Nonparametric Modeling of Longitudinal Covariance Structure in Functional Mapping of Quantitative Trait Loci

John Stephen Yap[1], **Jianqing Fan**[2] **and Rongling Wu**[1]

[1]Department of Statistics, University of Florida, Gainesville, FL 32611 USA and

[2]Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA

Running Head: Nonparametric Covariance Estimation in Functional Mapping

Key Words: Functional Mapping, Quantitative Trait Loci, Covariance Estimation, Longitudinal Data, Multivariate Normal Mixture

Author for correspondence:

Rongling Wu

Department of Statistics

University of Florida

Gainesville, FL 32611

Phone: (352)392-3806

FAX: (352)392-8555

E-mail: rwu@stat.ufl.edu

Abstract: Estimation of the covariance structure of longitudinal processes is a fundamental prerequisite for the practical deployment of functional mapping designed to study the genetic regulation and network of quantitative variation in dynamic complex traits. We present a nonparametric approach for estimating the covariance structure of a quantitative trait measured repeatedly at a series of time points. Specifically, we adopt Huang et al.'s (2006a) approach of invoking the modified Cholesky decomposition and converting the problem into modeling a sequence of regressions of responses. A regularized covariance estimator is obtained using a normal penalized likelihood with an L_2 penalty. This approach, embedded within a mixture likelihood framework, leads to enhanced accuracy, precision and flexibility of functional mapping while preserving its biological relevance. Simulation studies are performed to reveal the statistical properties and advantages of the proposed method. A real example from a mouse genome project is analyzed to illustrate the utilization of the methodology. The new method will provide a useful tool for genome-wide scanning for the existence and distribution of quantitative trait loci underlying a dynamic trait important to agriculture, biology and health sciences.

1 INTRODUCTION

The past two decades have witnessed extensive growth in an effort to map quantitative trait loci (QTLs) in a variety of organisms using statistical methodologies. A QTL refers to a gene or a region of chromosome that is associated with a quantitative trait, such as height, weight, or body mass (Lander & Botstein, 1989; Zeng, 1994; Kao et al., 1999; Lynch & Walsh, 1998; Wu et al., 2007). However, most mapping existing strategies, such as simple, composite, and multiple interval mapping, can only make use of phenotypic measurements at a single time point to estimate the genetic effects of QTLs. While many traits undergo developmental or dynamic changes in time course, these strategies fall short in capturing the temporal pattern of QTL expression. Many attempts to model this type of phenomenon are hindered by complexity in structure and intensive computation. Fortunately, a novel approach, called functional mapping by Ma et al. (2002), provides a useful framework for genetic mapping through mean and covariance modeling of multi- or longitudinal traits. The mean is typically modeled using a biologically relevant parametric function, such as a logistic curve for growth data (Bertalanffy, 1957; West et al., 2001), and the covariance is assumed to follow an AR(1) structure – a common choice in longitudinal data modeling (Diggle et al., 2002). The EM algorithm is used to estimate the model parameters. Functional mapping has the advantage of fully capturing the temporal change of effects of a QTL on an organism's trait. Because it requires a small number of model parameters to estimate, it is computationally efficient and can be used on data that have limited sample sizes. Functional mapping has shown potential as a powerful statistical method in QTL mapping. It has been used as a modeling tool in a number of areas such as allometric scaling (Wu et al., 2002; Long et al., 2006), thermal reaction norm (Yap et al., 2007), HIV-1 dynamics (Wang & Wu, 2004), tumor progression (Li et al., 2006), biological clock (Liu et al., 2007), and drug response (Lin et al., 2007).

In this paper, we investigate the covariance structure of functional mapping. The covariance is assumed to be identical among different genotypes or segregating groups of a QTL. The assumption of an AR(1) structure in functional mapping, like many longitudinal models, is more of a convenience issue rather than a meaningful approximation. The AR(1) has a simple structure, with only two parameters, and its inverse and determinant have closed forms. This makes computation easier and faster. Furthermore, the EM algorithm formulas for all model parameters at the M-step are easily derived. However, an AR(1) model assumes the data has variance and covariance stationarity. Approximate variance stationarity can usually be achieved by making use of the so-called transform-both-sides (Carroll & Ruppert, 1984) method which does an optimal power transformation of the data (Wu et al., 2004b). The AR(1) model can then be used on the transformed data. But covariance stationarity will still be a problem. Another parametric approach is by using structured antedependence models (SADs) (Zimmerman & Núñez-Antón, 2001; Zhao et al., 2005) which can model both nonstationary variance and correlation functions. Zhao et al. (2005) incorporated SAD in functional mapping and recommended using it in conjunction with an AR(1)-structured model.

The problem with assuming a parametric structure for the covariance matrix in likelihood-based models is that the underlying structure can be significantly different which can lead to considerable bias in parameter estimates. An alternative is to model the covariance matrix nonparametrically. We adopt the method proposed by Huang et al. (2006a) in estimating longitudinal covariance matrices. Their approach is based on the modified Cholesky decomposition (Newton, 1988) wherein the positive-definite covariance matrix Σ of a zero-mean random longitudinal vector $\mathbf{y} = (y_1, ..., y_m)'$, can be uniquely diagonalized as

$$T\mathbf{\Sigma}T' = D, \tag{1}$$

where T is a lower triangular matrix with ones in the diagonal, D is a diagonal matrix, and $'$ denotes matrix transpose. This diagonalization allows modeling of T and D instead of $\mathbf{\Sigma}$ directly. That is, if we can find estimates $\hat{T}$ and $\hat{D}$ of T and D, respectively, then an estimator of $\mathbf{\Sigma}$ is $\hat{\mathbf{\Sigma}} = \hat{T}^{-1}\hat{D}(\hat{T}^{-1})'$ which is positive-definite. It is possible to model T and D because their nonredundant entries have statistical interpretation (Pourahmadi, 1999): the subdiagonal entries of T are the regressions coefficients when each y_t $(t = 2, ..., m)$ is regressed on its predecessors $y_{t-1}, ..., y_1$ and the entries of D are the corresponding prediction error variances. More precisely, $y_1 = \epsilon_1$ and for $t = 2, ..., m$,

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_j + \epsilon_t \tag{2}$$

where $-\phi_{tj}$ is the (t, j)th entry (for $j < t$) of T, and σ_t^2 =var(ϵ_t) is the tth diagonal element of D. $\{\phi_{tj}, j = 1, ..., t-1; t = 2, ..., m\}$ and $\{\sigma_t^2, t = 1, ..., m\}$ are referred to as generalized autoregressive parameters (GARPs) and innovation variances (IVs), respectively. This implies that modeling the covariance matrix, through T and D, is equivalent to modeling a sequence of regressions. Therefore, variable selection and ridge regression types of procedures can be employed to shrink the regression coefficients to produce a regularized covariance estimator. These techniques are built within a normal penalized likelihood framework by using L_1, SCAD, and L_2 penalties, respectively (Fan and Li, 2001). In this paper, we adopt the L_2 penalty approach and propose an extension of this method to covariance estimation in the mixture likelihood framework of functional mapping. Such an extension is possible by capitalizing on the posterior probability representation of the mixture log-likelihood used in the implementation of the EM algorithm, as will be seen in Section 3. Estimation is then carried out by using the ECM algorithm (Meng & Rubin, 1993) with two CM-steps.

This paper will be organized in the following way. In Section 2, we briefly describe functional mapping. In Section 3, we discuss the nonparametric procedure by Huang et al. (2006a) and describe how it can be integrated into functional mapping. Sections 4 and 5 are devoted to simulation results and analysis of a real data, respectively. Section 6 concludes with a discussion.

2 FUNCTIONAL MAPPING

2.1 Model Formulation

Suppose there is a mapping population of n individuals. Each individual is typed for a panel of molecular markers used to construct a genetic linkage map for the genome. The genetic and statistical principles for linkage analysis and map construction with molecular markers were given in Wu et al. (2007). The mapping population is measured for a phenotypic trait at m time points, with a phenotypic observation vector for individual i expressed as $\mathbf{y}_i = (y_{i1}, ..., y_{im})'$. Assume that the trait is controlled by a set of QTLs that form a total of J genotypes. Under the assumption of a multivariate normal density, the phenotype of individual i that carries a QTL genotype k ($k = 1, ..., J$) is expressed as

$$f_k(\mathbf{y}_i) = (2\pi)^{-m/2}|\mathbf{\Sigma}|^{-1/2}\exp\{-(\mathbf{y}_i - \mathbf{g}_k)'\Sigma^{-1}(\mathbf{y}_i - \mathbf{g}_k)/2\}, \tag{3}$$

where the mean genotype value $\mathbf{g}_k$ is modeled by a logistic curve

$$\mathbf{g}_k = [g_k(t)]_{m\times 1} = \left[\frac{a_k}{1 + b_k e^{-r_k t}}\right]_{m\times 1} \tag{4}$$

and $\mathbf{\Sigma}$ is modeled accordingly, such as by an AR(1), SAD, etc.

The likelihood function can be represented by a multivariate mixture model

$$L(\mathbf{\Omega}) = \prod_{i=1}^{n}\left[\sum_{k=1}^{J} p_{ik} f_k(\mathbf{y}_i)\right] \tag{5}$$

where $\mathbf{\Omega}$ is the parameter vector which we will specify shortly, and p_{ij} is the conditional probability of a QTL genotype given the genotypes of flanking markers with $\sum_{k=1}^{J} p_{ik} = 1$. The conditional probability is expressed in terms of the recombination fraction between a putative QTL and the flanking markers that bracket the QTL. Its value is known if the position of a QTL between the two flanking markers is given. In practical computations, a QTL is searched at every 1 or 2 centi-Morgans (cM) on each marker interval throughout a linkage map so that p_{ij} is known beforehand. Thus, $\mathbf{\Omega}$ consists of the mean parameters $\mathbf{\Omega}_\mu = \{a_k, b_k, r_k\}_{k=1}^{J}$ plus the parameters for $\mathbf{\Sigma}$, $\mathbf{\Omega}_{\mathbf{\Sigma}}$. That is, $\mathbf{\Omega} = (\mathbf{\Omega}_\mu, \mathbf{\Omega}_{\mathbf{\Sigma}})$. The reader is referred to Wu et al. (2007) for more about QTL interval mapping.

2.2 Parameter Estimation

The log-likelihood function can be written as

$$\log L(\mathbf{\Omega}) = \sum_{i=1}^{n}\log\left[\sum_{k=1}^{J} p_{ik} f_k(\mathbf{y}_i)\right]. \tag{6}$$

Taking derivatives on equation (6) yields

$$\frac{\partial}{\partial\theta}\log L(\mathbf{\Omega}) = \sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\frac{\partial}{\partial\theta}\log f_k(\mathbf{y}_i) \tag{7}$$

where

$$P_{ik} = \frac{p_{ik} f_k(\mathbf{y}_i)}{\sum_{k=1}^{J} p_{ik} f_k(\mathbf{y}_i)}$$

is interpreted as the posterior probability that individual i has QTL genotype k and $\theta \in \mathbf{\Omega}$.

Let $\mathbf{P} = \{P_{ik}, k = 1, ..., J; i = 1, ..., n\}$. The maximum likelihood estimates (MLEs) are computed using the EM algorithm (Dempster et al., 1977; Lander & Botstein, 1989; Zeng, 1994; Ma et al., 2002) on the expanded parameter set $\{\mathbf{\Omega}, \mathbf{P}\}$ as follows:

1. Initialize $\mathbf{\Omega}$.

2. E-Step: Update $\mathbf{P}$.

3. M-Step: Conditional on $\mathbf{P}$, solve for $\mathbf{\Omega}$ in

$$\frac{\partial}{\partial \theta} \log L(\mathbf{\Omega}) = 0.$$

4. Repeat steps (2)-(3) until some convergence criterion is met.

The values at convergence are the MLEs of $\mathbf{\Omega}$. Ma et al. (2002) and Yap et al. (2007) provide formulas for updating $\mathbf{\Omega}$ in the case when the mean is modeled by logistic and rational curves, respectively, and $\mathbf{\Sigma}$ has an AR(1) structure in a backcross population.

After obtaining the MLEs, we can formulate a hypothesis about the existence of a QTL affecting genotype mean patterns as

$$\begin{aligned} H_0 &: \quad a_1 = ... = a_J, \quad b_1 = ... = b_J, \quad r_1 = ... = r_J \\ H_1 &: \quad \text{at least one of the inequalities above does not hold,} \end{aligned}$$

where H_0 is the reduced (or null) model so that only a single logistic curve can fit the phenotype data and H_1 is the full (or alternative) model in which case there exist more than one logistic curves that fit the phenotype data due to the existence of a QTL. A number of other important hypotheses can be tested, as outlined in Wu et al. (2004a).

The evidence for the the existence of a QTL can be displayed graphically using the log-likelihood ratio (LR) test statistic

$$\mathrm{LR} = -2\log\left[\frac{L(\tilde{\mathbf{\Omega}})}{L(\hat{\mathbf{\Omega}})}\right]$$

plotted over the entire linkage map, where $\tilde{\mathbf{\Omega}}$ and $\hat{\mathbf{\Omega}}$ denote the MLEs under H_0 and H_1, respectively. The peak of the LR plot, which we shall from hereon refer to as maxLR, would suggest a putative QTL because this corresponds to when H_1 is the mostly likely over H_0. The distribution of LR is difficult to determine. However, a nonparametric method called permutation tests by Doerge and Churchill (1996) can be used to find an approximate distribution and a significance threshold for the existence of a QTL. In permutation tests, the functional mapping model is applied to several random permutations of the phenotype data on the markers and a threshold is determined from the set of maxLR values obtained from each permutation test run. The idea here is to disassociate the markers and phenotypes so that repeated application of the model on permuted data will produce an approximate empirical null distribution.

3 COVARIANCE ESTIMATION

3.1 Modified Cholesky Decomposition and Penalized Likelihood

If

$$\begin{aligned} f_k(\mathbf{y}_i) &= (2\pi)^{-m/2}|\mathbf{\Sigma}|^{-1/2}\exp\left\{-(\mathbf{y}_i-\mathbf{g}_k)'\mathbf{\Sigma}^{-1}(\mathbf{y}_i-\mathbf{g}_k)/2\right\} \\ &= (2\pi)^{-m/2}|\mathbf{\Sigma}|^{-1/2}\exp\left\{-\mathbf{y}_i^{k\prime}\mathbf{\Sigma}^{-1}\mathbf{y}_i{}^k/2\right\} \end{aligned}$$

where $\mathbf{y}_i^k = \mathbf{y}_i - \mathbf{g}_k$, then equation (7) becomes

$$\begin{aligned}\frac{\partial}{\partial\theta}\log L(\mathbf{\Omega}) &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\frac{\partial}{\partial\theta}\left[\log|\mathbf{\Sigma}| + \mathbf{y}_i^{k\prime}\mathbf{\Sigma}^{-1}\mathbf{y}_i^k\right] \\ &= -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\frac{\partial}{\partial\theta}\left[\sum_{t=1}^{m}\log\sigma_t^2 + \sum_{t=1}^{m}\frac{\epsilon_{it}^{k\,2}}{\sigma_t^2}\right]\end{aligned}$$

by equation (1) where $\epsilon_{i1}^k = y_{i1}^k$ and $\epsilon_{it}^k = y_{it}^k - \sum_{j=1}^{t-1}\phi_{tj}y_{ij}^k$ for $t = 2, ..., m$. It is implicitly assumed, therefore, that σ_t^2 =var(ϵ_{it}^k) for $k = 1, ..., J$. Note that if $\epsilon^k = (\epsilon_1^k, ..., \epsilon_m^k)'$ and $\mathbf{y}^k = (y_1^k, ..., y_m^k)'$ then $\epsilon^k = T\mathbf{y}^k$ so that $\text{var}(\epsilon^k) = T\mathbf{\Sigma}T'$=$D$.

For a given tuning parameter $\lambda > 0$, define the *penalized negative log-likelihood* as

$$-2\log L(\mathbf{\Omega}) + \lambda p(\{\phi_{tj}\}) \tag{8}$$

where $p(\{\phi_{tj}\}) = \sum_{t=2}^{m}\sum_{j=1}^{t-1}\phi_{tj}^2$ is the L_2 penalty function. Conditional on $\mathbf{P}$ and $\mathbf{\Omega}_\mu$, minimization of (8) gives the penalized likelihood estimators of T and D and consequently, $\mathbf{\Sigma}$. The case when $\lambda = 0$ gives the maximum likelihood estimator. Other penalty functions can also be used (lam and fan, 2007), but we use the L_2-penalty to facilitate the computation.

3.2 ECM Algorithm

If no structure is imposed on the covariance matrix, it is difficult to find closed form M-step solutions in the EM algorithm for the mean parameters in functional mapping. Hence, estimation of the mean parameters is carried out by using an optimization procedure such as the simplex method (Nelder & Mead, 1965) which can be implemented by a built-in function in Matlab. We partition the parameter space according to mean and covariance parameters ($\mathbf{\Omega}_\mu$ and $\mathbf{\Omega}_\mathbf{\Sigma}$) and then use the ECM algorithm (Meng & Rubin, 1993) with two CM-steps. Our general algorithm is outlined as follows:

1. Initialize $\mathbf{\Omega} = (\mathbf{\Omega}_\mu, \mathbf{\Omega}_\mathbf{\Sigma})$.

2. E-Step: Update $\mathbf{P}$.

3. CM-Steps:

   - Conditional on $\mathbf{P}$ and $\mathbf{\Omega}_\mu$, solve for $\mathbf{\Omega}_\mathbf{\Sigma}$ using equations (11) – (13) (Section 3.3) to get $\mathbf{\Omega}'_\mathbf{\Sigma}$.
   - Conditional on $\mathbf{P}$ and $\mathbf{\Omega}'_\mathbf{\Sigma}$, estimate $\mathbf{\Omega}_\mu$ using an optimization procedure.

4. Repeat steps (2) – (3) until some convergence criterion is met.

3.3 Computing the Penalized Likelihood Estimates

The penalty likelihood, where $\mathbf{P}$ and $\mathbf{\Omega}_\mu$ are given as in the first CM step of the ECM algorithm, can be written as

$$\begin{aligned}
-2\log L(\mathbf{\Omega}) + \lambda p(\{\phi_{tj}\}) &= \sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\left(\sum_{t=1}^{m}\log\sigma_t^2 + \sum_{t=1}^{m}\frac{{\epsilon_{it}^{k}}^2}{\sigma_t^2}\right) + \lambda\sum_{t=2}^{m}\sum_{j=1}^{t-1}\phi_{tj}^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\left(\log\sigma_1^2 + \frac{{\epsilon_{i1}^{k}}^2}{\sigma_1^2}\right) + \\
&\quad \sum_{t=2}^{m}\left[\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\left(\log\sigma_t^2 + \frac{{\epsilon_{it}^{k}}^2}{\sigma_t^2}\right) + \lambda\sum_{j=1}^{t-1}\phi_{tj}^2\right].
\end{aligned}$$

Thus, we need to minimize

$$\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\left(\log\sigma_1^2 + \frac{{\epsilon_{i1}^{k}}^2}{\sigma_1^2}\right) \tag{9}$$

and

$$\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\left(\log\sigma_t^2 + \frac{{\epsilon_{it}^{k}}^2}{\sigma_t^2}\right) + \lambda\sum_{j=1}^{t-1}\phi_{tj}^2 \tag{10}$$

for $t = 2, ..., m$.

The minimizer of (9) is simply

$$\sigma_1^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}{y_{i1}^{k}}^2 \tag{11}$$

For $t = 2, ..., m$, (10) can be minimized by an alternating minimization over σ_t^2 and ϕ_{tj}, $j = 1, ..., t-1$:

- For fixed ϕ_{tj}, $j = 1, ..., t-1$, (10) is minimized with respect to σ_t^2 by

$$\sigma_t^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}(y_{it}^k - \sum_{j=1}^{t-1}\phi_{tj}y_{ij}^k)^2 \tag{12}$$

- Letting $\phi_{t(t)} = (\phi_{t1}, \phi_{t2}, ..., \phi_{t,t-1})'$ and $\mathbf{y}_{i(t)}^k = (y_{i1}^k, y_{i2}^k, ..., y_{i,t-1}^k)'$, minimization of (10) for fixed σ_t^2, leads to the closed form solution

$$\phi_{t(t)} = (H_t + \lambda I_t)^{-1}\mathbf{g}_t \tag{13}$$

  where

$$H_t = \frac{1}{\sigma_t^2}\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\mathbf{y}_{i(t)}^k\mathbf{y}_{i(t)}^k{}', \qquad \mathbf{g}_t = \frac{1}{\sigma_t^2}\sum_{i=1}^{n}\sum_{k=1}^{J} P_{ik}\mathbf{y}_{it}^k\mathbf{y}_{i(t)}^k,$$

  and I_t is a $(t-1) \times (t-1)$ identity matrix.

Note that in formulas (11)–(13), the posterior probabilities, P_{ik}'s, are the weights for the genotype groups, $k = 1, ..., J$.

The preceding calculations were based on the L_2 penalty, $p(\{\phi_{tj}\}) = \sum_{t=2}^{m}\sum_{j=1}^{t-1}\phi_{tj}^2$. If the L_1 penalty, $p(\{\phi_{tj}\}) = \sum_{t=2}^{m}\sum_{j=1}^{t-1}|\phi_{tj}|$, is used instead, a closed form solution like (13) cannot be obtained and an iterative algorithm is needed. This is carried out by using an iterative local quadratic approximation of $\sum_{j=1}^{t-1}|\phi_{tj}|$ (Fan and Li, 2001; Öjelund et al., 2001). The reader is referred to Huang et al.(2006a) for additional details.

3.4 Selection of the Tuning Parameter

The tuning parameter λ is selected using a K-fold cross-validation procedure, where $K = 5$ or 10, but generalized cross-validation can alternatively be used. The criterion is the log-likelihood function (6). The full data set Z is randomly split into K subsets of about the same size. Each subset, say Z^s $(s = 1, ..., K)$, is used to validate the log-likelihood based on the parameters estimated using the data $Z \setminus Z^s$. The value of λ that maximizes the average of all cross-validated log-likelihoods is used to select an estimate for $\mathbf{\Sigma}$.

Note that there really are two sets of tuning parameters in our setting - one under the null model and another under the alternative. However, because the log-likelihood under the null model is constant throughout a marker interval, we shall assume that the corresponding tuning parameter has been estimated accordingly and in the succeeding sections simply refer to the tuning parameters as the ones for the alternative model. This is important for constructing a meaningful and valid test as demonstrated in the generalized likelihood tests by Fan et al. (2001).

4 SIMULATIONS

In this section, the performance of the nonparametric covariance estimator is assessed and compared to an AR(1)-structured estimator. We investigate data generated from both multivariate normal and t-distributions. We begin with the former.

Consider an F_2 population in which there are three different genotypes at a single marker or QTL. Since the purpose of the simulation is to investigate the statistical properties of nonparametric modeling for the covariance structure in functional mapping, we will simulate only one linkage group in which a single QTL for a longitudinal trait is located. The simulated linkage group of length 100 cM contains six equally-spaced markers. A QTL is located between the second and third markers, 12 cM from the second marker. Each phenotype associated with the simulated QTL had $m = 10$ measurements and was sampled from a multivariate normal distribution, using logistic curves as expected mean vectors for three different QTL genotypes and three different types of covariance structure as given below. The curve parameters for three genotypes were $a_1 = 30, a_2 = 28.5, a_3 = 27.5$ for QQ, $b_1 = b_2 = b_3 = 5$ for Qq, and $r_1 = r_2 = r_3 = 0.5$ for qq and the covariance structures were assumed as

(1) $\mathbf{\Sigma}_1 = \mathrm{AR}(1)$ with $\sigma^2 = 3, \rho = 0.6$;

(2) $\mathbf{\Sigma}_2 = \sigma^2\{(1-\rho)\mathbf{I} + \rho\mathbf{1})\}$, with $\sigma^2 = 3$, $\rho = 0.5$, where $\mathbf{1}$ is a matrix of 1's, and $\mathbf{I}$ is the identity matrix (Compound Symmetry);

(3) An unstructured covariance matrix

$$\mathbf{\Sigma}_3 = \begin{pmatrix} 0.72 & 0.39 & 0.45 & 0.48 & 0.50 & 0.53 & 0.60 & 0.64 & 0.68 & 0.68 \\ 0.39 & 1.06 & 1.61 & 1.60 & 1.50 & 1.48 & 1.55 & 1.47 & 1.35 & 1.29 \\ 0.45 & 1.61 & 3.29 & 3.29 & 3.17 & 3.09 & 3.19 & 3.04 & 2.78 & 2.53 \\ 0.48 & 1.60 & 3.29 & 3.98 & 4.07 & 4.01 & 4.17 & 4.18 & 4.00 & 3.69 \\ 0.50 & 1.50 & 3.17 & 4.07 & 4.70 & 4.68 & 4.66 & 4.78 & 4.70 & 4.36 \\ 0.53 & 1.48 & 3.09 & 4.07 & 4.68 & 5.56 & 6.23 & 6.87 & 7.11 & 6.92 \\ 0.60 & 1.55 & 3.19 & 4.17 & 4.66 & 6.23 & 8.59 & 10.16 & 10.80 & 10.70 \\ 0.64 & 1.47 & 3.04 & 4.18 & 4.78 & 6.87 & 10.16 & 12.74 & 13.80 & 13.80 \\ 0.68 & 1.35 & 2.78 & 4.00 & 4.70 & 7.11 & 10.80 & 13.80 & 15.33 & 15.35 \\ 0.68 & 1.29 & 2.53 & 3.69 & 4.36 & 6.92 & 10.70 & 13.80 & 15.35 & 15.77 \end{pmatrix}.$$

$\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were considered previously by Huang et al. (2006a) and $\mathbf{\Sigma}_1$ by Levina et al. (2008). $\mathbf{\Sigma}_3$ has increasing diagonal elements and decreasing long term dependence which is typical of longitudinal growth data. It is based on the sample covariance matrix of a real data set.

Functional mapping was applied to the simulated data, with $n = 100$ and 400 samples, using a logistic model for the mean, and the proposed nonparametric estimator and an AR(1) structured estimator for the covariance matrix. The simulated linkage group was searched at every 4 cM (i.e. $0, 4, 8, ..., 100$) for a total of 26 search points across 5 marker intervals. The estimated model parameters at each point were used to construct an LR plot for the QTL linkage map. For the nonparametric covariance estimator, the LR plot is constructed from parameter estimates obtained out of individual tuning parameters λ_c ($c = 1, ..., 26$), that are separately cross-validated. However, we focused our attention only on those λ's corresponding to the maximum LR at each marker interval. An initial LR plot was constructed using an arbitrary λ_0 ($\lambda_c = \lambda_0$ for all $c = 1, ..., 26$), and the maximum on each marker interval was located. At the point corresponding to each maximum, $\lambda = \hat{\lambda}_d$ ($d = 1, ..., 5$) was selected using 5-fold cross-validation. The final model parameter estimates were based on the $\hat{\lambda}_d$ that produced the maximum LR or maxLR. In Figure 1, the broken line LR plot is the result of our procedure while the solid one is based on individual λ_c's that have each been separately cross-validated. For $n = 400$, these two plots are indistinguishable. The reason for this is that, the cross-validated λ's at each search point within a marker interval are not that different from one another. Thus, using one λ for each marker interval (the one that produces the maximum LR) will not significantly alter the general shape of the LR plot. The two dotted line plots were based on λ_c, for all $c = 1, 2, ..., 26$, set to two different arbitrary values of λ. They all have the same location of the maximizer.

To measure the fit of the estimate $\hat{\mathbf{\Sigma}}_l$ $(l = 1, 2, 3)$ of the true covariance structure $\mathbf{\Sigma}_l$, we used the nonnegative functions

$$\begin{aligned} L_E(\mathbf{\Sigma}_l, \hat{\mathbf{\Sigma}}_l) &= \operatorname{tr}(\mathbf{\Sigma}_l^{-1}\hat{\mathbf{\Sigma}}_l) - \log|\mathbf{\Sigma}_l^{-1}\hat{\mathbf{\Sigma}}_l| - m, \\ L_Q(\mathbf{\Sigma}_l, \hat{\mathbf{\Sigma}}_l) &= \operatorname{tr}(\mathbf{\Sigma}_l^{-1}\hat{\mathbf{\Sigma}}_l - \mathbf{I})^2, \end{aligned}$$

which correspond, respectively, to entropy and quadratic losses. Each of these is 0 when $\hat{\mathbf{\Sigma}}_l = \mathbf{\Sigma}_l$ and large values suggest significant bias. These functions were also used by Wu & Pourahmadi (2003), Huang et al. (2006a & b), and Levina et al. (2008) to assess the performance of covariance estimators.

A hundred of simulation runs were carried out and the averages on all runs of the estimated QTL location, logistic mean parameter estimates, maxLR, entropy and quadratic losses, including the respective Monte carlo standard errors (SE), were recorded. The results are shown in Tables 1 and 2. For $\mathbf{\Sigma}_1$, the AR(1) estimator performs well as expected, but the nonparametric estimator also does a good job. Both provide better precision with increased sample size. The maxLR values are comparable, i.e., 38.52 and 112.03 from Table 1 versus 37.78 and 128.21 from Table 2, respectively, are not too different from each other.

For $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$, the nonparametric estimator performs better than the AR(1) estimator. The AR(1) estimator shows high values for both averaged losses which translates to significantly biased estimates in QTL location and poor mean parameter estimates, particularly for $\mathbf{\Sigma}_3$ at the second and third genotype group. Increased sample size does not help and even makes mean parameter estimates worse in the case of $\mathbf{\Sigma}_3$. Values of maxLR for nonparametric and AR(1) estimators are very different in these cases. But because the averaged losses for the nonparametric estimator are much smaller, we would expect that the corresponding maxLR values must be close to the true ones.

To assess the robustness of our proposed nonparametric estimator, we modeled simulated data from a t-distribution with five degrees of freedom. The results are presented in Tables 3 and 4. The results show that despite inflated average losses, the nonparametric estimator still outperforms the AR(1) estimator. Notice that the quadratic loss is severely inflated because of the fat tails of the t-distribution. It may not be a reliable measure of performance but we present the results here for illustration.

5 DATA ANALYSIS

We study a real mouse data set from an experiment by Vaughn et al. (1999). Briefly, the data consists of an F_2 population of 259 male and 243 female progeny with 96 markers located on a total of 19 chromosomes. The mice were measured for their body mass at 10 weekly intervals starting at age 7 days. Corrections were made for the effects due to dam, litter size at birth, parity, and sex (Cheverud et al., 1996; Kramer et al., 1998).

Functional mapping was first used to analyze this data in Zhao et al. (2004), who investigated QTL × sex interaction. They used a logistic curve to model the genotype means and employed the transform-both-sides (TBS) technique for variance stabilization in order to utilize an AR(1) structure. Their method identified 4 of 19 chromosomes that each had significant QTLs and they concluded that there were sex differences of body mass growth in mice. However, Zhao et al. (2005) applied an SAD covariance structure in functional mapping and found three QTLs. Liu and Wu (2007) likewise analyzed the same data using a Bayesian approach in functional mapping and detected only three significant QTLs.

Here, we applied our proposed nonparametric model in a genome-wide scan for growth QTLs without regard to sex. We scanned the linkage map at intervals of 4 cM. Figure 2 shows the LR plots for all 19 chromosomes. They were obtained using λ's that were cross-validated at each search point. We conducted a permutation test (Doerge and Churchill, 1996; also briefly described in Section 2.2) to identify significant QTLs. For every permutation run, we calculated $\max\text{LR}_e$ for chromosome $e = 1, ..., 19$ using the same general procedure as in the simulations (section 4). In this mouse data set, however, some markers were either missing or not genotyped and we used only the available markers (Table 5). Thus, every marker interval had different sets of available phenotype data. But we believe this did not affect the results because of the large sample size of the available data. We looked at chromosomes 6 and 7 and found this to be the case. Figure 3 shows LR plots based on tuning parameters cross-validated at each search point (solid line) and using the same tuning parameter for each search point as the one corresponding to the maximum LR in each marker interval (broken line; our procedure). The dotted line plots were again based on arbitrary tuning parameters and presented here to illustrate shape consistency. Each permutation run yielded the maximum $\max\text{LR}_e$, for all $e = 1, ..., 19$, or the genome-wide maxLR. The two horizontal lines in Fig. 2 correspond to 95% (broken) and 99% (solid) thresholds based on 100 permutation test runs. There were nine chromosomes with significant QTLs ($1, 4, 6, 7, 9, 10, 11, 14$ and 15) based on the 95% threshold but only seven above the 99% threshold ($1, 4, 6, 7, 10, 11$ and 15). The two chromosomes that did not make the 99% threshold (9 and 14) barely made the 95%. For this mouse data set, we recommend using the 99% threshold because there were only 100 permutation test runs. Zhao et al. (2004) identified QTLs in chromosomes $6, 7, 11$ and 15, and Zhao et al. (2005) and Liu and Wu (2007) found QTLs in chromosomes $6, 7$ and 10. These were all at the 95% threshold. Our findings verified the results of these previous studies that made use of the functional mapping method and even detected more QTLs. Although there is a discrepancy in our results and others, it is inconclusive to say that these additional QTLs that our proposed model detected are nonexistent. In fact, Vaughn et al. (1999) identified 17 QTLs, although most of them are suggestive, using a simple interval mapping.

The estimated genotype mean curves for the detected QTLs are shown in Figure 4. Three genotypes at a QTL have different growth curves, indicating the temporal genetic effects of this QTL on growth processes for mouse body mass. Some QTLs, like those on chromosomes 6, 7 and 10, act in an additive manner because the heterozygote (Qq, broken curves) are intermediate between the two homozygotes (QQ, solid curves and qq, dot curves). Some QTL such as one on chromosome 11 are operational in a dominant way since the heterozygote is very close to one of the homozygotes.

6 DISCUSSION

Covariance estimation is an important aspect in modeling longitudinal data. It is difficult, however, because of a large number of parameters to estimate and the positive-definite constraint. Many longitudinal data models resort to structured covariances which, although positive-definite and computationally favorable due to a reduced number of parameters, are possibly highly biased. However, Pourahmadi (1999, 2000) recognized that a positive-definite estimator can be found if modeling is done through the components of the modified Cholesky decomposition of the covariance matrix which converts the problem into modeling a set of regression equations. Wu & Pourahmadi (2003) and Huang et al. (2006b) proposed banding T, noting that terms in the regression farther away in time are negligible and can therefore be set to zero. Huang et al. (2006a) employed LASSO (Tibshirani, 1996) and ridge regression (Hoerl & Kennard, 1970) techniques through L_1 and L_2 penalties, respectively, in a normal penalized likelihood framework. Lam and Fan (2007) proposed a general penalized likelihood method on the covariance matrix, or precision matrix, or its generalized Cholesky decomposition and showed that the difficulty in estimating a large covariance matrix due to dimensionality increases merely by a logarithmic factor of the dimensionality. They also showed that the biases due to the use of L_1-penalty can be significantly reduced by the SCAD penalty. Using these penalties allows shrinkage in the elements of T, even setting some of them to zero in the case of the L_1 penalty. Levina et al. (2008) proposed using a nested lasso penalty instead. This type of penalty produces a sparse estimator for the inverse of the covariance matrix by adaptively banding each row of T. Their estimator provides better precision when the dimension is large. Smith & Kohn (2002) proposed a Bayesian approach by using hierarchical priors to allow zero elements in T.

In this paper, we adopted Huang's L_2 penalty approach to produce a regularized nonparametric covariance estimator in functional mapping. This penalty works best when the true T matrix has many small elements. Using the L_1 or SCAD penalty gives a better estimator when some of the elements of T are actually zero. However, we believe that the differences in results between using either penalties will not be significant unless the dimension is very large. Nonetheless, the L_1 or SCAD penalty can be easily incorporated into our scheme. We have shown how to integrate Huang's procedure into the mixture likelihood framework of functional mapping. The key was to utilize the posterior probability representation of the derivative of the log-likelihood in (7) and apply an L_2 penalty to the negative log-likelihood. Estimation was then carried out using the ECM algorithm with two CM-steps, based on a partition of the mean and covariance parameters. Our simulations have shown better accuracy and precision in estimates for genotype mean parameters, QTL location, and maxLR values, compared to using an AR(1) covariance structure. The maxLR values are important because the complete LR plot provides the amount of evidence for the existence of a QTL. LR values noticeably change when very different covariance structures are used. This is of course under the assumption of multivariate normal data. In our analysis of the mice data, although there were a few chromosomes that were found to have significant QTLs, chromosomes 6 and 7 seemed to have the largest evidence for QTL existence. The LR plots are also used in permutation tests to find a significance threshold. More precise estimates of the covariance structure means better estimates of the the peak of the LR plot and therefore more reliable permutation tests results.

With regards to the utilization of our proposed model, we suggest a preliminary analysis of the data by checking variance and covariance stationarity. If these latter conditions are satisfied then an AR(1) covariance structure may be appropriate. If covariance stationarity is not an issue then a TBS method coupled with an AR(1) model is applicable. If no stationarity is detected then an SAD or the nonparametric model may be more useful. Although we did not assess the comparative performance of these two models, we think that SAD becomes more computationally intensive if the data exhibits long-term dependence, in which case the nonparametric approach may be more appropriate. The nonparametric method should also be considered if other parametric structures are suspect. It can also be used to validate or suggest a family of parametric models.

A recent paper by Yang et al. (2007) proposed a model called composite functional mapping (Zeng 1994) which is an integration of composite interval mapping and functional mapping. Original functional mapping was based on simple interval mapping which searches for QTLs within one marker interval at a time and ignores potential marker effects beyond the marker interval. Composite functional mapping allows modeling of other markers by using partial regression analysis. This significantly improves the precision of functional mapping in QTL detection. However, composite functional mapping assumes an AR(1) covariance structure. It would be advantageous to incorporate our proposed method into this newly developed approach.

Acknowledgments

We thank Dr. James Cheverud for providing us with his mouse data to test our model. The preparation of this manuscript is partially supported by NSF grant (0540745) to R. Wu and NIH R01-GM072611 for NIGMS and NSF grant DMS-0714554 to J. Fan.

References

Bertalanffy, von L. (1957). Quantitative laws for metabolism and growth. *Quart. Rev. Biol.* **32**, 217–231.

de Boor, C. (2001). *A Practical Guide to Splines* (Revised ed.). Springer New York.

Carrol, R. J. & Rupert, D. (1984). Power transformations when fitting theoretical models to data. *J. Am. Statist. Assoc.* **79**, 321–328.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* B **39**, 1–38.

Diggle, P. J., Heagerty, P., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data.* Oxford University Press, UK.

Doerge, R. W. & Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–1360.

Fan, J., Zhang, C. & Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153–193.

Green, P. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc.* B **52**, 443–452.

Hoerl, A. & Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Huang, J., Liu, N., Pourahmadi, M. & Liu, L. (2006a). Covariance selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.

Huang, J., Liu, L. & Liu, N. (2006b). Estimation of large covariance matrices of longitudinal data with basis function approximations. *J. Comput. Graph. Statist.* **16**, 189–209.

Kao, C.-H., Zeng, Z.-B. & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Lam, C. & Fan, J. (2007). Sparsistency and rates of convergence in large covariance matrices estimation. Manuscript (under review).

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

Levina, E., Rothman, A. & Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Statist.* **2**, 245–263.

Li, H. Y., Kim, B.-R. & Wu, R. L. (2006). Identification of quantitative trait nucleotides that regulate cancer growth: A simulation approach. *J. Theor. Biol.* **242**, 426–439.

Lin, M., Hou, W., Li, H. Y., Johnson, J. A. & Wu, R. L. (2007). Modeling sequence-sequence interactions for drug response. *Bioinformatics* **23**, 1251–1257.

Liu, T., Liu, X. L., Chen, Y. M. & Wu, R. L. (2007). A unifying differential equation model for functional genetic mapping of circadian rhythms. *Theor. Biol. Medical Model.* **4**, 5.

Liu, T., & Wu, R. L. (2007). A general Bayesian framework for functional mapping of dynamic complex traits. *Genetics* (tentatively accepted).

Long, F., Chen, Y. Q., Cheverud, J. M. & Wu, R. L. (2006). Genetic mapping of allometric scaling laws. *Genet. Res.* **87**, 207–216.

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

Ma, C., Casella, G. & Wu, R. L. (2002). Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics* **161**, 1751–1762.

Meng, X-L. & Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267–278.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Comput. J.* **7**, 308–313.

Newton, H. J. (1988). *TIMESLAB: A Time Series Analysis Laboratory.* Wadsworth & Brooks/Cole, Pacific Grove, CA.

Ö, H. Madsen, H., & Thyregod, P. (2001). Calibration with absolute shrinkage. *J. Chemomet.* **15**, 497-509.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425–435.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc.* B **58**, 267–288.

Vaughn, T., Pletscher, S., Peripato, A., King-Ellison, K., Adams, E., Erikson, C. & Cheverud, J. (1999). Mapping of quantitative trait loci for murine growth: A closer look at genetic architecture. *Genet. Res.* **74**, 313–322.

Wang, Z. H. & Wu, R. L. (2004). A statistical model for high-resolution mapping of quantitative trait loci determining human HIV-1 dynamics. *Statist. Med.* **23**, 3033–3051.

West, G. B., Brown, J. H. & Enquist, B. J. (2001). A general model for ontogenetic growth. *Nature* **413**, 628–631.

Wu, R. L., Ma, C.-X. & Casella, G. (2007). *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL*. Springer-Verlag, New York.

Wu, R. L., Ma, C., Lin, M. & Casella, G. (2004a). A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* **166**, 1541–1551.

Wu, R. L., Ma, C., Lin, M., Wang, Z. & Casella, G. (2004b). Functional mapping of quantitative trait loci underlying growth trajectories using a transform-both-sides logistic model. *Biometrics* **60**, 729–738.

Wu, R. L., Ma, C., Littell, R. & Casella, G. (2002). A statistical model for the genetic origin of allometric scaling laws in biology. *J. Theor. Biol.* **217**, 275–287.

Wu, W. B. & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831–844.

Yang, R. Q., Gao, H. J., Wang, X., Zhang, J., Zeng, Z.-B. & Wu, R. L. (2007). A semiparametric model for composite functional mapping of dynamic quantitative traits. *Genetics* **177**, 1859–1870.

Yap, J. S., Wang, C. G. & Wu, R. L. (2007). A simulation approach for functional mapping of quantitative trait loci that regulate thermal performance curves. *PLoS ONE* **2(6)**, e554.

Zeng, Z. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

Zhao, W., Ma, C., Cheverud, J. M. & Wu, R. L. (2004). A unifying statistical model for QTL mapping of genotype × sex interaction for developmental trajectories. *Physiol. Genomics* **19**, 218–227.

Zhao, W., Chen, Y., Casella, G., Cheverud, J. M. & Wu, R. L. (2005). A non-stationary model for functional mapping of complex traits. *Bioinformatics* **21**, 2469–2477.

Zimmerman, D. & Núñez-Antón, V. (2001). Parametric modeling of growth curve data: An overview (with discussions). *Test* **10**, 1–73.

Table 1: The averaged QTL position, mean curve parameters, maximum log-likelihood ratios (maxLR), entropy and quadratic losses and their standard errors (given in parentheses) for three QTL genotypes in an F$_2$ population under different sample sizes (n) based on 100 simulation replicates (Nonparametric Estimator, Normal Data).

		QTL	QTL genotype 1			QTL genotype 2			QTL genotype 3					
Covariance	n	Location	$\hat{a}_1$	$\hat{b}_1$	$\hat{r}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{r}_2$	$\hat{a}_3$	$\hat{b}_3$	$\hat{r}_3$	maxLR	L_E	L_Q
$\mathbf{\Sigma}_1$	100	32.84	30.11	5.04	0.50	28.52	4.97	0.50	27.47	5.06	0.50	38.52	0.53	1.00
		(0.99)	(0.07)	(0.04)	(0.00)	(0.05)	(0.03)	(0.00)	(0.07)	(0.04)	(0.00)	(1.27)	(0.01)	(0.02)
	400	31.52	30.00	4.99	0.50	28.49	5.01	0.50	27.52	4.97	0.50	112.03	0.14	0.28
		(0.28)	(0.03)	(0.01)	(0.00)	(0.02)	(0.01)	(0.00)	(0.03)	(0.02)	(0.00)	(1.80)	(0.00)	(0.01)
$\mathbf{\Sigma}_2$	100	32.56	30.07	4.98	0.50	28.55	4.99	0.50	27.38	5.07	0.51	47.05	0.44	0.83
		(0.76)	(0.06)	(0.03)	(0.00)	(0.04)	(0.02)	(0.00)	(0.06)	(0.04)	(0.00)	(1.32)	(0.01)	(0.02)
	400	31.68	30.04	4.97	0.50	28.48	5.01	0.50	27.54	4.98	0.50	145.83	0.13	0.26
		(0.26)	(0.02)	(0.01)	(0.00)	(0.02)	(0.01)	(0.00)	(0.02)	(0.01)	(0.00)	(2.00)	(0.00)	(0.01)
$\mathbf{\Sigma}_3$	100	33.24	30.07	5.04	0.50	28.59	5.01	0.50	27.66	5.01	0.50	19.57	0.56	1.09
		(2.22)	(0.10)	(0.03)	(0.00)	(0.06)	(0.02)	(0.00)	(0.09)	(0.02)	(0.00)	(0.59)	(0.01)	(0.02)
	400	32.32	29.99	5.00	0.50	28.50	5.00	0.50	27.62	5.01	0.50	38.90	0.14	0.29
		(1.19)	(0.04)	(0.01)	(0.00)	(0.03)	(0.01)	(0.00)	(0.05)	(0.01)	(0.00)	(1.06)	(0.00)	(0.01)
True values:		32	30	5	0.5	28.5	5	0.5	27.5	5	0.5			

Table 2: The averaged QTL position, mean curve parameters, maximum log-likelihood ratios (maxLR), entropy and quadratic losses and their standard errors (given in parentheses) for three QTL genotypes in an F_2 population under different sample sizes (n) based on 100 simulation replicates (AR(1) Estimator, Normal Data).

		QTL	QTL genotype 1			QTL genotype 2			QTL genotype 3					
Covariance	n	Location	$\hat{a}_1$	$\hat{b}_1$	$\hat{r}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{r}_2$	$\hat{a}_3$	$\hat{b}_3$	$\hat{r}_3$	maxLR	L_E	L_Q
$\mathbf{\Sigma}_1$	100	33.24	29.99	5.03	0.50	28.48	4.99	0.50	27.57	5.04	0.50	37.78	0.02	0.04
		(0.77)	(0.06)	(0.04)	(0.00)	(0.05)	(0.03)	(0.00)	(0.07)	(0.05)	(0.00)	(1.09)	(0.00)	(0.00)
	400	31.80	30.01	4.97	0.50	28.50	5.02	0.50	27.51	4.98	0.50	128.21	0.01	0.01
		(0.32)	(0.03)	(0.02)	(0.00)	(0.02)	(0.01)	(0.00)	(0.03)	(0.02)	(0.00)	(1.98)	(0.00)	(0.00)
$\mathbf{\Sigma}_2$	100	35.28	30.36	4.63	0.48	28.54	5.04	0.50	27.12	5.51	0.52	64.68	2.15	6.57
		(1.57)	(0.09)	(0.05)	(0.00)	(0.07)	(0.04)	(0.00)	(0.09)	(0.07)	(0.00)	(2.53)	(0.06)	(0.38)
	400	31.96	30.51	4.62	0.48	28.42	5.08	0.50	27.14	5.35	0.51	193.84	2.66	9.94
		(0.54)	(0.04)	(0.02)	(0.00)	(0.03)	(0.02)	(0.00)	(0.04)	(0.03)	(0.00)	(4.65)	(0.04)	(0.25)
$\mathbf{\Sigma}_3$	100	46.48	30.39	5.33	0.51	28.01	4.99	0.52	27.85	5.20	0.51	112.66	9.64	73.15
		(2.74)	(0.38)	(0.09)	(0.00)	(0.35)	(0.07)	(0.00)	(0.39)	(0.09)	(0.00)	(2.83)	(0.13)	(2.06)
	400	43.64	30.60	5.28	0.51	27.64	4.93	0.52	28.38	5.34	0.50	288.87	10.14	80.36
		(2.64)	(0.30)	(0.06)	(0.00)	(0.34)	(0.07)	(0.00)	(0.33)	(0.08)	(0.00)	(6.09)	(0.07)	(1.12)
True values:		32	30	5	0.5	28.5	5	0.5	27.5	5	0.5			

Table 3: The averaged QTL position, mean curve parameters, maximum log-likelihood ratios (maxLR), entropy and quadratic losses and their standard errors (given in parentheses) for three QTL genotypes in an F_2 population under different sample sizes (n) based on 100 simulation replicates (Nonparametric Estimator, Data from t-distribution).

		QTL	QTL genotype 1			QTL genotype 2			QTL genotype 3				
Covariance	n	Location	$\hat{a}_1$	$\hat{b}_1$	$\hat{r}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{r}_2$	$\hat{a}_3$	$\hat{b}_3$	$\hat{r}_3$	L_E	L_Q
$\mathbf{\Sigma}_1$	100	32.52	30.07	5.02	0.50	28.58	5.02	0.50	27.53	5.07	0.50	2.56	10.51
		(1.34)	(0.08)	(0.04)	(0.00)	(0.06)	(0.04)	(0.00)	(0.09)	(0.06)	(0.00)	(0.12)	(0.75)
	400	32.88	30.03	5.01	0.50	28.46	5.00	0.50	27.59	4.99	0.50	1.84	6.24
		(0.49)	(0.04)	(0.02)	(0.00)	(0.03)	(0.02)	(0.00)	(0.03)	(0.02)	(0.00)	(0.06)	(0.25)
$\mathbf{\Sigma}_2$	100	32.56	30.15	4.94	0.50	28.54	5.02	0.50	27.47	5.09	0.50	2.27	8.81
		(1.08)	(0.07)	(0.03)	(0.00)	(0.05)	(0.03)	(0.00)	(0.08)	(0.04)	(0.00)	(0.11)	(0.66)
	400	32.84	30.06	4.97	0.50	28.48	5.01	0.50	27.53	5.02	0.50	1.78	5.86
		(0.03)	(0.03)	(0.02)	(0.00)	(0.03)	(0.01)	(0.00)	(0.03)	(0.02)	(0.00)	(0.05)	(0.22)
$\mathbf{\Sigma}_3$	100	40.92	29.95	5.03	0.50	28.63	5.00	0.50	27.78	5.05	0.50	2.68	11.82
		(2.76)	(0.13)	(0.03)	(0.00)	(0.09)	(0.02)	(0.00)	(0.14)	(0.04)	(0.00)	(0.14)	(1.26)
	400	33.08	29.95	5.00	0.50	28.56	5.02	0.50	27.51	4.99	0.50	1.90	6.55
		(1.37)	(0.06)	(0.01)	(0.00)	(0.04)	(0.01)	(0.00)	(0.06)	(0.02)	(0.00)	(0.06)	(0.26)
True values:		32	30	5	0.5	28.5	5	0.5	27.5	5	0.5		

Table 4: The averaged QTL position, mean curve parameters, maximum log-likelihood ratios (maxLR), entropy and quadratic losses and their standard errors (given in parentheses) for three QTL genotypes in an F_2 population under different sample sizes (n) based on 100 simulation replicates (AR(1) Estimator, Data from t-distribution).

		QTL	QTL genotype 1			QTL genotype 2			QTL genotype 3				
Covariance	n	Location	$\hat{a}_1$	$\hat{b}_1$	$\hat{r}_1$	$\hat{a}_2$	$\hat{b}_2$	$\hat{r}_2$	$\hat{a}_3$	$\hat{b}_3$	$\hat{r}_3$	L_E	L_Q
$\mathbf{\Sigma}_1$	100	34.00	30.04	5.01	0.50	28.61	5.00	0.50	27.51	5.06	0.51	1.65	5.03
		(1.12)	(0.08)	(0.04)	(0.00)	(0.06)	(0.03)	(0.00)	(0.08)	(0.06)	(0.00)	(0.10)	(0.39)
	400	33.04	29.98	4.99	0.50	28.48	5.01	0.50	27.61	4.98	0.50	1.61	4.75
		(0.40)	(0.03)	(0.02)	(0.00)	(0.03)	(0.02)	(0.00)	(0.03)	(0.02)	(0.00)	(0.07)	(0.28)
$\mathbf{\Sigma}_2$	100	38.92	30.57	4.62	0.48	28.48	5.09	0.50	27.13	5.58	0.52	6.24	35.25
		(1.91)	(0.13)	(0.06)	(0.00)	(0.09)	(0.05)	(0.00)	(0.14)	(0.09)	(0.00)	(0.25)	(2.50)
	400	32.16	30.61	4.55	0.48	28.35	5.13	0.51	27.22	5.30	0.51	7.35	45.86
		(0.48)	(0.05)	(0.02)	(0.00)	(0.04)	(0.02)	(0.00)	(0.05)	(0.03)	(0.00)	(0.17)	(1.75)
$\mathbf{\Sigma}_3$	100	49.12	29.71	5.23	0.58	28.80	5.21	0.51	27.04	5.37	0.53	22.04	301.53
		(2.96)	(0.50)	(0.11)	(0.06)	(0.38)	(0.08)	(0.00)	(0.49)	(0.19)	(0.01)	(0.56)	(14.94)
	400	42.64	30.78	5.38	0.51	28.21	5.08	0.52	27.12	5.05	0.52	24.45	366.54
		(2.39)	(0.38)	(0.09)	(0.00)	(0.35)	(0.08)	(0.00)	(0.36)	(0.08)	(0.00)	(0.49)	(15.65)
True values:		32	30	5	0.5	28.5	5	0.5	27.5	5	0.5		

Table 5: Available markers and phenotype data of a linkage map in an F_2 population of mice (data from Vaughn et al. (1999)).

	Marker Intervals							
Chromosome	1	2	3	4	5	6	7	8
1	378	433	483	467	450	440	466	
2	414	404	453	465	430			
3	477	491	489	476	475			
4	461	475	481	481	491			
5	441	439	449	381	385			
6	467	483	485	481				
7	407	424	459	452	378	372	428	415
8	395	453	472					
9	498	496	498					
10	401	406	481	490	497			
11	431	451	468	464	446			
12	497	489	483	488				
13	450	443	466					
14	443	475	495					
15	491	494	468					
16	498							
17	371	394						
18	487	479	420					
19	445	468	468					

LEGENDS

Figure 1. Log-likelihood ratio (LR) plots based on simulated data under three different covariance structures. The solid line plot is based on cross-validated (CV) tuning parameters at each search point (individual λ's). The broken line plot is based on cross-validated tuning parameters (max λ's) corresponding to the maximum LR in each marker interval. The dotted line plot is based on two different arbitrary tuning parameter values, each assumed at all search points.

Figure 2. The profile of the log-likelihood ratios (LR) between the full model (there is a QTL) and reduced (there is no QTL) model for body mass growth trajectories across the genome in a mouse F_2 population. The genomic position corresponding to the peak of the curve is the optimal likelihood estimate of the QTL localization indicated by vertical broken lines. The ticks on the x-axis indicate the positions of markers on the chromosome. The map distances (in centi-Morgan) between two markers are calculated using the Haldane mapping function. The thresholds for claiming the genome-wide existence of a QTL are shown by horizontal lines.

Figure 3. Log-likelihood ratio (LR) plots for chromosomes 6 and 7 of the mice data. The solid line plot is based on cross-validated (CV) tuning parameters at each search point (individual λ's). The broken line plot is based on cross-validated tuning parameters (max λ's) corresponding to the maximum LR in each marker interval. The dotted line plot is based on two different arbitrary tuning parameter values, each assumed at all search points. Slight

differences between the solid and broken line plots may be due to different sample sizes among marker intervals (see Table 5).

Figure 4. Three growth curves each presenting a genotype at each of seven QTLs detected on mouse chromosomes 1, 4, 6, 7, 10, 11, and 15 for growth trajectories of mice in an F_2 population.

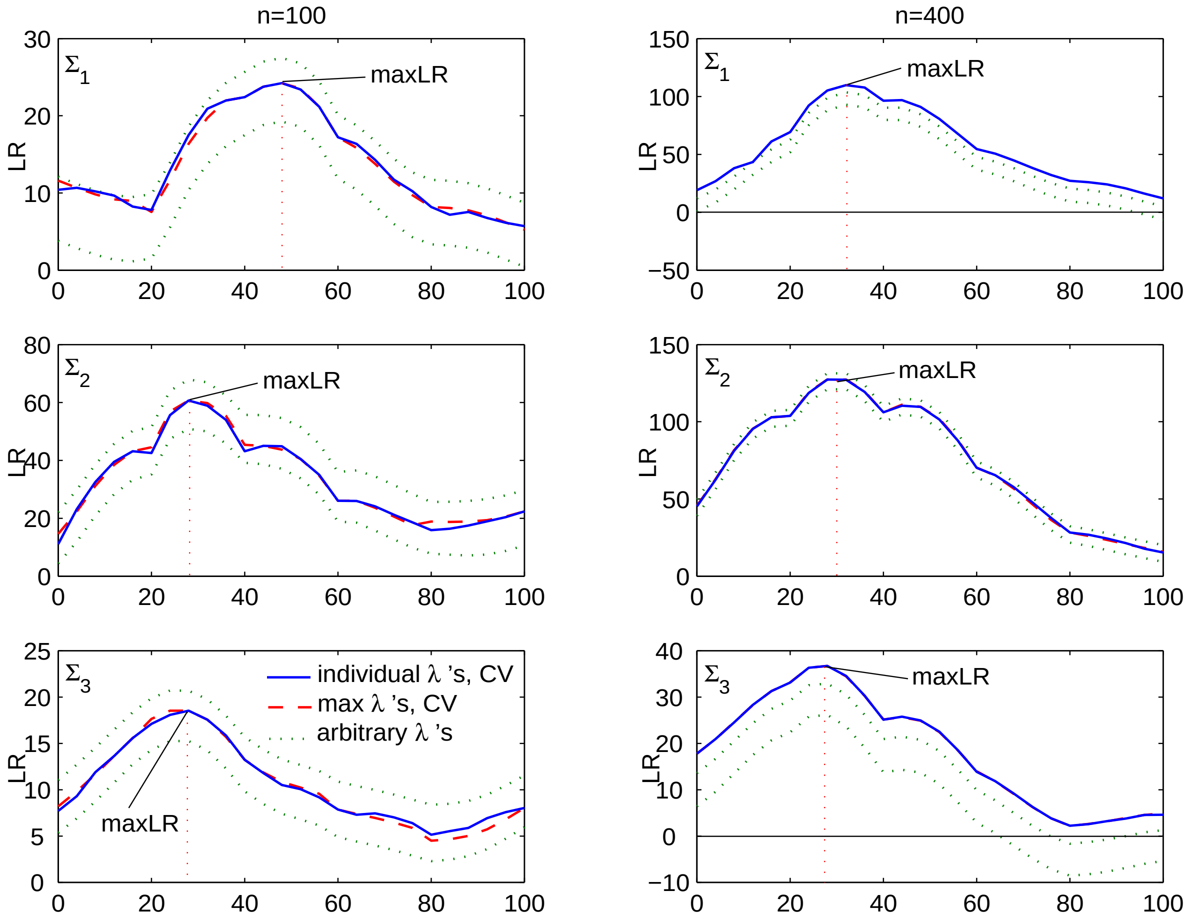
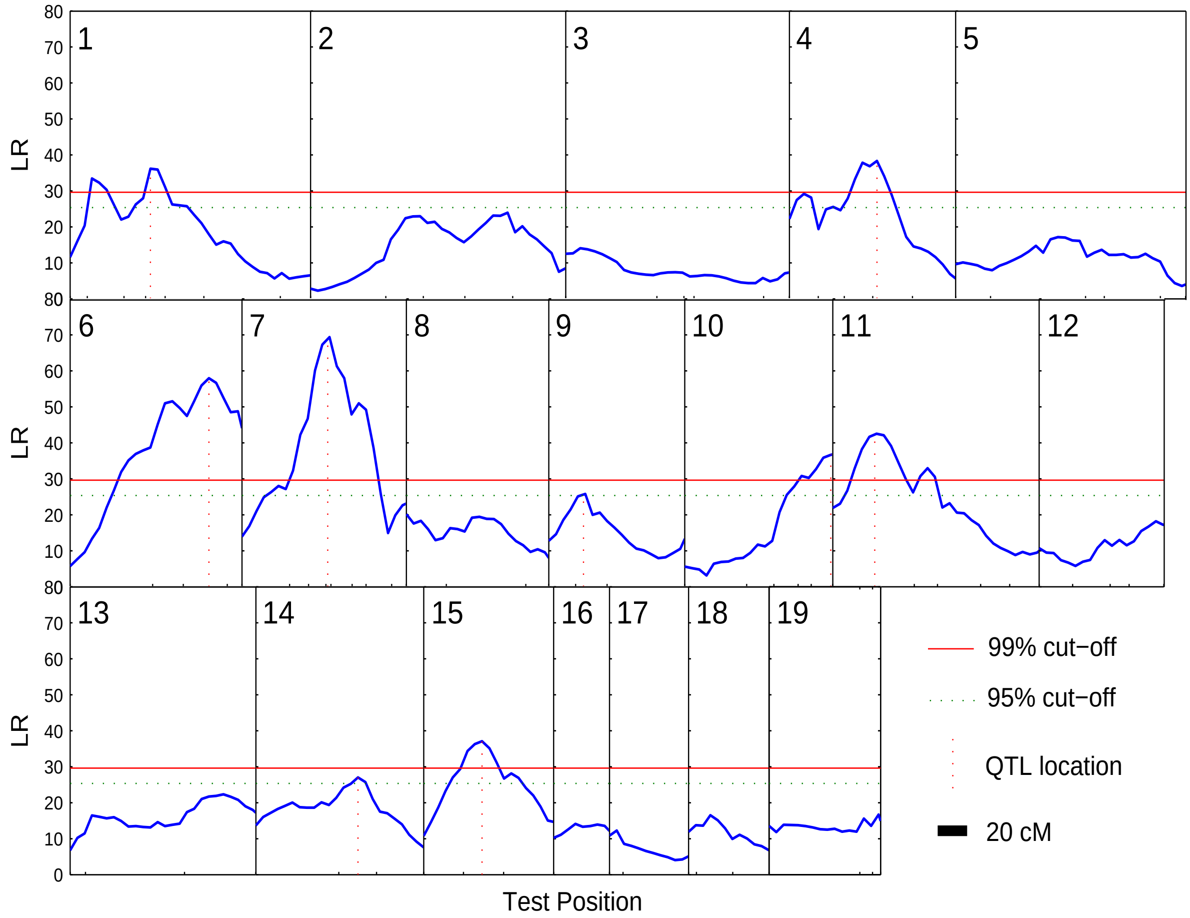

n=100
n=400
Σ_1
Σ_2
Σ_3
LR
maxLR
individual λ 's, CV
max λ 's, CV
arbitrary λ 's
0
20
40
60
80
100
30
20
10
0
150
100
50
0
−50
80
60
40
20
0
150
100
50
0
25
20
15
10
5
0
40
30
20
10
0
−10

FIGURE 1



LR
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
99% cut−off
95% cut−off
QTL location
20 cM
Test Position

FIGURE 2

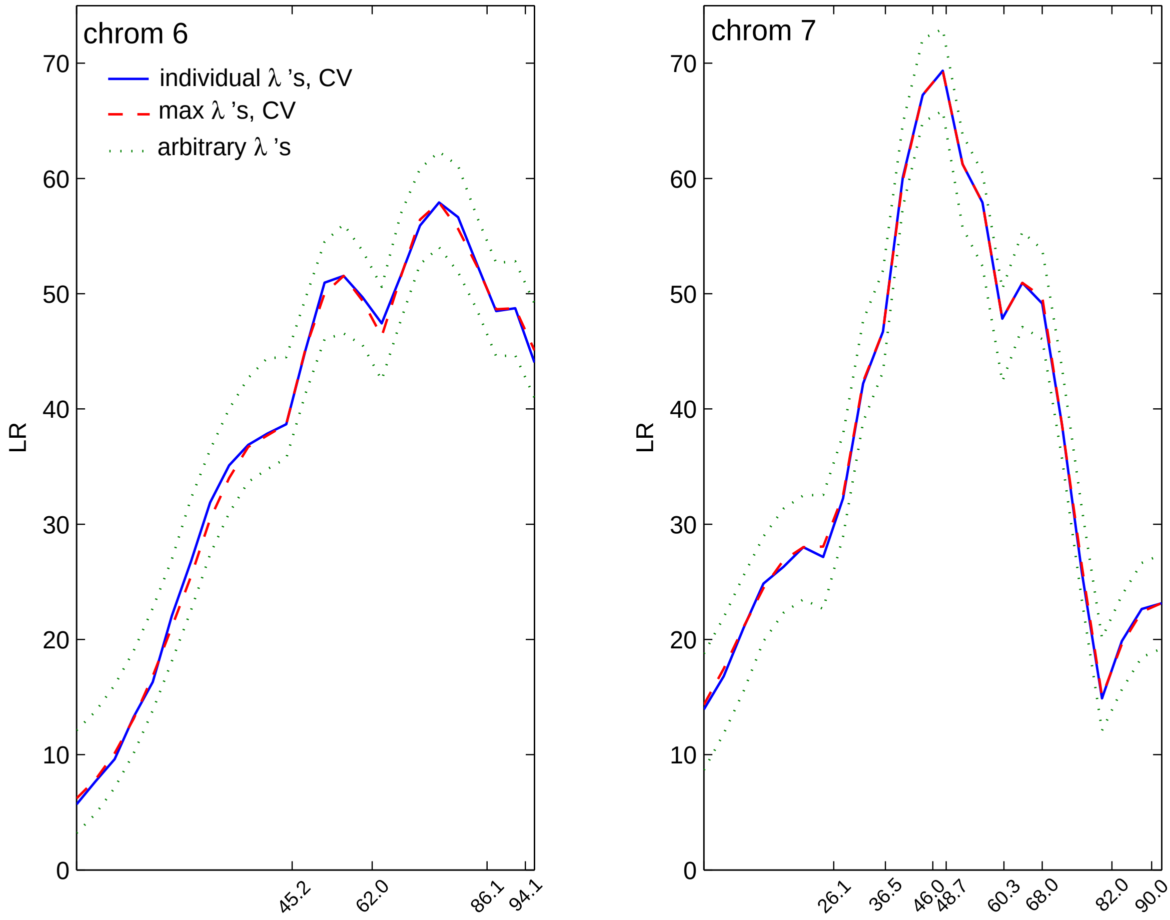

chrom 6
individual λ 's, CV
max λ 's, CV
arbitrary λ 's
LR
0
10
20
30
40
50
60
70
45.2
62.0
86.1
94.1
chrom 7
LR
0
10
20
30
40
50
60
70
26.1
36.5
46.0
48.7
60.3
68.0
82.0
90.0

FIGURE 3

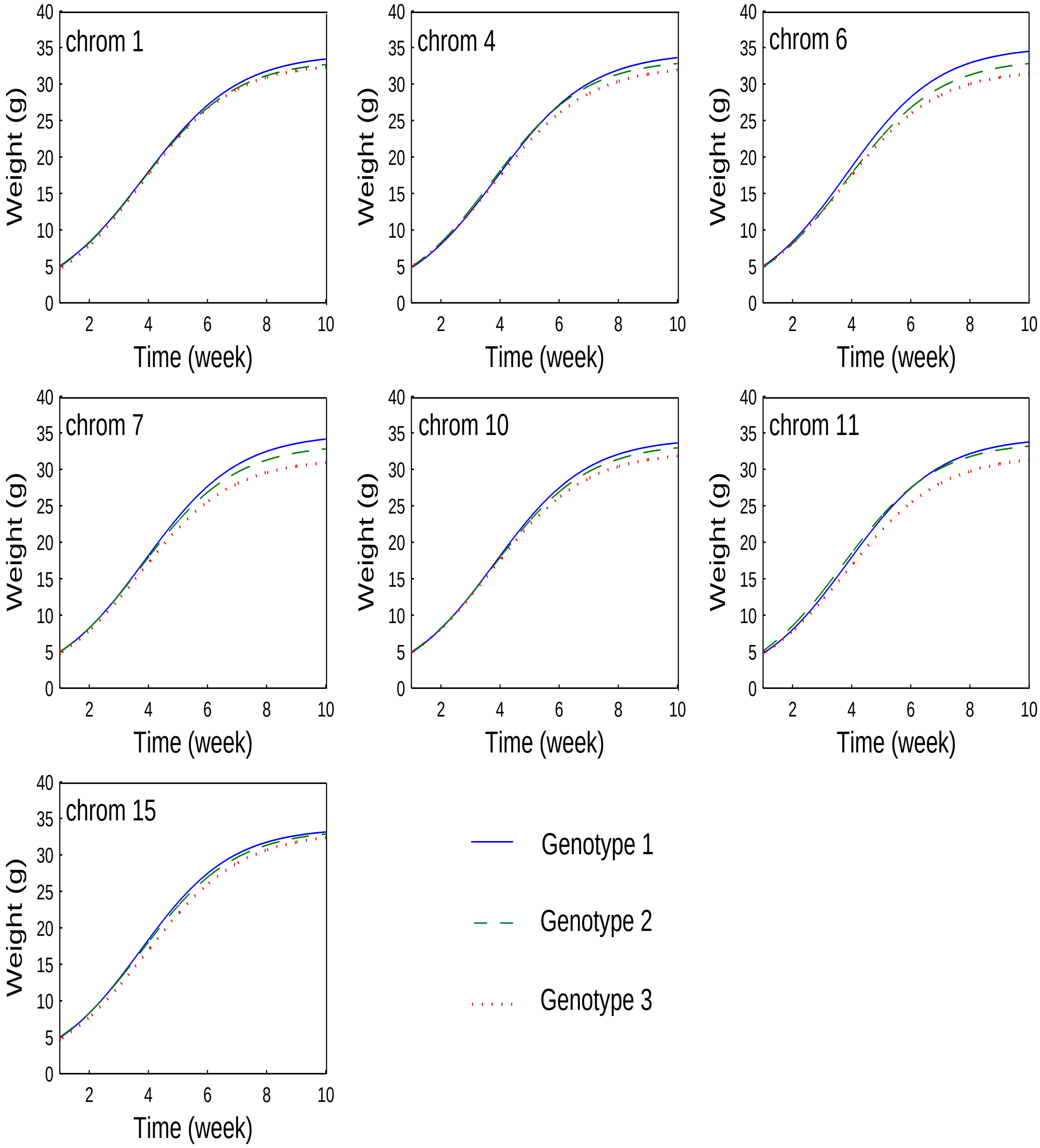

chrom 1
chrom 4
chrom 6
chrom 7
chrom 10
chrom 11
chrom 15
Weight (g)
Time (week)
0
5
10
15
20
25
30
35
40
2
4
6
8
10
Genotype 1
Genotype 2
Genotype 3

FIGURE 4