Fine mapping of QTL using linkage disequilibrium and linkage analysis for non-normal distributed traits

Rodrigo Labouriau * Mogens S. Lund *

September 2004 - 55^{th} EAAP, Section G6

Abstract

Fine mapping of QTL regions using linkage disequilibrium and linkage analyses can be accomplished by using appropriate multivariate normal linear mixed models (Lund *et al.*, 2003). We present extensions of these models based on the analogous generalized linear mixed models. In the new class of models introduced, traits are assumed distributed according to an "Exponential Dispersion Model" (Jørgensen *et al.*, 1996), which includes classic families of distributions such as the normal, gamma, binomial, negative binomial and the Poisson. Here an interesting case is the compound Poisson that is a family of positive valued distributions that present positive probability of taking the value zero. This can be used to represent a QTL situated in a switch-regulated chromosomal region or a QTL related to genes with very low penetrance with their effects sometimes being below a minimum action threshold.

1 Introduction

This text briefly discusses some ideas for extending statistical models currently used in animal genetics to detect, locate and estimate the effect of Quantitative Trait Loci (QTL). The models we have in mind are classically based on the assumption that the data can be reasonably described by a suitable multivariate normal distribution. We show that these models can be naturally extended to a richer class of parametric families of distributions: the exponential dispersion models. Since the exponential dispersion models contain many classic parametric families of distributions, other than the normal distribution (and many less know, but still interesting families), the proposed extension might provide a flexible tool which can have a strong impact in animal genetic applications. The techniques we describe are known in the statistical literature as "Generalized Linear Mixed Models" (GLIMM) (see Fahrmeir and Tutz, 1994 and the references therein).

2 The genetic scenario

We chose the following scenario to illustrate the use of the generalized linear mixed models in animal genetics. Consider a model for fine mapping of QTL as in Lund *et al.* (2003), where the QTL effects are treated as random components with covariance matrix proportional to the so called IBD matrix, i.e. a matrix where each element is the probability

*Department of Animal Breeding and Genetics; Danish Institute of Agricultural Sciences.
e-mail: rodrigo.labouriau@agrsci.dk

that a pair of alleles is identical by descent, given markers information and a putative QTL position in the chromosome region in play (for example calculated by the so called "gene dropping" method). Here typically a mixed model as described below is used. For simplicity we assume that there is only one QTL in the genome region studied and that the trait of interest is one-dimensional. The model can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Wq} + \boldsymbol{e} \ , \qquad (1)$$

where $\boldsymbol{y}$ is a $(n \times 1)$ vector of values of the trait of interest, $\boldsymbol{\beta}$ is a $(b \times 1)$ vector with b fixed effects for which we want to correct for, $\boldsymbol{X}$ is a $(n \times b)$ design matrix for the fixed effects, $\boldsymbol{u}$ is the $(l \times 1)$ vector of the l polygenic effects, $\boldsymbol{Z}$ is the $(n \times l)$ matrix relating individuals to the polygenic effects, q is the (1×1) QTL effect, $\boldsymbol{W}$ is the $(n \times 1)$ vector relating individuals to the QTL effect and $\boldsymbol{e}$ is the $(n \times 1)$ vector of residuals. It is assumed that $\boldsymbol{u}$, $\boldsymbol{q}$ and $\boldsymbol{e}$ are uncorrelated, multivariate normally distributed as specified below:

$$\boldsymbol{u} \ \sim \ N\left(\boldsymbol{0}\,,\, \sigma_g^2\, \boldsymbol{A}\right) \qquad (2)$$

$$\boldsymbol{q} \ \sim \ N\left(\boldsymbol{0}\,,\, \sigma_q^2\, \mathrm{IBD}_{/M,p}\right) \qquad (3)$$

$$\boldsymbol{e} \ \sim \ N\left(\boldsymbol{0}\,,\, \sigma_r^2 I\right) \ , \qquad (4)$$

where I is the $n \times n$ identity matrix, $\boldsymbol{A}$ is the genetic additive relationship matrix and $\mathrm{IBD}_{/M,p}$ is the IBD matrix for the putative QTL, conditional on the marker data M and the assumed QTL position on the chromosome.

Note that this is equivalent to assume that the trait is multivariate normal distributed with mean and covariance matrix given as below,

$$\boldsymbol{y} \sim N\Big(\boldsymbol{X\beta}, \sigma_g^2\, \boldsymbol{A} + \sigma_q^2 \mathrm{IBD}_{/M,p} + \sigma_r^2 I\Big). \qquad (5)$$

The idea of the generalization we propose is to replace the normal distribution in (5) by another distribution with similar properties (an exponential dispersion model). This substitution is made in such a way that the mean and the covariance structure described above is kept unchanged, preserving in this way the genetic interpretation of the model.

3 Exponential dispersion models and generalized linear mixed models

We introduce now the notion of exponential dispersion models (EDM), which will be used to generalize the model discussed in section 2. The following families of distributions are examples of EDM: normal, gamma, inverse gaussian, Poisson, binomial, negative binomial, among many others. EDM are parametric families of distributions which have many features in common with the family of normal distributions. For example, the EDM can be parametrized by the mean and the variance. Other nice mathematical properties of the EDM, that are probably not visible to the reader, made it possible to develop a detailed asymptotic theory of EDM that resembles very much the well known inference theory for the normal family of distributions. This was the basis of the extension of the classic linear models based on the normal distribution to the so called "generalized linear models" in the 1980s. This new statistical technique, not only put under the same hat several known techniques as linear regression, ANOVA, logistic and probit regression, but also proposed several new models. A further extension occurred in the 1990s by incorporating random effects in generalized linear models, generating what is now called "Generalized Linear Mixed Models" (GLIMM). The main idea of this text is that GLIMM provide the tools for extending many models commonly used in animal genetics, allowing the use of distributions other than the

normal distribution.

Formally, a family of distributions for which the density or probability function that can be expressed in the following form:

$$p(y;\theta,\lambda) = \exp\left[\,\lambda\left\{y\theta - b(\theta)\right\} + c(y,\lambda)\,\right], \quad (6)$$

for suitable functions b and c, is called an *exponential dispersion model* (Jørgensen, 1987). Here θ and λ are parameters indexing the family.

It is well known from the classic theory of exponential dispersion models (see Jørgensen, 1987 or Jørgensen, 1998) that the expectation of a random variable Y with distribution given by (6) is

$$\mathrm{E}(Y) = \mu = b'(\theta)\,. \quad (7)$$

The function b' is called the mean value mapping, since it connects the parameter θ with the mean. The function $V(\mu) = b''\left\{(b')^{-1}(\mu)\right\}$ is called the *variance function*, because

$$\mathrm{Var}(Y) = \frac{1}{\lambda}b''(\theta) = \frac{1}{\lambda}V(\mu) = \sigma^2 V(\mu)\,, \quad (8)$$

where $\sigma^2 = 1/\lambda$. The parameter σ^2 is called the *scale parameter*. The variance function plays an important role in the theory of exponential dispersion models, not only because it relates the mean to the variance, but also because it completely characterizes the exponential dispersion model. It is easy to shown that the mean value mapping is injective. Hence we can use the parameters μ and σ^2 to parametrize the model instead of θ and λ. We use the notation

$$Y \sim \mathrm{ED}(\mu, \sigma^2) \quad (9)$$

to denote that the random variable Y is distributed according to an exponential dispersion model with mean μ and scale parameter σ^2.

Now we have all the ingredients to generalize the models described in section 2. Assume that for the i^{th} animal, for $i = 1, \ldots, n$, we have

$$y_i \sim \mathrm{ED}(\upsilon_i, \sigma^2)\,, \quad (10)$$

where y_i is the value of the trait of interest for the i^{th} animal, σ^2 is a general scale parameter and the mean υ_i is given by

$$\upsilon_i = X_i\boldsymbol{\beta} + \boldsymbol{Z}u_i + \boldsymbol{W}q_i + e_i\ .$$

Here ED represents a given EDM. X_i is the i^{th} row of $\boldsymbol{X}$ and u_i, q_i and e_i are the i^{th} element of the vectors $\boldsymbol{u}$, $\boldsymbol{q}$ and $\boldsymbol{e}$ respectively. Furthermore, it is assumed that the vectors $\boldsymbol{u}$, $\boldsymbol{q}$ and $\boldsymbol{e}$ are uncorrelated and multivariate normal distributed with distribution described by (2)-(4). Note that the model described above is a GLIMM. Moreover, this model has the same mean and a similar covariance structure of the classical model defined in section 2, therefore we preserved the genetic interpretation. However, changing the EDM, ED, in (10) changes completely the nature of the model.

4 Some examples

Here is a selected list of exponential dispersion models that are potentially interesting for application in conjunction with the model formulated above.

Gamma The gamma is a classic family of continuous positive left skewed distributions with probability density

$$p(y;\Phi,\lambda) = \Gamma(\lambda)^{-1} y^{\lambda-1}\exp(-\Phi z + \lambda\log\Phi),$$

where y, λ and $\Phi > 0$. It has a variance function given by $V(\mu) = \mu^2$ which means that the variance is proportional to the square of the mean, or equivalently the coefficient of variation is constant. An example of practical use of the gamma distribution in animal sciences is the modelling of pig growth (see Labouriau *et al.*, 2000).

Inverse Gaussian The inverse gaussian is a positive left skewed distribution with density of the form

$$p(y;\theta,\lambda) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left[\frac{-\lambda}{2y}+\lambda\left\{\theta y+(-2\theta)^{1/2}\right\}\right],$$

where y, λ and $\theta > 0$. The inverse gaussian family is an EDM with variance function $V(\mu) = \mu^3$ and therefore is appropriate for modelling data where the variance increases rapidly with the mean.

It can be shown that the waiting time for the first passage of a Brownian motion (with drift) through a fixed barrier follows an inverse gaussian distribution. Since the Brownian motion is a classic model for describing the spatial distribution of molecules under agitation, this distribution has been used to model substance percolation through membranes. An example in animal sciences is the modelling of meat drip loss in chicken.

von Misses - Fisher This is a continuous distribution taking values in the interval $[0\,,\,2\pi]$ with probability distribution given by

$$p(y;\mu,\lambda) = \frac{1}{2\pi I_0(\lambda)}\exp\{\lambda\cos(y-\mu)\},$$

where y and μ are in the interval $[0\,,\,2\pi]$, $\lambda > 0$ and I_0 is the modified Bessel function. The von Misses-Fisher distributions form an EDM with the variance function $V(\mu) = 1$ (the same as the normal distribution, but with domain $(0, 2\pi)$) and is classically used to model angles (e.g. wind directions in meteorology). In animal sciences, apart from the uses for modelling some anatomic features, one can think in characteristics that are necessarily distributed in a limited region (from below and from above), since it is possible to re-scale such a variable to fit the interval $[0\,,\,2\pi]$.

Binomial The binomial distribution arises when the data is obtained by counting the number of successes in essays (called Bernoully essays) with a fixed probability of success. Classical models associated with this distribution are logistic regression and probit models. In animal sciences immediate uses of this family of distribution can be find in modelling occurrences of diseases and in fertility studies, among many others that can be easily found.

Poisson The Poisson is a classic discrete distribution used to model counting data. It plays a central role in modelling counting data.

Negative Binomial The negative binomial is a discrete distribution describing the number of Bernoully essays performed until obtaining the first success. This is a distribution proper for counting data, specially useful when the Poisson cannot be used due to data under- or over-dispersion, i.e. when the data dispersion does not fit the dispersion one would expect for a Poisson distributed data.

Compound Poisson The compound Poisson is a positive continuous distributions that attributes a positive probability to the value zero. This distribution can be obtained by the following construction. Let N and $X_1, X_2, \ldots$ be a sequence of independent random variables with N Poisson distributed and the X_is identically distributed according to a gamma distribution. Define a new variable Z by putting $Z = 0$ if $N = 0$ and if $N \neq 0$ then putting

$$Z = \sum_{i=1}^{N} X_i\,.$$

The distribution of Z is a compound Poisson. It can be shown that the compound Poisson distributions are EDM with variance function $V(\mu) = \mu^p$ for p between 1 and 2 (excluding both).

The compound Poisson yields an interesting model since it can be used to model data with positive probability of a zero outcome, but otherwise following a continuous distribution taking positive values. This distribution has been used in insurance for modelling the value of the yearly claim for an individual insurance holder (here N would describe the number of claims and the X_is would represent the values of the different claims). Other classic applications of this family of distributions is in meteorology for modelling the daily amount of rain. In animal genetics we can think on a trait which values are close a detection lower bound (that some times is detected and sometimes is not), or an additive effect on a trait that is switch regulated by an environmental factor or a gene that is not linked with the chromosomal region under study.

5 Discussion

The main advantage of GLIMM is that it makes it possible to vary the distribution used without changing the regression structure of the model. In the case of the genetic applications we illustrated here it is essential to keep the linearity of the regression function, preserving in this way the additivity of the random components. This is not accomplished if for example the random variable representing the trait of interest is transformed to reach normality.

The usual claim on robustness of normal distribution based methods against deviations from the normality is only valid in some examples and should be verified case by case. Indeed, relatively simple calculations based on the notion of Hampel's influence function (see Huber, 1981 and Hampel *et al*, 1986) and derived classic measures of robustness show that not all deviations from the normality are innocuous for the models we consider here,even with very large samples (see Labouriau, 1989). This is in accordance with our simulation studies (not presented here).

The statistical inference for generalized linear mixed models (GLMM) is not as straightforward as it seems at first glance. There are several proposals for statistical inference in GLMM in the literature, some of them implemented in standard software yet. However, general implementations might present serious limitations when highly complex covariance structures are present in the model, as it is the case in the applications we described here. We are currently developing an implementation based on iterated recursive solutions of re-weighted normal mixed problems. The advantage of this method is that it allows to use software specifically developed for classic QTL detection in animal genetics, as for example the highly optimized program DMU (Madsen and Jensen, 2003).

There are no difficulties to extend to multitraits when all the traits follow the same exponential dispersion model. Essentially the same technique (with the obvious adaptations) can be used. However, specific methods and software are necessary to be developed if the traits of interest do not follow the same probability law. We are currently incorporating this feature in a version of the program DMU.

In conclusion, we presented a technique to generate alternative models in animal genetics that does not assume the trait of interest to be normally distributed. Surely, many other similar examples from different areas of animal genetics can be generated using the techniques presented. The use of these techniques in animal genetics will depend on the capacity to interact and on the imagination of animal geneticists and statisticians.

References

[1] Hampel, F. R., Ronchetti, E. M., Rousseaauw, P. J. and Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions*. Wiley, New York.

[2] Huber P. J. (1981). *Robust Statistics*. John Wiley and Sons, New York.

[3] Labouriau, R (1989). Tópicos de Robustez Quantitativa em Famílias Exponenciais de Distribuições a um Parâmetro. (On quantitative robustness in exponential families of distributions with one-dimensional parameter) "Série D", **32**, IMPA, Rio de Janeiro, Brazil, 89 pp.

[4] Labouriau, R., Schulin-Zeuthen, M. and Danfær, A. (2000). Statistical analysis of pigs development: An application of Richards regression models. Internal Report of the Biometry Research Unit Nr. **14**, pp 13.

[5] Lund, M.,S., Sørensen,P., Guldbrandtsen, B. and Sorensen, D.A. (2003) Multitrait fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. *Genetics* **163**, 405-410.

[6] Jørgensen, B. (1997). *The Theory of Exponential Dispersion Models*. Chapman and Hall, London.

[7] Jørgensen, B., Labouriau, R. and Lundbye-Christensen, S.(1996). Linear Growth curve analysis based on exponential dispersion models. *Journal of the Royal Statistics Society*, B **58**, 3, 573-592.

[8] Madsen, P. and Jensen, J. (2003). *A User's Guide to DMU: A Package for Analysing Multivariate Mixed Models*. Version 6, release 4. Danish Institute of Agricultural Sciences, Research Centre Foulum.

[9] Tutz, G. and Fahrmeir, L. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-verlag, New York.