two individuals are different, and the difference in age if sexes are the same:

```
function d=disfun(XI,XJ)
d=abs(XI(:,2)-XJ(:,2)).*strcmp(XI(:,1),XJ(:,1));
```

After pressing 'Go', you then enter the multi-line description, and the name of the measure, in a dialog box. To make sure that the new measure has been correctly entered, I advise that you display it after it has been entered (8.2.5), and check one or two of the calculations.

8.2.2 *Transform association measure*

This takes the association measure highlighted on the list at the left (the first highlighted if several are highlighted), and allows you to transform it almost any way, into a new association measure. The 'Transform measure ...' window allows you to define the type of transformation. Examples include:

sqrt(X)
1-X
acos(X) [inverse of cosine]

You can also put any value you like on the diagonal, and give the transformed measure whatever short name you choose, but not including spaces. The original untransformed measure is retained in the list and the transformed measure added.

8.2.3 *Restrict association measure*

This takes the association measure highlighted on the list at the left (the first if several are highlighted), and allows you to restrict to certain individuals using the supplemental fields making a new association measure. See section 4.2 for more information on setting restrictions, including examples. Press 's' to put names of selected supplemental fields into the restriction editable box. You can give the restricted measure whatever short name you choose, but not including spaces. The original unrestricted measure is retained in the list and the restricted measure added.

8.2.4 *Remove association measure(s)*

To remove association measures from the list at the left, click on them to highlight, and then press this pushbutton. This action only removes measures from the list on the window, not the files from memory, so you can reenter them if you made a mistake or need them later.

8.2.5 *Display association measure(s)*

If one or more association measures are highlighted in the listbox at the left, they can be listed in the command window, by clicking on this button. The numeric format used is that given in the editable string at the top right: '4.2' means four characters with two decimal places, etc. You can use the select, copy and paste features of Windows to put the output measure or measures

into other documents.

8.2.6 *Change measure name(s)*

You can change the short name of one or more association measures by highlighting them on the list at the left and then pushing this button (no spaces in short names).

8.2.7 *Save as SOCPROG association measure(s)*

This allows you to save any association measures highlighted in the list box at the left. Each highlighted measure is saved as a separate MATLAB (.mat) file. In this option, the short name of the measure is also saved.

8.2.8 *Save as SOCPROG set of association measures*

This saves all the measures in the list at the left as one MATLAB file. They can later be reentered as a set (8.1.2). The file name used is placed in the title of the list box of association measures at the left.

8.3 Analyzing single association measures

At the bottom right of the 'Analyses of multiple association measures' window is a pushbutton 'Analyze single association measure'. This takes the association measure highlighted on the list at the left (the first if several are highlighted), and inputs it as the association measure in the 'Analyzing association indices' screen, so many (but not all) of the analyses described in chapter 6 are available: listing and saving association indices (but not standard errors), examining their distribution, sociograms, multidimensional scaling, principal coordinates analysis, hierarchical cluster analysis, and tests for reciprocity/unidirectionality as well as analysis of dominance hierarchies (if the matrix is asymmetric). If supplemental data are available (8.1.5), these are also transferred to the new screen and can be used to define classes (6.3.1) or labels (6.3.3).

8.4 Analyzing multiple association measures

In the bottom left quarter of the 'Analyses of multiple association measures' window are various options for analyzing multiple association measures. The potential of these analyses is discussed by Whitehead (2008b, 214-222). For all these analyses, the routine finds the individuals whose names that are common to all selected association measures, discards the others, and rearranges the association matrices so that they are all the same size and indexed by the same names in the same order. If all the selected matrices are symmetric, then only one value is used for each dyad (i.e. $x(A,B)=x(B,A)$). If any are asymmetric then both $x(A,B)$ and $x(B,A)$ are used. Self associations (i.e. $x(A,A)$) are not used.

8.4.1 *Dyadic plots*

Each dyad can be represented by a point in multivariate space, the axes of which are different association measures. To make such plots, select one or more of the association measures from the list on the left, and click on the 'Dyadic plots' pushbutton. The display you get will depend on the number of association measures that you have selected from the list:

8.4.1.1 Plot of one measure

If you just selected one measure, then a histogram of the dyadic values in that measure is displayed. If you entered a class variable (8.4.1.4), then separate bars are shown for each pair of classes of dyadic association (e.g. 'M-M', 'M-F', and 'F-F' if the matrix is symmetric, or 'M-M', 'M-F', 'F-M' and 'F-F' if it is not).

8.4.1.2 Plot of two measures

If you selected two measures, then the dyadic values are plotted against one another. If you entered a class variable (8.4.1.4), then each pair of classes of dyadic association is indicated by a different colour (e.g. 'M-M', 'M-F', and 'F-F' if the matrices are symmetric, or 'M-M', 'M-F', 'F-M' and 'F-F' if they are not). If you entered an individual label (8.4.1.5), pairs of labels for each dyad are shown beside each point.

8.4.1.3 Plot of three or more measures

If you selected more than two measures, then the dyadic values of each pair of measures are plotted against one another in a matrix arrangement, with histograms of each measure along the diagonal. With three or more measures you cannot label points or colour groups.

8.4.1.4 Defining class variables

You can define a class variable using the editable string at the top right, separating the individuals into classes using the supplemental fields or functions of them. Remember, you must enter supplemental data (8.1.5) for this to be possible. If you click on the small pushbutton marked 's' at the right side of the editable string, then the supplemental field which is currently selected in the list of supplemental fields is entered (the uppermost field is entered if several are selected) into the editable string used to define class variables. Plots can then distinguish these classes (8.4.1.1, 8.4.1.2). e.g.

Sex	<- Supplemental field *Sex*
10*(floor(*Age*/10))+5	<- *Age* in 10 unit divisions: 5, 15,25, ...

8.4.1.5 Labels for individuals

You can change how the individuals are labelled on plots using the editable string at the top right. You should write a combination of supplemental field names, '**ID**' (the name of the individual), and other characters. If an individual is not present in the supplemental data, the name is used instead. You must enter supplemental data (8.1.5) to enter anything other than '**ID**'. Here are some possibilities:

ID	<-Just the name of each individual (the default)
***ID*(*Sex*)**	<- This gives ID with Sex in parentheses, e.g. Moe(M)
	<- Nothing: leaves labels blank; may be useful for cluttered plots

8.4.2 *Principal components analysis of dyadic associations*

With several association measures, the dyadic plots (8.4.1) become hard to interpret; if they are correlated, they are, to some degree, redundant. Principal components analysis functions to reduce the dimensionality of multivariate data sets (e.g. Manly 1994), and it can be a most useful technique in the analysis of several association measures (Whitehead 2008b, 215, 218). To carry out a principal components analysis of the dyadic associations, select the association measures that you are interested in, and click on the 'Principal Components Analysis' button. A small window comes up, allowing you to choose various options for the analysis.

8.4.2.1 Correlation or covariance matrix

Principal components analyses can be carried out on either the correlation matrix or covariance matrix calculated from the data matrix (here the data matrix is indexed by dyads and association measures). Normally the correlation matrix should be used. However, if all the association measures are in the same units (say, they are all association indices but calculated in different behavioural states, such as foraging and resting) then the covariance matrix can be used. With use of the covariance matrix, association measures that have the greatest variance will tend to dominate the analysis, whereas with the correlation matrix, all measures are standardized using their standard deviations, and all will have equal impact on the output.

8.4.2.2 Number of components

A principal components analysis reduces the dimensionality of the data matrix, so if the data matrix has 5 association measures, the principal components analysis squeezes as much as possible of this information into 1, 2, 3, or 4 new, uncorrelated, variables: the principal components. How many of these components to use is up to the user. In the options window, you can set the number of components to be retained (bottom right) or retain those components with an eigenvalue greater than some value (bottom centre). The eigenvalue of a principal component indicates what proportion of the variance of the original data matrix it explains. If the correlation matrix is used for the analysis (8.4.2.1), then components with an eigenvalue greater than 1.0 (the default) explain more of the variance in the data set than the average of the original association measures–this is not true if the covariance matrix is used, and a value different from one should usually be selected. You may wish to retain most of the components for a first look, and then repeat the analysis retaining just a subset which explain much of the variance.

8.4.2.3 Output of principal components analysis of dyadic association measures

After clicking on the 'Make' button in the 'Principal Components Analysis' options window, the principal components analysis is performed. There are several outputs in the command window:

- The *coefficients* of the principal components which define how the new variables are obtained from the original dyadic association measures;
- The *loadings* which are the correlations between the principal components and the original association measures;
- A list of the principal components, and with each, the percent of the variance of the

original data set that they explain, the cumulative variance explained, and the eigenvalue. In addition, the principal components analysis produces a *scores plot*. This plots the positions of the dyads in the space defined by the principal components. The type of plot depends on how many principal components are produced:

- With just one component, the distribution of scores is shown using a histogram;
- With two components, the scores of the dyads are plotted against one another on axes with the same scale. The options for two variable dyadic plots (8.4.1.2) are also available when two principal components are output: labelling dyads and plotting classes of dyads in different colours. Additionally you can add *loadings vectors* to the plots by checking the box in the centre of the principal components options window. Loadings vectors show how the original association measures correlate with the components.
- With three or more components produced, they are plotted against each other in a matrix arrangement, with histograms of each measure along the diagonal as with the dyadic plots (8.4.1.3).

8.4.2.4 Saving principal components scores

The components produced by the principal components analysis can be saved back into the multiple association measure window by checking the box at the lower left of the principal components options window. New variables, called 'PC1', 'PC2', etc. are added to the list of association measures in the main window. These contain the scores of the dyads on the principal components arranged in square association matrices. If the matrices that are input to the principal components analyses are symmetric, then so are these output matrices; if any is asymmetric then the output matrices are also asymmetric. If the principal components scores are saved, then the options at the right of the main window are available, for instance to save them as SOCPROG association matrix files (8.2.7), list them (8.2.5), or rename them (8.2.6). They can also be compared with other association measures, in particular those not used as input for the principal components analysis, using dyadic plots (8.4.1) or tests (8.4.4). So, for instance, you can compare the first principal component of several behavioural association measures with an association measure of genetic kinship.

8.4.3 *Output of dyadic values*

SOCPROG only has a limited range of analyses of multiple association measures. However, the 'dyadic output' pushbutton allows you to export the data, and then use it in other programs or other ways. After selecting the association measures of interest in the list at the top left and clicking on this button, you may be asked whether you want to display information on those individuals 'in all measures' or those 'in any measures'. This question appears if there are some individuals in some of the selected association measures which are not in all the selected association measures. If you choose 'in all measures' you will only get data on individuals which are in all the selected association measures; if you choose 'in any measures' you will get all the data but there will be 'NaN's in cases where data for one or both of the members of the dyad are missing for that measure. You are then asked for an ASCII text file in which to store the data.

The output in this chosen file, as well as in the command window, consists of a table in which each row corresponds to one dyad. First are the names of the two individuals making up

that dyad, then their classes if classes have been selected (8.4.1.4), and then their values on each of the selected measures (with 'NaN' if data for one or both of the individuals were not entered for that measure, and 'in any measures' was chosen). Self-associations (individuals with themselves) are omitted. If all the selected association matrices are symmetric, there is only one entry for each dyad; if any are asymmetric, then rows for both 'A B' and 'B A' are shown. The numeric format used for the values of the measures is that given in the editable string at the top right: '4.2' means four characters with two decimal places, etc.

Here is an example of part of such an output file for five association measures (with sex classes and 'in any measures' selected):

```
Dyadic output for asymmetric association measures:
 Dyad       Class            AA     BB     CC     DD     EE
  100 101   M M            0.20    NaN    NaN   0.50   0.50
  100 102   M M            0.40   0.10    NaN   0.10    NaN
  100 103   M -            0.70    NaN    NaN    NaN    NaN
  100 104   M F            0.00   0.20    NaN   0.20   0.20
  100 105   M F            0.00   0.50    NaN   0.50   0.30
  100 106   M M            0.60   0.30    NaN   0.30   0.20
  100 107   M F            0.20    NaN    NaN    NaN    NaN
```

8.4.4 *Testing for relationships between association matrices*

SOCPROG allows you to test for correlations between two association matrices using the Mantel permutation test and its relatives (Mantel 1967; see Schnell *et al.* 1985; Smouse *et al.* 1986; Hemelrijk 1990a; Hemelrijk 1990b). These are permutation tests which test whether the dyadic values of one association measure are significantly correlated with those on another. So we can test whether two association measures are correlated, whether genetic relatives have higher association indices, whether individuals of similar age interact more frequently than those whose ages differ more, and so on. SOCPROG implements two general types of tests: bivariate Mantel tests and its nonparametric versions, and MRQAP tests. Both types of test use the same 'Test' command in the 'analyses of multiple association measures' window.

8.4.4.1 Mantel and related tests

The options for these tests are given in the upper part of a frame towards the bottom left of the 'analyses of multiple association measures' window. You must enter one of the association measures in the editable string 'Measure 1:' and another in 'Measure 2:'. If you click on the small pushbutton marked 's' at the right side of either editable string, then the association measure which is currently selected in the list of association measures is entered (the uppermost field is entered if several are selected) into the editable string. Click on 'Test' to run the tests. There are a few options available for Mantel tests:

These analyses can be performed on parts of the association matrix, between classes of individual, if a class variable has been set (8.4.1.4), so you can look for relationships between the two variables only among females or between females and males (Hemelrijk 1990b). To use classes, enter a class variable (8.4.1.4), click on the 'Between classes?' checkbox, and then two dropdown lists will appear from which you can select the classes that you are interested in.

SOCPROG carries out four tests for correlations between 'Measure 1'and 'Measure 2', although the last three are only carried out if the 'Other tests?' box is checked :

- Mantel Z-test–this is the standard test for correlations between two association measures, but can be strongly affected by large (or small) outlying values (Mantel 1967).
- Dietz (1983) R-test–this is the same as a Mantel test but values of the matrices are replaced by their ranks, so this is analogous to Spearman's rank correlation coefficient. This is much less strongly affected by large (or small) outlying values than the Mantel test.
- Rr-test of the correlation of 'Measure 1' with 'Measure 2' (Hemelrijk 1990b). In this case, the values in each row of the matrix are replaced by their within-row ranks.
- Kr-test of the correlation of 'Measure 1'with 'Measure 2' (Hemelrijk 1990b). In this case, the values in each row of the matrix are compared with all other values within the row (as in Kendall's τ non-parametric correlation coefficient).

SOCPROG does all these tests using the number of permutations set on the top right of the 'analyses of multiple association measures' window, and gives two-sided p-values. In the first three cases it also produces matrix correlation coefficients, the correlation between non-diagonal elements of the test matrices (ranked in the second and third cases).

If there are 'NaN's ('not-a-number') in either association matrix, the corresponding dyads are ignored when performing the tests. This can be very useful when data are unavailable for some dyads, or should be ignored for other reasons (e.g. mother-offspring).

8.4.4.2 MRQAP tests

MRQAP (Multiple Regression Quadratic Assignment Procedure) is related to the Mantel test but one of the association matrices is considered 'dependent' and the others 'predictors'. The test considers whether each of the predictor matrices makes a significant contribution towards explaining the dependent matrix while controlling for the presence of the other predictors. It uses permutation tests (the 'double-semi-partialing' technique of Dekker et al. (2007)), while measuring the effective contribution of each predictor using partial correlation coefficients. You must include a dependent variable and at least two predictor variables to run MRQAP tests. If you click on the small pushbutton marked 's' at the right side of either the 'Dependent:' editable string or 'Predictors:' editable list, then the association measure which is currently selected in the list of association measures is entered (the uppermost field is entered if several are selected) into the editable string or list. If you click on the small pushbutton marked '/' at the right side of the 'Predictors:' editable list then the association measure which is currently selected in the 'Predictors:' list is removed from it. Click on 'Test' to run the MRQAP tests.

9 POPULATION ANALYSES

This module carries out mark-recapture population analyses. There are a number of good references for mark-recapture methods of estimating population parameters (e.g. Seber 1982; Cormack 1985; Seber 1992; White and Burnham 1999). The range of options and analyses available in SOCPROG is much smaller than some other programs which are available (e.g. 'MARK', http://www.cnr.colostate.edu/~gwhite/mark/mark.htm (White and Burnham 1999), 11). However, the analyses included are those I find most useful for cetacean data, and, if you are already using SOCPROG to look at social structure, you do not need to recompile your data to do

population analyses. I have found some of these analyses hard to implement in MARK. For open population models, an important difference seems to be that MARK estimates capture probabilities for each sampling period using likelihood, whereas I take the shortcut of assuming it is the number of individuals identified divided by the estimated population size (as described mathematically in Appendix I of Gowans *et al.* 2000). While the MARK approach is theoretically better, it is much harder to implement in practice and may be impossible if the data set includes many sampling periods. Additionally all SOCPROG's data assembly and restriction sections are available, so you can easily estimate population parameters, for example, on just those females born after 1985.

To implement the population analyses, you must enter a data set into SOCPROG and set the sampling period using the sampling-restrictions-associations window (4.1). Sampling periods for population analyses are often larger (e.g. years) than is typical for social analysis. In this window, you can also set restrictions if you like (4.2). With data entered and sampling period set, click on 'Population analyses' in the master SOCPROG window.

9.1 Choosing models

You come to a 'Mark-recapture model' screen. Choose (top left) from the eight different models:

a) '*Closed (Petersen)*' This assumes a closed population (no immigration, emigration, birth, death) between each pair of consecutive sampling intervals. The size of the population in each pair of consecutive sampling intervals is estimated using the Chapman modification of Petersen's two-sample estimate (Seber 1982). The standard error of the estimate is from the formula in Seber (1982). If several Petersen estimates are generated, and none have zero or infinite estimated standard errors, then an inverse variance weighted mean overall estimate, and standard error is also produced.

b) *'Closed (Schnabel)'* This assumes a closed population, whose size is estimated by maximum likelihood.*

c) *'Jolly-Seber'* This is the Jolly-Seber model in which mortality/emigration rates and birth/immigration numbers vary between sampling periods. Unlike the other models (except Petersen) which use iterations to find the maximum likelihood, the estimates are derived analytically (Seber 1965; Jolly 1965). No estimates of precision are calculated.

d) *'Mortality'* This assumes a population of constant size, where mortality (which may include permanent emigration) is balanced by birth (which may include immigration). The population size and mortality rate (per sampling period) are estimated by maximum likelihood.*

e) *'Mortality + trend'* This assumes a population growing or declining at a constant rate. The population size, mortality rate (per sampling period) and growth or decline of the population (instantaneous proportional rate per sampling period) are estimated by maximum likelihood.*

f) *'Reimmigration'* This is the model of Whitehead (1990) in which members of a closed population move from (emigration rate) and into (reimmigration rate) a study area. The population size in the study area, the total population size, the emigration and

reimmigration rates are estimated by maximum likelihood.*

g) *'Reimmigration + mortality'* This is model f) with the exception that mortality (which may include permanent emigration from the total population) is balanced by birth (which may include immigration).*

h) *'General'* This allows you to describe more generally how population changes with time (indexed by sampling period). See 9.1.1.*

For the models solved iteratively (b, d-h), you can set the initial values of the parameters (on right of the 'population analysis' window). This is useful for speeding up slow analyses, and checking that you have reached a global optimum.

When you have chosen the model, initial parameters and options (see 9.2 below), click on 'Run'.

The likelihood models (marked * in the list above) condition on first identification of each individual. All assume that identification rates may vary between sampling periods.

9.1.1 *General models*

This option allows you to set a custom population model by using the editable box which appears at the bottom of the window. These models include mortality and the function that you type in is multiplied by **N**, the population size at the midpoint of the study to give the population size at other times. Use '**t**' for sampling period, and '**a1**', '**a2**', etc. for the parameters. You can set initial values of these parameters in the editable box marked 'Params', separated by spaces. There should be as many initial values in this box as a1, a2's etc in the editable box below. Initial values not specified are assumed to be zero. For instance:

'**exp(a1*t)**' is the same as the 'Mortality + trend' model (e)

'**1+a1*t**' has a linear trend (a constant number being added or subtracted from the population each sampling period)

'**1+a1*t+a2*t.*t**' gives a quadratic trend

'**1**' is the same as the 'Mortality' model (d)

'**1+a1*t*(t>a2)**' gives a linear trend starting part way through the study in period a2

9.2 Population analysis options

Options are available for some models (choose by checking box):

i. *'95% likelihood c.i.s'* (models a, b, d-h) 95% confidence intervals for each parameter (except total population size in models f and g) are estimated using the likelihood support function (see Edwards 1992). The interval for a parameter *p* is such that the difference between the overall maximum likelihood, and the maximum likelihood given any value of *p* in this range is less than 2. In the Petersen model (a) 95% confidence intervals are calculated assuming the lognormal distribution.

ii. *'Bootstrap 95% c.i. + s.e.'* (models b, d-h) If you click on this checkbox, the program makes non-parametric bootstrap estimates (conditioning on first capture) of the s.e.'s and 95% confidence intervals of the parameters (see Buckland and Garthwaite 1991). An editable box opens up allowing you to enter the number of bootstrap replicates. At least 100 are recommended. This option may take some

time if one of the more complex models is being used.

iii. *'Jackknife s.e.'* (models b, d-h) If you click on this checkbox, the program makes non-parametric jackknife estimates (conditioning on first capture) of the standard errors of the parameters, by omitting each sampling period in turn (Efron and Gong 1983). This option may take some time if there are many sampling periods. This method of estimating s.e.'s of the parameters has not been fully justified. It seems to work reasonably well in practice, although, like most jackknife applications, it also seems to be conservative (overestimating s.e.'s see Efron and Stein 1981). In general bootstrap (ii) or likelihood (iii) methods of obtaining estimates of the precision of parameter estimates are preferred to the jackknife for mark-recapture analyses (Buckland and Garthwaite 1991). However, when the probabilities of capturing individuals are not independent (as when they travel in fairly closed groups), then bootstrap and likelihood methods are invalidated, but this version of the jackknife is probably still roughly OK.

iv. *'Support function plot'* (models b, d-f) This plots values of the likelihood support function: the difference between the overall maximum likelihood and the maximum likelihood given that combination of parameters (see Edwards 1992). Contours where the support is about two indicate the approximate boundaries of a 95% confidence region.

v. *'Residuals plot'* (models b, d, e) This plots, for each catch history (combination of caught/not caught on each sampling period), the difference between the number of individuals with the catch history and the expected number given the fitted model and estimated parameters, against the number of sampling periods in which an individual was captured. The overall sum of residuals for each number of years identified is also shown. This plot is useful for examining assumptions of heterogeneity of capture, failures of which are indicated by a 'U-shaped' pattern of points (i.e. more individuals with very many and very few captures than would be expected; see Cormack (1985)). This takes a very long time to execute if you have more than about 15 sampling intervals, and will not run with more than 20 sampling periods.

vi. *'Tolerance'* (models b, d-h) This sets the tolerance for the minimization function. Smaller tolerance will give you more accurate parameter estimates, but take longer.

vii. *'Max. iterations'* (models b, d-h) This sets the maximum iterations for the minimization function. You can try increasing this if you get the message that 'Maximum number of function evaluations has been exceeded'.

9.3 Output of population analyses

The screen output gives the name of the data file, the sampling period, the number of sampling periods and individuals captured, the estimated parameters (and 95% confidence intervals, if option i and/or ii was chosen; s.e.'s for ii and iii), the log likelihood, and the Akaike Information Criterion. Better models increase the log-likelihood (an increase of about 2 suggests a statistically significant increase for one extra parameter), and decrease the AIC. In some respects ('Kullback-Leibler information' to be technical), the 'best' model is that with lowest AIC (Burnham and Anderson 2002). For the Jolly-Seber model, only parameter estimates are given

(for each sampling period where appropriate). Support function and residuals plots are given as requested when possible.

Plots of estimated population size against time are given for the following models: Petersen (a); Jolly-Seber (c); Mortality + trend (e); General (h). 95% error bars are given for the Petersen model if these have been selected. If several models are selected in turn, the plot from each model is added to those from previous models.

Note: These programs may behave strangely in degenerate situations as when all individuals in the population are identified in most sampling periods.

10 ESTIMATION OF MORTALITY USING SOCIALITY

Estimates of mortality from mark recapture methods (as in 9.3; or from Programme MARK, 11) become biased when there is heterogeneity in identification among individuals, so that some individuals are more likely to be identified than others, and especially when patterns of heterogeneity change with time. If animals have long-term social partners, then identifications of these social partners without the animal in question may indicate its mortality. This module implements a likelihood method described by Whitehead and Gero (2014) for estimating mortality using such social data. The technique can produce less biased and more precise estimates of mortality than standard methods when individuals are almost always identified with some associates, and when there are time-varying patterns of heterogeneity in identifiability. I suggest you read this paper before using the method.

To estimate mortality using sociality, import the data into SOCPROG (3), in *linear mode* only (3.3.1), and then, in the 'set sampling period, association, restrictions' window, set any restrictions on the data that you wish (4.2), choose how you want to define association (4.3), and define the major sampling periods (4.1), used as units in the population analysis—these are often years. They are also the units of the mortality estimate. Then press the 'mortality using sociality' pushbutton on the main SOCPROG window.

The 'estimate mortality using social structure' screen appears. At the top, in italics, are listed the definition of association and the major sampling period that you have just set. Next is an editable string where you set the minor sampling period. The minor sampling period (e.g. days) is used to estimate the amount of identification effort directed at associates of an individual during a major sampling period, so if associates of individual X were identified during 10 minor sampling periods in major sampling period T, but individual X was not, then it is likely that individual X has switched associates or died. You can set the minor sampling period using primary data fields, or combinations of them (as in 4.1 for major sampling periods), listed at the left of the screen.

You can also set initial values for the parameters of the model: the rate (per major sampling period) that individuals switch associates, the mortality rate (per sampling period), the sighting parameter *f* which is the probability of identifying an individual during a minor sampling period during which its associates were identified, and the sighting parameter, and the sighting parameter *g* which relates the probability that an individual, without its associates, is identified in a major sampling period to the total number of individuals identified in that major sampling period. You can also ask for 95% confidence intervals for just the mortality parameter, or for all parameters. These are calculated using the support function of the likelihood. Finally you can change the maximum number of iterations (increase if you get a message 'Maximum number of function evaluations has been exceeded'), and the tolerance (higher for greater speed, lower for

greater precision).

The screen output gives the data file, the major and minor sampling periods, the estimated parameters using maximum likelihood and 95% confidence intervals if selected. The log likelihood of the model with the selected parameters is listed, together with the AIC. However you must not use this AIC to compare the sociality model with the standard population models (9)—they use fundamentally different data.

11 PREPARE DATA FOR 'MARK'

This pushbutton on the main SOCPROG window leads to a window which helps you prepare data for the population analysis program 'MARK' (White and Burnham 1999) (http://www.cnr.colostate.edu/~gwhite/mark/mark.htm). SOCPROG format data are converted into an 'Encounter Histories File' used by MARK (in 'LLLL' format). After clicking on 'Run', listed in the MATLAB command window is the information that MARK needs to be entered: the number of 'Encounter occasions' (equivalent to sampling periods in SOCPROG notation), the time intervals between encounter occasions, and, if there are less than 12 encounter occasions, a printout of the 'Encounter histories file' which has been saved in a '.inp' file for entry into MARK. You can copy the 'time intervals between encounter occasions' data to the clipboard and then paste them into MARK's 'Set Time Intervals' window. You can also define 'attribute groups' of individuals (equivalent to 'classes' in SOCPROG notation) using supplemental fields as described in 6.3.1. Press 's' to enter the first highlighted supplemental field into the editable string which describes classes. If you do enter 'attribute groups', then the number of these groups and their names are listed in the command window–the latter can be copied and pasted into MARK's 'Enter Group Labels' window.

12 MOVEMENT ANALYSES

SOCPROG can carry out two general types of movement analysis: those for studying situations in which we want to know about individuals moving into or out of specific study areas, or between them; and the modelling of movement in continuous space.

12.1 Movements among areas

These analyses look at how individuals move among study areas. Here SOCPROG allows two related types of analysis. In the simplest situation we simply consider movement out of, and perhaps back into, one area. This can be examined using the 'lagged identification rate'. The lagged identification rate is the probability that if an individual is identified in the area at any time, it is identified during any single identification made in the area some time lag later (Whitehead 2001). If the population is closed, and identifications are independent, then the lagged identification rate is the inverse of the population size. If there is emigration or mortality, then lagged identification rates typically fall with time lag. Cyclical movements can generate lagged identification rates that may fall and then rise with time lag.

If data have been collected in more than one area, we can also look at lagged identification rates between them: the probability that if an individual is identified in area A at any time, it is

identified during any single identification made in area B some time lag later (Whitehead 2001). Between-area lagged identification rates typically rise over short time intervals as individuals move between them. It is also possible to consider overall lagged identification rates within and between areas: indicating the general probabilities that individuals are in the same study area or in different ones after particular time lags.

The second type of analysis of movements among areas implemented by SOCPROG is a parameterized Markov model, in which, at each time unit, an individual has a certain probability of moving from the area it is in to another. SOCPROG can estimate these probabilities, as well as the population size, and a mortality rate (which includes permanent emigration from all study areas).

To implement the analyses of movements between areas, you must enter a data set into SOCPROG (only *linear mode*) and set the sampling period using the sampling-restrictions-associations window (4.1), where you can also set restrictions if you like (4.2). With data entered and sampling period set, click on 'Movement between areas' in the master SOCPROG window. The new window has three frames, 'Options' on the right, 'Lagged Identification Rates' on the left, and 'Movement Models' beneath. You set the options first and the either run the lagged identification rates or the Markov movement models.

12.1.1 *Options*

'Area variable' At the top of the 'Options' part of the window is an editable box where you can set the area variable. You need to set this for all further analyses of movements between areas, except lagged identification rates for the whole study area. The area variable should be a primary data field, or an expression based upon one or more of them. The primary data fields are listed in a popup menu. You can choose the one you want and press 's' to put it in the 'Area variable' editable box. Some examples of area variables are:

Area <- A numeric integer variable giving, say, 4 locations: 1, 2, 3, 4

***Area*>1** <- Lumps locations 2, 3 and 4

Bay <- A string variable called *Bay*, e.g.: 'SF', 'SB', 'MO'

'Max time lag' You can set a maximum time lag to be considered in units of sampling periods. This is useful if, for instance, you have data from several field seasons but you only want to consider movements within field seasons. Shortening the maximum time lag speeds up calculations.

'Log x-axis' Logs the x-axis (time lag) of the lagged identification rate plots–useful if a large range of time lags are being considered.

'Bootstrap' Check this, and set the number of bootstrap replications, to get bootstrap-estimated standard errors of the lagged identification rates, and parameter estimates for fitted models of lagged identification rates and movement parameters. These are valid if individuals move independently. If you set a large number of bootstrap replicates, especially if you are fitting models, the analysis will be slow, so it makes sense just to use a few replicates (maybe 10) while you are exploring, and a larger number (maybe 1,000) to get more precise confidence intervals, when you have decided on the chosen analyses.

'Max evaluations' Models are fit by iteration. When fitting models with many parameters, especially Markov movement models, the default number of evaluations may be insufficient to get maximize the summed log likelihood. This will be indicated by a

message in the command window such as

Exiting: Maximum number of function evaluations has been exceeded

If this happens, you can increase the maximum number of evaluations in this editable box.

12.1.2 *Lagged identification rates*

On the 'Lagged Identification Rates' part of the window, the different forms of analysis are implemented by pushbuttons. Each analysis gives a plot, or plots, of lagged identification rate against time lag (in units of the sampling period), and displays the lags, rates and raw data in the MATLAB command window (see Whitehead 2001 for notation). The forms of analysis available are:

'Whole study:' Lagged identification rates for the whole study area, so the analysis is examining emigration or mortality from the whole study area.

'Among all areas:' The study is divided into areas based upon an expression using the primary data fields, entered in the editable text box in the 'Options' area of the window (12.1.1). Lagged identification rates are calculated and plotted for each area alone and between each pair of areas.

'Within/Between:' Calculated and plotted are the overall lagged identification rates for individuals staying in the same area (it can be any area), or moving between any pair of areas, with areas determined using the editable text box in the 'Options' area of the window (12.1.1).

'Select:' Here you can select an area, or pair of areas, for the analysis, by typing their names in the editable strings, and pressing 'Go'. So:

'From area **SF** to area **SF**' calculates and plots the lagged identification rates for area 'SF'

'From area **SF** to area **SB**' calculates and plots the lagged identification rates for movements from area 'SF' to area 'SB'

12.1.2.1 Fitting models to lagged identification rates

NOTE: In contrast to the fitting of models to lagged association rates (7.5), when fitting models to lagged identification rates, you select the model(s) to be fitted FIRST, as described below, and THEN press 'Whole study', 'Among all areas', etc. The lagged identification rates are plotted, and the model(s) are fitted at the same time.

You can fit mathematical models of lagged identification rates to the data by pressing the 'Model fitting for one area' (relevant for lagged identification rates for one area) or 'Model fitting for two areas' (relevant for lagged identification rates between two areas) buttons. After pressing either of these, you see windows with a number of models of how the lagged identification rate changes with time lag, and an informal explanation of each model (as in Table 9). The models are of the exponential form proposed and used by Whitehead (2001). In these models the time lag is represented by 'td' and the parameters of the models by 'a1', 'a2, 'a3', ... (Table 9). Check beside one, or more, of these models to fit them. Alternatively, or additionally, you can specify your own, custom, model at the bottom. In specifying this model, use the standard MATLAB (and 'C') mathematical notation, '**td**' for the lag, and '**a1**', '**a2**', etc for the parameters. If you have three parameters use '**a1**', '**a2**', and '**a3**'. Do not skip any in this sequence (so do not use '**a1- a3*td**'). You can also use functions like '**sin**' and '**cos**' to give cyclical lagged identification rates. e.g. to

represent a seasonally-returning population, you might try:
a1*cos(a2*td)+a3

Table 9. Models that can be fitted to lagged identification rates. Note: the following pairs of models are structurally identical but parameterized differently: A&B, C&D, E&F, G&H, J&K, L&M.

	Equation	Explanation
	With one study area (*N* is population size in study area)*:*	
A	*a1*	Closed (1/*a1=N*)
B	1/*a1*	Closed (*a1=N*)
C	*a2**exp(-*a1***td*)	Emigration/mortality (*a1*=emigration rate; 1/*a2=N*)
D	(1/*a1*)*exp(-*td*/*a2*)	Emigration/mortality (*a1=N*; *a2*=Mean residence time)
E	*a2*+*a3**exp(-*a1***td*)	Emigration + reimmigration (*a1*=emigration rate; *a2*/(*a2*+*a3*)=proportion of population in study area at any time)
F	(1/*a1*)*((1/*a3*)+(1/*a2*)*exp(-(1/*a3*+1/*a2*)**td*))/(1/*a3*+1/*a2*)	Emigration + reimmigration (*a1=N*; *a2*=Mean time in study area; *a3*=Mean time out of study area)
G	*a3**exp(-*a1***td*)+*a4**exp(-*a2***td*)	Emigration + reimmigration + mortality
H	(exp(-*a4***td*)/*a1*)*((1/*a3*)+(1/*a2*)*exp(-(1/*a3*+1/*a2*)**td*))/(1/*a3*+1/*a2*)	Emigration + reimmigration + mortality (*a1=N*; *a2*=Mean time in study area; *a3*=Mean time out of study area; *a4*=Mortality rate)
I	...	Custom
	With two study areas: area 1 to area 2 (*N* is total population size)*:*	
J	*a1*	Fully mixed (1/*a1=N*)
K	1/*a1*	Fully mixed (*a1=N*)
L	*a2**(1-exp(-*a1***td*))	Migration—full interchange (*a1*=diffusion rate from area 1 to area 2; *a2*=1/N)

M	(1/*a1*)*(1-exp(-*td*/*a2*))	Migration—full interchange (*a1=N*; *a2*=Mean residence time in area 1)
N	...	Custom

On the right of the screen you can set the start values of each parameter (**a1**, **a2**, ...) for each model. This is useful for checking whether the fitting, which is iterative, has reached a global optimum. So change the start values of the parameters, and see if you get the same result (or at least a worse one). Also changing the start values of the parameters is often a good way to resolve problems of model fitting (indicated by a MATLAB error message in the command window or a resultant fitted curve that looks little like the data curve). The initial values of all parameters are set to 0.5 unless the parameter is preceded by a forward slash, division sign, (e.g. '1/**a1**') in which case the initial value is 20.

The models of lagged identification rates are fitted using maximum likelihood and binomial loss to the full data set being used, not simply the estimated lagged identification rates (see Whitehead 2001).

Once you have chosen the models to be fitted, press 'Whole study', 'Among all areas', etc. The lagged identification rates are plotted, and the model(s) are fitted. The fitted model is plotted on the graph of lagged identification rates, and in the MATLAB command window are given the model type, the informal explanation, the fitted parameters (with estimated standard errors if the bootstrap has been used) and the summed log-likelihood. Also given are the results of a goodness-of-fit chi-squared test of the number of identifications of the same individual in each time lag bin against the expected number given the model, and lumping time lags so that the expected number is at least six in each bin, as well as the variance inflation factor (chi-squared statistic divided by degrees of freedom).

Because of a lack of independence of data points (identifications), the summed log-likelihoods from different models cannot be used for formal likelihood ratio tests. However, I have done some simulations (Whitehead 2007) which suggest that the Akaike Information Criterion (AIC) or quasi-AIC (QAIC which tries to account for overdispersion of count data) provide some basis for selection among models of lagged association rates:

$$\text{AIC} = -2 \cdot \text{Summed log-likelihood} + 2 \cdot K$$
$$\text{QAIC} = -[2 \cdot \text{Summed log-likelihood} / \hat{c}] + 2 \cdot K$$

where K is the number of parameters being estimated (plus one if QAIC is being used) and $\hat{c}$ is the variance inflation factor for the most general of the models being compared (Burnham and Anderson 2002). QAIC should be used in place of AIC if $\hat{c}>1$. The model with the minimum AIC or QAIC is selected, and the difference between the AIC or QAIC of any other model and the selected one, ΔAIC or ΔQAIC, gives an indication of how well the data support the less favoured model (Burnham and Anderson 2002):

ΔQAIC or ΔQAIC: 0-2	substantial support for model
ΔQAIC or ΔQAIC: 4-7	considerably less support
ΔQAIC or ΔQAIC: >10	essentially no support

If using the QAIC for model selection, you must run all the models being considered simultaneously, so that the variance inflation factor can be taken from the most general model, and applied to all of them. DO NOT COMPARE QAIC's FOR MODELS RUN SEPARATELY.

12.1.3 *Movement models*

This area at the bottom of the 'movement between areas' window allows you to estimate rates of movement between study areas. There is one further option in this part of the window: 'Extra (outside) area'. It may make sense to consider that there is one additional area from which you have not collected any data, and that individuals may move to or from it. If so, check this option.

Clicking on 'Fit Movement Model' brings up a new window of possible movement parameters. There are editable strings, with checkboxes beside them, for $a \cdot (a\text{-}1)+2$ parameters, where a is the number of areas. At the top is N, the total population size (assumed constant through the study), and the mortality rate (the probability of an individual dying or otherwise permanently leaving the population per sampling period). Beneath are transition probabilities between the areas; these are the probabilities that an individual, say, in area A moves to area B between sampling periods. Then the probability that an individual remains in a study area one sampling period later is one minus the sum of the transition probabilities on its corresponding row.

The values set on this window are those used at the start of the optimization process–the process of finding the set of these parameters that maximizes the likelihood of the identification data. The checked parameters are those that are optimized, the unchecked ones are left fixed. Reasons for not checking boxes include:

- You have independent knowledge of the population size, N. If so enter it and clear the neighbouring checkbox–one less parameter to be estimated makes the estimation faster and more precise.
- If the study is of sufficiently short duration that mortality is not an issue, leave the mortality at zero and the neighbouring checkbox cleared (the default).
- The study areas may be arranged, perhaps linearly, so that movement between some pairs is not practical over one sampling period. If so, enter zero in the corresponding editable strings and clear the neighbouring checkboxes. This will speed up the calculation and make it more precise.

When the parameters are initialized, and those to be optimized are chosen, press 'Estimate'. The program finds a set of parameters that maximizes the likelihood of the identification data using the Poisson approximation (see Hilborn 1990). The calculation may take some time, especially if you have chosen to do bootstrap replicates or there are many parameters to be estimated. When finished, the following information will appear in the command window: the summed log-likelihood; the estimated (marked by a '*'), and fixed, parameter values, with their estimated standard errors if bootstrap replicates were chosen; and estimated equilibrium population sizes in each of the study areas (given the total population size and estimated movement parameters), with estimated standard errors if bootstrap replicates were chosen.

Some notes on this procedure:

i) The routine needs a fair amount of data (numbers of individuals observed moving between areas) to produce even moderately precise estimates of parameters.

ii) With many parameters being estimated (>~4) the routine is slow, and may be inaccurate. You may need to increase the maximum number of evaluations, if you receive the message:

Maximum number of function evaluations has been exceeded

Try to reduce numbers of parameters by strategies such as combining areas, or setting a 'Max time lag' such that mortality is not an issue, and so can be set to zero.

iii) Working with simulated data suggests that the estimated equilibrium populations in each study area are particularly inaccurate.

iv) Especially when working with many parameters, the program may not converge on the global optimum. You can check this by changing the starting values of the parameters to be estimated, and seeing whether you get the same answers. If you don't, you may wish to try quite a range of combinations of starting values to explore. The answer with the highest summed log-likelihood is theoretically the best.

12.2 Movement in continuous space

This module analyzes the movements of individually identified individuals in continuous space. In particular, diffusion rates and mean squared displacements with time lag are estimated using the standard methods (sum of squared displacements divided by time lag see Turchin 1998, 254, 258), and a modified maximum likelihood method which accounts for the distribution of effort (Whitehead 2001). The standard method is appropriate when the probability of reidentifying an individual does not depend on its movement, either because the individual was being tracked or if all the areas where it was likely to move were being sampled uniformly or randomly. The modified maximum likelihood method can be used when the probability of reidentification varies with movement (i.e. when the study areas do not include all the range of the individuals) but is invalid when individuals are being tracked.

12.2.1 *Setting up continuous movement analysis*

The data should be entered in *linear mode* (3.1.1, see also 3.3.1), and 1-3 primary data fields should indicate position (e.g. *lat*, *long*; *x*, *y*, *z*), depending on how many dimensions are being considered. So each record gives information on where one identified individual was at one time. You must set a sampling period using the sampling-restrictions-association window (4.1) and can enter restrictions (4.2). Then, from the master SOCPROG window, press the 'Movements in continuous space' button, and you reach the 'Movements through continuous space' window.

Select your options for distance variables and scales in the upper left frame:

'x-variable', 'y-variable', 'z-variable' Make sure these contain the names of the variables in your data file which give geographical position. If the x- and y-variables are entered and the z-variable is blank, a two dimensional movement analysis will be carried out, etc. If you are using latitudes and longitudes, make sure that the longitude variable is the x-variable, and the latitude variable the y-variable.

'Range:' If you are using latitudes and longitudes, make sure the 'Rhumb line (lat-lon)' option is selected (this corrects for the difference in scale between latitudes and longitudes at high latitudes). Otherwise use the 'Euclidean' option which assumes variables are on the same scale.

'Distance multiplier' This allows you to change distance scales. For instance, if your position variables are latitudes and longitudes in degrees, one unit of distance will be 60 nautical miles. To change the scale to nautical miles multiply by 60, to change to kilometres multiply by 111.12.

'Distance units' Name these. e.g. **km**, **nautical mile**. The distance units are the units of your position variables multiplied by the distance multiplier.

Select your options for time variables in the lower left frame:

'Time units' Name these. e.g. **day**, **hour**. The time units are those of your sampling periods, selected earlier on the sampling-restrictions-associations window.

'Minimum, maximum lag' By default, the minimum time lag considered is one sampling period, the maximum is the time between the first and last of your sampling periods. You can change these. Logistical reasons to do so are that the calculation time and storage space used can be reduced by restricting the range of time lags used. Theoretical reasons to restrict the range of time lags, are that data over long or short lags are not reliable (e.g. positions are sufficiently inaccurate that short times should not be considered, or sampling at long ranges was sufficiently sparse that estimates over long time lags are poor, or time lags are potentially greater than the lifetime of the organism).

'Include lag of 1?' This checkbox allows you to omit lags of one sampling unit. If sampling periods are contiguous (i.e. identifications in adjacent periods may be only a small fraction of a time unit apart), then I suggest you omit lags of one time unit; if they are not contiguous (e.g. sampling units are days and identifications were only collected during daylight) then you can include these short duration lags.

In the upper right frame, you can decide on the type of analysis:

'Likelihood estimates?' If this box is checked the likelihood estimates which correct for the distribution of sampling effort (Whitehead 2001) are carried out together with the standard, direct, estimates (Turchin 1998, 254, 258) that do not correct for sampling effort. If you suspect that your sampling was neither random nor uniform with respect to where the individuals might have moved, use the likelihood estimates. In contrast if your sampling was uniform or random or the individuals were tracked, then remove the check on this box (as likelihood estimates use much more computer time and are less precise than the direct estimates). Direct estimates are always calculated (but may not be valid!).

'Method of calculation' You can select one of three methods of calculation:

'Linear method (slow)'. Here each identification (or mean of set of identifications within a sampling period, if more than one) of each individual in a sampling period is compared with each other identification (or mean of set of identifications) in all other sampling periods, as long as the time lag between the sampling periods is within that chosen. This method is theoretically the simplest, but with large data sets, and, especially, a great range of time lags being considered, it may be very slow, and use up an

unacceptable amount of computer memory.

'Logarithmic method (medium)'. This method (the default) is faster and less demanding of computer memory than the linear method. It does this by setting up logarithmically spaced time bins (1,2-3,4-7,8-15, ... sampling units) and, only considering lags of twice the bin size (except lags of one unit, if selected). Thus it makes many fewer comparisons than the linear method and is faster.

'Binning method (fast)'. This is the fastest, and least memory demanding method, and may be the only method that really works for very large data sets. In this method, time is binned as in the logarithmic method, but range is also binned.

Experimentation suggests that the three methods produce similar results. Although the linear method uses more detailed data than the logarithmic, and the logarithmic more than the binning, the detail discarded by the faster methods is largely redundant, and the rounding into bins is of a lesser order than the natural variation within a data set.

In the centre right frame are options which affect the look of the results:

'Number of time intervals:', 'or, interval dividers' In addition to the overall analysis, you can run the analysis separately for different time lag intervals. Specify these: either by entering a number of time intervals (e.g. **3**), in which case the intervals are calculated automatically by putting roughly equal amounts of data into each (e.g. 1-2hr, 3-7hr, 8-25hr); or by specifying dividers for the intervals (e.g. **1.5 5 9** might give intervals of 1-1hr, 2-4hr, 5-8hr and 9-25hr). Sometimes you will get fewer intervals than you specified if some turn out to contain no data.

'Log x-axis?' If you enter more than one time interval, plots are produced of the diffusion rate and mean squared displacement against mean time lag for each interval. Check this box to have logged x-axes (time) on these plots. Logging the x-axis makes sense if the maximum time lag being considered is a few orders of magnitude larger than the minimum lag.

At the bottom right are options which allow you to produce measures of confidence in your output, using the jackknife method:

'Individual jackknife?' If you click on this checkbox, the program makes jackknife estimates of the standard errors of the parameters by omitting individuals in turn (see Efron and Gong 1983). Jackknifing can take a lot of computer time, especially if there are many individuals, so before you do jackknife runs, I suggest you run the analysis without jackknifing, and see how long it takes. The elapsed time with jackknifing will be roughly n+1 times the elapsed time without it, where n is the number of individuals. Individual jackknifing is not valid if individuals move around in fairly permanent groups and may take too much computer time if there are many individuals. In these cases use the 'Time jackknife'.

'Time jackknife?' If you click on this checkbox, the program makes jackknife estimates of the standard errors of the parameters by omitting sequences of sampling periods in turn. This method of estimating the standard errors of the parameters, which will probably mainly be used if individuals move around in closed groups or there are too many individuals for the individual jackknife to be valid, has not been fully justified. It seems to work reasonably well in practice, although, like most

jackknife applications, it also seems to be conservative (overestimating standard errors, see Efron and Stein 1981). The elapsed time of the analysis with jackknifing will be roughly n+1 times the elapsed time without it, where n is the number of jackknife sequences (shown at bottom right of window). You can alter the number of jackknife sequences by changing the 'Jackknife grouping factor' which groups sampling periods for jackknifing. The default is 1 (jackknifing on each sampling period), but if this produces too many jackknife sequences, or if the sampling periods are not independent, increase the factor. For instance, if the sampling period is 1 day, and you put the jackknife grouping factor equal to 30, then you are jackknifing (approximately) on months of data, so each month is omitted in turn. After resetting the jackknife grouping factor and pressing 'Enter', the display of the number of jackknife sequences on the bottom right is updated.

When you have chosen the model, initial parameters and options, click on 'Calculate'. If using the 'Binning method (slow)' of calculation a window appears ('Range categories') on which you can set the minimum range being considered (set at roughly the distance resolution of your positioning data) and the factor for logarithmic range categories. This should be set to a number greater than 1. Smaller numbers give greater resolution (e.g. 1.2), larger ones (e.g. 2) greater speed. So if the minimum range is 0.5, and the factor is 3, the range categories considered are: 0.5-1.5, 1.5-4.5, 4.5-13.5, ... Press 'OK' when you have chosen these.

12.2.2 *Output of continuous movement analysis*

The command window output first gives the number of dimensions and calculation method used (together with whether the jackknife method was used to obtain standard errors), any restrictions entered, and the sampling interval. It then gives the time lags being considered (these may be approximate if binning methods are used), and estimates of the diffusion rate (see Turchin 1998) from the likelihood method if selected (Whitehead 2001) as well as the standard calculations (sum of displacements squared divided by sum of time lags, where sums are over pairs of identifications of the same individual, Turchin 1998, 258). An estimate of density is also output by the likelihood method if selected, although simulations suggest that it is not particularly accurate (Whitehead 2001). If you requested jackknife standard errors, these are given for all estimates, when possible.

If you requested estimates for different intervals of time lag, then the estimated diffusion rate, by both standard and likelihood methods if selected, are also given separately for each interval, together with a plot of diffusion rate against mean lag, and a plot of mean squared displacement against mean lag for each interval ((this is a standard method of displaying movement patterns, see Turchin 1998, 254). Standard estimates as well as those from the likelihood method are plotted, and if s.e.'s were calculated, these are shown by vertical bars ($\pm$ 1 s.e.). These may not be the plots you most wanted so another window comes up. It allows you to make one (or more) custom plots, and includes the possibility of plotting root-mean-squared displacement on the y-axis, a quantity which is easily interpretable (roughly 'average distance moved') but less theoretically justifiable than mean squared displacement. It also allows you to plot several graphs on the same figure—'Subplot' (type **help subplot** in the command window to get an idea of how this works).

Notes:

i) The programs may behave strangely in degenerate situations. e.g. when most individuals are only identified in one sampling period.
ii) Errors in the input file, so that there appear to be occasional long-distance movements over short time periods, can seriously bias the likelihood method.

13 APPENDIX: POSSIBLE M-FILES FOR DEFINING ASSOCIATION

Here are some m-files which can be used to define different kinds of association. They are all part of the SOCPROG2 package. You can edit them as you wish.

13.1 Association as a declining function of time interval between associations

This program considers association in an analogous fashion to Whitehead and Arnbom (1987) as a declining function of time interval between associations.

```
%tomass.m
%calculates association as in Whitehead and Arnbom (1987)
assstr='Association using m-file tomass.m';
qq=zeros(nid,nid);
for i1=1:nid %list of from inds
   i1tim=60*Hour(rl{i1});
   for i2=1:nid %list of to inds
      i2tim=60*Hour(rl{i2});
      if i1==i2
         qq(i1,i2)=length(rl{i1});%diagonal elements are number of sightings
      else
         for rr=1:length(rl{i1})
            tdifferent=abs(i1tim(rr)-i2tim);
            qq(i1,i2)=qq(i1,i2)+5/(5+min(tdifferent));
         end
      end
   end
end
```

13.2 Association from nearest-neighbour data

Here the positions of the individuals are given in two dimensions by xx and yy, as in this Excel file.

Date	xx	yy	ID
12/9/89 9:00	230.6	67.7	1
12/9/89 9:00	285.7	40.6	9
12/9/89 9:00	251.2	78.4	14
12/9/89 9:00	258.2	46.3	15
12/9/89 12:00	240.1	88.9	8
12/9/89 12:00	288.9	33.2	11
12/9/89 12:00	285.2	68.2	13
12/9/89 12:00	268.7	29.2	20
12/9/89 15:00	273.1	44.7	4
12/9/89 15:00	248.9	52.6	7

```
    12/9/89 15:00          275.8     65.0       12
    12/9/89 15:00          272.3     39.7       17
    12/9/89 15:00          274.0     84.2       19
```

This m-file gives a 1 for nearest neighbour pairs, 0 otherwise. This is not necessarily symmetric as A may be B's nearest neighbour but C may be A's.

```
%nnass.m
%calculates association as nearest neighbour or not
assstr='Nearest-neighbour association using m-file nnass.m';
qq=zeros(nid,nid);
if nid==1;
 qq=1;
else
 for i1=1:nid %list of from inds
  lisi2=[1:(i1-1) (i1+1):nid];%list of to inds
   for i2=lisi2;
      dis=[];
    for r1=rl{i1};
     dis=[dis sqrt((xx(r1)-xx(rl{i2})).^2 +(yy(r1)-yy(rl{i2})).^2)];
    end
   qq(i1,i2)=min(dis);
  end
  qq(i1,lisi2)=(qq(i1,lisi2)==min(qq(i1,lisi2)));
  qq(i1,i1)=1;
 end
end
```

Association standardizing on effort on focal animals

14 REFERENCES

Barrat, A., M. Barthélemy, R. Pastor-Satorras and A. Vespignani. 2004. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America 101:3747-3752.

Bayly, K. L., C. S. Evans and A. Taylor. 2006. Measuring social structure: a comparison of eight dominance indices. Behavioural Processes 73:1-12.

Beilharz, R. G., and D. F. Cox. 1967. Social dominance in swine. Animal Behaviour 15:117-122.

Bejder, L., D. Fletcher and S. Bräger. 1998. A method for testing association patterns of social animals. Animal Behaviour 56:719-725.

Bridge, P. D. 1993. Classification. Pages 219-242 *in* J. C. Fry, ed. Biological data analysis. Oxford University Press, Oxford, UK.

Brown, J. L. 1975. The evolution of behavior. Norton, New York.

Buckland, S. T., and P. H. Garthwaite. 1991. Quantifying precision of mark-recapture estimates using the bootstrap and related methods. Biometrics 47:255-268.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer-Verlag, New York.

Christal, J., and H. Whitehead. 2001. Social affiliations within sperm whale (*Physeter macrocephalus*) groups. Ethology 107:323-340.

Cormack, R. M. 1985. Examples of the use of GLIM to analyse capture-recapture studies. Pages 243-273 *in* B. J. T. Morgan, and P. M. North, eds. Statistics in ornithology. Springer-Verlag, Berlin.

Croft, D. P., J. R. Madden, D. W. Franks and R. James. 2011. Hypothesis testing in animal social networks. Trends in Ecology and Evolution 26:502-507.

Croft, D. P., R. James, and J. Krause. 2008. Exploring animal social networks. Princeton University Press, Princeton, NJ.

Crow, E. L. 1990. Ranking paired contestants. Communications in Statistics: Simulation and Computation 19:749-769.

David, H. A. 1987. Ranking from unbalanced paired-comparison data. Biometrika 74:432-436.

de Vries, H. 1995. An improved test of linearity in dominance hierarchies containing unknown or tied relationships. Animal Behaviour 50:1375-1389.

de Vries, H. 1998. Finding a dominance order most consistent with a linear hierarchy: a new procedure and review. Animal Behaviour 55:827-843.

de Vries, H., J. M. G. Stevens and H. Vervaecke. 2006. Measuring and testing the steepness of dominance hierarchies. Animal Behaviour 71:585-592.

Dekker, D., D. Krackhardt and T. A. Snijders. 2007. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psychometrika 72:563-581.

Dietz, E. J. 1983. Permutation tests for association between two distance matrices. Systematic Zoology 32:21-26.

Edwards, A. W. F. 1992. Likelihood. Johns Hopkins University Press, Baltimore.

Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician 37:36-48.

Efron, B., and C. Stein. 1981. The jackknife estimate of variance. Annals of Statistics 9:586-596.

Gammell, M. P., H. de Vries, D. J. Jennings, C. M. Carlin and T. J. Hayden. 2003. David's score: a more appropriate dominance ranking method than Clutton-Brock et al.'s index. Animal Behaviour 66:601-605.

Ginsberg, J. R., and T. P. Young. 1992. Measuring association between individuals or groups in behavioural studies. Animal Behaviour 44:377-379.

Godde, S., L. Humbert, S. D. Côté, D. Réale and H. Whitehead. 2013. Correcting for the impact of gregariousness in social network analyses. Animal Behaviour 85:553-558.

Gowans, S., H. Whitehead, J. K. Arch and S. K. Hooker. 2000. Population size and residency patterns of northern bottlenose whales (*Hyperoodon ampullatus*) using the Gully, Nova Scotia. Journal of Cetacean Research and Management 2:201-210.

Hemelrijk, C. K. 1990a. A matrix partial correlation test used in investigations of reciprocity and other social interaction patterns at group level. Journal of Theoretical Biology 143:405-420.

Hemelrijk, C. K. 1990b. Models of, and tests for, reciprocity, unidirectionality and other social interaction patterns at a group level. Animal Behaviour 39:1013-1029.

Hilborn, R. 1990. Determination of fish movement patterns from tag recoveries using maximum likelihood estimators. Canadian Journal of Fisheries and Aquatic Sciences 47:635-643.

Holme, P., S. M. Park, B. J. Kim and C. R. Edling. 2007. Korean university life in a network perspective: dynamics of a large affiliation network. Physica A 373:821-830.

Jarman, P. J. 1974. The social organization of antelope in relation to their ecology. Behaviour 48:215-267.

Jolly, G. M. 1965. Explicit estimates from capture-recapture data with both death and dilution--stochastic model. Biometrika 52:225-247.

Landau, H. G. 1951. On dominance relations and the structure of animal societies: I Effect of inherent characteristics. Bulletin of Mathematical Biophysics 13:1-19.

Manly, B. F. J. 1994. Multivariate statistical methods. Chapman & Hall, New York.

Manly, B. F. J. 1995. A note on the analysis of species co-occurrences. Ecology 76:1109-1115.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27:209-220.

Miklós, I., and J. Podani. 2004. Randomization of presence-absence matrices: comments and new algorithms. Ecology 85:86-92.

Milligan, G. W., and M. C. Cooper. 1987. Methodology review: clustering methods. Applied Psychological Measurement 11:329-354.

Morgan, B. J. T., M. J. A. Simpson, J. P. Hanby and J. Hall-Craggs. 1976. Visualizing interaction and sequential data in animal behaviour: theory and application of cluster-analysis methods. Behaviour 56:1-43.

Newman, M. E. J. 2004. Analysis of weighted networks. Physical Review E 70:056131.

Newman, M. E. J. 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America 103:8577-8582.

Schnell, G. D., D. J. Watt and M. E. Douglas. 1985. Statistical comparison of proximity matrices: applications in animal behaviour. Animal Behaviour 33:239-253.

Seber, G. A. F. 1965. A note on the multiple recapture census. Biometrika 52:249-259.

Seber, G. A. F. 1982. The estimation of animal abundance and related parameters. Griffin, London.

Seber, G. A. F. 1992. A review of estimating animal abundance II. International Statistical Review 60:129-166.

Slater, P. 1961. Inconsistencies in a schedule of paired comparisons. Biometrika 48:303-312.

Smouse, P. E., J. C. Long and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Systematic Zoology 35:627-632.

Turchin, P. 1998. Quantitative analysis of movement. Sinauer Associates, Sunderland, Massachusetts, U.S.A.

van Hooff, J. A. R. A. M., and J. A. B. Wensing. 1987. Dominance and its behavioral measures in a captive wolf pack. Pages 219-252 *in* H. Frank, ed. Man and wolf. Junk, Dordrecht.

Wey, T., D. T. Blumstein, W. Shen and F. Jordán. 2008. Social network analysis of animal behaviour: a promising tool for the study of sociality. Animal Behaviour 75:333-344.

White, G. C., and K. P. Burnham. 1999. Program MARK: survival estimation from populations of marked animals. Bird Study (Supplement) 46:120-138.

Whitehead, H. 2008a. Precision and power in the analysis of social structure using associations. Animal Behaviour 75:1093-1099.

Whitehead, H., and S. Gero. 2014. Using social structure to improve mortality estimates; an example with sperm whales. Methods in Ecology and Evolution 5:36.

Whitehead, H., and R. James. 2015. Generalized affiliation indices extract affiliations from social network data. Methods in Ecology and Evolution .

Whitehead, H. 1990. Mark-recapture estimates with emigration and re-immigration. Biometrics 46:473-479.

Whitehead, H. 1995. Investigating structure and temporal scale in social organizations using identified individuals. Behavioral Ecology 6:199-208.

Whitehead, H. 2001. Analysis of animal movement using opportunistic individual-identifications: application to sperm whales. Ecology 82:1417-1432.

Whitehead, H. 2007. Selection of models of lagged identification rates and lagged association rates using AIC and QAIC. Communications in Statistics--Simulation and Computation 36:1233-1246.

Whitehead, H. 2008b. Analyzing animal societies: quantitative methods for vertebrate social analysis. Chicago University Press, Chicago, IL.

Whitehead, H., and T. Arnbom. 1987. Social organization of sperm whales off the Galápagos Islands, February-April 1985. Canadian Journal of Zoology 65:913-919.

Whitehead, H., L. Bejder and A. C. Ottensmeyer. 2005. Testing association patterns: issues arising and extensions. Animal Behaviour 69:e1-e6.

Wittemyer, G., I. Douglas-Hamilton and W. M. Getz. 2005. The socio-ecology of elephants: analysis of the processes creating multi-tiered social structures. Animal Behaviour 69:1357-1371.