Heredity (2006) 97, 409–417

www.nature.com/hdy

A multivariate model for ordinal trait analysis

S Xu and C Xu[1]
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Many economically important characteristics of agricultural crops are measured as ordinal traits. Statistical analysis of the genetic basis of ordinal traits appears to be quite different from regular quantitative traits. The generalized linear model methodology implemented via the Newton–Raphson algorithm offers improved efficiency in the analysis of such data, but does not take full advantage of the extensive theory developed in the linear model arena. Instead, we develop a multivariate model for ordinal trait analysis and implement an EM algorithm for parameter estimation. We also propose a method for calculating the variance-covariance matrix of the estimated parameters. The EM equations turn out to be extremely similar to formulae seen in standard linear model analysis. Computer simulations are performed to validate the EM algorithm. A real data set is analyzed to demonstrate the application of the method. The advantages of the EM algorithm over other methods are addressed. Application of the method to QTL mapping for ordinal traits is demonstrated using a simulated baclcross (BC) population.
Heredity (2006) **97,** 409–417. doi:10.1038/sj.hdy.6800885; published online 16 August 2006

Keywords: EM algorithm; generalized linear model; ordinal traits; QTL; truncated normal distribution

Introduction

Many disease resistance traits in plants are scored in several ordered categories based on the magnitude of the disease symptoms. For example, this approach has been used for sheath blight resistance in rice (Zou *et al*, 2000), clubroot resistance in *Brassica napus* (Manzanares-Dauleux *et al*, 2000) and cucumber mosaic virus resistance in pepper (Caranta *et al*, 2002). Similarly, many characters in animals and humans are expressed as binary or ordinal traits, including the score for calving difficulty, expression of congential malformations, numbers of reproductive events and so on. In other cases traits are actually continuously distributed, but, for technical reasons, measured as ordinal traits.

Special statistical methods are required to analyze traits measured on an ordinal scale (McCullagh and Nelder, 1989). The probability model of McIntyre *et al* (2001) used the trait penetrances directly as the genetic parameters of interest. The method can only be applied to QTL mapping for binary traits. A generalized linear model is currently considered to be the most appropriate for ordinal data analysis method because, using a simple link function, we can adopt theory and methodology developed extensively in linear model for continuously distributed data. The key to using a generalized linear model is the use of a hypothetic continuous latent variable (known as liability). The observed categorical phenotype depends on whether the liability exceeds one or more of an ordered threshold. This generalized linear model is, therefore, also called the threshold model (Lynch and Walsh, 1998).

Many statistical methods of estimation and hypothesis testing have been developed under the threshold model. They include a maximum likelihood method (Aitchison and Silvery, 1957; Ashford, 1959) and a Bayesian method (Albert and Chib, 1993; Sorensen *et al*, 1995). Under the maximum likelihood framework, parameters are often estimated iteratively via the Fisher scoring algorithm or the Newton–Raphson ridge algorithm (Ashford, 1959). When the Bayesian method is applied, the posterior means or modes of parameters are often inferred from a posterior sample generated from a Markov chain Monte Carlo process (Sorensen *et al*, 1995). The Bayesian method is more versatile than the ML method because it can handle more complicated models. However, the ML method is more cost effective because no MCMC sampling is required. Both methods are, therefore, currently being used in ordinal data analysis. A thorough description of the statistical methods for ordinal data analysis may be found in McCullagh and Nelder (1989) and Fahrmeir and Tutz (1994).

When parameters are estimated using the Bayesian method via MCMC sampling, realizations of the latent variable are sampled from its conditional posterior distribution (a truncated normal distribution if the probit link function is used). Once the latent variable has been sampled, the problem of parameter estimation becomes that of parameter estimation in the usual linear model. In this way, the latent variable is treated as a missing value in the Bayesian analysis.

For binary data, Xu *et al* (2003) developed an EM algorithm to search for the MLE of parameters, also by treating the latent variable as a missing value. Xu *et al* (2005a, b) recently extended the EM algorithm for binary data to handle ordinal traits and multiple binary traits. Compared to the Fisher scoring or the Newton–Raphson

Correspondence: S Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.
E-mail: xu@genetics.ucr.edu
The work was done when Dr Chenwu Xu was a postdoctoral research associate at UC Riverside. He has returned to China.
[1]*Current address: Department of Agronomy, Yangzhou University, Yangzhou, Jiangshu Province, PR China*
Received 18 December 2005; accepted 12 July 2006; published online 16 August 2006

algorithm, the EM algorithm has the following desirable properties: (1) it is numerically stable, in stark contrast to the Newton–Raphson algorithm that crashes easily when the thresholds are not well separated; (2) the steps of EM iterations are transparent and intuitive, and thus easily understood by biologists; (3) the EM algorithm takes full advantage of the results developed in the usual linear model analysis. Unfortunately, the EM algorithm also has two undesirable properties: (1) the convergence process may be slow and (2) it does not automatically provide an estimate of the variance–covariance matrix of the parameters. The first undesirable property is no longer a problem, thanks to the ever-growing computing power. The second problem has been circumvented by using the Louis' (1982) information matrix for EM. Xu *et al* (2003) developed the information matrix of parameters for binary trait analysis, which were extended recently by Xu *et al* (2005a) to handle ordinal traits. However, the method of Xu *et al* (2005a) estimates parameters in two steps: estimating the regression coefficients conditional on the thresholds and estimating the thresholds conditional on the regression coefficients. They called the method ECM algorithm (expectation and conditional maximization, Meng and Rubin (1993)). Although the ECM method is a convenient approach for finding the ML solution, the information matrix of the parameters are hard to derive. Therefore, Xu *et al* (2005a) did not provide an estimate of the variance–covariance matrix.

In this study, we propose an EM algorithm for parameter estimation in a single step, that is, solving for the thresholds and regression coefficients simultaneously. As a consequence, the variance–covariance matrix of the parameters can be found on the basis of the Louis' information matrix.

Statistical models

Univariate model

Let s_j be the ordinal data observed for subject j, $\forall j = 1, \ldots, n$, where n is the sample size. Let C be the number of ordered categories in the data set. Variable s_j is defined as $S_j = k$ if j belongs to category k, $\forall k = 1, \ldots, C$. A set of fixed thresholds, $t_1, t_2, \ldots, t_{C-1}$, on an underlying scale define the observed categories on the ordinal scale. Further define y_j as an underlying latent variable for individual j. The relationship between the latent variable and the thresholds is $S_j = k$ if $t_{k-1} < y_1 \le t_k$ where $t_0 = -\infty$ and $t_C = \infty$. Here, we define $C+1$ thresholds but only $C-1$ of them are parameters of interest and these thresholds are denoted by a vector $\mathbf{t} = [t_1\, t_2 \ldots t_{C-1}]^T$. For notational convenience, let $m = C-1$ so that $\mathbf{t}$ is an $m \times 1$ vector.

The natural choice for the distribution of y is the normal distribution, under which the model is called the probit model. The latent variable is described by the following linear model

$$y_j = \mathbf{X}_j\mathbf{b} + e_j \tag{1}$$

where $\mathbf{b}$ is a $p \times 1$ vector for the model effects, $\mathbf{X}_j$ is a $1 \times p$ known design matrix, and e_j is a residual error assumed to be $N(0, 1)$ distributed. This model is commonly used in ordinal data analysis, especially in QTL mapping for ordinal traits (Hackett and Weller, 1995; Rao and Xu, 1998; Xu *et al*, 2005a). It is called the univariate model or the threshold model. As mentioned in the introduction, the variance–covariance matrix of the estimated regression coefficients is hard to derive under the EM algorithm, although estimation of the regression coefficients themselves is relatively straightforward (Xu *et al*, 2005a).

Multivariate model

An alternative model for ordinal traits is the so-called multivariate model or cumulative threshold model, in which we formulate the ordinal trait analysis as a multivariate problem. The ordinal trait with C categories can be described by $C-1$ binary traits. Each binary trait is controlled by its own latent variable with its own threshold. These binary observations are defined as

$$w_{jk} = \begin{cases} 0 & \text{if } s_j \le k \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

For each binary trait, we define a latent variable y_{jk} that is linked to w_{jk} by $w_{jk} = 0$ for $y_{jk} \le t_k$ and $w_{jk} = 1$ for $y_{jk} > t_k$. So, each subject is described by a vector $\mathbf{w}_j = [w_{j1}\, w_{j2} \ldots w_{jm}]^T$ which is controlled by a vector $\mathbf{y}_j = [y_{j1}\, y_{j2} \ldots y_{jm}]^T$. The relationship between S_j under the univariate model and $\mathbf{w}_j$ under the multivariate model is shown in Table 1 for the case where $C = 6$.

The multivariate liabilities are described by the following multivariate linear model,

$$\begin{bmatrix} y_{j1} \\ y_{j2} \\ \vdots \\ y_{jm} \end{bmatrix} = -\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix} + \begin{bmatrix} x_{j1} & x_{j2} & \cdots & x_{jp} \\ x_{j1} & x_{j2} & \cdots & x_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jp} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_{j1} \\ e_{j2} \\ \vdots \\ e_{jm} \end{bmatrix} \tag{3}$$

In matrix notation, we have

$$\mathbf{y}_j = -\mathbf{I}_m\mathbf{t} + (\mathbf{1}_m \otimes \mathbf{X}_j)\mathbf{b} + \mathbf{e}_j \tag{4}$$

where $\mathbf{I}_m$ is a $m \times m$ identity matrix, $\mathbf{1}_m$ is an $m \times 1$ vector with all elements equal to 1 and $\otimes$ represents the Kronecker matrix product. Let $\boldsymbol{\theta} = \mathbf{t}//\mathbf{b}$ be an $(m+p) \times 1$ vector for the parameters and $\mathbf{Z}_j = (-\mathbf{I}_m)\,||\,(\mathbf{1}_m \otimes \mathbf{X}_j)$ be an $m \times (m+p)$ design matrix, where the symbols '//' and '||' represent vertical and horizontal matrix concatenations, respectively, a notation adopted from SAS/IML

Table 1 Relationship between s_j under the univariate model and $\mathbf{w}_j$ under the multivariate model for $C = 6$

Univariate (s_j)	*Multivariate* ($\mathbf{w}_j$)				
	w_{j1}	w_{j2}	w_{j3}	w_{j4}	w_{j5}
1	0	0	0	0	0
2	1	0	0	0	0
3	1	1	0	0	0
4	1	1	1	0	0
5	1	1	1	1	0
6	1	1	1	1	1

(SAS Institute, 1999a). The above model (Equation (4)) can be rewritten as

$$\mathbf{y}_j = \mathbf{Z}_j\boldsymbol{\theta} + \mathbf{e}_j \tag{5}$$

The residual errors are assumed to be distributed as an m-dimensional independent normal, that is, $\mathbf{e}_j \sim N_m(\mathbf{0}, \mathbf{I}_m)$.

Parameter estimation

Under the multivariate model, we have formulated the thresholds as a subset of the regression coefficients. As a result, we are able to estimate the entire parameter vector (including the thresholds and the original regression coefficients) simultaneously in a single step. If we treat the liability vector as data, we have the following complete-data log likelihood

$$L(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{j=1}^{n}(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\theta}) \tag{6}$$

Note that the variance–covariance matrix of the residual errors is an identity matrix (constant), and thus it does not play a role in the maximum likelihood analysis. In addition, this log likelihood function is not the observed likelihood function, which is a function of $\mathbf{w}_j$ with $\mathbf{y}_j$ integrated out (the actual form of the likelihood function is dealt with in the discussion).

The EM algorithm requires maximization of $eL(\boldsymbol{\theta})$, the expected $L(\boldsymbol{\theta})$, with respect to the parameters because the latent variable $\mathbf{y}$ is missing (not observable). The expectation is taken with respect to $\mathbf{y}$ conditional on the parameters ($\boldsymbol{\theta}$) and the observed data ($\mathbf{w}$), and has the following form,

$$\begin{aligned} eL(\boldsymbol{\theta}) &= E[L(\boldsymbol{\theta})|\mathbf{w}, \boldsymbol{\theta}^{(t)}] \\ &= -\frac{1}{2}\sum_{j=1}^{n} E\left[(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{Z}_j\boldsymbol{\theta})|\mathbf{w}_j, \boldsymbol{\theta}^{(t)}\right] \end{aligned} \tag{7}$$

The parameter values, however, are the quantities to be estimated. To calculate the conditional expectation, we need to choose an arbitrary value of $\boldsymbol{\theta}$ from its legal domain to start the maximization process. Once $\boldsymbol{\theta} = \boldsymbol{\theta}^{(\tau)}$ is chosen for, $\tau = 0$, we can use $\boldsymbol{\theta}^{(\tau)}$ to compute the conditional expectation of the complete-data log likelihood (still a function of $\boldsymbol{\theta}$). We then maximize the conditional expectation of the log likelihood (Equation (7)) with respect to the parameters and obtain the following EM iteration formula for $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^{(\tau+1)} = \left[\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\right]^{-1}\left[\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}E(\mathbf{y}_j|\mathbf{w}_j, \boldsymbol{\theta}^{(\tau)})\right] \tag{8}$$

where

$$E(\mathbf{y}_j|\mathbf{w}_j, \boldsymbol{\theta}^{(\tau)}) = \begin{bmatrix} E(y_{j1}|w_{j1}, \boldsymbol{\theta}^{(\tau)}) \\ \vdots \\ E(y_{jm}|w_{jm}, \boldsymbol{\theta}^{(\tau)}) \end{bmatrix} \tag{9}$$

is an $m \times 1$ vector for the conditional expectation of $\mathbf{y}_j$ given the parameter values at the τth iteration and the observed ordinal trait. The kth element of the above vector is

$$\begin{aligned} &E(y_{jk}|w_{jk}, \boldsymbol{\theta}^{(\tau)}) \\ &= \begin{cases} -t_k^{(\tau)} + \mathbf{X}_j\mathbf{b}^{(\tau)} + \dfrac{\phi(t_k^{(\tau)} - \mathbf{X}_j\mathbf{b}^{(\tau)})}{\Phi(\mathbf{X}_j\mathbf{b}^{(\tau)} - t_k^{(\tau)})} \text{for } w_{jk} = 1 \\ -t_k^{(\tau)} + \mathbf{X}_j\mathbf{b}^{(\tau)} - \dfrac{\phi(t_k^{(\tau)} - \mathbf{X}_j\mathbf{b}^{(\tau)})}{\Phi(t_k^{(\tau)} - \mathbf{X}_j\mathbf{b}^{(\tau)})} \text{for } w_{jk} = 0 \end{cases} \end{aligned} \tag{10}$$

where $\phi(x)$ and $\Phi(x)$ represent the standardized normal density function and the cumulative standardized normal distribution function, respectively. Equation (10) is the expectation of a truncated normal variable (Cohen, 1991).

In summary, the EM algorithm requires initialization of the parameters with $\boldsymbol{\theta} = \boldsymbol{\theta}^{(\tau)}$ for $\tau = 0$ and the following two steps:

E-Step:
Calculate $E(\mathbf{y}_j \mid \mathbf{w}_j, \boldsymbol{\theta}^{(\tau)})$ using Equation (10);

M-Step:
Update parameter $\boldsymbol{\theta}$ with $\boldsymbol{\theta} = \boldsymbol{\theta}^{(\tau+1)}$ using Equation (8).

The E- and M-Steps are repeated several times until some criterion of convergence is satisfied. Let $\tau_{\max}$ be the number of iterations taken for the EM algorithm to converge. The MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(\tau_{\max})}$.

Information matrix

The observed information matrix for $\boldsymbol{\theta}$ can be found using the method of Louis (1982), which requires the first and the second partial derivatives of the complete-data log likelihood. The first partial derivative is

$$\frac{\partial}{\partial\boldsymbol{\theta}}L(\boldsymbol{\theta}) = S(\boldsymbol{\theta}, \mathbf{y}) = \sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{y}_j - \sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\boldsymbol{\theta} \tag{11}$$

and the second partial derivative is

$$\frac{\partial^2}{\partial\boldsymbol{\theta}^2}L(\boldsymbol{\theta}) = B(\boldsymbol{\theta}, \mathbf{y}) = -\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j \tag{12}$$

The observed information matrix of Louis (1982) is

$$\begin{aligned} I(\boldsymbol{\theta}) &= \underset{\mathbf{y}}{E}[-B(\boldsymbol{\theta}, \mathbf{y})] - \underset{\mathbf{y}}{E}[S(\boldsymbol{\theta}, \mathbf{y})S^{\mathrm{T}}(\boldsymbol{\theta}, \mathbf{y})] \\ &= \sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j - \sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j \\ &\quad + \sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}E(\mathbf{y}_j)\boldsymbol{\theta}^{\mathrm{T}}\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j \\ &\quad + \sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\boldsymbol{\theta}\sum_{j=1}^{n}E(\mathbf{y}_j^{\mathrm{T}})\mathbf{Z}_j - E\left[\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{y}_j\sum_{j=1}^{n}\mathbf{y}_j^{\mathrm{T}}\mathbf{Z}_j\right] \end{aligned} \tag{13}$$

where $E(\mathbf{y}_j)$ is a short notation for $E(\mathbf{y}_j \mid \mathbf{w}_j, \boldsymbol{\theta})$. After extensive algebraic manipulation, we can show that

$$\begin{aligned} I(\boldsymbol{\theta}) &= \underset{\mathbf{y}}{E}[-B(\boldsymbol{\theta}, \mathbf{y})] - \underset{\mathbf{y}}{E}[S(\boldsymbol{\theta}, \mathbf{y})S^{\mathrm{T}}(\boldsymbol{\theta}, \mathbf{y})] \\ &= \sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\Big(\mathbf{I} - V(\mathbf{y}_j)\Big)\mathbf{Z}_j \end{aligned} \tag{14}$$

where $V(\mathbf{y}_j)$ is a short notation for $V(\mathbf{y}_j \mid \mathbf{w}_j, \boldsymbol{\theta})$ the conditional variance–covariance matrix of $\mathbf{y}_j$. It is an $m \times m$ diagonal matrix with the kth diagonal element defined as

$$V(y_{jk}|w_{jk}, \boldsymbol{\theta}) = \begin{cases} 1 - Q_{j0}(Q_{j0} + \xi_{jk}) \text{ for } w_{jk} = 0 \\ 1 - Q_{j1}(Q_{j1} - \xi_{jk}) \text{ for } w_{jk} = 1 \end{cases} \quad (15)$$

where

$$\begin{cases} \xi_{jk} = t_k - \mathbf{X}_j\mathbf{b} \\ Q_{j0} = \phi(\xi_{jk})/\Phi(\xi_{jk}) \\ Q_{j1} = \phi(\xi_{jk})/\Phi(-\xi_{jk}) = \phi(\xi_{jk})/[1 - \Phi(\xi_{jk})] \end{cases} \quad (16)$$

(see Cohen, 1991 for the variance of a truncated normal distribution).

Hypothesis tests

An analytical form of the likelihood function for the observed data is needed only if a likelihood ratio test is to be performed. Instead, under the multivariate model, we are able to formulate the variance-covariance matrix for the estimated parameters by taking the inverse of the information matrix,

$$V(\boldsymbol{\theta}) \approx I^{-1}(\boldsymbol{\theta}) \quad (17)$$

which can be used to derive the Wald test statistic (Fahrmeir and Tutz, 1994). Therefore, the likelihood ratio test statistic is no longer required. The Wald test for the null hypothesis of H_0: $\mathbf{b} = \mathbf{0}$ is

$$W = \hat{\mathbf{b}}^{\mathrm{T}}[V(\mathbf{b})]^{-1}\hat{\mathbf{b}}, \quad (18)$$

where $V(\mathbf{b})$ is simply a subset of matrix $V(\boldsymbol{\theta})$. Under the null hypothesis, W will asymptotically follow a χ^2-distribution with p degrees of freedom. Each individual regression coefficient can also be tested separately with

$$W_k = \frac{\hat{b}_k^2}{V(b_k)}. \quad (19)$$

Under the null hypothesis H_0: $b_k = 0$, $\boldsymbol{W}_k$ will asymptotically follow a χ^2-distribution with one degree of freedom.

Interval mapping of QTL

The particular reason for developing this EM algorithm was to solve the problem of QTL mapping for ordinal traits. The algorithm is sufficiently general, that we have been able to present it in a more general way. We now demonstrate application of the model to QTL mapping. Suppose that we collect the phenotypes of n backcross (BC) progeny and genotyped all the markers for these individuals. We can scan QTL along the genome using the idea of interval mapping (Lander and Botstein, 1989; Haley and Knott, 1992). The model in the context of interval mapping is

$$\begin{bmatrix} y_{j1} \\ y_{j2} \\ \vdots \\ y_{jm} \end{bmatrix} = -\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix} + \begin{bmatrix} x_j \\ x_j \\ \vdots \\ x_j \end{bmatrix} b + \begin{bmatrix} e_{j1} \\ e_{j2} \\ \vdots \\ e_{jm} \end{bmatrix} \quad (20)$$

where x_j is the genotype indicator variable for individual j at a putative position of the genome and it is defined as $x_j = 1$ for one genotype and $x_j = -1$ for the alternative genotype (only two genotypes are present in an BC population at any particular locus). The regression coefficient $b = \mu_{AA} - \mu_{Aa}$ is the difference between the average values of the two genotypes. Variable x is not observable but its probability distribution can be inferred from genotypes of flanking markers (ie interval mapping).

Two approaches can be taken to incorporate the conditional probability distribution of variable x. One is the mixture model approach (Lander and Botstein, 1989) by treating x_j as a missing value. A detailed algorithm for the mixture model for binary traits has been developed by Xu *et al* (2003), which can be directly adopted here for ordinal traits without much modification. The other approach is to adopt the idea of Haley and Knott (1992) who replaced x_j by $\hat{x}_j = p_j - q_j = 2p_j - 1$, the conditional expectation of x_j given marker information, where $p_j = \Pr(x_j = 1 \mid \text{flanking marker})$ and $q_j = \Pr(x_j = -1 \mid \text{flanking marker}) = 1 - p_j$.

To map QTL in F2 populations, the model should be written as

$$\begin{bmatrix} y_{j1} \\ y_{j2} \\ \vdots \\ y_{jm} \end{bmatrix} = -\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix} + \begin{bmatrix} x_{j1} & x_{j2} \\ x_{j1} & x_{j2} \\ \vdots & \vdots \\ x_{j1} & x_{j2} \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_{j1} \\ e_{j2} \\ \vdots \\ e_{jm} \end{bmatrix} \quad (21)$$

where $x_{j1} = \{+1, 0, -1\}$ and $x_{j2} = \{-1, +1, -1\}$ for genotype array $\{AA, Aa, aa\}$, and $b_1 = a = \frac{1}{2}(\mu_{AA} - \mu_{aa})$ (additive effect) and $b_2 = d = \mu_{Aa} - \frac{1}{2}(\mu_{AA} + \mu_{aa})$ (dominance effect). Let

$$\begin{cases} p_j(+1) = \Pr(x_j = +1 | \text{flanking marker}) \\ p_j(0) = \Pr(x_j = 0 | \text{flanking marker}) \\ p_j(-1) = \Pr(x_j = -1 | \text{flanking marker}) \end{cases}$$

The conditional expectations of the two x variables are

$$\begin{cases} \hat{x}_{j1} = p_j(+1) - p_j(-1) \\ \hat{x}_{j2} = p_j(0) - [p_j(+1) + p_j(-1)] \end{cases}$$

which will replace x_{j1} and x_{j2} in model (21) for QTL mapping in F2 populations.

Illustrations

Example 1: This example shows the analysis of a simulated data set with multiple replications. We simulated four explanatory variables $\mathbf{X}_j = [x_{j1}\ x_{j2}\ x_{j3}\ x_{j4}]$. The values of the four variables for each subject were generated from a multivariate normal distribution, that is, $\mathbf{X}_j$: $N_4(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 & 0 & 0.1 \\ 0.5 & 1 & 0.5 & -0.2 \\ 0 & 0.5 & 1 & 0 \\ 0.1 & -0.2 & 0 & 1 \end{bmatrix}$$

The true values of the regression coefficients were

$$\mathbf{b} = [\,b_1 \quad b_2 \quad b_3 \quad b_4\,]^{\mathrm{T}} = [\,0.5 \quad 0 \quad -0.5 \quad -1.0\,]^{\mathrm{T}}$$

The liability for subject j was generated by $y_j = \mathbf{X}_j\mathbf{b} + e_j$ where e_j was simulated from $N(0, 1)$. The observed

ordinal measurement was converted from y_j using thresholds

$$\mathbf{t} = [t_1 \quad t_2 \quad t_3]^{\mathrm{T}} = [-1.5 \quad 0 \quad 1.5]^{\mathrm{T}}$$

There were three thresholds (excluding $t_0 = -\infty$ and $t_C = +\infty$) and four ordinal categories ($C = 4$). The sample size (n) was simulated at the following four levels: 100, 250, 500 and 750.

For each simulated data set, the proposed EM algorithm was used to estimate the parameters. Meanwhile, the variance-covariance matrix of the EM estimates was calculated using the inverse of the Louis' (1982) information matrix. The simulation was replicated 1000 times, from which we were able to evaluate the property of the EM algorithm. Note that the data were simulated under the univariate model, but analyzed under the proposed multivariate model. We demonstrated that the multivariate model is a good approximation to the univariate model.

For comparison, we also analyzed each data set using the SAS procedure, PROC LOGISTIC (SAS Institute, 1999b), with the logit link function replaced by the probit link function. We were also able to use PROC PROBIT (SAS Institute, 1999b), but for some reason the estimated parameters had signs opposite to the ones obtained from our EM estimates. The PROC LOGISTIC program uses the univariate model but finds the MLE with the Newton–Raphson ridge algorithm. The program also calculates the variance–covariance matrix using the observed information matrix. Results of the Newton–Raphson algorithm are considered as exact because the data were simulated using exactly the same univariate model.

Results from the simulated data analysis are listed in Table 2 for the estimated parameters and the variances of the estimated parameters. Tables 3 and 4 give the covariances of the estimated parameters. Table 2 shows (A) the average $\hat{\boldsymbol{\theta}} = [\hat{t}_1\ \hat{t}_2\ \hat{t}_3\ \hat{b}_1\ \hat{b}_2\ \hat{b}_3\ \hat{b}_4]^{\mathrm{T}}$ calculated across the 1000 replicates, denoted by $\bar{\hat{\boldsymbol{\theta}}} = \frac{1}{1000}\sum_{i=1}^{1000} \hat{\boldsymbol{\theta}}$, (B) the variance of $\hat{\boldsymbol{\theta}}$ calculated across the 1000 replicates, denoted by $V(\hat{\boldsymbol{\theta}}) = \frac{1}{1000-1}\sum_{i=1}^{1000} (\hat{\boldsymbol{\theta}} - \bar{\hat{\boldsymbol{\theta}}})^2$, and (C) the average $V(\boldsymbol{\theta})$ calculated across the 1000 replicates, denoted by $\overline{V(\boldsymbol{\theta})} = \frac{1}{1000}\sum_{i=1}^{1000} V(\boldsymbol{\theta})$.

The bias of parameter estimation for each method can be evaluated by comparing $\hat{\boldsymbol{\theta}}$ with the true value of $\boldsymbol{\theta}$. When the sample size was small, say 100–250, slight bias has been observed for each estimated parameter. There is no clear trend on which method has a larger bias than the

Table 2 Estimated parameters and the variances of the estimates for the EM algorithm and the Newton–Raphson algorithm in the simulated data analysis (example 1)

Sample size	*Method*	t_1	t_2	t_3	b_1	b_2	b_3	b_4
100	EM							
	A	−1.5735	−0.0060	1.5831	0.5317	0.0078	−0.5318	−1.0452
	B	0.0520	0.0256	0.0520	0.0271	0.0434	0.0287	0.0295
	C	0.0480	0.0244	0.0510	0.0189	0.0293	0.0191	0.0241
	Newton							
	A	−1.5569	0.0028	1.5584	0.5226	0.0030	−0.5263	−1.0463
	B	0.0442	0.0239	0.0529	0.0252	0.0417	0.0265	0.0316
	C	0.0445	0.0241	0.0471	0.0237	0.0375	0.0237	0.0285
250	EM							
	A	−1.5279	0.0056	1.5297	0.5080	0.0015	−0.5116	−1.0188
	B	0.0186	0.0095	0.0175	0.0100	0.0142	0.0087	0.0097
	C	0.0189	0.0092	0.0179	0.0079	0.0103	0.0067	0.0078
	Newton							
	A	−1.5290	0.0000	1.5300	0.5124	0.0052	−0.5139	−1.0269
	B	0.0181	0.0094	0.0188	0.0102	0.0136	0.0081	0.0094
	C	0.0179	0.0091	0.0170	0.0103	0.0136	0.0086	0.0093
500	EM							
	A	−1.5080	0.0019	1.5108	0.5016	−0.0003	−0.5028	−1.0039
	B	0.0082	0.0043	0.0084	0.0049	0.0061	0.0047	0.0044
	C	0.0088	0.0045	0.0089	0.0038	0.0048	0.0037	0.0038
	Newton							
	A	−1.5181	0.0008	1.5113	0.5091	−0.0023	−0.5011	−1.0136
	B	0.0086	0.0044	0.0090	0.0054	0.0067	0.0047	0.0048
	C	0.0084	0.0044	0.0084	0.0050	0.0063	0.0048	0.0046
750	EM							
	A	−1.5063	−0.0008	1.5054	0.4979	0.0062	−0.5031	−1.0006
	B	0.0054	0.0028	0.0058	0.0033	0.0040	0.0030	0.0035
	C	0.0059	0.0030	0.0058	0.0024	0.0029	0.0022	0.0026
	Newton							
	A	−1.5139	0.0003	1.5078	0.5014	0.0007	−0.5038	−1.0073
	B	0.0059	0.0029	0.0057	0.0029	0.0038	0.0029	0.0033
	C	0.0056	0.0029	0.0055	0.0031	0.0039	0.0029	0.0032

A: The average $\hat{\boldsymbol{\theta}}$ calculated across the 1000 replicates using $\bar{\hat{\boldsymbol{\theta}}} = \frac{1}{1000}\sum_{i=1}^{1000} \hat{\boldsymbol{\theta}}$.
B: The sample variance of $\hat{\boldsymbol{\theta}}$ calculated across the 1000 replicates using $V(\hat{\boldsymbol{\theta}}) = \frac{1}{1000-1}\sum_{i=1}^{1000} (\hat{\boldsymbol{\theta}} - \bar{\hat{\boldsymbol{\theta}}})^2$.
C: The average $V(\hat{\boldsymbol{\theta}})$ calculated across the 1000 replicates using $\overline{V(\boldsymbol{\theta})} = \frac{1}{1000}\sum_{i=1}^{1000} V(\boldsymbol{\theta})$.

Table 3 Covariances of estimated parameters calculated from 1000 replicates for the EM algorithm (upper triangles) and the Newton–Raphson algorithm (lower triangles) in the simulated data analysis (example 1)

Sample size	t_1	t_2	t_3	b_1	b_2	b_3	b_4
100							
t_1		0.0115	−0.0118	−0.0052	−0.0033	0.0092	0.0133
t_2	0.0073		0.0062	0.0017	−0.0037	0.0025	0.0000
t_3	−0.0072	0.0064		0.0094	−0.0036	−0.0052	−0.0154
b_1	−0.0041	0.0024	0.0092		−0.0116	0.0023	−0.0073
b_2	−0.0041	−0.0048	−0.0047	−0.0151		−0.0144	0.0073
b_3	0.0092	0.0031	−0.0039	0.0038	−0.0184		−0.0001
b_4	0.0122	−0.0004	−0.0147	−0.0079	0.0097	−0.0013	
250							
t_1		0.0020	−0.0040	−0.0022	−0.0002	0.0027	0.0049
t_2	0.0024		0.0022	0.0003	−0.0004	0.0001	−0.0004
t_3	−0.0025	0.0027		0.0028	−0.0003	−0.0022	−0.0051
b_1	−0.0021	0.0004	0.0028		−0.0054	0.0018	−0.0027
b_2	−0.0003	−0.0005	−0.0004	−0.0072		−0.0050	0.0014
b_3	0.0027	0.0002	−0.0020	0.0027	−0.0066		0.0012
b_4	0.0047	−0.0004	−0.0050	−0.0030	0.0019	0.0011	
500							
t_1		0.0010	−0.0018	−0.0011	−0.0002	0.0015	0.0024
t_2	0.0013		0.0010	0.0002	−0.0003	0.0003	0.0001
t_3	−0.0011	0.0013		0.0013	−0.0002	−0.0009	−0.0024
b_1	−0.0010	0.0002	0.0013		−0.0025	0.0009	−0.0018
b_2	−0.0003	−0.0004	−0.0003	−0.0033		−0.0026	0.0013
b_3	0.0015	0.0004	−0.0008	0.0013	−0.0035		0.0002
b_4	0.0023	0.0000	−0.0023	−0.0022	0.0018	0.0001	
750							
t_1		0.0000	−0.0012	−0.0008	0.0001	0.0008	0.0016
t_2	0.0008		0.0000	0.0000	0.0001	−0.0001	0.0000
t_3	−0.0007	0.0009		0.0007	0.0001	−0.0008	−0.0015
b_1	−0.0008	0.0000	0.0007		−0.0015	0.0004	−0.0012
b_2	0.0001	0.0002	0.0001	−0.0021		−0.0014	0.0008
b_3	0.0008	−0.0001	−0.0008	0.0006	−0.0019		0.0002
b_4	0.0016	0.0000	−0.0015	−0.0014	0.0011	0.0001	

A covariance between two estimated parameters was calculated from 1000 replicates using $\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j) = \frac{1}{1000-1}\sum_{k=1}^{1000}(\hat{\theta}_i - \bar{\hat{\theta}}_i)(\hat{\theta}_j - \bar{\hat{\theta}}_j)$.

other for the regression coefficients, but the estimated thresholds, for example, t_1, appear to have a larger bias for the Newton method than for the EM algorithm.

The variance of each estimated parameter calculated from the sample of 1000 replicates, $V(\hat{\boldsymbol{\theta}})$, is a good indication of the precision of the estimate. When the sample size was small, say 100, the Newton method tends to have a consistently smaller variance than the EM algorithm, although the difference is barely noticeable. The multivariate model appears to be a good approximation of the univariate model.

Recall that $V(\boldsymbol{\theta})$ is the variance of parameters calculated from the information matrix for each replicate and $\overline{V(\boldsymbol{\theta})}$ is the average of $V(\boldsymbol{\theta})$ calculated across the 1000 replicates. If $\overline{V(\boldsymbol{\theta})}$ is close to $V(\hat{\boldsymbol{\theta}})$, it means that the method for calculating $V(\boldsymbol{\theta})$ is reasonable. When the sample size was small ($n = 100$), the EM algorithm has a smaller $\overline{V(\boldsymbol{\theta})}$ than $V(\hat{\boldsymbol{\theta}})$, but the bias goes away quickly as the sample size increases. The Newton method, however, always behave well, regardless of the sample size.

Table 3 gives the covariances between estimated parameters for both the EM algorithm (upper triangular elements) and the Newton method (lower triangular elements) using $\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j) = \frac{1}{1000-1}\sum_{k=1}^{1000}(\hat{\theta}_i - \bar{\hat{\theta}}_i)(\hat{\theta}_j - \bar{\hat{\theta}}_j)$. Again, when the sample size was small ($n = 100$), the two methods are slightly different for the covariances, but the differences diminish quickly as the sample size increases. Table 4 lists the averages of the covariances between parameters calculated from the 1000 replicates for both the EM algorithm (upper triangular elements) and the Newton method (lower triangular elements) using $\overline{\mathrm{cov}(\theta_i, \theta_j)} = \frac{1}{1000}\sum_{k=1}^{1000}\mathrm{cov}(\theta_i, \theta_j)$. The conclusion of Table 3 also applies to Table 4. Comparing Table 3 with Table 4, we conclude that the methods for calculating $\mathrm{cov}(\theta_i, \theta_j)$ in both the EM algorithm and the Newton method are reasonably good.

Example 2: This example shows the analysis of a real data set. The data were obtained from Koch and Edwards (1988) for a double-blind clinical trial investigating a new treatment for rheumatoid arthritis. In this data set, there were $n = 84$ subjects with different ages who received an active or placebo treatment for their arthritis pain, and the subsequent extent of improvement was recorded as marked, some, or none. The dependent variable was an ordinal categorical observation with three categories (1 = none, 2 = some and 3 = marked). The three explanatory variables were treatment (active or placebo), sex (male or female), and age (recorded as a continuous variable), respectively. The design matrix was $\mathbf{X}_j = [x_{j1}\ x_{j2}\ x_{j3}]$ where the three variables in the vector correspond to treatment, sex and age, respectively. We analyzed this real data set using both the proposed EM

Table 4 Averages of covariances among parameters calculated from 1000 replicates for the EM algorithm (upper triangles) and the Newton–Raphson algorithm (lower triangles) in the simulated data analysis (example 1)

Sample size	t_1	t_2	t_3	b_1	b_2	b_3	b_4
100							
t_1		0.0088	−0.0077	−0.0037	−0.0009	0.0152	0.0131
t_2	0.0058		0.0085	0.0038	−0.0066	0.0050	−0.0009
t_3	−0.0067	0.0073		0.0113	−0.0049	−0.0046	−0.0147
b_1	−0.0045	0.0038	0.0086		−0.0167	0.0051	−0.0095
b_2	−0.0011	−0.0058	−0.0058	−0.0174		−0.0230	0.0104
b_3	0.0083	0.0036	−0.0030	0.0058	−0.0202		−0.0008
b_4	0.0137	−0.0012	−0.0169	−0.0075	0.0125	−0.0024	
250							
t_1		0.0022	−0.0024	0.0018	−0.0007	0.0029	0.0045
t_2	0.0028		0.0033	0.0009	−0.0011	0.0004	−0.0008
t_3	−0.0024	0.0028		0.0034	−0.0006	−0.0020	−0.0050
b_1	−0.0015	0.0001	0.0036		−0.0072	0.0027	−0.0035
b_2	−0.0002	−0.0001	−0.0004	−0.0069		−0.0066	0.0022
b_3	0.0020	0.0002	−0.0015	0.0025	−0.0062		0.0010
b_4	0.0048	−0.0005	−0.0056	−0.0035	0.0022	0.0006	
500							
t_1		0.0011	−0.0013	−0.0011	0.0001	0.0014	0.0024
t_2	0.0012		0.0015	0.0002	−0.0001	0.0000	−0.0001
t_3	−0.0013	0.0015		0.0016	−0.0004	−0.0007	−0.0024
b_1	−0.0016	0.0005	0.0017		−0.0031	0.0012	−0.0021
b_2	0.0002	−0.0006	−0.0004	−0.0036		−0.0033	0.0017
b_3	0.0014	0.0004	−0.0007	0.0014	−0.0035		0.0001
b_4	0.0027	−0.0002	−0.0026	−0.0025	0.0018	0.0001	
750							
t_1		0.0005	−0.0010	−0.0009	0.0000	0.0009	0.0017
t_2	0.0008		0.0008	0.0000	0.0001	0.0001	0.0000
t_3	−0.0011	0.0008		0.0009	0.0002	−0.0008	−0.0016
b_1	−0.0009	−0.0001	0.0008		−0.0021	0.0004	−0.0016
b_2	0.0000	0.0001	0.0001	−0.0020		−0.0018	0.0012
b_3	0.0009	0.0001	−0.0008	0.0007	−0.0019		0.0002
b_4	0.0017	0.0001	−0.0015	−0.0014	0.0010	0.0002	

The average of a covariance was calculated using $\overline{\mathrm{cov}(\theta_i, \theta_j)} = \frac{1}{1000}\sum_{k=1}^{1000} \mathrm{cov}(\theta_i, \theta_j)$.

Table 5 Estimated parameters and their variances for the EM algorithm and the Newton–Raphson algorithm in the real data analysis (example 2)

Method	*Statistic*	t_1	t_2	b_1	b_2	b_3
EM	$\hat{\boldsymbol{\theta}}$	2.3189	2.8528	−1.0600	−0.7199	−0.0243
	Var($\boldsymbol{\theta}$)	0.2776	0.3030	0.0487	0.0598	0.0001
Newton	$\hat{\boldsymbol{\theta}}$	2.2910	2.8257	−1.0689	−0.7429	−0.0232
	Var($\boldsymbol{\theta}$)	0.4246	0.4473	0.0789	0.0997	0.0001

algorithm and the Newton method implemented in the PROC LOGISTIC procedure of SAS.

Table 5 gives the estimated parameters and their variances and Table 6 give the covariances of the estimated parameters for both methods. The estimated parameters are very similar for the two methods. The variances of parameter estimates are different for the two methods. The EM algorithm produced smaller variances than the Newton method, implying that the sample size ($n = 84$) was not sufficiently large for the EM to provide accurate estimates for the variances, although the estimates of parameters are remarkably close to those of the Newton method.

Example 3: This example shows the analysis of a simulated BC population for interval mapping of quantitative trait loci. This simulation serves as a working example to demonstrate the method for QTL mapping. We assumed that the liability has a zero mean and a unit residual variance. A single QTL was placed at position 25 cM of a 100 cM long chromosome covered by

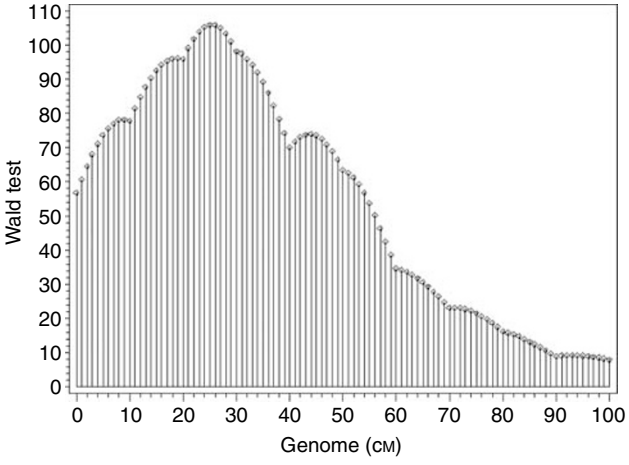

Figure 1 Wald test statistic profile for the simulated chromosome. The true position of the QTL was 25 cM.

Table 6 Covariances of estimated parameters for the EM algorithm (upper traingles) and the Newton–Raphson algorithm (lower triangles) obtained from the real data analysis (example 2)

	t_1	t_2	b_1	b_2	b_3
t_1		0.2669	−0.0305	−0.0477	−0.0039
t_2	0.4274		−0.0353	−0.0508	−0.0040
b_1	−0.0490	−0.0536		0.0117	0.0000
b_2	−0.0750	−0.0779	0.0194		−0.0000
b_3	−0.0060	−0.0061	−0.0001	−0.0001	

Table 7 Estimated parameters (column 2) and their covariance matrix (columns 3–7, upper triangular elements) at position 26 cM of the simulated chromosome (example 3)

	Estimate	t_1	t_2	t_3	t_4	b
t_1	−1.560049	0.01303	0.00044	−0.00023	−0.00076	−0.00160
t_2	−0.65383		0.00668	−0.00010	−0.00035	−0.00075
t_3	0.56573			0.00646	0.00019	0.00040
t_4	1.46154				0.01218	−0.00000
b	0.53811					0.00273

11 evenly placed markers. The effect of the QTL was $a=0.50$, which explained $h^2=a^2/(a^2+1)=20\%$ of the liability variance (see Xu *et al* (2005b) for the definition of h^2). We simulated five ordered categories ($C=5$) with four threshold values. The four thresholds were chosen by trial and error so that the expected frequencies of the five categories occurring in the BC population had a ratio of 1:2:4:2:1. The threshold values generating this ratio were

$$\mathbf{t}=[\,-1.4394 \quad -0.5932 \quad 0.5932 \quad 1.4394\,]^{\mathrm{T}}$$

The population size was $n=300$. Only one sample was generated and analyzed as a working example to demonstrate the method.

The chromosome was scanned from one end to the other with a one cM increment. Figure 1 shows the Wald test statistic profile across the genome. The peak of the test statistic profile occurs at position 26 cM (the true position was 25 cM). The estimated QTL parameters at position 26 cM are given in Table 7, which are quite close to the true values. Table 7 also gives the covariance matrix.

Discussion

The log likelihood function given in Equation (6) is the complete-data log likelihood function, which is only used to derive the EM algorithm. The actual observed log likelihood function that can be used for deriving the likelihood ratio test statistic must have the latent variable $\mathbf{y}$ integrated out. Such an observed log likelihood function has the following form,

$$L_{\mathrm{o}}(\boldsymbol{\theta})=\sum_{j=1}^{n}\sum_{k=1}^{C}\left[w_{jk}\log_e\Phi(\mathbf{x}_j\mathbf{b}-t_k)+(1-w_{jk})\log_e\Phi(t_k-\mathbf{x}_j\mathbf{b})\right]. \tag{22}$$

As we have proposed using the Wald test statistic for testing hypotheses, this observed log likelihood function is not required. For people who prefer the likelihood ratio test, Equation (22) must be used. The focus of this paper has been to derive the EM algorithm and the information matrix under the EM method. As a result, hypothesis tests were only briefly mentioned in the manuscript and the actual tests were carried out in neither the simulation experiment nor the real data analysis.

Some advantages of the EM algorithm over the Newton–Raphson algorithm were discussed in the introduction. One of them is that the EM formulae are more transparent and intuitive to biologists who have little knowledge in advanced statistics except some basic background in linear modeling. The derivations of the EM algorithm and the variance–covariance matrix of the parameters are demanding to some extent, but the final results are extremely simple. The estimated parameters and the variance–covariance matrix of the parameters have the following expressions,

$$\begin{cases}\hat{\boldsymbol{\theta}}=\left[\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\right]^{-1}\left[\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}E(\mathbf{y}_j)\right]\\ V(\boldsymbol{\theta})=\left\{\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\left[\mathbf{I}-V(\mathbf{y}_j)\right]\mathbf{Z}_j\right\}^{-1}\end{cases} \tag{23}$$

People will immediately recognize the similarity between this set of equations and those commonly seen in the usual linear model analysis. If $\mathbf{y}_j$ were observed variables, as in the usual linear regression analysis, then $E(\mathbf{y}_j)=\mathbf{y}_j$ and $V(\mathbf{y}_j)=\mathbf{0}$ would hold. The above equations would be

$$\begin{cases}\hat{\boldsymbol{\theta}}=\left[\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\right]^{-1}\left[\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{y}_j\right]\\ V(\boldsymbol{\theta})=\left[\sum_{j=1}^{n}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\right]^{-1}\end{cases} \tag{24}$$

which are exactly the least square estimates of the parameters and the variance–covariance matrix of the estimates in the usual regression analysis. The EM algorithm only requires substitutions of $\mathbf{y}_j$ by $E(\mathbf{y}_j)$ and $\mathbf{I}$ by $\mathbf{I}-V(\mathbf{y}_j)$, where $E(\mathbf{y}_j)$ and $V(\mathbf{y}_j)$ are the means and variances of truncated normal variables (Cohen, 1991). This is a very desirable property of the proposed EM algorithm.

The exact form of the variance–covariance matrix, $V(\mathbf{y}_j)$, is still unknown. We ignored the covariance elements (assumed to be zero) of the matrix and used a diagonal approximation. This approximation has caused biased (downward) estimates for the variances of the regression coefficients when the sample size was small. As a result, the Wald test statistic is biased upward. Therefore, the test statistic under the null model may not follow the assumed χ^2-distribution. In QTL mapping, however, the exact form of the distribution for the test statistic is not important because the critical value of the test statistic used to declare statistical significance is often drawn from a permutation test (Churchill and Doerge, 1994). Therefore, the slightly biased Wald test will not alter the conclusion of QTL mapping relative to an unbiased test.

Developing a Bayesian method for ordinal traits is straightforward. If an uninformative prior is assigned to the parameters, the conditional posterior distribution

of $\boldsymbol{\theta}$ given $\mathbf{y}$ is multivariate normal with mean $\hat{\boldsymbol{\theta}}$ and variance–covariance matrix $V(\boldsymbol{\theta})$. Given $\boldsymbol{\theta}$ sampled from the normal distribution, $\boldsymbol{y}_{jk}$ is a truncated normal variable with mean $-t_k + \mathbf{x}_j\mathbf{b}$ and variance one. Standard algorithms for sampling a truncated normal variable are available (Devroye, 1986). Given the fact that the Bayesian method implemented via the MCMC is so simple, how do we justify the need of an EM algorithm? The reason is twofold, the cost effectiveness and the prevention of MCMC error.

Acknowledgements

The research was supported by the National Institute of Health Grant R01-GM55321 and the National Science Foundation Grant DBI-0345205 to SX.

References

Aitchison J, Silvery SD (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika* **44**: 131–140.

Albert JH, Chib S (1993). Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc* **88**: 669–679.

Ashford JR (1959). An approach to the analysis of data for semi-quantal responses in biology response. *Biometrics* **15**: 573–581.

Caranta C, Plieger S, Lefebvre V, Daubeze AM, Thabuis A, Palloix A (2002). QTLs involved in the restriction of cucumber mosaic virus (CMV) long-distance movement in pepper. *Theor Appl Genet* **104**: 586–591.

Cohen AC (1991). *Truncated and Censored Samples*. Marcel Dekker Inc.: New York.

Churchill GA, Doerge RW (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.

Devroye L (1986). *Non-Uniform Random Variable Generation*. Springer-Verlag: New York.

Fahrmeir L, Tutz G (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Spring-Verlag Inc.: New York.

Hackett CA, Weller JI (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**: 1252–1263.

Haley CS, Knott SA (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.

Koch G, Edwards S (1988). Clinical efficiency trials with categorical data. In: Peace KE (ed) *Biopharmaceutical Statistics for Drug Development*. Marcel Dekker: New York. pp 403–451.

Lander ES, Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.

Louis TA (1982). Finding the observed information matrix when using the EM algorithm. *J Roy Statist Soc B* **44**: 226–233.

Lynch M, Walsh B (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates: Sunderland, MA.

Manzanares-Dauleux MJ, Delourme R, Baron F, Thomas G (2000). Mapping of one major gene and of QTLs involved in resistance to clubroot in Brassica napus. *Theor Appl Genet* **101**: 885–891.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman and Hall: New York.

McIntyre LM, Coffman CJ, Doerge RW (2001). Detecting and localization of a single binary trait locus in experimental populations. *Genet Res* **78**: 79–92.

Meng XL, Rubin DB (1993). Maximum likelihood estimation via the ECM algorithm. *Biometrika* **80**: 267–278.

Rao S, Xu S (1998). Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity* **81**: 214–224.

SAS Institute (1999a). *SAS/IML User's Guide, Version 8*. SAS Institute Inc.: Cary, North Carolina.

SAS Institute (1999b). *SAS/SATA User's Guide, Version 8*. SAS Institute Inc.: Cary, North Carolina.

Sorensen DA, Andersen S, Gianola D, Korsgaard I (1995). Bayesian inference in threshold models using Gibbs sampling. *Genet Sel Evol* **27**: 229–249.

Xu C, Li Z, Xu S (2005a). Joint mapping of quantitative trait loci for multiple binary characters. *Genetics* **169**: 1045–1059.

Xu C, Zhang YM, Xu S (2005b). An EM algorithm for mapping quantitative resistance loci. *Heredity* **94**: 119–128.

Xu S, Yi N, Burke D, Galecki A, Miller RA (2003). An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. *Genet Res* **82**: 127–138.

Zou J, Pan XB, Chen ZX, Xu JY, Lu JF, Zhai WX *et al* (2000). Mapping quantitative trait loci controlling sheath blight resistance in two rice cultivars. *Theor Appl Genet* **101**: 569–573.