Vignette for R Package Sped

Charles J. Geyer Elizabeth A. Thompson

March 25, 2020

1 R

```
> library(sped)
```

- The version of R used to make this document is 3.6.3.
- The version of the `sped` package used to make this document is 0.2.1.

2 History

R package `sped` is a re-implementation of some of the functionality of an S package of the same name for old S (Becker and Chambers, 1984, 1985). That old S package, described in Geyer (1988), was obsolete almost as soon as it was created because S version 2 (Becker, et al., 1988) came along and broke all extension packages like `sped`.

Although we have the code for the old `sped` package, there is no point to trying to port it to S or R because the extension interface to old S was so crazy, that it is much easier to rewrite the package than to port it.

We took Thompson (1983, 1986) and Geyer (1988) as definitions of what we had to re-implement. For details see the design document, which can be found in the `DesignDoc` directory of any installation of the `sped` package.

3 Pedigrees

We work relative to a pedigree in which every individual has either two parents or none specified. Those with parents unspecified are called *founders*. A pedigree may be specified by a *triplets matrix* having three columns and each row gives the names of a non-founder individual, its father, and its mother, in that order. We check that no individual is its own ancestor. Optionally, we check that sexes are consistent (no individual is both father and mother). This check is optional so that we can deal with hermaphroditic organisms.

Any ancestors of individuals in the pedigree that are outside the pedigree, parents of founders, grandparents of founders, great-grandparents of founders, and so forth, are assumed to not be individuals in the pedigree. That is, we are assuming that known individuals and unknown individuals are disjoint sets.

4 Multigene Descent Probabilities

At each autosomal locus of the genome there are two genes, one inherited from the father and one inherited from the mother. Thompson (1983) defines *multigene descent probabilities* $g_S(B_1, \ldots, B_n)$ to be the probability that one gene randomly chosen from the two genes at a particular autosomal locus for each of the individuals $B_1, \ldots, B_n$ are all descended from genes (not necessarily the same gene) in some set S of genes in individuals in the pedigree. The individuals $B_1, \ldots, B_n$ need not be distinct. The set S can be specified by giving for each individual in the pedigree an integer 0, 1, or 2 that says how many of its genes (at the autosomal locus in question) are in the set S.

Multigene descent probabilities have many interesting applications (Thompson, 1983, 1986). Here we just look at some examples taken from Geyer (1988).

```
> data(alberta)
> head(alberta)

     ind pa  ma
[1,]  58 11  12
[2,] 100 39  40
[3,] 101 39  40
[4,] 103 39  40
[5,] 107 39 100
[6,] 113 17  18

> descent(1260, alberta, c("52"=1))

[1] 0.03125
```

The `alberta` pedigree, included in the package, is of individuals of the species *Equus przewalskii*, common names Asian wild horse, Mongolian wild horse, and Przewalski's horse, who are individuals living in Alberta, Canada in 1988, and their known ancestors. The "names" of individuals in the pedigree are just numbers, their numbers in the *E. przewalskii* studbook.

Here $B_1, \ldots, B_n$ are just the one individual 1260. The set S has just one gene in individual 52 (a founder). We have to quote 52 in the call of R function `descent` because we are making it the name of an element of the vector `c("52"=1)` that specifies the set S. We only need to specify individuals that have genes in S. Other individuals are assumed to have zero genes in S.

Here is a more complicated example.

```
> descent(c(1085, 1094, 1180, 1260), alberta, c("52"=1))

[1] 8.773804e-05
```

This is the probability that one gene (as a specified autosomal locus) drawn at random from each of the four individuals 1085, 1094, 1180, and 1260 are all descended from one gene in individual 52.

The old S package had a "feature" that when the first argument of this function was a "column vector" then it calculated something else. This was probably bad user interface design (tricky nonsense). We can use R function `Vectorize` to easily obtain this functionality when wanted.

```
> vescent <- Vectorize(descent, vectorize.args = "individuals")
> b <- c(1085, 1094, 1180, 1260)
> names(b) <- b
> vescent(b, alberta, c("52"=1))
```

```
     1085      1094      1180      1260
0.0156250 0.0156250 0.0078125 0.0312500
```

We gave the first argument to `vescent` names (with `names(b) <- b`) so that the output would also have names.

5 Alphas, Betas, and Gammas

The Greek letters in the section title refer to particular multigene descent probabilities that are useful in particular applications Thompson (1986).

The fraction of genes (at the autosomal locus in question) in individual B that comes from founder A is

$$\gamma(A, B) = g_{S_A}(B)$$

where S_A is the set of genes that contains the two genes of A (at the autosomal locus in question) and no other genes.

Here is an example.

```
> data(thompson)
> gammas(c("U", "V", "Q", "R", "W"), thompson)
```

```
      U      V    Q     R       W
A 0.000 0.0000 0.25 0.125 0.06250
B 0.375 0.3125 0.25 0.250 0.28125
C 0.375 0.3125 0.00 0.125 0.21875
F 0.250 0.1250 0.25 0.500 0.31250
K 0.000 0.2500 0.00 0.000 0.12500
L 0.000 0.0000 0.25 0.000 0.00000
```

This function gives the gamma for each individual in its first argument and each founder in the pedigree. The founders are the row names of the result (which is a matrix).

If individual B has father F and mother M (in the given pedigree), then

$$\beta(A, B) = g_{S_A}(F, M)$$

is the bilineal contribution of founder A to individual B, the probability that both genes of B are descended from genes of founder A.

Here is an example of that.

```
> foo <- betas(c("U", "V", "Q", "R", "W"), thompson)
> foo
```

```
      U       V      Q      R         W
A 0.000 0.00000 0.0625 0.0000 0.0000000
B 0.125 0.09375 0.0625 0.0625 0.0781250
C 0.125 0.09375 0.0000 0.0000 0.0390625
F 0.000 0.00000 0.0000 0.2500 0.0625000
K 0.000 0.00000 0.0000 0.0000 0.0000000
L 0.000 0.00000 0.0000 0.0000 0.0000000
```

The output matrix has the same form as for gammas

```
> foo["B", "Q"]
```

```
[1] 0.0625
```

is the bilineal contribution of founder B to individual Q.

Now let T_A be the set of genes that contains one gene of founder A and no other genes, and let F and M be as above, then

$$\alpha(A, B) = 2g_{T_A}(F, M)$$

is the inbreeding of individual B relative to founder A, the probability that both genes of individual B come from the same gene in founder A.

Here is an example of that.

```
> foo <- alphas(c("U", "V", "Q", "R", "W"), thompson)
> foo
```

```
       U        V       Q       R          W
A 0.0000 0.000000 0.03125 0.00000 0.00000000
B 0.0625 0.046875 0.03125 0.03125 0.04296875
C 0.0625 0.046875 0.00000 0.00000 0.02343750
F 0.0000 0.000000 0.00000 0.12500 0.03125000
K 0.0000 0.000000 0.00000 0.00000 0.00000000
L 0.0000 0.000000 0.00000 0.00000 0.00000000
```

6 Inbreeding Coefficients

When the alphas are summed over all founders

$$\alpha(B) = \sum_{A \in \text{Founders}} \alpha(A, B)$$

this gives the *inbreeding coefficients* of the individuals.

Here is an example of that.

```
> colSums(foo)
```

```
         U          V          Q          R          W
0.12500000 0.09375000 0.06250000 0.15625000 0.09765625
```

Because this is a widely used concept, we give it its own function

```
> inbreeding(c("U", "V", "Q", "R", "W"), thompson)
```

```
         U          V          Q          R          W
0.12500000 0.09375000 0.06250000 0.15625000 0.09765625
```

(this does exactly the same thing as the preceding example, it just does it in one step).

7 Kinship Coefficients

The *kinship coefficient* of individuals B_i and B_j is

$$\phi(B_i, B_j) = 2 \sum_{A \in \text{Founders}} g_{T_A}(B_i, B_j)$$

Here is an example of that.

```
> foo <- kinship(c("U", "V", "Q", "R", "W"), thompson)
> foo
```

```
          Q         W          R          V         U
Q 0.5312500 0.1210938 0.18750000 0.05468750 0.0781250
W 0.1210938 0.5488281 0.33789062 0.32226562 0.2382812
R 0.1875000 0.3378906 0.57812500 0.09765625 0.1484375
V 0.0546875 0.3222656 0.09765625 0.54687500 0.3281250
U 0.0781250 0.2382812 0.14843750 0.32812500 0.5625000
```

Here

```
> foo["Q", "R"]
```

```
[1] 0.1875
```

gives the kinship coefficient of individuals Q and R. And

```
> foo["Q", "Q"]
```

```
[1] 0.53125
```

gives the kinship coefficient of individual Q with itself. Every non-inbred individual has kinship coefficient 1/2 with itself.

These individuals are inbred.

```
> inbreeding("Q", thompson)
```

```
     Q
0.0625
```

so their kinship coefficients with themselves are greater than 1/2.

8 Numerator Relationship Matrix

When we find the matrix of kinship coefficients of all individuals and multiply by two, this is called the *numerator relationship matrix*. It has an important use in quantitative genetics, but we won't explain that here.

```
> foo <- 2 * kinship(unique(thompson), thompson)
```

References

Becker, R. A., and Chambers, J. M. (1984). *S : An Interactive Environment for Data Analysis and Graphics*. Wadsworth, Belmont, CA.

Becker, R. A., and Chambers, J. M. (1985). *Extending the S System*. Wadsworth, Monterey, CA.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Geyer, C. J. (1988). Software for Calculating Gene Survival and Multigene Descent Probabilities and for Pedigree Manipulation and Drawing. Technical Report No. 153, Department of Statistics, University of Washington. https://www.stat.washington.edu/article/tech-report/software-calculating-gene-survival-and-multigene-descent-probabilities-and

Thompson, E. A. (1983). Gene extinction and allelic origins in complex genealogies (with discussion). *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **219**, 241–251. https://doi.org/10.1098/rspb.1983.0072.

Thompson, E. A. (1986). Ancestry of alleles and extinction of genes in populations with defined pedigrees. *Zoo Biology*, **5**, 161–170. https://doi.org/10.1002/zoo.1430050210.