Marker-assisted selection using ridge regression

JOHN C. WHITTAKER[1]*, ROBIN THOMPSON[2] AND MIKE C. DENHAM[1]

[1] *Department of Applied Statistics, University of Reading, PO Box 240, Earley Gate, Reading RG6 2FN, UK*
[2] *IACR Rothamsted, Harpenden, Herts AL5 2JQ, UK, and Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK*

(*Received 2 March 1999 and in revised form 29 July 1999*)

Summary

In crosses between inbred lines, linear regression can be used to estimate the correlation of markers with a trait of interest; these marker effects then allow marker assisted selection (MAS) for quantitative traits. Usually a subset of markers to include in the model must be selected: no completely satisfactory method of doing this exists. We show that replacing this selection of markers by ridge regression can improve the mean response to selection and reduce the variability of selection response.

1. Introduction

When two inbred lines are crossed, linkage disequilibrium is generated between genetic markers and quantitative trait loci (QTL). Lande & Thompson (1990) noted that this made marker assisted selection (MAS) possible, by using regression of phenotype on markertype to determine the markers associated with the trait because of linkage to QTL and to estimate the correlation between each marker and the trait caused by this linkage. Combination of these marker effects with phenotypic information using a selection index gives a procedure which has been shown by computer simulation to be more effective than selection on phenotype alone when sample sizes are large and heritability low (Zhang & Smith, 1992; Gimelfarb & Lande, 1994*a*, *b*; Whittaker *et al*., 1995). However, Xie & Xu (1998) have pointed out that if the cost of markertyping individuals is allowed for, phenotypic selection may be more cost-effective than MAS at present: widespread adoption of MAS may therefore require further reductions in markertyping costs.

In general, we cannot include all markers in the regression model and so must select a subset of markers to fit. No entirely satisfactory way of doing this exists. Here we evaluate a method which replaces the subset selection procedure by *ridge regression*, a method which often performs better than subset selection in regression problems when prediction is of primary interest (Breiman, 1995).

2. Methods

We shall consider an F2 population derived from a cross between two inbred lines. We label the alleles at the ith QTL in the first line Q_i, and the alleles at the jth marker locus M_j. The corresponding alleles in the second line are labeled q_i and m_j. For each individual in the population we know the phenotype y and the number of M_i alleles at the ith marker locus, $x_i \in \{0, 1, 2\}$, so that the marker genotype of an individual is described by $\mathbf{x} = (x_1, \mathrm{x}_2, \dots x_k)$. From these we wish to construct an estimate $\hat{z}$ of the genetic value of the individual, z. Lande & Thompson (1990) suggested using the linear estimator $\hat{z} = b_0 y + b_1 s$ where, for any individual, the marker score s is given by

$$s = \hat{\beta}_0 + \sum_{i \in \mathscr{A}} \hat{\beta}_i x_i.$$

Here $\mathscr{A}$ is the set of markers for which effects have been fitted. We have considered the problems involved in calculating b_0 and b_1 elsewhere (Whittaker *et al*., 1997); in this paper we are concerned with the calculation of the marker score s and so shall concentrate on the model

$$\hat{z} = \sum_{\mathrm{i} \in \mathscr{A}} \hat{\beta}_{\mathrm{i}} x_i,$$

* Corresponding author. Tel: +44 (0)118 9318023. e-mail: j.c.whittaker@reading.ac.uk

https://doi.org/10.1017/S0016672399004462 Published online by Cambridge University Press

where the constant term $\hat{\beta}_0$ has been suppressed for notational convenience. The marker effects $\hat{\beta}_i$ are estimated by the usual least squares estimators

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y},$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ for the n individuals in the population and $\mathbf{X}$ is a $n \times k$ matrix whose ith row gives the markertype of the ith individual in the population.

(i) Subset selection

Usually we cannot include all markers in $\mathscr{A}$. In an extreme case there may be more markers than individuals in the population, but even if this is not so, fitting too many markers increases the variance of the $\hat{\beta}$ and so results in bad estimates of z (Zhang & Smith, 1992). The problem of choosing a linear model so as to trade off the variance and bias of the estimator of interest (here $\hat{z}$), which increase and decrease respectively as the number of variables included in the model increases, and thus to minimize prediction error, has been much discussed in the statistical literature (Miller, 1990). This remains an active research area.

A number of methods of selecting $\mathscr{A}$ have been suggested. Gimelfarb & Lande (1994*a*, *b*) used a forward selection procedure to select the p markers giving the largest reduction in residual sum of squares (RSS). The forward selection procedure works well at selecting the best p markers, in the sense of performing similarly to examination of all possible models whilst being much cheaper computationally, but does not help us choose p: rather, p must be chosen in advance by the researcher. As the optimum number of markers to include varies with the number and distribution of QTL, marker map, etc., and is therefore unknown, this seems unsatisfactory. Accordingly Whittaker *et al.* (1995) developed an automatic procedure based on Mallows' C_p (Mallows, 1973).

For a linear model with p parameters, Mallows' C_p is defined as

$$C_p = \frac{RSS}{\hat{\sigma}_e^2} - n + 2p,$$

where RSS is the residual sum of squares from the model under consideration, n is the number of observations and $\hat{\sigma}_e^2$ is an estimate of the error variance $\hat{\sigma}_e^2$. Choosing p to minimise C_p is a commonly used way of selecting the number of variables to include in a linear model. However, there are well-known problems with this approach. The theoretical justification of C_p relies on comparing a *single* model with p parameters with another model with p^* parameters. In most applications, including the current one, we compare the best fitting model with p parameters with the best fitting model with p^* parameters. The fact that these models have been selected from a number of possible p and p^* parameter models invalidates this underlying theory and can lead to the selection of far from optimal models. An alternative would be to base a stopping rule for forward selection on F-tests, but it is not clear what the appropriate significance threshold is in these subset selection problems (Miller, 1990).

(ii) Ridge regression

An alternative to subset selection is *ridge regression* (e.g. Myers, 1992). Here *all* variables are included in the model, but the normal least squares estimators given above are replaced by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mathbf{I}$ is the identity matrix. That is, the usual estimates are shrunk towards zero, with the degree of shrinkage determined by the parameter λ. This shrinking of all estimates towards zero by a constant factor is sensible only if all variables have mean zero and the same variance: therefore it is usual to centre and standardize the variables so that we work with the transformed variables $x_i \rightarrow \frac{x_i - \bar{x}_i}{sd_i}$ where $\bar{x}_\mathrm{i}$ and sd_i are the sample mean and standard deviation of the variable x_i.

There has been much debate about the advantages and disadvantages of ridge regression (Breiman, 1995), but there is evidence that it outperforms subset selection in some problems. Here we compare the performance of subset selection and ridge regression in MAS by simulation.

To perform ridge regression, we need some way of choosing λ; a number of ways of doing this exist (Myers, 1992), but we are restricted by our need for an automatic procedure to allow the use of simulation. Write the (unknown) genetic values of the individuals as $\mathbf{z} = (z_1, z_2, \ldots z_n)$ and suppose that for $i = 1, 2, \ldots n$, $y_i = z_i + \epsilon_i$ with ϵ_i independent and identically distributed with mean zero and variance σ_e^2. We want to choose λ to minimize what Breiman (1992) called the *model error* (ME); here

$$ME = \sum_{i=1}^{n} (z_i - \hat{z}_i)^2,$$

where $\hat{z}_i = \sum_{\mathrm{j=1}}^{\mathrm{k}} \hat{\beta}_j x_{ij}$ and $\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\,\mathbf{X}^T\,\mathbf{y}$ as above. It can be shown (Mallows, 1973; Breiman, 1992) that a good estimator of the model error is

$$\hat{ME} = \mathrm{RSS} - \mathrm{n}\sigma_e^2 + 2\,\sigma_e^2\mathrm{tr}[\mathbf{X}^\mathrm{T}\mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X} + \lambda\mathbf{I})^{-1}],$$

where RSS is the residual sum of squares from the model under consideration and tr($\mathbf{M}$) denotes the trace of the $n \times n$ matrix $\mathbf{M}$. Thus we can evaluate $\hat{ME}$

https://doi.org/10.1017/S0016672399004462 Published online by Cambridge University Press

for a range of λ and choose the λ which minimizes $\hat{ME}$. In practice, of course, σ_e^2 must be replaced by an estimate of the environmental variance, as in the formula for Mallows' C_p.

Note a further advantage of ridge regression: the addition of the $\lambda\mathbf{I}$ term reduces collinearity and prevents the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ becoming singular or near-singular. Singular $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ often occur after several generations of selection as the population becomes increasingly inbred so this is potentially important here.

3. Simulations

We compared the following three methods: selection based solely on an individual's phenotypic value (PHENO), selection based on $\hat{z} = \Sigma_{i \in \mathscr{A}} \hat{\beta}_{\mathrm{i}} x_i$ where $\mathscr{A}$ is chosen using the Mallow's C_p scheme described in Whittaker *et al.* (1995) (MARKER) and selection based on ridge regression (RIDGE) as outlined above.

In all simulations, QTL are assumed to combine additively both between and within loci. The genetic map from which the data were simulated had 10 chromosomes, each of length 1 morgan; 5 marker loci were spaced evenly along each chromosome. Locations for 50 QTL were chosen from a uniform distribution, with the QTL effects a_i generated as in Lande & Thompson (1990). Positive and negative alleles were allocated at random between the two lines. Simulations were run with heritabilities 0·1 and 0·2 and with 100 and 400 individuals of each sex. The number of replicates was varied with the population size, with the minimum number used being 300. In every generation the top 20% of individuals of each sex were selected and paired at random; each pair was then assumed to produce exactly five offspring of each sex.

4. Results and discussion

The results obtained are shown in Table 1 as percentages of the maximum genetic value obtainable, that is the genetic value of an individual possessing all favourable alleles. As usual, marker-assisted methods are increasingly favoured, in comparison with selection on phenotype, by increasing population size and decreasing heritability (e.g. Moreau *et al.*, 1998; Van Berloo & Stam, 1999). In all cases RIDGE performs slightly better than both MARKER and PHENO. Also, the standard errors of selection response are smaller for RIDGE than for MARKER, so that RIDGE is the more reliable selection method. This is probably because ridge regression is a very stable procedure in the sense that small changes in the data do not produce large changes in the estimated regression coefficients: subset selection is unstable and so produces more variable response to selection.

Note that in practice we would combine the marker score s with phenotypic information via a selection index, so that $\hat{z} = b_0 y + b_1 s$. Calculation of b_0 and b_1 requires estimation of $\mathrm{Cov}(z; s)$. This is difficult when subset selection is used, but an approach based on cross-validation appears to work reasonably well (Whittaker *et al.*, 1997). Cross-validation could be used again here; alternatively the estimator of ME described above is easily modified to give an estimator for $\mathrm{Cov}(z; s)$.

A number of other subset selection approaches, such as the *non-negative garrote* (Breiman, 1995) or the *lasso* (Tibshirani, 1994), have been suggested; it

Table 1. *Accumulated selection response* (*standard error*) *after* 1, 2, 4, 8 *generations of selection*, *for* (a) $n = 100$, $h^2 = 0{\cdot}1$; (b) $n = 100$, $h^2 = 0{\cdot}2$; (c) $n = 400$, $h^2 = 0{\cdot}1$ and (d) $n = 400$, $h^2 = 0{\cdot}2$

	Generation	PHENO	MARKER	RIDGE
(a)	1	0·0721 (0·0181)	0·1114 (0·0241)	0·1268 (0·0214)
	2	0·1364 (0·0253)	0·2056 (0·0340)	0·2228 (0·0287)
	4	0·2525 (0·0342)	0·3289 (0·0477)	0·3561 (0·0402)
	8	0·4166 (0·0412)	0·4681 (0·0588)	0·5024 (0·0474)
(b)	1	0·1033 (0·0236)	0·1413 (0·0278)	0·1528 (0·0250)
	2	0·1944 (0·0330)	0·2503 (0·0386)	0·2663 (0·0337)
	4	0·3349 (0·0397)	0·3907 (0·0449)	0·4185 (0·0354)
	8	0·5037 (0·0362)	0·5348 (0·0440)	0·5669 (0·0354)
(c)	1	0·0745 (0·0128)	0·1621 (0·0162)	0·1650 (0·0151)
	2	0·1435 (0·0181)	0·2686 (0·0227)	0·2749 (0·0185)
	4	0·2589 (0·0215)	0·3945 (0·0283)	0·4182 (0·0233)
	8	0·4301 (0·0218)	0·5112 (0·0309)	0·5445 (0·0244)
(d)	1	0·1031 (0·0128)	0·1810 (0·0162)	0·1806 (0·0151)
	2	0·1944 (0·0181)	0·2923 (0·0227)	0·2979 (0·0185)
	4	0·3375 (0·0215)	0·4319 (0·0283)	0·4537 (0·0233)
	8	0·5094 (0·0218)	0·5537 (0·0309)	0·5787 (0·0244)

https://doi.org/10.1017/S0016672399004462 Published online by Cambridge University Press

would be interesting to examine their performance in MAS. The Bayesian interpretation of ridge regression also suggests a number of promising alternative methods. In addition, reversible jump MCMC (Green, 1995; Sillanpää & Arjas, 1998) is another approach that has great potential.

References

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: *X*-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

Breiman, L. (1995). Better subset selection using the nonnegative garrote. *Technometrics* **37**, 373–384.

Gimelfarb, A. & Lande, R. (1994*a*). Simulation of marker-assisted selection in hybrid populations. *Genetical Research* **63**, 39–47.

Gimelfarb, A. & Lande, R. (1994*b*). Simulation of marker assisted selection for non-additive traits. *Genetical Research* **64**, 127–136.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Lande, R. & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743–756.

Mallows, C. L. (1973) Some comments on C_p. *Technometrics* **15**, 661–675.

Miller, A. J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

Moreau, L., Charcosset, A., Hospital, F. & Gallais, A. (1998) Marker assisted selection efficiency in populations of finite size. *Genetics* **148**, 1353–1365.

Myers, R. L. (1992). *Classical and Modern Regression Analysis*, 2nd edn. New York: Wiley.

Sillanpää, M. J. & Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line data. *Genetics* **148**, 1373–1388.

Tibshirani, R. (1994). *Regression Selection and Shrinkage Via the Lasso*. Technical Report 9401. Toronto University of Toronto, Department of Statistics.

Van Berloo, R. & Stam, P. (1999) Comparison between marker-assisted selection and phenotypic selection in a set of *Arabidopsis thaliana* recombinant inbred lines. *Theoretical and Applied Genetics* **98**, 113–118.

Whittaker, J. C., Curnow, R. N., Haley, C. S. & Thompson, R. (1995). Using marker-maps in marker-assisted selection. *Genetical Research* **66**, 255–265.

Whittaker, J. C., Haley, C. S. & Thompson, R. (1997) Optimal weighting of information in MAS. *Genetical Research* **69**, 137–144.

Xie, C. Q. & Xu, S. Z. (1998) Efficiency of multistage marker-assisted selection in the improvement of multiple quantitative traits. *Heredity* **80**, 489–498.

Zhang, W. & Smith, C. (1992). Computer simulation of marker-assisted selection utilizing linkage disequilibrium. *Theoretical and Applied Genetics* **83**, 813–820.

https://doi.org/10.1017/S0016672399004462 Published online by Cambridge University Press