SECATEURS

to 'prune' your pedigrees

Karin Meyer

Animal Genetics and Breeding Unit[1],
University of New England,
Armidale, NSW 2351

1 Purpose

SECATEURS is a simple program to 'prune' a list of pedigrees, elimating any individuals which are uninformative. It is meant as preliminary step for analyses estimating variance components. In addition, SECATEURS carries out some consistency checks on the pedigree, provides the options to compute inbreeding coefficients and set up the inverse of the numerator relationship (NRM) matrix for pruned pedigree, and provides summary information on the pedigree structure.

1.1 What is 'pruning' ?

When fitting additive genetic effecs, we include animals which do not have records themselves but are parents of animals represented in the data. If further pedigree information is available, we generally also include effects for parents of parents, parents of grand-parents, etc. in the model of analysis.

Any individual without records connected to only one other animal in the pedigree does not add any information, for example, a parent with a single offspring only and unknown parents him/herself. For the purpose of analysis, this parent can be treated as unknown, and thus eliminated from the analysis. This is equivalent to 'absorbing' the equation for this animal, and referred to as 'pruning', as the procedure is similar to the removal of 'dead wood' in plants. For pedigrees spanning several generations backwards, this should be done repeatedly, as pruning of a parent may make an animal a candidate for pruning because its known parent has become 'unknown'

[1] AGBU is a joint venture of NSW Agriculture and UNE

in the process. For direct genetic effects, pruning is done downwards, i.e. scanning the pedigree from oldest to youngest animals.

Frequently, we fit parental genetic effects, most commonly maternal effects. If these are assumed uncorrelated to the direct genetic effects, the parental pedigree can be pruned separately. For maternal (or paternal) effects, only animals which have offspring in the data have a 'record'. Hence animals in the data which do not have progeny are candidates for pruning, i.e. pruning needs to be carried out upwards as well as downwards the pedigree. This can lead to substantial numbers of parental animal effects which can be disregarded. If the software used for the mixed model analysis allows separate pedigrees or inverses of the numerator relationship matrix to be used for different genetic efffects, pedigrees for maternal of paternal effects should thus be pruned separately.

1.2 Why use SECATEURS ?

- 'Pruning' reduces the number of levels to be fitted for genetic effects, thus reducing the computational requirements of subsequent analyses.
- Looking at the structure of 'pruned' pedigrees gives a better indication of the amount of relationship information available.
- SECATEURS incorporates a fast procedure to calculate inbreeding coefficients (Tier [1]; code written by Bruce Tier). Hence, setting up the inverse of the numerator relationship matrix for large pedigrees may be considerably faster than doing so using proprietary variance component estimation software.
- SECATEURS provides some, albeit limited, consistency checks of the pedigree information, and gives some summary information on the structure after 'pruning'. Hence, running SECATEURS may be a quick and easy way to have a more detailed look at the pedigree file.

Caveat : SECATEURS is *not* suitable for pre-processing of pedigrees for analyses fitting a non-zero genetic covariance between direct and parental genetic effects !

2 Input files

SECATEURS requires both the pedigree and the data file for the analysis planned.

2.1 Pedigree file

The form of the pedigree file required is the same as that for most variance component estimation programs.

1. The file should be a standard ASCII text file with one record per animal and 3 variables, separated by spaces, per record :
   (a) The animal ID (identity)
   (b) The sire ID, using a code of 0 for an unknown sire
   (c) The dam ID, using a code of 0 for an unknown dam
2. There must be a pedigree record for each animal with data, even if its parents are unknown.
3. All IDs must be numeric and, in a FORTRAN sense, valid positive INTEGERs. The maximum ID allowed is $2,147,483,646$.
4. The animal ID must be numerically bigger than the sire and dam IDs.
5. SECATEURS is written with animal pedigrees in mind. Hence, selfing is not allowed, i.e. known sire and dam IDs must be different.

It is desirable, though not strictly necessary, that the pedigree file is sorted in ascending order of animal IDs.

2.2 Data file

Similarly, the data file is expected to be a text file with space-separated variables. Each record in the data file must have the code(s) for the genetic effect(s) to be modelled represented, and these must be identical to the corresponding code(s) in the pedigree file. For analyses including maternal effects, this implies that any 'dummy' dam IDs inserted for unknown dams must have been inserted in the pedigree file as well as the data file.

3 Running SECATEURS

When running SECATEURS, information on the input files and their layout, and the tasks to be carried out needs to be supplied. This is expected in the form of *command line arguments* or options. The syntax used is UNIX (LINUX)-like, and the program fussily requires things to be specified *just right* :

- Each argument must begin with a "–" sign, immediately followed by a letter and, if applicable, the value of the option chosen.
- Blanks between the – sign, the letter and the option are not allowed.
- Multiple options after a – sign are not recognised.
- Options have to be separated by spaces.

Alternatively, SECATEURS can read the options required from a parameter file. This should have a separate line for each option. The name of this file must have extension "`.par`". In specifying the file name, the extension

can be omitted. The name of the parameter file must be the *last* command line argument. If not given, SECATEURS will look for a parameter file with standard name `prune.par`.

Valid command line arguments are :

–P : followed by the name of the file containing pedigree information. Default : `pedigrees.dat`

–D : followed by the name of the data file. Default `records.dat`

–I : followed by an integer number, which gives the position of the animal ID (column no.) in the *data file*. Default : 1.

–M : or –MAT followed by an integer number, which gives the position of the dam ID (column no.) in the *data file*. Default : no second genetic effect.

–N : or –NRM to specify that output of the NRM inverse is required. Default : no NRM.

–B : or –BIN to select output of the NRM to a binary file (ignored if –N is not specified). Default : formatted output.

–L : followed by an integer number request additional counts of animal and relative numbers with at least that number of records (useful for repeated records analyses). Multiple –L options can be given (up to 10). Default : 0.

–R : followed by an integer number giving the record length in the data file as number of characters. This requests recoding of the code(s) for genetic effect(s) in running order from 1 to number of levels.

–V : to select verbose run time mode. Default : non-verbose

Examples of use

`secateurs` : specifies a run with all options to be read from `prune.par`,

`secateurs PP` : is similar, but options are to be read from file `PP.par`,

`secateurs -Dwgt.dat -Pped.dat -I1 -MAT5 -N` : specifies a run with data file `wgt.dat` and pedigree file `ped.dat`, with the animal code represented by the first variable in the data file. It is chosen to prune pedigrees for maternal effects in addition to direct genetic effects, with the dam code the $5-$th variable in the data file. Further, files with the NRM inverse are to be written out.

4 Output files

4.1 Program log file

Details about the run carried out and summary information are written to file `PRUNE.log`.

4.2 New pedigree files

SECATEURS writes out file(s) with the pedigrees after pruning. These have the same form as the input pedigree file, and can be used to replace the latter. If it has been chosen to create the NRM inverse, the new file contains a fourth variable, giving the animals' inbreeding coefficient ($\times 100$). Output files have standard names

- `pednew.dat` for animal effects, and
- `pednewmat.dat` for maternal effects.

4.3 NRM inverse files

Non-zero elements of the NRM inverse created are written out to file(s)

- `nrminv.dat` for animal effects, and
- `nrminvmat.dat` for maternal effects.

with one record for each element, consisting of row number, column number and element. For formatted output, the three variables are separated by spaces. If binary output has been selected, the extension '.dat' is replaced by '.bin'.

Only the elements of the lower triangle (including the diagonal) of the symmetric NRM inverse are written out. This is done in ascending row number, with column numbers for each row in ascending order between 1 and row number [e.g. elements (1,1), (2,2), (3,3), (4,1), (4,4), (5,5), (6,1), (6,4), (6,6), ...].

4.4 Recoded data file

To use the NRM inverse(s) created by SECATEURS in conjunction with a variance component estimation program, it may be necessary to have the code(s) of the genetic effects renumbered in running order corresponding to the row and column numbers of the NRM inverse, i.e. from 1 to number of levels. If requested through the "–R" option, SECATEURS will write out a file `recordsnew.dat` which is the original data file augmented by the respective codes, as the last variable(s) on each record. If two genetic effects are recoded, the first new variable is the recoded animal ID, the second the recoded dam ID.

References

[1] Tier B., 1990. Computing inbreeding coefficients quickly. Genet. Select. Evol. 22:419–425.

Appendix : How to use SECATEURS together with ASReml

1. Read section 9.6 "Reading a user defined inverse relationship matrix" of the ASReml manual !
2. Run SECATEURS using the "–N" and "–R" options to obtain NRM inverse files and animal IDs recoded in running order.
3. Rename `nrminv.dat` to `nrminv.giv`. If maternal effects are fitted, also rename `nrminvmat.dat` to `nrminvmat.giv`.
4. Set up your `.as` parameter file following the rules given in section 9.6 of the ASReml manual, using `recordsnew.dat` as the data file and the recoded rather than the original IDs as codes. Remember not to give the "!I" qualifier for the recoded IDs.
5. Run ASReml as usual (and marvel at the speed-up).

Example for an ASReml parameter file using pruned pedigrees and NRM inverses for direct and maternal effects created by SECATEURS.

```
Example parameter file
  animal                  Original animal ID
  gendam                  Original dam ID
  cgrp !I
  pedam !I
  wgt
  newan 1576              Recoded animal ID with no. of animals after pruning
  newgdam 656             Recoded dam ID with no. of dams after pruning
nrminv.giv                NRM inverse for direct effects from SECATEURS
nrminvmat.giv             NRM inverse for maternal effects from SECATEURS
recordsnew.dat            Data file with recoded animal and dam ID
wgt ~ mu cgrp !r giv(newan,1) 0.2 giv(newgdam,2) 0.1 pedam
```
