RESEARCH NOTE

# A simple route to maximum-likelihood estimates of two-locus recombination fractions under inequality restrictions

IAIN L. MACDONALD* and PHILASANDE NKALASHE

*Actuarial Science, University of Cape Town, 7701 Rondebosch, South Africa*

[MacDonald I. L. and Nkalashe P. 2015 A simple route to maximum-likelihood estimates of two-locus recombination fractions under inequality restrictions. *J. Genet.* **94**, 479–481]

## Introduction, and statement of the problem

Two methods have been recently described to find the maximum-likelihood estimates of two-locus recombination fractions for the phase-unknown triple backcross with two offspring in each family. We present here an approach which seems much easier to implement than those two methods, and (like those methods) respects the relevant constraints. It requires fewer than 15 lines of R code.

Zhou *et al.* (2008, 2011) describe, and present two solutions of, a statistical estimation problem arising in genetics which can be summarized briefly as follows. Consider three biallelic loci $A$, $B$ and $C$, in that order. Let the two-locus recombination fractions be denoted by $\theta_{AB}$, $\theta_{BC}$ and $\theta_{AC}$, which must satisfy

$$\theta_{AB} \le \theta_{AC}, \quad \theta_{BC} \le \theta_{AC}, \quad \theta_{AC} \le \theta_{AB} + \theta_{BC}, \quad \theta_{AC} \le 1/2. \tag{1}$$

Zhou *et al.* (2008, p. 3, col. 2) give the rationale of these constraints, which they describe as 'natural and necessary'.

The joint recombination fraction, denoted by $g_{ij}$, is the proportion of recombination events that occurred between loci $A$ and $B$ if $i = 1$, and between loci $B$ and $C$ if $j = 1$. For instance, $g_{10}$ represents the probability of recombination between loci $A$ and $B$, but not between loci $B$ and $C$. The $\theta$ s and the $g$ s are related thus:

$$\theta_{AB} = g_{11} + g_{10}, \quad \theta_{BC} = g_{11} + g_{01}, \quad \theta_{AC} = g_{10} + g_{01}.$$

We seek to estimate the three probabilities $g_{11}$, $g_{10}$ and $g_{01}$, and hence $\theta_{AB}$, $\theta_{BC}$ and $\theta_{BC}$, subject to the four constraints

$$g_{11} \le g_{01}, \quad g_{11} \le g_{10}, \quad g_{11} \ge 0, \quad g_{01} + g_{10} \le 1/2. \tag{2}$$

It can be shown, and indeed is stated by Zhou *et al.* (2008, p. 4, col. 1) that the constraints (2) are equivalent to the constraints (1) on the marginal recombination fractions. With $g_{00} = 1 - g_{01} - g_{10} - g_{11}$, which by the constraints (2) is $\in [0, 1]$, we write $\boldsymbol{g} = (g_{11}, g_{10}, g_{01}, g_{00})$.

Ott (1991) divides all possible two-offspring haplotype pairs into four phenotype classes with probabilities $p_k$ ($k = 1, 2, 3, 4$) given as follows in terms of the joint recombination fractions $g_{..}$; see Ott (1991, table 6.4, p. 119) or Zhou *et al.* (2008, table 2).

$$\begin{aligned} p_1(\boldsymbol{g}) &= g_{11}^2 + g_{10}^2 + g_{01}^2 + g_{00}^2; \\ p_2(\boldsymbol{g}) &= 2(g_{11}g_{10} + g_{01}g_{00}); \\ p_3(\boldsymbol{g}) &= 2(g_{11}g_{01} + g_{10}g_{00}); \\ p_4(\boldsymbol{g}) &= 2(g_{11}g_{00} + g_{10}g_{01}). \end{aligned}$$

Given the number of observations $n_k$ in each phenotype class ($k = 1, 2, 3, 4$), the relevant log-likelihood is $l = \sum_{k=1}^{4} n_k \log p_k(\boldsymbol{g})$. The maximum-likelihood estimate of $\boldsymbol{g}$ is obtained by maximizing $l$ with respect to $g_{11}$, $g_{10}$ and $g_{01}$, subject to the constraints (2).

## Two solutions and a simpler alternative

It is rightly pointed out by Zhou *et al.* (2008) that any 'solution' of an estimation problem that does not satisfy constraints on the parameters which are both natural and necessary is not in fact a solution. Zhou *et al.* (2008) therefore develop a 'restricted EM algorithm' that does respect the constraints, describe several generalizations, carry out a simulation study of their proposed algorithm, and apply it to a dataset given by Clemens *et al.* (2000).

Subsequently, Zhou *et al.* (2011) have described another proposal, a restricted projection algorithm, which also solves the problem subject to the constraints, and converges at quadratic rate. Both methods appear to require for their application considerable mathematical and coding effort; see, for instance, the lengthy and detailed derivations of Zhou *et al.* (2008, p. 4–5) and Zhou *et al.* (2011, p. 277).

*For correspondence. E-mail: iain.macdonald@uct.ac.za.

**Keywords.** recombination fractions; maximum likelihood estimates; inequality restrictions; constrained numerical optimization.

It is our purpose here to describe a simpler route to the MLEs, an approach which we believe can easily be applied, and if necessary modified, by those who lack expertise or interest in the mathematics. We implement our proposal on the same data as considered by Zhou *et al.* (2008, 2011), and compare our parameter estimates and maximized log-likelihood with their results.

## Use of a general-purpose constrained optimizer

The problem as described above is the maximization of a (nonlinear) function of three variables, subject to four linear inequality constraints. Such a constrained optimization problem is well within the capabilities of freely available, easy-to-use constrained optimizers, for instance the standard routine `constrOptim` in R (R Core Team 2014). The method used by `constrOptim` to impose the constraints is the addition of a logarithmic barrier to the objective, after which an unconstrained optimizer is called.

With the starting values $g_{11} = 0.1$ and $g_{10} = g_{01} = 0.2$, we used `constrOptim` to fit the model to the data on a backcross of mice appearing in Clemens *et al.* (2000) and analysed by Zhou *et al.* (2008, 2011); i.e., to the counts $\boldsymbol{n} = (21, 17, 14, 15)$. (Note that `constrOptim` requires starting values to lie in the interior of the feasible set.) No derivatives were supplied to `constrOptim`, and default settings were used. On a standard machine, the results were essentially instantaneous. A range of starting values other than the above were then also attempted, with convergence in all cases to the same log-likelihood value to six significant figures, with some minor variation in the parameter estimates. But this is not surprising in view of the conclusion of Zhou *et al.* (2011, p. 277, col. 2) that their restricted estimate is unique.

We provide in table 1 the resulting parameter estimates and log-likelihood values, plus comparable results from the two methods described by Zhou *et al.* (2008, 2011). What is noticeable is that the log-likelihood values achieved by those two methods appear to be very slightly inferior to that found here by `constrOptim`, although it has to be pointed out that those apparently inferior values were computed from the estimates given to four significant figures by Zhou *et al.* (2008, 2011); likelihood values are not given by Zhou *et al.* (2008, 2011). By tightening the convergence criterion in all three cases, one could probably achieve closer agreement. For comparison, we state the 'unrestricted' estimates supplied by Zhou *et al.* (2008, 2011), along with the associated log-likelihood. The unrestricted estimates violate the constraint $\theta_{BC} \leq \theta_{AC}$, and the resulting log-likelihood is the largest of the four, as one might expect.

**Table 1.** Recombination fractions as estimated by the three methods, plus the unrestricted estimates. The bracketed figures in the log-likelihood column do not appear in the references cited, but were computed from the (presumably rounded) estimates appearing there.

| Source | $\widehat{\theta}_{AB}$ | $\widehat{\theta}_{BC}$ | $\widehat{\theta}_{AC}$ | $l$ |
|---|---|---|---|---|
| Unrestricted, Zhou *et al.* (2008, 2011) | 0.3167 | 0.3942 | 0.3634 | (–92.04725) |
| Zhou *et al.* (2008) | 0.3166 | 0.3738 | 0.3738 | (–92.06545) |
| Zhou *et al.* (2011) | 0.3162 | 0.3744 | 0.3744 | (–92.06519) |
| This work | 0.3168 | 0.3778 | 0.3778 | –92.06450 |

We give below the self-contained `R` code (fewer than 15 lines) that was used to find these estimates. No 'packages' are needed, just the standard optimizer `constrOptim`. Perhaps it is worth stressing that the underlying unconstrained optimizer used here by `constrOptim` is the default, Nelder–Mead simplex, which is relatively crude. Nevertheless, the resulting performance was certainly not worse than that of the published, more sophisticated, alternatives.

## The R code used to find the MLEs

```
minusl = function(g123)
{g=c(g123,1-sum(g123))
p=c(sum(g^2), 2*(g[1]*g[2]+g[3]*g[4]),
       2*(g[1]*g[3]+g[2]*g[4]),
       2*(g[1]*g[4]+g[2]*g[3]))
-sum(n*log(p))
}
g123totheta=function(g)
    c(g[1]+g[2],g[1]+g[3],g[2]+g[3])
n=c(21,17,14,15)
st=c(0.1, 0.2, 0.2)
U=matrix(c(-1,1,0, -1,0,1, 1,0,0,
    0,-1,-1),byr=T,nrow=4)
c=c(0,0,0,-0.5)
model=constrOptim(st,minusl,
    grad=NULL,ui=U,ci=c)
model
g123totheta(model$par)
```

## Conclusion

Much insight into the nature of the optimization problem is to be found in Zhou *et al.* (2008, 2011), but the restricted EM and restricted projection algorithms seem to us an unnecessarily long way round of solving the estimation problem in practice. We suggest that, if the aim is simply to find the MLEs of the recombination fractions, the use of a general-purpose linearly constrained optimizer such as `constrOptim` is much the easier route. We invite readers to run our code and decide for themselves whether our claim is reasonable. More generally, we believe that unnecessarily complicated routes to MLEs are sometimes used in genetic applications when a simpler method, e.g. direct numerical maximization of the likelihood, is entirely adequate; see the examples in MacDonald (2014, Section 4).

**Acknowledgements**

The editor, the referee and Dr Miguel Lacerda are thanked for their very helpful suggestions.

## References

Clemens K. E., Churchill G., Bhatt N., Richardson K. and Noonan F. P. 2000 Genetic control of susceptibility to UV-induced immunosuppression by interacting quantitative trait loci. *Genes Immun.* **1**, 251–259.

MacDonald I. L. 2014 Numerical maximisation of likelihood: A neglected alternative to EM? *Int. Stat. Rev.* **82**, 296–308.

Ott J. 1991 *Analysis of human genetic linkage*. Johns Hopkins University Press, Baltimore, USA, revised edition.

R Core Team 2014 R: a language and environment for statistical computing, version 3.1.2. R Foundation for Statistical Computing, Vienna, Austria.

Zhou Y., Shi N.-Z., Fung W.-K. and Guo J. 2008 Maximum likelihood estimates of two-locus recombination fractions under some natural inequality restrictions. *BMC Genet.* **9**.

Zhou Y., Ma W., Sheng X. and Wang H. 2011 A new strategy for estimating two-locus recombination fractions under some natural inequality restrictions. *J. Genet.* **90**, 275–282.

Received 24 September 2014, in revised form 27 January 2015; accepted 10 February 2015
Unedited version published online: 16 February 2015
Final version published online: 22 July 2015