The LASSO Linear Mixed Model for Mapping Quantitative Trait Loci

Scott Foster

Doctor of Philosophy
May 2006

Supervisors: Arūnas Verbyla and Wayne Pitchford

THE UNIVERSITY OF ADELAIDE
School of Agriculture and Wine
BiometricsSA and Agricultural and Animal Sciences

Contents

Abstract

This thesis concerns the identification of quantitative trait loci (QTL) for important traits in cattle line crosses. One of these traits is birth weight of calves, which affects both animal production and welfare through correlated effects on parturition and subsequent growth. Birth weight was one of the traits measured in the Davies' Gene Mapping Project. These data form the motivation for the methods presented in this thesis.

Multiple QTL models have been previously proposed and are likely to be superior to single QTL models. The multiple QTL models can be loosely divided into two categories: 1) model building methods that aim to generate good models that contain only a subset of all the potential QTL; and 2) methods that consider all the observed marker explanatory variables. The first set of methods can be misleading if an incorrect model is chosen. The second set of methods does not have this limitation. However, a full fixed effect analysis is generally not possible as the number of marker explanatory variables is typically large with respect to the number of observations. This can be overcome by using constrained estimation methods or by making the marker effects random.

One method of constrained estimation is the least absolute selection and shrinkage operator (LASSO). This method has the appealing ability to produce predictions of effects that are identically zero. The LASSO can also be specified as a random model where the effects follow a double exponential distribution.

In this thesis, the LASSO is investigated from a random effects model perspective. Two methods to approximate the marginal likelihood are presented. The first uses the standard form for the double exponential distribution and requires adjustment of the score equations for unbiased estimation. The second is based on an alternative probability model for the double exponential distribution. It was developed late in the candidature and gives similar dispersion parameter estimates to the first approximation, but does so in a more direct manner.

The alternative LASSO model suggests some novel types of predictors. Methods for a number of different types of predictors are specified and are compared for statistical efficiency.

Initially, inference for the LASSO effects is performed using simulation. Essentially, this treats the random effects as fixed effects and tests the null hypothesis that the effect is zero. In simulation studies, it is shown to be a useful method to identify important effects. However, the effects are random, so such a test is not strictly appropriate. After the specification of the alternative LASSO model, a method for making probability statements about the random effects being above or below zero is developed. This method is based on the predictive distribution of the random effects (posterior in Bayesian terminology).

The random LASSO model is not sufficiently flexible to model most QTL mapping data.

Typically, these data arise from large experiments and require models containing terms for experimental design. For example, the Davies' Gene Mapping experiment requires fixed effects for different sires, a covariate for birthdate within season and random normal effects for management group. To accommodate these sources of variation a mixed model is employed. The marker effects are included into this model as random LASSO effects. Estimation of the dispersion parameters is based on an approximate restricted likelihood (an extension of the first method of estimation for the simple random effects model). Prediction of the random effects is performed using a generalisation of Henderson's mixed model equations.

The performance of the LASSO linear mixed model for QTL identification is assessed via simulation. It performs well against other commonly used methods but it may lack power for lowly heritable traits in small experiments. However, the rate of false positives in such situations is much lower. Also, the LASSO method is more precise in locating the correct marker rather than a marker in its vicinity. Analysis of the Davies' Gene Mapping Data using the methods described in this thesis identified five non-zero marker-within-sire effects (there were 570 such effects). This analysis clearly shows that most of the genome does not affect the trait of interest.

The simulation results and the analysis of the Davies' Gene Mapping Project Data show that the LASSO linear mixed model is a competitive method for QTL identification. It provides a flexible method to model the genetic and experimental effects simultaneously.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made.

I give consent to this copy of my thesis, when deposited in the University Library, being available in all forms of media, now or hereafter known.

SIGNED: ..... **DATE:** .....

Acknowledgements

Firstly and most importantly, I would like to thank my supervisors Ari and Wayne. Not only have they been very generous with their advice, knowledge and patience, but also with their friendship. I have learnt both technical and personal skills from them that I will endeavour to carry through life. It has been a real pleasure.

The seminal idea for this thesis was passed on from Robin Thompson. I am indebted as it has proved interesting, challenging and rewarding.

The data set from the Davies' Gene Mapping Project was generously made available by Wayne and Cindy Bottema. It represents many years of hard work by their team and I appreciate the chance to use it.

I would like to thank the staff and students at BiometricsSA and in the Discipline of Agricultural and Animal Sciences. They have provided support, technical advice and stimulating discussions. Julian Taylor has been especially helpful (and has been very gracious about giving time for 'dumb' computing questions).

My scholarship was provided in part by the University of Adelaide and in part by the Co-operative Research Centre for Beef and Cattle Quality. I am grateful for their support, without which I would not have even started.

My partner Zoë deserves special thanks. Our time in Adelaide has been truly happy.