A SAS Macro “HAPEM” – a Quick Reference Sheet

- by JINGXIA LIU -

I. Macro Interpretation

This is a SAS Macro for computing haplotype frequencies at multiple linked biallelic marker (e.g., SNP) loci from their observed unphased genotype data. The EM algorithm proposed by Excoffier and Slatkin (1995) is implemented based on the assumption of Hardy-Weinberg Equilibria at these marker loci.

(1) Initialization: the haplotype frequencies p1,...,pn are initialized from direct counting of all possible haplotype pairs per subject.

(2) Iteration: In each step, the current estimates are used as if they were the unknown true frequencies. In the E step, the conditional probability of all possible haplotype pairs that are compatible with the observed marker genotypes are computed per subject. Estimates of haplotype frequencies are then derived from direct counting and used in the next iteration. Maximum likelihood estimates of the haplotype frequencies are obtained when the convergence is achieved.

(3) Convergence criterion: Euclidean distance between the current and previous estimated haplotype frequencies is less than the pre-determined value (default=10^(-7)).

II. Macro Loading

HAPEM MACRO is loaded in the program via the "%INCLUDE 'filename'" statement where 'filename' is the name of the HAPEM MACRO file you saved.

III. Macro Invocation

HAPEM MACRO is invoked in the form of

%HAPEM(indata, nloci=, nallele= <,option 1 <...,option n>>).

(1) indata: The input data set name

(i) The input data set consists of genotypes at a number of markers. With each row being genotypes of a subject, the column variables are in the following order:

Genotype1,....,Genotype&nloci

(ii) The input data set may have other variables following marker genotypes. However this macro will ignore them.

(iii) The genotype value needs to be specified as follows: 0-aa(/AA); 1-Aa; 2-AA(/aa); .-missing data.

(2) nloci: The number of marker loci you want to analyze. It should

be less than 20.

(3) nallele: The number of alleles at each locus. It may be a vector or a number if all loci have the same number of alleles. This macro currently works only for nallele=2 (i.e., biallelic markers)

These three arguments must be given.

(4) The symbol of angle brackets identifies optional arguments. There are at most 7 options available and they can be given in any order after the first three. Each option is specified as a keyword followed by an equal sign and a value (not case-sensitive).

The options are:

(i) dreport=Y: Prints the data set after cleaning.

(ii) converr=value: Set convergent criterion in EM algorithm. Its default value is 10##(-7).

(iii) convmstp=integer value: Set the maximum step in EM algorithm. Its default value is 500.

(iv) haptop=integer value: Set the length of the top haplotype frequencies. Those top haplotype frequencies will be used later and others will be collapsed to one. Its default value is 20.

(v) itsumm=Y: Summary of the EM iteration history option

(vi) outdata=value: Output data set option. If the value is specified as a legal SAS data set name, then that data set will contain variables "decimal" (number to record the 'position' of haplotype), "Probmarkers" (the haplotype frequency of markers) and "hap1",. . ., "hap&nloci"(the corresponding haplotype).

(vii) Error=value: Default as 0. If error happens in the input dataset, its value will return as 1.

(5) The output in the SAS output file ".lst" includes variables "hap1",. . ., "hap&nloci"(the corresponding haplotype) and "Probmarkers" (the haplotype frequency of markers).

IV. Reference:

L. Excoffier & M. Slatkin (1995) Maximum Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. Mol Biol Evol 12: 921-927.

V. Example

1) True value setting:

```
Hap=[0 0 0 0 0 0;
     0 0 1 0 1 0;
     0 0 0 1 0 0;
     0 1 1 1 0 1;
     1 0 0 0 1 0;
     1 1 1 0 0 1;
     1 0 0 1 0 0;
     1 1 1 1 1 0];

hapfreq=[0.25 0.20 0.20 0.10 0.10 0.05 0.05 0.05];
```

2) Macro Result:

```
data test2;
input GENO1-GENO6;
cards;
     1     1     2     1     2     0
     1     1     2     1     2     0
     1     0     1     0     2     0
     1     0     1     0     2     0
     0     0     1     1     1     0
     0     0     1     1     1     0
     0     0     0     0     0     0
     0     0     0     0     0     0
     0     0     1     1     1     0
     1     1     1     1     0     1
     1     0     0     1     0     0
     0     0     2     0     2     0
     1     1     1     0     0     1
     2     2     2     1     1     1
     0     0     1     0     1     0
     0     0     0     2     0     0
     1     2     2     2     1     1
     0     1     1     2     0     1
     2     1     1     0     1     1
     0     0     0     1     0     0
     0     0     1     0     1     0
     2     1     1     1     2     0
     0     0     1     0     1     0
     1     1     1     0     0     1
     1     1     1     1     1     1
     0     1     1     1     0     1
     0     1     2     1     1     1
     0     0     0     1     0     0
     0     1     1     1     0     1
     0     0     0     0     0     0
     0     0     2     0     2     0
     1     1     1     0     0     1
     0     0     1     1     1     0
     1     0     1     0     2     0
     0     0     0     1     0     0
     0     0     1     1     1     0
     0     0     0     1     0     0
     0     0     1     0     1     0
     2     1     1     1     0     1
```

0	0	0	1	0	0
0	0	1	0	1	0
0	1	1	1	0	1
0	1	1	2	0	1
0	1	1	1	0	1
1	1	1	1	1	0
0	1	1	2	0	1
1	0	1	0	2	0
0	1	1	1	0	1
0	0	0	2	0	0
0	1	2	1	1	1
0	0	2	0	2	0
0	0	0	1	0	0
1	0	1	0	2	0
1	0	0	0	1	0
2	0	0	1	1	0
1	0	0	1	0	0
1	1	1	1	1	1
0	0	1	0	1	0
0	0	0	2	0	0
1	1	1	1	1	1
1	1	1	1	1	0
1	0	0	2	0	0
0	0	1	1	1	0
0	0	0	2	0	0
0	0	0	0	0	0
0	1	1	2	0	1
2	1	1	1	0	1
0	0	2	0	2	0
1	2	2	2	1	1
0	0	1	0	1	0
0	1	1	2	0	1
0	2	2	2	0	2
0	0	1	0	1	0
0	0	1	0	1	0
0	1	2	1	1	1
1	0	0	1	1	0
0	0	2	0	2	0
0	0	2	0	2	0
1	0	0	1	1	0
1	0	0	1	1	0
0	1	1	1	0	1
0	0	0	2	0	0
1	1	1	1	1	1
0	0	0	2	0	0
0	0	0	0	0	0
0	1	1	2	0	1
1	2	2	2	1	1
1	0	1	0	2	0
1	2	2	1	0	2
0	1	1	2	0	1
0	0	0	0	0	0
0	1	1	1	0	1
0	0	1	0	1	0
1	0	0	1	0	0
0	1	1	2	0	1
2	1	1	0	1	1

0	1	1	2	0	1
1	1	2	1	2	0
1	1	1	2	1	0
1	0	0	1	0	0
0	0	2	0	2	0
0	0	0	1	0	0
0	0	1	0	1	0
1	1	1	2	1	0
0	1	1	1	0	1
0	0	1	1	1	0
0	1	1	2	0	1
1	1	2	1	2	0
1	0	1	0	2	0
1	1	1	1	1	1
1	0	1	0	2	0
1	0	0	1	1	0
2	0	0	0	2	0
0	0	0	0	0	0
1	0	1	1	1	0
1	0	0	2	0	0
1	0	0	1	1	0
0	0	1	1	1	0
1	1	1	1	0	1
0	0	0	2	0	0
0	1	2	1	1	1
0	0	1	1	1	0
1	1	1	2	1	0
0	0	0	1	0	0
0	0	0	1	0	0
1	0	1	0	2	0
2	1	1	0	1	1
1	1	2	0	1	1
0	1	1	1	0	1
0	0	1	0	1	0
0	0	1	1	1	0
0	0	1	1	1	0
0	0	2	0	2	0
0	0	0	1	0	0
0	0	0	1	0	0
1	1	1	0	0	1
2	2	2	0	0	2
1	0	1	0	2	0
1	0	0	0	1	0
1	1	1	1	1	1
1	1	1	1	1	1
0	0	0	0	0	0
0	0	1	0	1	0
0	0	0	1	0	0
0	0	0	2	0	0
0	0	0	1	0	0
0	0	1	0	1	0
0	0	1	1	1	0
0	1	2	1	1	1
1	0	0	2	0	0
1	0	0	1	0	0
0	0	1	1	1	0
0	0	1	1	1	0

```
     0     0     0     1     0     0
     1     0     0     1     1     0
     0     0     0     0     0     0
     2     1     1     0     1     1
     1     1     2     1     2     0
     0     0     0     1     0     0
     0     0     0     1     0     0
     1     0     0     0     1     0
     0     0     0     2     0     0
     0     2     2     2     0     2
     0     0     0     1     0     0
     0     0     1     0     1     0
     0     0     1     1     1     0
     0     1     1     1     0     1
     1     0     0     1     1     0
     1     0     1     0     2     0
     0     0     1     1     1     0
     1     1     2     1     2     0
     0     0     0     0     0     0
     2     1     1     0     1     1
     1     0     0     0     1     0
     1     1     2     0     1     1
     0     1     1     2     0     1
     1     1     2     1     2     0
     1     0     0     2     0     0
     0     0     0     0     0     0
     0     0     1     1     1     0
     0     0     0     0     0     0
     0     0     0     2     0     0
     0     1     2     1     1     1
     0     0     1     1     1     0
     0     0     1     1     1     0
     0     0     0     0     0     0
     1     1     2     0     1     1
     0     0     0     2     0     0
     1     1     1     1     1     0
     0     0     0     2     0     0
     0     0     1     0     1     0
     0     1     1     2     0     1
     0     0     1     1     1     0
     1     0     0     2     0     0
     0     1     1     2     0     1
     1     0     1     1     1     0
     0     0     2     0     2     0
     0     0     1     1     1     0
     0     1     2     1     1     1
     0     0     1     0     1     0
;
run;

%HAPEM(test2, nloci=6, nallele=2, haptop=8, outdata=out, itsumm=Y,
converr=10##(-3))

proc print data=out;
title "Print Outdata Set";
```

~~~~~~

Checking Missing or Non-integer Genotype Values

~~~~~~

No non-integer on genotype.

==> All 200 observations will be used in the analysis.

The Convergent Criterion Used Here is 0.0010000

The Maximum Step Used Here is 500

True Iteration Step is 7

The length of the top haplotype frequency used here is 8

100 means the haplotypes not shown above have been collapsed

Marker Haplotype & Haplotype Frequency

0	0	0	0	0	0	0.2149424
0	0	0	1	0	0	0.2300870
0	0	1	0	1	0	0.2149993
0	1	1	1	0	1	0.1124639
1	0	0	0	1	0	0.0949594
1	0	0	1	0	0	0.0374369
1	1	1	0	0	1	0.0499910
1	1	1	1	1	0	0.0449995
100	100	100	100	100	100	0.0001206

Print Outdata Set
3

Obs	decimal	probmarkers	hap1	hap2	hap3	hap4	hap5	hap6
1	0	0.21494	0	0	0	0	0	0
2	4	0.23009	0	0	0	1	0	0
3	10	0.21500	0	0	1	0	1	0
4	29	0.11246	0	1	1	1	0	1
5	34	0.09496	1	0	0	0	1	0
6	36	0.03744	1	0	0	1	0	0
7	57	0.04999	1	1	1	0	0	1
8	62	0.04500	1	1	1	1	1	0
9	100	0.00012	100	100	100	100	100	100