A FLEXIBLE ESTIMATING EQUATIONS APPROACH FOR
MAPPING FUNCTION-VALUED TRAITS

Hao Xiong*, Evan H Goulding†, Elaine J Carlson†, Laurence H Tecott‡, Charles E McCulloch*, and Śaunak Sen*

June 8, 2011

*Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143.
†Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL 60611.
‡Department of Psychiatry, University of California, San Francisco, CA 94143.
Running head: Flexible mapping of function-valued traits

Key words: QTL, functional mapping, circadian, dynamic traits, estimating equations

Corresponding author:

Šaunak Sen
Department of Epidemiology and Biostatistics
University of California San Francisco, CA 94143-0560
Phone: 415-514-8035
Fax: 415-514-8150
Email: sen@biostat.ucsf.edu
In genetic studies, many interesting traits, including growth curves and skeletal shape, have temporal or spatial structure. They are better treated as curves or function-valued traits. Identification of genetic loci contributing to such traits is facilitated by specialized methods that explicitly address the function-valued nature of the data. Current methods for mapping function-valued traits are mostly likelihood-based, requiring specification of the distribution and error structure. However, such specification is difficult or impractical in many scenarios. We propose a general functional regression approach based on estimating equations that is robust to misspecification of the covariance structure. Estimation is based on a two-step least squares algorithm, which is fast and applicable even when the number of time points exceeds the number of samples. It is also flexible due to a general linear functional model; changing the number of covariates does not necessitate a new set of formulas and programs. In addition, many meaningful extensions are straightforward. For example, we can accommodate incomplete genotype data, and the algorithm can be trivially parallelized. The framework is an attractive alternative to likelihood-based methods when the covariance structure of the data is not known. It provides a good compromise between model simplicity, statistical efficiency, and computational speed. We illustrate our method and its advantages using circadian mouse behavioral data.
Many phenotypes (or traits) in genetic studies are quantitative and can be summarized by a single number; examples include bone mineral density and body weight. Other traits of interest to geneticists such as growth curves (KRAMER ET AL., 1998), circadian rhythms (SHIMOMURA ET AL., 2001), and morphology (LEAMY ET AL., 2008), have spatial or temporal structure and cannot be reduced to a single number. Such traits may be viewed as curves or as function-valued (PLETCHER AND GEYER, 1999). Quantitative genetics of such traits benefit from generalizations tailored to their function-valued nature (KIRKPATRICK AND HECKMAN, 1989). In this paper we present a new framework for identifying regions of the genome, quantitative trait loci (QTL) (RAPP, 2000; BROMAN, 2001), contributing to variation in function-valued traits.

Acknowledging the function-valued nature of a trait has several advantages. We can naturally treat features such as rates of growth or periodic fluctuations. We may also arrive at a more parsimonious representation of the data: a long sequence of correlated observations can often be described by a few values. From an evolutionary perspective, the function-valued nature of the data puts constraints on the patterns of variation possible, and how selection might operate on the trait (KINGSSOLVER ET AL., 2001). For these reasons, application of functional data analysis (FDA) (RAMSAY AND SILVERMAN, 2005) to genetic data has been fruitful. WU AND LIN (2006) called genetic mapping of function-valued traits functional mapping. Function-valued traits have been variously referred to as infinite-dimensional traits, longitudinal traits, repeated measures, functional traits, and dynamic traits (KINGSSOLVER ET AL., 2001).
Most existing methods for mapping function-valued traits use a likelihood framework where the analyst must specify the mean function, the error distribution, and the error structure. When those features can be specified, one can use parametric functional mapping (Ma et al., 2002; Wu et al., 2004; Wu and Lin, 2006). Here the analyst specifies the form of the mean function (say logistic, for modeling growth curves), and the error function (say autoregressive Gaussian errors). If we do not have enough information about the form of the mean function, one may model the mean function nonparametrically using different basis function families: Legendre polynomials (Lin and Wu, 2006), orthogonal polynomials (Yang et al., 2006), B-splines (Yap et al., 2009; Yang et al., 2009), wavelets (Zhao et al., 2007) have all been used in the past. The approach of Yang et al. (2009) and Yap et al. (2009) allow for an unstructured form of the variance-covariance matrix of the errors and use a multivariate Gaussian distribution. However, when the number of measurements per individual exceeds the number of samples and the empirical covariance matrix is thus singular, additional procedures such as wavelet dimension-reduction (Zhao et al., 2007) or regularized estimation of the covariance matrix must be employed. For likelihood based methods, incomplete genotype data is typically handled by the EM algorithm as in Lander and Botstein (1989). Bayesian formulations have been considered by Yang and Xu (2007) and Liu and Wu (2009).

In this report, we present an alternative framework for mapping function-valued traits based on estimating equations. It is especially suited for scenarios where the error structure is unknown or poorly specified, or when the number of time points exceeds the number of samples. Such data is becoming increasingly common with advances in automatic
phenotyping. We present a mouse behavioral dataset below that has these characteristics. Our simulation studies indicate that likelihood functional regression models, when mis-specified, may not have the desired false positive rates, and can have lower power than our estimating equations approach. Generalized estimating equations have been used for mapping non-normal traits (LANGE AND WHITTAKER, 2001), but, to our knowledge, they have not been applied to mapping function-valued traits. Note that for genetic mapping, the covariance may be considered a nuisance parameter; for evolutionary studies and animal breeding, the covariance function is a key parameter of interest (KINGSOLVER ET AL., 2001; MEZEY AND HOULE, 2005).

Central to our approach is a general linear functional model that can accommodate any number of covariates and use a single set of computer programs without modification for different number of covariates. We use a simple two-step least squares estimation procedure that is comparable in speed and memory to regular linear regression. That facilitates the use of commonly needed but computationally intensive procedures such as permutation tests for assessment of critical values or multiple testing correction, multiple imputation of incomplete genotypes, and model selection procedures.

Our paper is organized as follows. The next section describes a mouse behavior dataset that motivated our method (Mouse behavioral data). Then we outline the statistical theory underlying our method (Model and estimation), including the model, estimation, and testing methods. We examine the method’s characteristics in simulations, and on the mouse behavioral data (Simulation studies and data analysis). We conclude with a summary (Discussion).
Mouse behavioral data

The mouse experimental data come from a genetic analysis of mouse behavior using an automated home cage monitoring system. A backcross population was generated by crossing C57BL/6J and 129S1/SvImJ strains. The resulting F1 mice were then crossed back to the 129S1/SvImJ strain to generate the 89 N1 backcross mice that were phenotyped and genotyped. Genotyping was performed using Illumina's low density mouse SNP panel which provided 233 informative (polymorphic) autosomal markers between these two strains. The average marker distance was 9.7 Mb (median 8.0 Mb, range 0.2 - 35 Mb). For phenotyping, mice were individually housed for 16 days in a home cage monitoring system which has been described in detail elsewhere (Goulding et al., 2008). Briefly, this system allowed automated collection of feeding (detection of photobeam breaks when mice access food), drinking (detection of capacitance change with lick contacts at water spout), and movement (load cell position platform) data under a 12:12 light dark cycle (lights on 7am, off at 7pm).

Data was processed for errors in detection of feeding, drinking, and movement events as previously described (Goulding et al., 2008). Only mice with at least 4 days of complete data were used for subsequent analysis. Mice were allowed 4 days of acclimation to the monitoring cages and the final 12 days of data were used for analysis.

Mice in their home cages exhibit transitions between two major distinct states: an active state where the animals engage in bouts of feeding, drinking, and locomotion and an inactive state in which the animals engage in prolonged episodes of minimal movement at a discrete home base location (Goulding et al., 2008). Transitions between active and
inactive states were identified with a resolution of 20 milliseconds; the data then were aggregated into six-minute bins across all analysis days. Thus, each measurement represents the probability that an individual mouse was in the active state (active state probability, ASP) during any six-minute interval across all analysis days (GOULDING ET AL., 2008).

As indicated earlier, the data showed several features that motivated the development of new methodology:

1. The number of measurements per mouse (220) exceeds the number of mice (89), which presents problems for many methods as the empirical covariance matrix is singular.

2. There is considerable variation in individual mouse active state probabilities (Figure 1, top and middle panels).

3. The distribution is awkward to describe. There are mice with active state probabilities of zero or one at some time points.

4. The variance between individuals as a function of time is smooth with two peaks (Figure 1, bottom panel).

5. The average ASP pattern is also smooth with two peaks: one at 7-9pm and one at 6-8am. There is a third moderate peak between these two peaks. There is a minimum or nadir at night at 4-6am. Note, however, that usual transformations of the data (such as the arcsine transformation) failed to stabilize the variance (results not shown).
6. The correlation between time points was also smooth but without an easily-articulated structure (Figure 2).

MODEL AND ESTIMATION

We outline our method in this section. We start with the functional regression model. Using estimating equations, the parameter estimates and test statistics have a closed-form solution, assuming complete genotype data. Next, we present some extensions that broaden the method’s applicability and describe our strategy for handling incomplete genotypes. Finally, we discuss the computational requirements of our method, basis function choice, and shrinkage estimation of the covariance.

Model  When complete genotype data is available, we can express the trait data in terms of a regression model. For example, a linear regression model is traditionally used for mapping a quantitative trait:

\[ y = z_1 \beta_1 + z_2 \beta_2 + \cdots + z_p \beta_p + e, \]

where \( y \) is the quantitative trait of interest, and \( z_k, k = 1 \cdots p \), are covariates, \( \beta_k, k = 1 \cdots p \), are the effects of the covariates, and \( e \) is the random error. The covariates may include an intercept, the QTL genotypes, and any other factors such as age, sex, or body weight that may contribute to the trait. The focus of QTL mapping is to identify the genetic covariates contributing to the trait.

Analogously, we use a functional regression model for our function-valued trait when complete genotype data is available. Let \( y(t) \) be the function describing the observable
trait as a function of $t$; in the remainder of our treatment we consider $t$ to be time, but it can be extended to consider spatial position in one or more dimensions. We assume that the function can be represented as follows in terms of covariates $z_1, z_2, \ldots, z_p$ (that may include genotypes)

$$y(t) = z_1\beta_1(t) + z_2\beta_2(t) + \ldots + z_p\beta_p(t) + e(t) = \sum_{k=1}^{p} z_k\beta_k(t) + e(t),$$

(1)

where $\beta_k(t), k = 1 \cdots p$, are unknown functions, and $e(t)$ is random error. We assume that each of the $\beta_k(t)$ functions can be represented as a finite linear combination of a family of basis functions (say splines, polynomials, or Fourier series) as follows

$$\beta_k(t) = b_{k1}\psi_1(t) + b_{k2}\psi_2(t) + \ldots + b_{kq}\psi_q(t) = \sum_{l=1}^{q} b_{kl}\psi_l(t),$$

where $q$ is the number of basis functions. The number and nature of the covariates $z_1, z_2, \ldots, z_p$ will depend on intended use. They may include a single QTL, multiple QTL with or without interaction effects, or non-genetic covariates such as age and sex.

Suppose there are $n$ individuals with $y_i(t)$ denoting the function-valued trait for the $i$-th individual. Further suppose that the trait is observed at $m$ times $t_1, t_2, \ldots, t_m$. If we denote by $y_{ij}$ the trait value for the $i$-th individual at $t_j$, then we can represent the observed trait data as an $n \times m$ matrix

$$Y = [y_{ij}]_{n \times m}.$$

Let $z_{ik}$ denote the value of the $k$-th covariate in the $i$-th individual, let $\psi_{jl} = \psi_l(t_j)$ be the value of the $l$-th basis function at the $j$-th time point $t_j$, and let $e_{ij}$ be the random error for the $i$-th individual at the $j$-th time point. In matrix notation, we write

$$Z = [z_{ik}]_{n \times p}, \quad \Psi = [\psi_{jl}]_{m \times q} B = [b_{kl}]_{p \times q}, \quad \text{and} \quad E = [e_{ij}]_{n \times m}.$$
Then, it can be shown that

\[ Y = Z B \Psi^T + E. \]  (2)

Writing \( y = \text{vec}(Y^T), \epsilon = \text{vec}(E^T), \beta = \text{vec}(B^T), \) and \( X = Z \otimes \Psi, \) where \( \otimes \) denotes the Kronecker product and \( \text{vec} \) is an operator that stacks columns of a matrix into a single, large column vector, the above equation can be written in the familiar form

\[ y = X \beta + \epsilon. \]

If the \( n \) individuals are independent with common covariance matrix then \( \Lambda = \text{var}(\epsilon) = \text{var}(\text{vec}(E)) = I_n \otimes \Sigma, \) where \( \Sigma \) is the covariance matrix of \((e(t_1), e(t_2), \ldots, e(t_m))^T.\) Note that \( \Sigma \) is, in general, unknown and must be estimated from the data.

**Estimation**  As stated earlier, without a distributional assumption for the random error, we cannot write out a likelihood function. Instead we choose \( B \) (or equivalently \( \beta \)) to minimize the residual sum of squares,

\[ S(\beta) = (y - \hat{y})^T(y - \hat{y}) = \text{vec}(Y - \hat{Y}) \text{vec}(Y - \hat{Y})^T, \]  (3)

where \( \hat{y} = \text{vec}(\hat{Y}^T) \) is an estimate depending on \( \hat{\beta} = \text{vec}(\hat{B}^T). \) The criterion, equation (3), is an approximation to the integrated squared error of the estimate

\[ \int_{t_*}^{t^*} (y(t) - \hat{y}(t))^2 dt, \]

where \([t_*, t^*]\) is the range of the interval we wish to consider. Assuming that the range of observed time points is close (or identical) to the range we wish to consider, the integrated squared error may be approximated by the sum

\[ \sum_{j=1}^{m} (y(t_j) - \hat{y}(t_j))^2. \]
Optimizing this criterion is equivalent to solving a least squares problem which has a closed form solution,

\[ \hat{B}^T = (\Psi^T \Psi)^{-1} \Psi^T Y^T Z (Z^T Z)^{-1}. \]  \hspace{1cm} (4)

This can be seen as combination of two least squares: one on basis functions (the \( \Psi \) part) and the other on genotypes (the \( Z \) part). The least squares equations are thus our estimating equations; the resulting solution is asymptotically unbiased (Liang and Zeger, 1986). Further, a consistent estimate of the variance of the estimated coefficients is obtainable if a consistent estimate of the residual covariance matrix, \( \Sigma \), is available. Note that estimation of the coefficients does not require knowledge or estimation of \( \Sigma \).

The variance of the estimate is

\[ \text{var}(\hat{\beta}) = (Z^T Z)^{-1} \otimes \left( (\Psi^T \Psi)^{-1} \Psi^T \Sigma \Psi (\Psi^T \Psi)^{-1} \right) \equiv \tau. \]  \hspace{1cm} (5)

The predictions are

\[ \hat{Y} = \left( Z \left( Z^T Z \right)^{-1} Z^T \right) Y \left( \Psi \left( \Psi^T \Psi \right)^{-1} \Psi^T \right). \]

In the context of statistical methods for longitudinal data, our approach corresponds to using estimating equations with a working independent correlation structure and assuming a Gaussian distribution. In general, these estimates will be less efficient compared to a correctly-specified likelihood model (Godambe, 1960). However, significant loss of efficiency seems to be the exception rather than the rule (McCulloch et al., 2008; Diggle et al., 2002; Chandler and Bate, 2007). We explored this in our simulations (see Data analysis and simulations).
**Testing** A genome scan considers every locus on the genome, and asks the question: is trait variation explained by genetic variation at this locus? The evidence is traditionally summarized by the LOD score, which is the logarithm (base 10) of a likelihood ratio statistic. The likelihood ratio compares a trait regression model with and without the genetic locus under consideration. Since we do not specify a probability model for the data, we cannot calculate a traditional LOD score. Instead, we consider two related quantities as explained below. These can be used as test statistics to test the null hypothesis at each locus, i.e., genetic variation at that locus does not contribute to (function-valued) trait variation.

Consider the trait model in equation (1). The standard null hypothesis is no effect, that is, some $\beta_k(t)$ are identically zero, which is same as the coefficients of their basis expansion being zero. Suppose the null hypothesis is $B = 0$ (testing select entries by replacing $B$ with $SB$ in equation (4), where $S$ is a selection matrix so that $SB$ selects specific columns of $B$, and subsequent derivations would be based on modified estimation).

**Wald statistic:** Let $\text{vec}(\hat{B}^T)$ be the estimate of coefficients in vector form and $\hat{\tau}$ the estimated covariance matrix of $\text{vec}(\hat{B}^T)$. The quadratic form

$$\text{vec}(\hat{B}^T)^T \hat{\tau}^{-1} \text{vec}(\hat{B}^T)$$  \hspace{1cm} (6)

follows a Hotelling’s $T^2$ statistic with parameters $d$ and $r$ where $d$ is the degrees of freedom of the estimate of $\tau$ (equation (5)) and $r$ is the length of $\text{vec}(B^T)$. For reasonably large sample sizes, this can be approximated by a $\chi^2$ statistic with $r$ degrees of freedom.
The estimate of the covariance matrix \( \tau \) is
\[
\hat{\tau} = (Z^T Z)^{-1} \otimes \left( (\Psi^T \Psi)^{-1} \Psi^T \hat{\Sigma} \Psi (\Psi^T \Psi)^{-1} \right),
\]
where \( \hat{\Sigma} \) is an estimate of \( \Sigma \). The simplest estimate one can consider is formed from the residuals:
\[
\hat{\Sigma} = \frac{1}{n - p} \sum_{i=1}^{n} (y_i - \hat{y}_i)(y_i - \hat{y}_i)^T.
\]
This estimate is unbiased, but one may also want to consider biased estimates as discussed later in this section.

Residual error statistic: An alternative statistic would be the difference in residual sum of squares between the model with the genetic locus and a null model corresponding to the null hypothesis. Thus, if \( \hat{Y}_0 \) denotes the fitted values from the null model, and \( \hat{Y}_1 \) denotes the fitted values from the model including the genetic locus under consideration, we would calculate \( S_i = \text{vec}(Y - \hat{Y}_i)^T \text{vec}(Y - \hat{Y}_i), i = [0, 1] \), and then use
\[
\frac{S_0 - S_1}{S_1}
\]
as a test statistic. This statistic is closely connected to the proportion of the variance explained by the locus. The asymptotic null distribution of this statistic is a mixture of \( \chi^2 \) variables (Rotnitzky and Jewell, 1990; Shen and Faraway, 2004). The mixing proportions depend on the eigenvalues of a matrix depending on \( \Psi \) and \( \Sigma \).

Assessing significance As stated earlier, the Wald statistic has an asymptotic \( \chi^2 \) distribution, while the distribution of the residual error statistic is more complicated. The genome-wide significance of either can be established by permutation test (Churchill
AND DOERGE, 1994). For GWAS, one may contemplate a Bonferroni correction or an FDR correction on point-wise p-values. In this context the Wald statistic would be more convenient.

**Incomplete genotypes** The functional regression model (1) assumed that we have complete genotypes, i.e. the genotype of every individual at every genomic location is known. In practice this is rarely the case, as there are gaps between typed markers, genotyping reactions for some individuals may fail, or selective genotyping may have been used. A key part of the QTL mapping problem is to accommodate incomplete genotype data. If typed markers are reasonably dense and no selective genotyping is used, one can use Haley-Knott regression (Haley and Knott, 1992). Here we replace the indicator variables corresponding to possible genotypes at a locus by their probabilities conditional on typed markers. Then we use functional regression as if we had complete genotypes. This method is very fast, and easily parallelized, but is susceptible to bias when selective genotyping is used or when the marker spacing is big (Kao, 2000; Sen et al., 2005). In such cases we can use multiple imputation (Sen and Churchill, 2001). For both multiple imputation and Haley-Knott regression the analyst can contemplate different functional regression models for the trait without needing any additional computational machinery.

**Computational considerations** The modularity of our algorithm also simplifies computation. If we assume $n$ and $m$ are greater than $p$ and $q$, then a naive application of least squares to the $y = X\beta + \epsilon$ problem, where $X \sim mn \times pq$, would be in the order of $O(p^2q^2nm + p^3q^3)$ (Golub and Van Loan, 1996). In our method, we solve two smaller
least squares problems. In equation (4), the two inverses are of order $O(mq^2 + q^3 + np^2 + p^3)$, while other matrix multiplications add $O(mnq + npq)$.

Furthermore, only the part involving $Z$ needs to be recomputed for different loci, while the rest can be computed once and saved for later use. Likelihood-based methods employing the EM algorithm cannot take advantage of this; every instance of the EM algorithm has to be run separately. Suppose we are performing $r$ permutations, and using $s$ imputations, the complexity will be $O(mq^2 + q^3 + mnq + lrs(np^2 + p^3 + npq))$ for $l$ loci. Use of the EM algorithm for maximum likelihood also has the disadvantage that it requires modification each time the function regression model is changed.

**Basis functions** The choice of basis functions and their tuning parameters (such as knot position and number of splines) are vitally important in functional regression and an active research area. As this is beyond the scope of this paper, we will only note that the family and number of basis functions are key choices and can be flexibly accommodated with our method. We used B-splines for the mouse behavior data, and natural splines for the simulations.

**Shrinkage estimation of $\Sigma$** When the number of individuals is greater than the number of time points, the obvious estimate of $\Sigma$ is the empirical covariance of residuals. It is unbiased, but may suffer from high variance (Kauermann and Carroll, 2001). If the number of time points ($m$) is large relative to the number of individuals ($n$) the analyst may consider using a biased estimator such as the shrinkage estimator proposed by Ledoit and Wolf (2004) and adapted by Schafer and Strimmer (2005). The result-
ing covariance estimate is non-singular (unlike the empirical estimate when the number of individuals is smaller than the number of time points). We evaluate this choice in a simulation study in the next section. Shrinkage estimation of the genetic covariance for function-valued data has been considered by Meyer and Kirkpatrick (2010).

Additional practical considerations In our development we assumed that there was no missing trait data. If there is a modest amount of missing trait data we can use local smoothing methods such as smoothing splines to fill in the missing data. Alternatively, one can incorporate estimation of missing data into the estimation procedure by using a basis-expansion on the left-hand-side of equation (2) and replacing $Y$ with the coefficients of the expansion.

Our estimating equations correspond to assuming homoscedasticity and independence between time points. We can easily accommodate heteroscedasticity by using weighted least squares instead of ordinary least squares. There can be weights on individuals and/or different time points. More generally, if the analyst has prior knowledge about the covariance between samples or time points, it can be used to increase efficiency, yet the method will remain robust to their misspecification.

Simulation studies and data analysis

We conducted three simulation studies and analyzed the mouse behavioral data presented earlier, to evaluate our method. We studied its Type-I error and power, and compared it to other methods (functional and non-functional). In the first simulation study,
we compared our method to a parametric likelihood-based functional mapping method based on MA ET AL. (2002). We computed the Type-I error under three covariance structures. In the second simulation study we compared our method to a likelihood-based method that estimates the covariance (YAP ET AL., 2009). In addition, we looked at the power of different methods. In the third simulation study, we compared our method with a non-functional method (a cross-sectional method), which compresses multiple observations within a sample into a summary statistics by averaging; this material is in Supplementary material. Finally, we applied both the Wald statistic and the integrated residual error statistic to our mouse behavioral data and compared it to a cross-sectional method (taking means over time intervals) that is currently used to analyze such data (NISHI ET AL., 2010).

**Simulation studies: Comparison with parametric functional approach**  Here we report simulation studies comparing our estimating equations approach to a likelihood-based parametric approach. Our objective was to compare the null distribution and power of our method against likelihood-based functional mapping under three conditions: when the likelihood model is correctly specified, the point-wise error distribution is incorrectly specified, and the covariance is misspecified.

Our simulations were loosely modeled after the poplar tree data in MA ET AL. (2002). We assumed that 13 equally spaced measurements between times 0 and 6 were made on 200 individuals; 100 individuals each had one of two possible genotypes 0 and 1. The mean of the observations was assumed to follow a logistic curve as in MA ET AL. (2002) with
the functional form

\[ y(t) = \frac{\theta_0}{1 + \theta_1 \exp(-\theta_2 t)} + \epsilon(t), \]  

(7)

where \( \theta = (\theta_0, \theta_1, \theta_2) \) are parameters describing the mean curve, and \( \epsilon(t) \) is a stationary stochastic process with mean 0 and point-wise variance \( \sigma^2 = 0.01 \). The functional parameters of individuals with genotypes equal to \( i \) is described by \( \theta^i \), \( i = 0, 1 \). Under the alternative, \( \theta^0 = (1.00, 9.0, 1) \), and \( \theta^1 = (0.95, 8.5, 1) \). Under the null, \( \theta^0 = \theta^1 = (0.975, 8.75, 1) \). The plots of three mean functions are in Figure S2 in the Supplementary material.

We considered three error structures:

**Gaussian, autoregressive** The marginal (point-wise) and all finite-dimensional distributions of the error are Gaussian, and the correlation structure is autoregressive so that the covariance between measurements separated by a time \( t \) is described by the exponential correlation function \( \rho(t) = \sigma^2 r^t \). We evaluated performance under \( r = 0.61, 0.83, 0.94 \), which are the correlations between successive time points for the Matérn correlation function (with smoothness parameters 0.5, 1, and 2 respectively) considered below (see Supplementary material for definition of the covariance function and references).

**t-distributed, autoregressive** The finite-dimensional distributions have a t-distribution with 4 degrees of freedom (the smallest for which the third moment exists). The correlation function is autoregressive, as above.

**Gaussian, Matérn** The finite-dimensional distributions are Gaussian, and the correlation function is Matérn with amplitude parameter \( \sigma^2 \), scale parameter 1, and smoothness parameters 0.5, 1, and 2. The three correlations functions are shown in Figure S3.
We compared three functional approaches:

**Likelihood (Lik)** We used the likelihood-based method presented in MA ET AL. (2002).

This assumes that the mean function has a logistic form, and that the errors are Gaussian with an autoregressive structure.

**Estimating equations (EE)** We used the Wald statistic without a shrinkage estimator for the error covariance matrix, $\Sigma$. The mean function was modeled using a natural spline basis with 7 degrees of freedom.

**Estimating equations with shrinkage (EES)** We used the Wald statistic with a natural spline basis with the SCHAFER AND STRIMMER (2005) method for shrinking the covariance matrix.

Note that the estimating equations approach is applied using no knowledge of the mean and correlation functions. The likelihood method is correctly specified for the first set of simulations (Gaussian, autoregressive). It misspecifies the finite-dimensional distributions in the second set (t-distributed, autoregressive), but the correlation is correctly specified. In the third set, the correlation functions are misspecified.

Table 1 gives tail probabilities of the test statistics under the null. It shows that the correctly-specified likelihood ratio statistic has $\chi^2$ tail probabilities as expected. This holds true when the marginal distribution of the errors has a t-distribution as well. However, when the correlation structure is misspecified, it no longer has a $\chi^2$ distribution and the Type-I error corresponding to all the critical values are higher than expected.

The simulations indicate that the regular estimating equations approach has near the expected Type-I error behavior under all circumstances. However, the estimating equations
approach with shrinkage does not have this property with slightly too low Type-I error rates, and thus its null distribution would need to be obtained empirically using permutations. The likelihood ratio statistic has the expected $\chi^2$ distribution only when the correlation structure is correctly specified; otherwise it may be way off target. Thus, in practice, if the covariance is hard to specify, the likelihood ratio statistic’s statistical significance should be established using a permutation distribution.

Since the null distribution of the statistics are not always as expected, we used the null distribution quantiles as the critical values to calculate power (Table 2). We find that the likelihood approach has the greatest power when the correlation structure is correctly specified (note that the likelihood method in this case also benefits from having the correct mean function). The estimating equations approaches have greater power when the correlation structure is incorrectly specified. In all situations, the estimating equations with shrinkage has slightly greater power than the approach without shrinkage. Thus, our simulations demonstrate that power of likelihood methods may be compromised if the correlation structure is misspecified (even if the Type-I error has been recalibrated).

**Simulation studies: Comparison with parametric functional method with unstructured covariance**  
YAP ET AL. (2009) proposed using regularized covariance estimation within the framework of parametric likelihood-based functional mapping. The regularization parameter was selected by k-fold cross-validation and assumed constant throughout a marker interval.

We performed simulations using the same scenario as YAP ET AL. (2009). We simulated a 100 cM linkage group with 6 equally spaced markers with a QTL at 32 cM. The asso-
ciated phenotypes was sampled from a multivariate Gaussian distribution with logistic function describing the mean function conditional on the QTL genotypes. There are three genotypes with three mean curves following logistic functions as in equation (7): the first genotype has $\theta_0 = 30, \theta_1 = 5, \theta_2 = 0.5$; the second genotype has $\theta_0 = 28.5, \theta_1 = 5, \theta_2 = 0.5$; the third genotype has $\theta_0 = 27.5, \theta_1 = 5, \theta_2 = 0.5$. Each individual is observed at 10 time points. The residual error was assumed to be multivariate normal, with three different covariance structures:

1. $\Sigma_1$ is auto-regressive with $\sigma^2 = 3, \rho = 0.6$.

2. $\Sigma_2$ is equi-correlated with $\sigma^2 = 3, \rho = 0.5$.

3. $\Sigma_3$ is an “unstructured” covariance matrix, as given in Yap et al. (2009) (reproduced in the Supplementary material). It does not have an easily-described structure.

All parameter values including the covariance matrices are taken from Yap et al. (2009). We ran 10,000 simulation replicates to obtain stable estimates; Yap et al. (2009) had 100 simulation replicates.

The LOD score (or linkage test statistic) was calculated every 4 cM resulting in 26 loci. As in the simulations reported by Yap et al. (2009), for each simulation, we estimated the location of the QTL as the location of the maximum LOD score. The average, standard deviation, and root mean squared error of QTL location estimates were compared. Lower root mean squared error indicates a better method. (Note that we are unable to compare the power of the methods as we do not have access to the software used in the Yap et al. (2009) study.)
We found that our method performed comparably to Yap et al. (2009)'s method for the most part, except for $\Sigma_3$ with a small sample size of 100, where Yap et al. (2009)'s method had better mean and slightly smaller standard deviation. But looking at the large values of standard deviations for all methods, it is clear that no method performed satisfactorily with this small sample size. For $\Sigma_3$ with sample size 400, the performance of both methods were again comparable. Note however, that our method is considerably simpler to implement, and has lower computational complexity.

Data analysis  Please refer to the section entitled “Mouse Behavioral Data” for the details of the mapping population, markers, and phenotypic data collection. Our analytic goal was to detect genetic loci that contribute to individual variation in the shape of the curve describing how active state probability (ASP) changes with time of day. We applied our method using the Wald statistic and the residual error statistic. We also applied a non-functional approach informed by prior knowledge and a visual exploration of the data. The approach tries to mimic what we might do if we did not have a functional method at our disposal. It parallels the approach taken recently by Nishi et al. (2010) for similar home-cage movement data in consomic strains. Note that a credible parametric model for mouse behavior cannot be easily specified, and thus we cannot use a parametric functional method.
Non-functional method  We constructed six measures based on the marginal distribution of the ASP curves as shown in Figure 1. Note that the change in active state probability with time of day exhibited by the mice appears to be divided into five phases. During the dark cycle, there is an onset peak (7-9pm) in active state probability which then decreases somewhat exhibiting a broad peak from around 9pm-4am followed by a dark cycle nadir from 4-6am. An offset peak (6-8am) in the active state probability then occurs near the end of the dark cycle. Finally, the active state probability is low throughout the majority of the light cycle. Based on these observations we segmented the day into the following phases to measure the probability that a mouse was in the active state:

1. **Daily**  This is the mean over all time points.

2. **Dark cycle onset peak**  This is the mean over 7-9pm.

3. **Mid-Dark cycle**  This is the mean over 9pm-4am.

4. **Dark cycle nadir**  This is the mean over 4-6am.

5. **Dark cycle offset peak**  This is the mean over 6-8am.

6. **Light cycle**  This is the mean over 8am-noon and 2-7pm.

We performed genome scans using the Haley-Knott method (Haley and Knott, 1992) for each of these measures. To correct for the fact that we used 6 correlated genome scans, we calculated a genome-wide threshold for the maximum of the 6 genome scans using 1000 permutations. Using the 5% multiple-scan corrected threshold we found only one locus on chromosome 1 for daily ASP.
Functional method  We used B-splines as our basis functions. We applied 10-fold cross-validation to the behavioral data, ignoring the genotype data, to select the number of basis functions to use for smoothing. We selected the smallest number of basis functions that gave a residual sum of squares within one standard deviation of the least value. This led us to select B-splines with 16 degrees of freedom with equally-spaced knots. These basis functions were used for all genome scans.

We performed genome scans using the Haley Knott method with both the Wald statistic and the residual error statistics. We used 1000 permutations to establish genome-wide statistical significance. Using the functional approach and the 5% genome-wide threshold, we found two loci (on chromosomes 1 and 9) using the residual error statistic, and one loci on chromosome 9 using the Wald statistic. Thus, in this example, the Wald statistic appears to be less sensitive than the residual error statistic.

**DISCUSSION**

We have presented a new approach based on estimating equations for mapping function-valued traits. It is an attractive alternative to likelihood-based methods, especially when we have incomplete knowledge of the covariance structure or when the number of time points exceeds the number of samples. While information about the covariance can be incorporated to improve efficiency, our method is robust to its misspecification. Relative to a correctly-specified likelihood model, the estimating equations approach may be slightly less efficient. However, such loss was modest in our simulations and other studies (CHANDLER AND BATE, 2007). We thus believe that our method would be a good
choice in most settings where the analyst is uncertain about the error distribution. The mouse behavioral data we presented is such an example. Likelihood methods may be preferred, say, for traits such as growth curves, which have a long history of study, lending confidence to the data model.

Misplaced confidence in the covariance structure may have a price, as our simulations indicate. Likelihood-based methods with a misspecified covariance may not have the target Type-I error, and may have lower power than the estimation equations method. Although the anti-conservative behavior of a misspecified likelihood observed in our simulations may be rectified by a permutation test, that has an additional computational cost. This cost is manageable for QTL mapping in experimental crosses, but less so in GWAS where a correctly-calibrated p-value is desired at each locus. Thus, the estimating equations Wald statistic has an edge in GWAS.

There are several ways to deal with more time points than the sample size, which engenders a singular covariance matrix. ZHAO ET AL. (2007) used a wavelet transform for time series dimension reduction as a precursor to parametric functional mapping. Our method, on the other hand, has dimension reduction built-in via the basis functions, which in fact, could be wavelets. While ZHAO ET AL. (2007) used wavelets for dimension reduction as a separate step we perform dimension reduction and functional mapping in one step.

Computational efficiency, as opposed to statistical efficiency, plays a bigger role for high dimensional data. If the number of time points is very large or if model selection, permutation testing or multiple imputation are needed, computational time increases sev-
eral fold. In such settings, our method has a clear advantage, as it is based on two low-dimensional least squares operations; it can be computed quickly and does not involve any complex optimization procedures. The computational advantage is particularly pronounced when covariance structure defies easy specification. While Yap et al. (2009)'s method does not rely on specification of covariance structure, it comes at the expense of a generalized EM algorithm that must numerically solve a nonlinear optimization problem as part of M-step. For large-scale studies with dense genotyping, computational efficiency is essential.

The estimating equation approach for mapping function-valued traits can be adapted for different situations. Our formulation assumes that the error variance is independent of the mean, which might not hold for traits whose point-wise marginal distribution is markedly non-Gaussian such as Poisson-distributed count data. A simple change to the sandwich estimator in equation (5) can be used to obtain a robust variance estimate in this scenario. This involves replacing the estimate by a sum of $n$ terms with the same form as in equation (5), with $\Sigma$ replaced by the empirical multivariate residual for each individual. Another extension is when trait data on all individuals are not observed on the same set of time points. While small amounts of missing trait data can be dealt by smoothing or imputation, the more general problem of uneven time points is more challenging. The approach of Yao et al. (2005) may offer a resolution under Gaussian modeling assumptions.

In summary, our estimating equation method based on a general functional linear model makes few distributional assumptions. It has broad applicability to genetic studies, and
has opened an attractive alternative avenue to functional mapping distinct from likelihood methods. The possible improvements and adaptations listed above suggest the method’s promise and flexibility. To facilitate applications, open source software implementing our method is available from http://www.biostat.ucsf.edu/sen/software/.

**ACKNOWLEDGMENTS**

Genotyping was carried out in the DNA Technologies Core at the UC Davis Genome Center. We thank Ethelyn Layco for animal care, sample preparation for Illumina genotyping, and assistance with behavioral experiments. We thank Rongling Wu for providing Matlab code for the likelihood method which we adapted for use in R, Karl Broman and Tom Juenger for helpful comments, and Sandra Erickson for suggesting we work on mapping function-valued traits. This research was supported by grants from the National Institutes of Health (GM078338, GM074244 and DK072187). Some computations were performed using the UCSF Biostatistics High Performance Computing System.

**LITERATURE CITED**


Table 1: Null distribution p-values from parametric and estimating equation approaches under three error distributions using 1000 simulations each. We considered three different covariance structures: (a) Gaussian with autoregressive correlations, (b) t-distributed with 4 degrees of freedom and an autocorrelated covariance, and (c) Gaussian with Matern correlation function. We consider three correlation strengths as measured by the correlation between nearest neighbors: 0.61, 0.83, and 0.94. The parametric likelihood method assumes autocorrelated error (Lik) and has the desired Type-I error for both Gaussian and t-distributed errors, when the correlation structure is correctly specified. However, it is substantially off target when the correlation is misspecified. The estimating equations approach with shrinkage (EES) has close to the desired Type-I error, but is slightly off target. The used estimating equations approach (EE) always has the correct Type-I error.

<table>
<thead>
<tr>
<th>Neighbor correlation</th>
<th>Covariance structure</th>
<th>Method</th>
<th>Method</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutoff</td>
<td>Lik</td>
<td>EES</td>
<td>EE</td>
<td>Lik</td>
</tr>
<tr>
<td>0.1</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>0.05</td>
<td>0.057</td>
<td>0.053</td>
<td>0.054</td>
<td>0.064</td>
</tr>
<tr>
<td>0.01</td>
<td>0.013</td>
<td>0.011</td>
<td>0.011</td>
<td>0.016</td>
</tr>
<tr>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutoff</td>
<td>Lik</td>
<td>EES</td>
<td>EE</td>
<td>Lik</td>
</tr>
<tr>
<td>0.1</td>
<td>0.099</td>
<td>0.099</td>
<td>0.098</td>
<td>0.099</td>
</tr>
<tr>
<td>0.05</td>
<td>0.048</td>
<td>0.052</td>
<td>0.058</td>
<td>0.044</td>
</tr>
<tr>
<td>0.01</td>
<td>0.015</td>
<td>0.010</td>
<td>0.013</td>
<td>0.006</td>
</tr>
<tr>
<td>0.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutoff</td>
<td>Lik</td>
<td>EES</td>
<td>EE</td>
<td>Lik</td>
</tr>
<tr>
<td>0.1</td>
<td>0.11</td>
<td>0.084</td>
<td>0.11</td>
<td>0.092</td>
</tr>
<tr>
<td>0.05</td>
<td>0.050</td>
<td>0.039</td>
<td>0.054</td>
<td>0.051</td>
</tr>
<tr>
<td>0.01</td>
<td>0.012</td>
<td>0.006</td>
<td>0.011</td>
<td>0.013</td>
</tr>
</tbody>
</table>
Table 2: Power of parametric and estimating equations approaches under three error distributions: Gaussian, autoregressive (GA); t, 4df, autoregressive (tA); Gaussian, Matérn (GM). The table shows the proportion of times the p-values under the alternative was smaller than the 5th percentile of the null. The parametric approach (Lik) has the highest power when the covariance structure of the error is correctly specified. This is true for both Gaussian and t-distributed errors. The estimating equations approaches have slightly less power than the likelihood method with correctly-specified correlation. When the covariance structure is incorrectly specified, the estimating equations approaches have greater power. Estimating equations with shrinkage (EES) appears to have slightly higher power than estimating equations sans shrinkage (EES).

<table>
<thead>
<tr>
<th>Neighbor correlation</th>
<th>0.61</th>
<th>0.83</th>
<th>0.94</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GA</td>
<td>tA</td>
<td>GM</td>
</tr>
<tr>
<td>Lik</td>
<td>0.96</td>
<td>0.96</td>
<td>0.60</td>
</tr>
<tr>
<td>Methods</td>
<td>EES</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>EE</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Table 3: Mean, standard deviation, and root mean squared error of the genome scan peak in simulations using our method and YAP ET AL. (2009). The true QTL is at 32cM. The columns labeled “Yap(nonpara)” and “Yap(autoreg)” are derived from Table 1 and Table 2 of YAP ET AL. (2009). They refer to the results of using the likelihood-based methods of YAP ET AL. (2009) with a estimated regularized covariance and autocorrelated covariance respectively. The columns labeled “EE(Wald)” and “EE(Residual)” refer to our estimating equations approach with the Wald statistic and the residual error statistic respectively. YAP ET AL. (2009) performed 100 simulation replicates, whereas we used 10,000 replicates which gave stable estimates. For each method we report the mean position of the genome scan maximum (“mean”) over simulation replicates, the standard deviation (“std”) and the root mean squared error (“rmse”). Note that YAP ET AL. (2009) had reported standard error, which we converted to standard deviation as the latter is independent of the number of simulation replicates. The simulations were performed with three error structures corresponding to an autocorrelated covariance ($\Sigma_1$), an equicorrelated covariance ($\Sigma_2$), and an unstructured covariance ($\Sigma_3$).

<table>
<thead>
<tr>
<th>Covariance</th>
<th>n</th>
<th>Yap(nonpara) mean</th>
<th>Yap(nonpara) std</th>
<th>Yap(nonpara) rmse</th>
<th>Yap(autoreg) mean</th>
<th>Yap(autoreg) std</th>
<th>Yap(autoreg) rmse</th>
<th>EE(Wald) mean</th>
<th>EE(Wald) std</th>
<th>EE(Wald) rmse</th>
<th>EE(Residual) mean</th>
<th>EE(Residual) std</th>
<th>EE(Residual) rmse</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Sigma_1$</td>
<td>100</td>
<td>32.8</td>
<td>9.9</td>
<td>9.9</td>
<td>33.2</td>
<td>7.7</td>
<td>7.8</td>
<td>33.0</td>
<td>10.4</td>
<td>10.5</td>
<td>32.8</td>
<td>9.09</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>31.5</td>
<td>2.8</td>
<td>2.8</td>
<td>31.8</td>
<td>3.2</td>
<td>3.2</td>
<td>32.1</td>
<td>2.9</td>
<td>2.9</td>
<td>32.1</td>
<td>2.9</td>
<td>2.9</td>
</tr>
<tr>
<td>$\Sigma_2$</td>
<td>100</td>
<td>32.6</td>
<td>7.6</td>
<td>7.6</td>
<td>35.3</td>
<td>15.7</td>
<td>16.0</td>
<td>32.7</td>
<td>7.7</td>
<td>7.7</td>
<td>33.1</td>
<td>12.0</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>31.7</td>
<td>2.6</td>
<td>2.6</td>
<td>32.0</td>
<td>5.4</td>
<td>5.4</td>
<td>32.1</td>
<td>2.4</td>
<td>2.4</td>
<td>32.2</td>
<td>3.7</td>
<td>3.7</td>
</tr>
<tr>
<td>$\Sigma_3$</td>
<td>100</td>
<td>33.2</td>
<td>22.2</td>
<td>22.2</td>
<td>46.5</td>
<td>27.4</td>
<td>31.0</td>
<td>40.6</td>
<td>26.5</td>
<td>27.9</td>
<td>39.3</td>
<td>25.3</td>
<td>26.3</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>32.3</td>
<td>11.9</td>
<td>11.9</td>
<td>43.6</td>
<td>26.4</td>
<td>28.9</td>
<td>33.1</td>
<td>10.8</td>
<td>10.9</td>
<td>33.1</td>
<td>10.7</td>
<td>10.8</td>
</tr>
</tbody>
</table>
Figure 1: Mouse active state probability data trajectories, mean, and variance. The top panel shows the raw trajectories of three randomly-chosen individual mice. The dark period (7pm-7am) is shown in gray. The middle panel shows all individual trajectories in gray, and the mean trajectory overlaid in maroon. There is a peak in activity between 7-8pm, which gradually dies out, followed by a second peak at about 7am. The dark period is shown in a box. The bottom panel shows the variance of the trajectories as a function of time. The variance function shows roughly the same pattern as the mean function.
Figure 2: Correlation between time points for mouse active state probability data. There is some structure in the correlation matrix, although it does not fall into an easily classifiable category. There is high local correlation around between 7pm and 11pm, and a negative correlation between 7am-9am and 7pm-5am.
**Figure 3:** Genome scans using non-functional approach. The top panel shows genome scans for total daily ASP (black), dark cycle onset peak (blue) and mid-dark cycle (red). The bottom panel shows genome scans for dark cycle nadir (black), dark cycle offset peak (blue) and light cycle activity (red). The solid horizontal lines are the 5% permutation threshold for the maximum of the 6 genome scans, that adjusts for the fact that we are examining 6 correlated genome scans.
Figure 4: Genome scans using functional approach. The top panel shows a genome scan using the residual error statistics expressed as a proportion of the integrated error explained by a locus. The horizontal line is the 5% permutation threshold. The bottom panel shows a genome scan using the Wald statistic, and the corresponding 5% permutation threshold.
Figure 5: Functional effect of two genetic loci on chromosome 1 (top panel) and chromosome 9 (bottom panel). The dark period is shaded gray. Both loci appear to have the greatest effect in the high activity periods; there is little difference between the two genotypes in the low activity period concentrated between hours 8am and 5am. The chromosome 1 locus effect is largely between 6pm and 1am. The chromosome 9 locus affects activity between 6pm and 7am. For both loci there is evidence of the dark cycle nadir shifting a little.