# Influence analysis in quantitative trait loci detection

**Xiaoling Dou***[1], **Satoshi Kuriki**[1], **Akiteru Maeno**[2], **Toyoyuki Takada**[2], and **Toshihiko Shiroishi**[2]

[1] The Institute of Statistical Mathematics, Research Organization of Information and Systems, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[2] Mammalian Genetics Laboratory, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan

Received 7 September 2012; revised 12 December 2013; accepted 20 December 2013

This paper presents systematic methods for the detection of influential individuals that affect the log odds (LOD) score curve. We derive general formulas of influence functions for profile likelihoods and introduce them into two standard quantitative trait locus detection methods—the interval mapping method and single marker analysis. Besides influence analysis on specific LOD scores, we also develop influence analysis methods on the shape of the LOD score curves. A simulation-based method is proposed to assess the significance of the influence of the individuals. These methods are shown useful in the influence analysis of a real dataset of an experimental population from an $F_2$ mouse cross. By receiver operating characteristic analysis, we confirm that the proposed methods show better performance than existing diagnostics.

*Keywords:* Influence score vector; Profile likelihood; ROC analysis; Shape of LOD score curve; Standardized empirical influence function.

Additional supporting information may be found in the online version of this article at the publisher's web-site

## 1 Introduction

Quantitative trait locus (QTL) analysis is a statistical method for detecting the precise location of chromosome regions associated with a particular phenotypic trait. To date, many models have been proposed for this analysis (e.g., Lander and Botstein, 1989; Zeng, 1993; Kao and Zeng, 1997; Sen and Churchill, 2001). Among them, single marker analysis and the interval mapping method are the most widely used. In QTL detection, log odds (LOD) scores are calculated for marker loci to indicate the plausibility of the existence of QTLs (e.g., Siegmund and Yakir, 2007; Wu et al., 2007).

Figure 1 shows the LOD scores computed using data on 170 $F_2$ mice. The LOD scores at 119 marker loci are plotted as small circles. The circles on the same chromosome are linked by solid segments, and the joint lines form the LOD score curve. The highest peak appears at chromosome 16; moreover, two moderate peaks located closely on chromosome 3 are observed. For geneticists planning to clone the genes responsible for QTLs, it is important to determine whether the two moderate peaks were caused by two QTLs or by stochastic fluctuations. However, the sample sizes of experimental populations generated from genetic crosses are usually not large and normally consist of a few hundred individuals at most. In such small populations, a few observations can have a major impact on LOD score curves and lead to unstable results.

*Corresponding author: e-mail: xiaoling@ism.ac.jp, Phone: +81-50-5533-8500, Fax: +81-42-526-4339

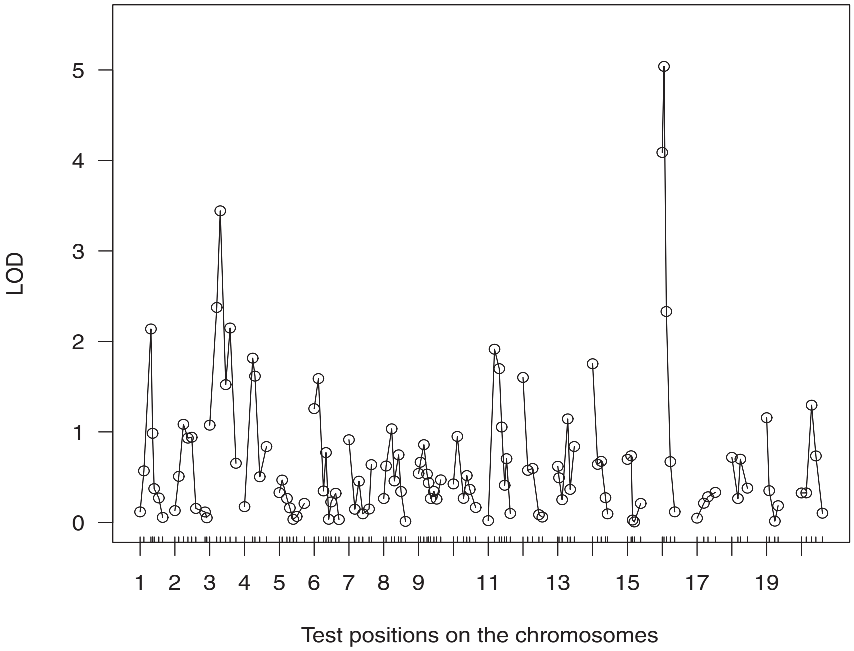

**Figure 1** LOD score curves for mouse chromosomes.

The importance of such sensitivity analysis is well recognized in experimental genetics. In genetic experiments, errors in genotyping and phenotyping are inevitable, which could cause unstable results. To solve this problem efficiently, we can find those individuals having large influence on the LOD score curve, to genotype them again, and to remeasure the phenotypes. Therefore, statistical methods for the identification of influential individuals are helpful in the process of QTL detection. In general, an observation is called influential if it has a large impact on the estimates of interest. It should be distinguished from outliers that are distinct from most of the data points in a sample but not necessarily influential (Bollen and Jackman, 1990). In fact, influential individuals are not necessarily bad and need not to be removed if reliable. They may contain the most interesting information and should be examined carefully. If the reexamination reveals that the original influential data are correct, and if the assumed model fits the data well, geneticists can go to the next stage for positional cloning of the causative gene for the phenotype. If the data are found to be unreliable, researchers need to conduct another QTL analysis based on the updated dataset. However, the conventional method currently used by geneticists to find influential individuals is inadequate from a statistical viewpoint (see Section 4). The purpose of this paper is to provide systematic methods for identifying individuals that influence LOD score curves.

In statistical terminology, the LOD score is nothing but a profile likelihood function based on a statistical model describing the relationship between genotypes and phenotypes. In this paper, we first present a general theory of influence functions for profile likelihoods and apply formulas to the models of QTL detection. Then, to identify individuals that affect the shape of the profile likelihood, we propose the use of the empirical influence functions (EIFs) for a linear combination of LOD scores, which is designed specifically for the shape of interest. We propose three methods for designing the linear combination coefficients: projection based on orthogonal polynomials, principal component analysis (PCA) based on the EIF matrix, and the quadratic form of the EIFs. To assess the significance of the EIFs, we also propose a simple method for providing $p$-values by a robustified parametric bootstrap. This method is confirmed to control false positive rates through numerical studies.

This paper is organized as follows. In Section 2, the general formulas of influence functions for profile likelihoods are derived. In Section 3, EIFs are introduced into QTL analysis. In Section 4, the three methods for designing the linear combination are proposed. After that, the method for calculating *p*-values for significance testing is proposed. In Section 5, we apply these methods to a real dataset of an experimental mouse population. In Section 6, we examine the validity and the statistical power of the proposed methods through simulations. Finally, in Section 7, we summarize the proposed methods, discuss the results of our data analysis, and provide guidelines for application. Mathematical details are provided in the Appendix.

## 2 Influence functions for the profile likelihood

As previously stated, the LOD score in QTL detection is defined as a profile likelihood, because the model describing the relationship between genotypes and phenotypes (and covariates, if available) includes an unknown location parameter $\gamma$ of the QTL. The details of the QTL models are given in Section 3. Here, we develop the theory of influence analysis for profile likelihood functions in a general setting, apart from the context of QTL detection.

We assume that for each individual $i = 1, \ldots, n$, its observation record $x_i$ is taken from a statistical model $f(\cdot; \gamma, \boldsymbol{\theta})$, where $\gamma \in \Gamma \subseteq \mathbb{R}$ is the parameter of interest, and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ is a vector consisting of the other parameters. In QTL detection, the parameter space $\Gamma$ is the search region for the QTLs and is a discrete or continuous set; we consider both cases.

Let $\ell(\gamma, \boldsymbol{\theta}; x) = \ell(\gamma, \boldsymbol{\theta}) = \log f(x; \gamma, \boldsymbol{\theta})$ be the log-likelihood function. The average log profile likelihood function $M_n$ with respect to parameter $\gamma$ is defined as

$$M_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(\gamma, \widehat{\boldsymbol{\theta}}_n(\gamma); x_i\big), \quad \text{where} \quad \widehat{\boldsymbol{\theta}}_n(\gamma) = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ell(\gamma, \boldsymbol{\theta}; x_i) \tag{1}$$

(e.g., Murphy and van der Vaart, 2000). To develop an influence analysis on the characteristics of the profile likelihood $M_n$, we focus on the following functionals of $M_n$:

(i) linear functional: $\int_\Gamma M_n(\gamma) dC(\gamma)$,
(ii) maximum of the profile likelihood: $\max_{\gamma \in \Gamma} M_n(\gamma)$,
(iii) maximizer of the profile likelihood: $\operatorname{argmax}_{\gamma \in \Gamma} M_n(\gamma)$.

Let $\phi(F)$ be a functional of the distribution function $F$ and let $\phi(F_n)$ be its sample version, where $F_n$ is the empirical distribution function defined by the data $\{x_i\}_{i=1,\ldots,n}$. Let $\delta_x$ be the distribution function that has probability mass 1 at $x$. Then, the influence function of $\phi(F)$ is defined as the directional derivative of $\phi$ at $F$ in the direction of $\delta_x$:

$$\mathsf{IF}(x; \phi, F) = \lim_{\epsilon \to 0} \frac{\phi((1-\epsilon)F + \epsilon\delta_x) - \phi(F)}{\epsilon} = \Big(\frac{d}{d\epsilon}\Big)_0 \phi\big(F_x^\epsilon\big),$$

where $F_x^\epsilon = (1-\epsilon)F + \epsilon\delta_x$, and $(\frac{d}{d\epsilon})_0$ denotes the derivative coefficient with respect to $\epsilon$ at $\epsilon = 0$. The EIF (Hampel et al., 1986) of $\phi(F_n)$ at the observation record $x_i$ is the influence function of $\phi$ at $F_n$ in the direction of $\delta_{x_i}$:

$$\mathsf{EIF}(i; \phi) = \mathsf{EIF}_i(\phi) = \mathsf{IF}(x_i; \phi, F_n). \tag{2}$$

Here, $\mathsf{EIF}(i; \phi)$ measures the influence of individual $i$ on the statistic $\phi(F_n)$. To simplify the notation, we use $\mathsf{EIF}_i(\phi)$ interchangeably. The basic strategy of influence analysis is identifying individuals with large absolute values of $\mathsf{EIF}_i(\phi)$.

To derive the EIFs for the functionals (i)–(iii) of $M_n$, we first express $M_n(\gamma)$ as a functional of the empirical distribution function $F_n$. Let $L(\gamma, \boldsymbol{\theta}; F) = \int \ell(\gamma, \boldsymbol{\theta}; y) dF(y)$. Since $\frac{1}{n}\sum_{i=1}^{n} \ell(\gamma, \boldsymbol{\theta}; x_i) =$

$L(\gamma, \boldsymbol{\theta}; F_n)$, the profile likelihood function and its maximizer in (1) can be rewritten as $M_n(\gamma) = M(\gamma, F_n)$ and $\widehat{\boldsymbol{\theta}}_n(\gamma) = \widehat{\boldsymbol{\theta}}(\gamma, F_n)$, respectively, where

$$M(\gamma, F) = L\big(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F\big) \quad \text{and} \quad \widehat{\boldsymbol{\theta}}(\gamma, F) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\gamma, \boldsymbol{\theta}; F).$$

Then, the EIFs of the functionals (i)–(iii) can be obtained by the following theorem. The regularity conditions and the proofs are provided in Appendix A.1.

**Theorem 2.1.** *(i) Let $C$ be a bounded variation function on $\Gamma$. The influence function of the linear functional $\phi_1(C; F) = \int_\Gamma M(\gamma, F) dC(\gamma)$ at $F$ is*

$$\mathsf{IF}(x; \phi_1(C; \cdot), F) = \int \Big\{ \ell\big(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); x\big) - L\big(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F\big) \Big\} dC(\gamma), \tag{3}$$

*provided that $\widehat{\boldsymbol{\theta}}(\gamma, F) = \operatorname{argmax}_{\theta \in \Theta} L(\gamma, \boldsymbol{\theta}; F)$ exists uniquely.*

*(ii) The influence function of the maximum profile likelihood $\phi_2(F) = \max_{\gamma \in \Gamma} M(\gamma, F)$ at $F$ is*

$$\mathsf{IF}(x; \phi_2, F) = \ell\big(\widehat{\gamma}(F), \widehat{\boldsymbol{\theta}}(F); x\big) - L\big(\widehat{\gamma}(F), \widehat{\boldsymbol{\theta}}(F); F\big), \tag{4}$$

*provided that $(\widehat{\gamma}(F), \widehat{\boldsymbol{\theta}}(F)) = \operatorname{argmax}_{(\gamma, \theta) \in \Gamma \times \Theta} L(\gamma, \boldsymbol{\theta}; F)$ exists uniquely.*

*(iii) Assume that $\Gamma$ is a continuous set. The influence function of the maximizer of the profile likelihood $\phi_3(F) = \operatorname{argmax}_{\gamma \in \Gamma} M(\gamma, F)$ at $F$ is*

$$\mathsf{IF}(x; \phi_3, F) = \frac{\ell_\gamma(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; x) - L_{\gamma\theta}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F) L_{\theta\theta}^{-1}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F) \ell_\theta(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; x)}{-L_{\gamma\gamma}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F) + L_{\gamma\theta}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F) L_{\theta\theta}^{-1}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F) L_{\theta\gamma}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F)}, \tag{5}$$

*where $(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}) = (\widehat{\gamma}(F), \widehat{\boldsymbol{\theta}}(F))$. Here, the subscripts indicate partial derivatives. For example, $L_\theta(\gamma, \boldsymbol{\theta}; F) = \partial L(\gamma, \boldsymbol{\theta}; F)/\partial \boldsymbol{\theta}$ is the gradient vector and $L_{\theta\theta}$ is the Hessian matrix.*

Actually, (5) is a special case of the influence function formula for $M$-estimators. Its numerator and denominator are the efficient score and the Fisher information for parameter $\gamma$ in the presence of the nuisance parameter $\boldsymbol{\theta}$, respectively (e.g., Murphy and van der Vaart, 2000).

## 3 Statistical models and influence functions

### 3.1 Experimental crossing data

We first review the statistical models of QTL analysis. For simplicity, we only consider the data from $F_2$ populations, although the statistical methods developed here are applicable to all other experimental design data.

In QTL analysis, data are taken from $n$ individuals (e.g., mice). The observation record of the $i$-th ($i = 1, \dots, n$) individual consists of a phenotype $y_i$, a genotype vector $\boldsymbol{z}_i = (z_i^{(1)}, \dots, z_i^{(m)})$ at $m$ marker loci located at $d_1, \dots, d_m$, and covariates $u_i$. (In this paper, we let $u_i$ be a scalar for simplicity.) Moreover, we assume that a putative QTL exists that has genotype $z_i^*$ at location $\gamma$ on a chromosome and affects trait $y_i$. Note that $\gamma$ is an unknown parameter, and $z_i^*$ is an unobserved variable in general. We assume that except for the QTL, none of the loci affect the phenotype. In the $F_2$ population from two strains A and B, each marker has one of the three genotypes: AA homozygous, AB heterozygous, and BB homozygous. Throughout this paper, we use $z_i^{(j)}, z_i^* = -1, 0, 1$ to denote genotypes AA, AB, and BB, respectively.

### 3.2 Statistical model for the interval mapping method

Various methods and computer programs are available for QTL detection. Among them, we consider two basic methods—the interval mapping method and single marker analysis. It should be noted that influence functions can be defined in any QTL detection method that is based on LOD score. In this paper, we pick these two methods as examples. Although both methods are formulated as single QTL models, they are often applied for constructing scan statistics to decide whether a QTL exists at a particular location, even though more than one QTL may exist (Wright and Kong, 1997).

Interval mapping is a classical method proposed by Lander and Botstein (1989) in the early stages of research on QTL analysis. Although many other procedures have been developed from this method, it is still used as a standard method and has been incorporated into many leading software packages (e.g., Manly and Olson, 1999; Broman et al., 2003). Here, we briefly review the statistical model for this method.

The statistical model for the interval mapping method consists of two parts. One part determines the joint distribution of the genotypes at the marker loci and the putative QTL. Because the positions of the marker loci are known, once the position $\gamma$ of the QTL is given (as an unknown parameter), we can calculate the joint distribution $\mathrm{P}(\boldsymbol{z}_i, z_i^*; \gamma)$ of the $m+1$ genotypes $(\boldsymbol{z}_i, z_i^*) = (z_i^{(1)}, \ldots, z_i^{(m)}, z_i^*) \in \{-1, 0, 1\}^{m+1}$ as a function of $\gamma$ by using the stochastic structure of the linkage. Similarly, we can calculate the joint distribution $\mathrm{P}(\boldsymbol{z}_i)$, and then the conditional distribution of $z_i^*$ given $\boldsymbol{z}_i$ can be obtained as

$$\mathrm{P}(z_i^* | \boldsymbol{z}_i; \gamma) = \frac{\mathrm{P}(\boldsymbol{z}_i, z_i^*; \gamma)}{\mathrm{P}(\boldsymbol{z}_i)}. \tag{6}$$

Appendix A.2 provides details of (6) under Haldane's linkage model.

The second part of the interval mapping method models the stochastic behavior of phenotype $y_i$ when genotype $z_i^*$ of the QTL and covariate $u_i$ are given. Various statistical models can be constructed for different types of $y_i$. Since the concept of influence analysis can easily be extended to other settings, we only consider real-valued traits and use the normal linear model.

For individuals $i = 1, \ldots, n$, assume that

$$z_i^* | \boldsymbol{z}_i \sim \mathrm{P}(\cdot | \boldsymbol{z}_i; \gamma), \qquad y_i | (\boldsymbol{z}_i, z_i^*, u_i) \sim N\big(\alpha z_i^* + \beta w(z_i^*) + \mu + \nu u_i, \sigma^2\big), \tag{7}$$

where

$$w(z) = \begin{cases} 1 & (z = \pm 1), \\ -1 & (z = 0), \end{cases}$$

and $\boldsymbol{\theta} = (\alpha, \beta, \mu, \nu, \sigma^2)$ are unknown parameters. Then, the probability density function of $y_i$, given $(\boldsymbol{z}_i, u_i)$, becomes a three-component finite mixture of normal distributions

$$f(y_i | \boldsymbol{z}_i, u_i; \gamma, \boldsymbol{\theta}) = \sum_{z_i^* = -1}^{1} g(y_i | z_i^*, u_i; \boldsymbol{\theta}) \mathrm{P}(z_i^* | \boldsymbol{z}_i; \gamma), \tag{8}$$

where $g(y_i | z_i^*, u_i; \boldsymbol{\theta})$ is the density function of the normal distribution in (7).

In model (7), the parameters $\alpha$ and $\beta$ indicate the additive and dominance effects of the QTL, respectively. The null hypothesis $H_0 : \alpha = \beta = 0$ means that no QTL affects the phenotype. Note that when $H_0$ is true, the density $g(y_i | z_i^*, u_i; \boldsymbol{\theta})$ in (8) becomes unrelated to $z_i^*$, and the QTL location parameter $\gamma$ is no longer identifiable. In the rest of the paper, we refer to the density (8) as $g(y_i | *, u_i; \boldsymbol{\theta})$ under $H_0$.

The LOD score is defined as the base 10 logarithm of the likelihood ratio (LR) for testing the null hypothesis $H_0$ when the QTL location is $\gamma$. The test statistic is defined as a function of $\gamma$. Let $\widehat{\boldsymbol{\theta}}(\gamma)$

be the maximum likelihood estimator (MLE) given $\gamma$, and let $\widetilde{\boldsymbol{\theta}}$ be the MLE under $H_0$. As mentioned above, $\widetilde{\boldsymbol{\theta}}$ is independent of $\gamma$. Thus, the LOD score in the interval mapping method is defined as

$$\mathsf{LOD}(\gamma) = \frac{n}{\log 10}\left\{L_n\big(\gamma, \widehat{\boldsymbol{\theta}}(\gamma)\big) - L_n\big(\widetilde{\boldsymbol{\theta}}\big)\right\}, \tag{9}$$

where $L_n\big(\gamma, \widehat{\boldsymbol{\theta}}(\gamma)\big) = \frac{1}{n}\sum_{i=1}^{n} \log f\big(y_i|\boldsymbol{z}_i, u_i; \gamma, \widehat{\boldsymbol{\theta}}(\gamma)\big)$, and $L_n(\widetilde{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \log g\big(y_i|*, u_i; \widetilde{\boldsymbol{\theta}}\big)$. Because $L_n(\gamma, \widetilde{\boldsymbol{\theta}})$ does not depend on $\gamma$, we denote it as $L_n(\widetilde{\boldsymbol{\theta}})$.

An EM algorithm for estimating $\widehat{\boldsymbol{\theta}}(\gamma)$ can be found in Lander and Botstein (1989), and $\widetilde{\boldsymbol{\theta}}$ can be obtained by using the ordinary least-square method. A LOD score curve obtained from the interval mapping method is shown in Figure 2A.

### 3.3 Influence functions in the interval mapping method

Inside the braces of (9), the first term is the profile likelihood with respect to $\gamma$, and the second term is the ordinary likelihood. Although both are conditional likelihoods given $(\boldsymbol{z}_i, u_i)$, the general theory discussed in Section 2 is valid without any alteration. The theorem below follows immediately from Theorem 2.1 and the definition of the EIF in (2).

**Theorem 3.1.**

*(i) The empirical influence function of a linear combination of k LOD scores* $\mathsf{LODC}(\boldsymbol{c}) = \sum_{j=1}^{k} c_j \mathsf{LOD}(\gamma_j)$ *with the coefficient vector* $\boldsymbol{c} = (c_j)$ *for individual i is*

$$\mathsf{EIFC}_i(\boldsymbol{c}) = \mathsf{EIF}\big(i; \mathsf{LODC}(\boldsymbol{c})\big) = \frac{n}{\log 10}\sum_{j=1}^{k} c_j\left\{\ell_i\big(\gamma_j, \widehat{\boldsymbol{\theta}}(\gamma_j)\big) - \ell_{i0}\big(\widetilde{\boldsymbol{\theta}}\big)\right\} - \mathsf{LODC}(\boldsymbol{c}), \tag{10}$$

*where* $\ell_i(\gamma, \boldsymbol{\theta}) = \log f(y_i|\boldsymbol{z}_i, u_i; \gamma, \boldsymbol{\theta})$ *and* $\ell_{i0}(\boldsymbol{\theta}) = \log g(y_i|*, u_i; \boldsymbol{\theta})$. *In particular,*

$$\mathsf{EIF}(i; \mathsf{LOD}(\gamma)) = n\{\ell_i(\gamma, \widehat{\boldsymbol{\theta}}(\gamma)) - \ell_{i0}(\widetilde{\boldsymbol{\theta}})\}/\log 10 - \mathsf{LOD}(\gamma).$$

*(ii) The empirical influence function of the maximum score* $\max_{\gamma\in\Gamma} \mathsf{LOD}(\gamma) = \mathsf{LOD}(\widehat{\gamma})$ *for individual i is*

$$\mathsf{EIF}\big(i; \mathsf{LOD}(\widehat{\gamma})\big) = \frac{n}{\log 10}\left\{\ell_i\big(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}\big) - \ell_{i0}\big(\widetilde{\boldsymbol{\theta}}\big)\right\} - \mathsf{LOD}(\widehat{\gamma}), \tag{11}$$

*where* $(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}) = \operatorname{argmax}_{(\gamma,\theta)\in\Gamma\times\Theta} L_n(\gamma, \boldsymbol{\theta})$.

*(iii) The empirical influence function of the maximizer* $\widehat{\gamma}$ *for individual i is*

$$\mathsf{EIF}(i; \widehat{\gamma}) = \frac{\ell_{i,\gamma}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}) - L_{n,\gamma\theta}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}})L_{n,\theta\theta}^{-1}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}})\ell_{i,\theta}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}})}{-L_{n,\gamma\gamma}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}) + L_{n,\gamma\theta}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}})L_{n,\theta\theta}^{-1}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}})L_{n,\theta\gamma}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}})}, \tag{12}$$

*where the subscripts indicate partial derivatives.*

The derivatives with respect to $\gamma$ and/or $\boldsymbol{\theta}$ can be computed numerically or analytically.

### 3.4 Statistical model and influence functions in single marker analysis

In the interval mapping method, the putative QTL is assumed to be located in a continuous region $\Gamma$. However, when the observed markers are sufficient in number and densely located, we can assume that the QTL is one of the marker loci and use a simpler method—single marker analysis. This method is actually a special case of the interval mapping method.

Assume that the possible QTL region $\Gamma$ is given by the locations of the marker loci $\{d_1, \dots, d_m\}$. When $\gamma = d_j$ for some $j$, the conditional probability (6) becomes

$$\mathrm{P}(z_i^* | \boldsymbol{z}_i; \gamma) = \begin{cases} 1 & \left(\text{if } z_i^* = z_i^{(j)}\right), \\ 0 & (\text{otherwise}). \end{cases}$$

Then, the statistical model in (7) is reduced to the following single marker analysis model:

$$y_i | (\boldsymbol{z}_i, u_i) \sim N\left(\alpha z_i^{(j)} + \beta w\left(z_i^{(j)}\right) + \mu + \nu u_i, \sigma^2\right), \quad i = 1, \dots, n, \tag{13}$$

where $z_i^{(j)}$ is the marker genotype observed at location $d_j \in \Gamma$.

As in the interval mapping method, the parameters $\alpha$ and $\beta$ in model (13) indicate the additive effect and the dominance effect, respectively. We again consider the null hypothesis $H_0 : \alpha = \beta = 0$. Let $\widehat{\boldsymbol{\theta}}(\gamma)$ be the MLE given $\gamma$, and let $\widetilde{\boldsymbol{\theta}}$ be the MLE under $H_0$. We can again use (9) to define the LOD score, but $\gamma$ is restricted to $\Gamma = \{d_1, \dots, d_m\}$, and $\widehat{\gamma} = \mathrm{argmax}_{\gamma \in \Gamma} \mathsf{LOD}(\gamma)$. Similarly, the EIFs of the score $\mathsf{LOD}(\gamma)$ at individual $i$ can be obtained from (10) and its maximum $\mathsf{LOD}(\widehat{\gamma})$ is given in (11). The EIFs for $\widehat{\gamma}$ cannot be defined since $\widehat{\gamma}$ takes discrete values.

## 4 Influence analysis on aspects of the shape

### 4.1 Influence analyses for the shape of LOD score curves

If we are only interested in the score $\mathsf{LOD}(\gamma)$ at a particular location $\gamma$, we can calculate $\mathsf{EIF}_i(\mathsf{LOD}(\gamma))$ for each individual $i$ from (10) and conclude that individuals having large absolute values $|\mathsf{EIF}_i(\mathsf{LOD}(\gamma))|$ influence the LOD score. However, as stated in Section 1, attention is paid not only to the value of the score at a particular $\gamma$ but also to the shape of the score curve in genetics studies.

A conventional approach to find influential individuals for the shape of the LOD score curve in an experimental population generated from a genetic cross is to examine the individual genotype data in the region in question. Because of the linkage, strong positive correlations exist among the genotypes $z_i^{(1)}, \dots, z_i^{(k)}$ $(1 \leq k \leq m)$ on a linked chromosomal region. The probability that flanking markers take the same genotypes (e.g., $z_i^{(j)} = z_i^{(j+1)}$) is close to 1. If some individuals have recombinant genotypes at some markers in the linked region of interest, they may influence the shape of the LOD score curve. Therefore, geneticists try to identify individuals that show recombinant genotypes near that region. However, this method has some difficulties. First, because recombinant genotypes may have many different patterns, identifying all the potentially influential patterns is difficult. Second, because this method does not take phenotype and covariates into consideration, the detected individuals are not necessarily influential.

In this section, to detect individuals having a large influence on a particular shape of the LOD score curve, we propose to use $\mathsf{EIFC}_i(\boldsymbol{c})$, the EIF of $\mathsf{LODC}(\boldsymbol{c}) = \sum_j c_j \mathsf{LOD}(\gamma_j)$ for individual $i$. Regardless of whether the set $\Gamma$ of possible QTL locations is continuous or discrete, we are interested in the shape of the LOD score curve in the region where $k$ chromosome positions $\gamma_1, \dots, \gamma_k$ are located. The influence matrix is defined as an $n \times k$ matrix

$$\mathbf{EIF} = \left(\mathbf{EIF}(\gamma_1), \dots, \mathbf{EIF}(\gamma_k)\right), \quad \text{where } \mathbf{EIF}(\gamma) = \left(\mathsf{EIF}_1(\mathsf{LOD}(\gamma)), \dots, \mathsf{EIF}_n(\mathsf{LOD}(\gamma))\right)'$$

is the $n \times 1$ EIF vector of $\mathsf{LOD}(\gamma)$. We also define $\mathbf{EIF}_i = \left(\mathsf{EIF}_i(\mathsf{LOD}(\gamma_1)), \dots, \mathsf{EIF}_i(\mathsf{LOD}(\gamma_k))\right)'$, the transpose of the $i$-th row vector of $\mathbf{EIF}$.

In this approach, the choice of the coefficient vector $\boldsymbol{c} = (c_j)_{1 \leq j \leq k}$ is crucial. For example, if we want to examine a linear increasing trend in LOD scores $(\mathsf{LOD}(\gamma_1), \mathsf{LOD}(\gamma_2), \mathsf{LOD}(\gamma_3))$ at 3 loci, then a monotone coefficient vector $\boldsymbol{c} = (c_1, c_2, c_3)'$ with $(c_1 < c_2 < c_3)$ can be used for this purpose. The following are three strategies for setting the coefficients $\boldsymbol{c}$ systematically, when we do not have a clear idea for choosing $\boldsymbol{c}$. Note that the coefficients $\boldsymbol{c}$ should be chosen by taking the selection of locations $\gamma_j$ into consideration. We let $\mathbf{Cov}$ be the covariance matrix of $\big(\mathsf{LOD}(\gamma_1), \ldots, \mathsf{LOD}(\gamma_k)\big)$, and use the matrix $\mathbf{Cov}$ as the metric for standardizing EIFs.

Our first proposal is the use of orthogonal polynomials. Let $\boldsymbol{c}_l^* = \mathbf{Cov}^{-1}((\gamma_1)^l, \ldots, (\gamma_k)^l)'$ ($l = 0, 1, \ldots$), where $(\cdot)^l$ is the $l$-th power. Let $\boldsymbol{c}_0 = \boldsymbol{c}_0^* / \sqrt{\boldsymbol{c}_0^{*\prime}\,\mathbf{Cov}\,\boldsymbol{c}_0^*}$. Applying the Gram–Schmidt orthonormalization process to $\boldsymbol{c}_l^*$'s, we define $\boldsymbol{c}_l$ sequentially by

$$\widetilde{\boldsymbol{c}}_l = \boldsymbol{c}_l^* - \sum_{k=0}^{l-1} (\boldsymbol{c}_k'\,\mathbf{Cov}\,\boldsymbol{c}_l^*)\,\boldsymbol{c}_k, \qquad \boldsymbol{c}_l = \frac{1}{\sqrt{\widetilde{\boldsymbol{c}}_l'\,\mathbf{Cov}\,\widetilde{\boldsymbol{c}}_l}}\widetilde{\boldsymbol{c}}_l \qquad (l = 1, 2, \ldots). \tag{14}$$

The process (14) is written as a recursion form

$$\boldsymbol{H}_l = \boldsymbol{H}_{l-1} - \mathbf{Cov}\,\boldsymbol{c}_{l-1}\boldsymbol{c}_{l-1}', \qquad \boldsymbol{c}_l = \frac{1}{\sqrt{\boldsymbol{c}_l^{*\prime}\boldsymbol{H}_l\,\mathbf{Cov}\,\boldsymbol{H}_l'\boldsymbol{c}_l^*}}\boldsymbol{H}_l'\boldsymbol{c}_l^* \qquad (l = 1, 2, \ldots),$$

where $\boldsymbol{H}_0 = \boldsymbol{I}_k$. Note that $\boldsymbol{H}_l$ is an orthogonal projection matrix. The vectors $\boldsymbol{c}_l$ are orthonormal in the sense that $\boldsymbol{c}_k'\,\mathbf{Cov}\,\boldsymbol{c}_l = 1$ ($k = l$), 0 ($k \neq l$). Hence, the $\mathsf{LODC}(\boldsymbol{c}_l)$ are uncorrelated between different $l$'s. Then $\mathsf{EIFC}_i(\boldsymbol{c}_l)$ ($i = 1, \ldots, n$) can be obtained from (10). It also can be written as $\mathsf{EIFC}_i(\boldsymbol{c}_l) = \boldsymbol{c}_l'\,\mathbf{EIF}_i$. One advantage of this method is that the coefficient vectors are easy to interpret. The vectors $\boldsymbol{c}_0$, $\boldsymbol{c}_1$, and $\boldsymbol{c}_2$ are corresponding to the grand mean, linearity, and curvature, respectively.

Alternatives to using specific contrasts are our second and third proposals given below. Note that the matrices $\boldsymbol{H}_l$ defined in (14) can be used for deleting the components that are *not* of interest. For example, for deleting the parallel shift or linear components, we can consider a class of linear combinations of EIFs defined by a projection $\mathsf{LODC}(\boldsymbol{H}'\boldsymbol{c}) = \boldsymbol{c}'\boldsymbol{H}(\mathsf{LOD}(\gamma_1), \ldots, \mathsf{LOD}(\gamma_k))'$ with $\boldsymbol{H} = \boldsymbol{H}_1$ or $\boldsymbol{H}_2$, respectively. In the following discussion, we start from this projected LOD score.

Our second proposal is based on the PCA of Lu et al. (1997) and Tanaka (1994). Noting that the variance of $\mathsf{LODC}(\boldsymbol{H}'\boldsymbol{c})$ is $\boldsymbol{c}'(\boldsymbol{H}\mathbf{Cov}\boldsymbol{H}')\boldsymbol{c}$, we consider the following singular value decomposition for the standardized influence matrix:

$$\mathbf{EIF}\boldsymbol{H}'\big(\boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'\big)^{-\frac{1}{2}} = \sum_l \sqrt{\lambda_l}\,\boldsymbol{h}_l\,\boldsymbol{u}_l', \tag{15}$$

where $\boldsymbol{h}_l$ and $\boldsymbol{u}_l$ are orthonormal $n$- and $k$-vectors, respectively, and $(\cdot)^{-\frac{1}{2}}$ is the symmetric square root matrix of the Moore–Penrose pseudoinverse matrix. Multiplying both sides of (15) by $\boldsymbol{u}_l$ and defining $\boldsymbol{c}_l := \big(\boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'\big)^{-1/2}\boldsymbol{u}_l$, we have $\mathbf{EIF}\boldsymbol{H}'\big(\boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'\big)^{-1/2}\boldsymbol{u}_l = \mathbf{EIF}(\boldsymbol{H}'\boldsymbol{c}_l) = \sqrt{\lambda_l}\,\boldsymbol{h}_l$. Then $\boldsymbol{H}'\boldsymbol{c}_l$ becomes the coefficient vector corresponding to the $l$-th component, and the principal component $\sqrt{\lambda_l}\boldsymbol{h}_l$ is the corresponding influence function. We refer to $\sqrt{\lambda_l}\boldsymbol{h}_l$ as the *influence score vector* in the context of influence analysis. For individual $i$, the EIF with respect to the $l$-th component can be written as $\mathsf{EIFC}_i(\boldsymbol{H}'\boldsymbol{c}_l) = \boldsymbol{c}_l'\boldsymbol{H}\mathbf{EIF}_i = \sqrt{\lambda_l}(\boldsymbol{h}_l)_i$.

This is an exploratory approach, and as in most cases of PCA, the result of this approach is not always easy to interpret. Thus, we recommend that the number of principal components, $\mathrm{rank}(\boldsymbol{H})$, not to be large.

Our third proposal is to use the coefficients $\boldsymbol{c}$ maximizing $\mathsf{EIFC}_i(\boldsymbol{H}'\boldsymbol{c}) = \boldsymbol{c}'\boldsymbol{H}\,\mathbf{EIF}_i$ under the condition $\mathrm{Var}(\mathsf{LODC}(\boldsymbol{H}'\boldsymbol{c})) = \boldsymbol{c}'\big(\boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'\big)\boldsymbol{c} = 1$. Define $\boldsymbol{a} = \boldsymbol{H}\,\mathbf{EIF}_i$, $\boldsymbol{Q} = \boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'$, and let $\boldsymbol{c}'\boldsymbol{Q}\boldsymbol{c} = 1$. Using the Cauchy–Schwarz inequality, we have

$$\boldsymbol{c}'\boldsymbol{a} = (\boldsymbol{Q}^{1/2}\boldsymbol{c})'\,(\boldsymbol{Q}^{-1/2}\boldsymbol{a}) \leq \sqrt{\boldsymbol{c}'\boldsymbol{Q}\boldsymbol{c}} \cdot \sqrt{\boldsymbol{a}'\boldsymbol{Q}^{-}\boldsymbol{a}} = \sqrt{\boldsymbol{a}'\boldsymbol{Q}^{-}\boldsymbol{a}}\,.$$

Because the maximizer is $\boldsymbol{c} = \kappa \boldsymbol{Q}^{-}\boldsymbol{a}$ for a constant $\kappa \geq 0$, we find that

$$\kappa = \frac{1}{\sqrt{\boldsymbol{a}'\boldsymbol{Q}^{-}\boldsymbol{a}}}, \qquad \boldsymbol{c} = \frac{\boldsymbol{Q}^{-}\boldsymbol{a}}{\sqrt{\boldsymbol{a}'\boldsymbol{Q}^{-}\boldsymbol{a}}} = \frac{\left(\boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'\right)^{-}\boldsymbol{H}\,\mathbf{EIF}_i}{\sqrt{\mathbf{EIF}_i'\boldsymbol{H}'\left(\boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'\right)^{-}\boldsymbol{H}\,\mathbf{EIF}_i}}$$

and the maximum is the square root of the quadratic form below:

$$\mathsf{QEIF}_i = \mathbf{EIF}_i'\boldsymbol{H}'\left(\boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'\right)^{-}\boldsymbol{H}\,\mathbf{EIF}_i = \left(\mathbf{EIF}\,\boldsymbol{H}'\left(\boldsymbol{H}\,\mathbf{Cov}\,\boldsymbol{H}'\right)^{-}\boldsymbol{H}\,\mathbf{EIF}'\right)_{ii}. \tag{16}$$

Note that $\mathsf{QEIF}_i$ can be rewritten in a way of the PCA-based method: $\mathsf{QEIF}_i = \sum_{l=1}^{k}(\mathsf{EIFC}_i(\boldsymbol{H}'\boldsymbol{c}_l))^2 = \sum_{l=1}^{k}\lambda_l\,(\boldsymbol{h}_l)_i^2$. We refer to this EIF as the quadratic EIF (QEIF) hereinafter. This alternative method can be used when PCA does not give an explicable result.

### 4.2 Approximation of Cov

For the covariance matrix **Cov**, we propose an approximate value evaluated under the null hypothesis $H_0$. It might seem better to use the covariance under the non-null model. However, in QTL analysis, the single-QTL models (7) or (13) are used for defining only the scanning statistic (LOD score), and each column vector $\mathsf{EIF}(\gamma_j)$ comes from a different statistical model for the specified $\gamma_j$. Hence, model-based approaches such as the use of observed information or Cook's local influence (Cook, 1986) cannot be applied to our problem. Therefore, we use the covariance under the null model as a second-best alternative.

To estimate **Cov**, we define two $n \times 2$ matrices: $\boldsymbol{Z}(\gamma) = \left(\bar{z}_i(\gamma), \bar{w}_i(\gamma)\right)_{1\leq i\leq n}$ and $\boldsymbol{U} = \left(1, u_i\right)_{1\leq i\leq n}$, with $\bar{z}_i(\gamma) = \sum_{z_i^*=-1}^{1} z_i^*\mathrm{P}(z_i^*|\boldsymbol{z}_i;\gamma)$ and $\bar{w}_i(\gamma) = \sum_{z_i^*=-1}^{1} w(z_i^*)\mathrm{P}(z_i^*|\boldsymbol{z}_i;\gamma)$. We propose an approximation for **Cov** as follows. The regularity conditions and the proof are provided in Appendix A.3.

**Theorem 4.1.** *When n is large, the covariance matrix* **Cov** *of the LOD scores under the assumption of no QTL can be approximated by*

$$(\mathbf{Cov})_{jl} = \mathrm{Cov}\big(\mathsf{LOD}(\gamma_j), \mathsf{LOD}(\gamma_l)\big) \doteq \frac{2}{(2\log 10)^2}\mathrm{tr}\big(\boldsymbol{R}(\gamma_j)\boldsymbol{R}(\gamma_l)\big), \quad j, l \in \{1, \ldots, k\}, \tag{17}$$

*where*

$$\boldsymbol{R}(\gamma) = \big(\boldsymbol{Z}(\gamma)\ \boldsymbol{U}\big)\left\{\begin{pmatrix}\boldsymbol{Z}(\gamma)'\\ \boldsymbol{U}'\end{pmatrix}\big(\boldsymbol{Z}(\gamma)\quad \boldsymbol{U}\big)\right\}^{-1}\begin{pmatrix}\boldsymbol{Z}(\gamma)'\\ \boldsymbol{U}'\end{pmatrix} - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'. \tag{18}$$

### 4.3 Significance of EIFs

To assess the statistical significance of the EIFs for suspected individuals, we calculate their $p$-values in the framework of multiple testing problem without identifying the unknown true QTL model. In this case, nonparametric resampling approaches look favorable. However, they have difficulty in estimating tail probabilities of extreme statistics. To address this problem, we propose a robustified parametric bootstrap below.

First, estimate the involved parameters from the given (observed) dataset by a robust regression (such as, using the Huber estimator) for the full regression model

$$y_i = \mu + \sum_{j\in J}\big(\alpha_j z_i^{(j)} + \beta_j w(z_i^{(j)})\big) + \nu u_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \tag{19}$$

where $J$ is a region including all possible QTLs of interest (see, e.g., Section 5.3). The effects from QTLs outside $J$ can be regarded as a part of $\varepsilon_i$. Second, generate datasets with the same sample size as in the dataset using the estimated model (19) and Haldane's linkage model. Then find the maximum absolute EIF from each simulated dataset. Finally, the $p$-value can be calculated as the upper probability of the empirical distribution of the maximum EIFs. Individuals with relatively small $p$-values can be declared influential.

The validity of this method is examined in Section 6.1. In the next section, we will use $\mathsf{SEIF}_i = \mathsf{EIFC}_i/\sqrt{\sum_{i=1}^{n}\mathsf{EIFC}_i^2}$, the standardized version of $\mathsf{EIFC}_i$, instead of the original ones for estimating $p$-values. This is because an additional numerical comparison study we conducted shows that the original EIFC sometimes gives too conservative $p$-values when no outlier exists. The proposed method is also applicable to the interval mapping method by generating datasets with model (19) containing all marker loci on the chromosome under consideration.

## 5 Analysis of $F_2$ mice data

### 5.1 $F_2$ mice data

In this section, we show how our proposed methods work with a real dataset, which is available from Supporting Information. Our data are taken from $n = 170$ $F_2$ progeny generated from the intercross of $F_1$ hybrids of C57BL/6J and MSM/Ms mouse strains. Extensive phenotypic variation and great genomic diversity are observed between these two strains (Takada et al., 2008). In this analysis, we consider blood adiponectin concentration ($\log_{10}$[ng/ml]) as a quantitative trait. Adiponectin is a key adipokine in metabolic syndrome and is important in mammalian metabolism. Genotypes of the $F_2$ progeny were observed at $m = 119$ marker loci (including 94 SNP markers and 25 microsatellite markers). The LOD score curves obtained by single marker analysis are depicted in Figure 1. The LOD score curve at chromosome 16 attains the maximum. In fact, the adiponectin coding gene (*Adipoq*, MGI:106675) is located on this chromosome. In our analysis below, we focus on chromosome 3 where six SNP marker loci are genotyped.

### 5.2 Data analysis on a single location

The LOD score curve obtained from the interval mapping method is shown in Figure 2A. We chose the point ($\widehat{\gamma} = 32.2$, $\mathsf{LOD}(\widehat{\gamma}) = 3.55$) for further analysis because it is the maximum peak point of the LOD score curve.

The EIF values of the maximum (11) and the maximizer (12) of the LOD score for the 170 individuals are plotted in Figure 2B. In this figure, each circle corresponds to one mouse. The horizontal axis indicates the influence of each mouse on the maximum LOD score, and the vertical axis measures the influence of all the mice on the location of the putative QTL.

In order to show the extent of the influence of individuals on the LOD score, we plot the LOD score curves obtained without the specific mice and compare them with the original LOD score curve. The extreme points in the horizontal direction are No. 13 and No. 63. Accordingly, in Figure 2C, the maximum LOD score decreases and increases over a wide range of values by deleting mice No. 13 and No. 63. Figure 2D shows that deletion of mice No. 62 or No. 94 leads to a peak location shift from side to side without changing the maximum LOD score substantially. In particular, the data for No. 94 moves the location of the putative QTL by 3.2 centi-Morgans (cM). Although these mice are not detected as influential using the method in Section 4.3, we can see from this example that influential individuals may mislead us when making decisions about the existence and the locations of the QTLs.

### 5.3 Data analysis on aspects of the shape

For the LOD score curve on chromosome 3 in Figure 1, let us focus on the locations of the peaks and valley $(\gamma_3, \gamma_4, \gamma_5) = (35.4, 53.8, 68.4)$ cM, and check whether any mice determine the two-peak shape.

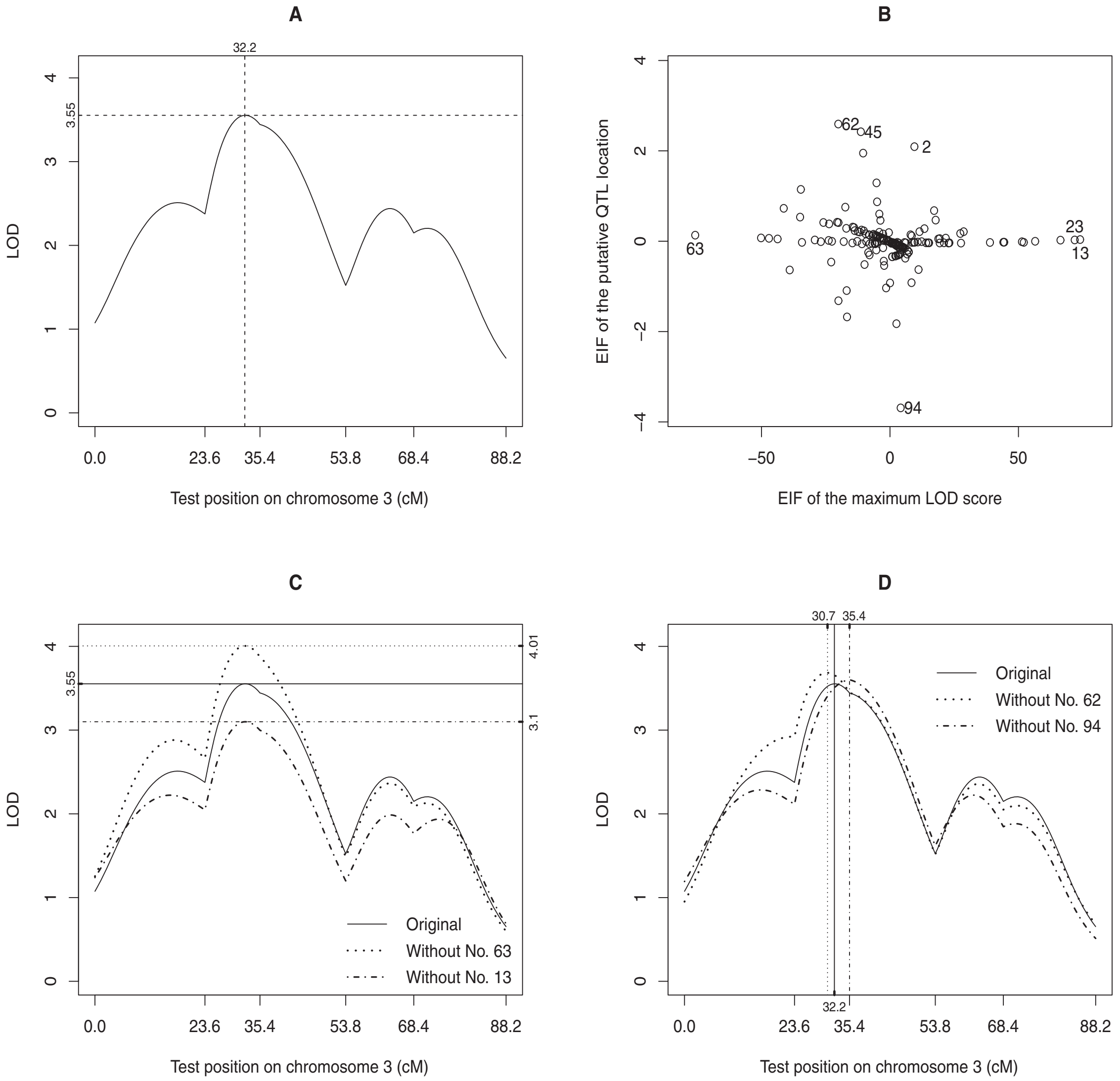

**Figure 2** The LOD score curve for chromosome 3 and influence analyses on the maximum point of the curve. (A) The solid line is the LOD score curve obtained from the interval mapping method. The LOD score curve attains maximum of 3.55 at 32.2 cM. (B) EIF of the maximum LOD score (11) and EIF of the putative QTL location (12) for each mouse. The mice with identification numbers are candidate influential individuals. (C) The maximum LOD score is changed by removing the two mice. (D) The location of the maximum LOD score is moved by removing the two mice.

From (14), we can calculate a vector of $\boldsymbol{c}_l$ to detect the influential mice that significantly affect the parallel shift, inclination, or curvature of the LOD score curve. For the curvature, vector $\boldsymbol{c}_2 = (1.04, -2.36, 1.31)'$ is obtained. Figure 3A depicts the $\mathsf{SEIF}_i$ obtained from the orthogonal polynomial method with $\boldsymbol{c}_2$ for the 170 individuals. Among the individuals, mouse No. 60 has the largest SEIF of 0.447. Its phenotype is 4.776, and genotypes at the 3 loci $(\gamma_3, \gamma_4, \gamma_5)$ are $(0, -1, 0)$. This mouse appears to greatly affect the curvature of the LOD score curve (Figure 3C), and thus is worthy of attention.

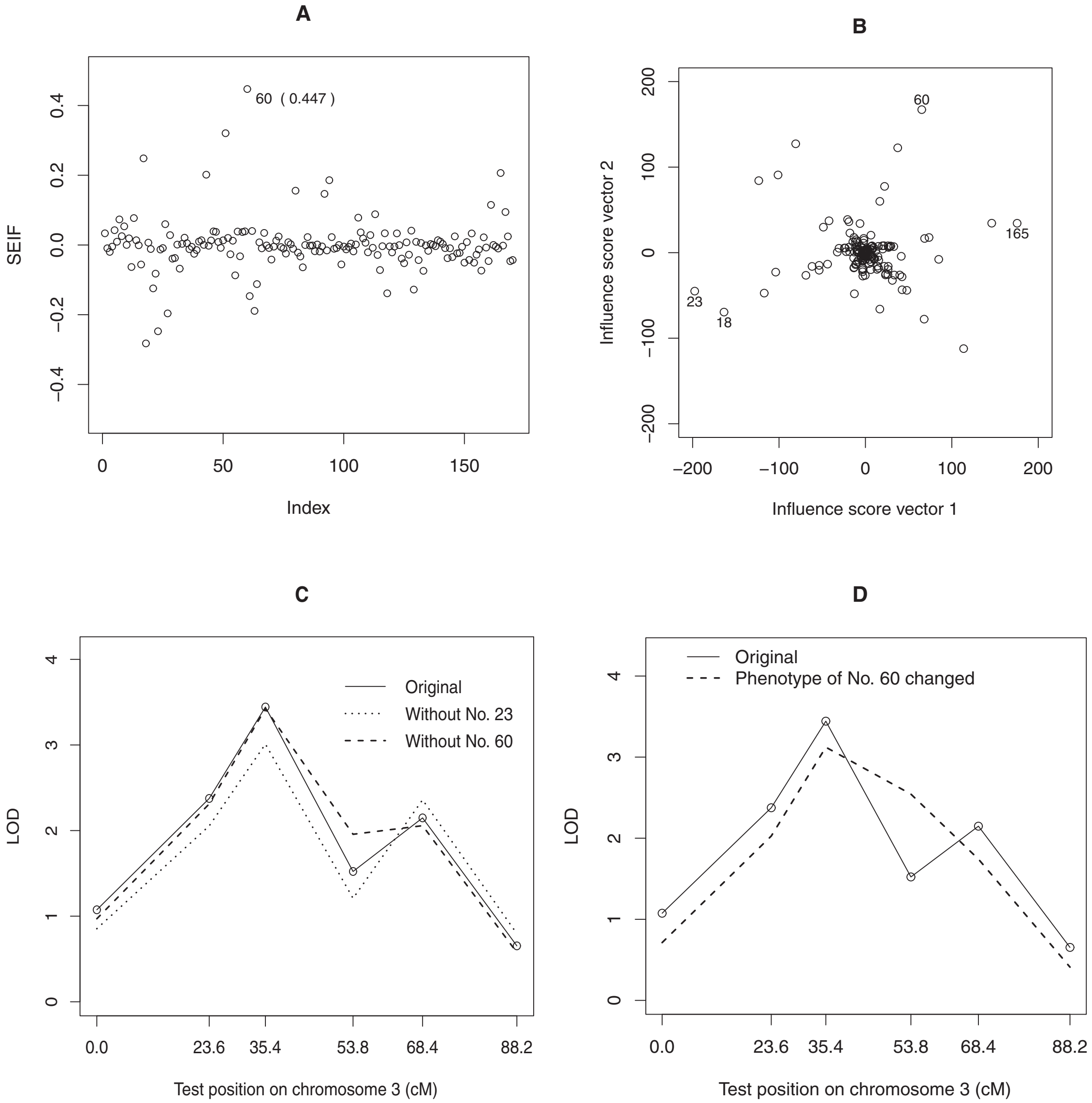

**Figure 3** Influence analyses for the shape of the LOD score curve. (A) SEIFs of the 170 mice obtained by the orthogonal polynomial method. Each circle indicates one mouse. (B) Scatter chart of the two influence score vectors obtained from the PCA method. The mice with identification numbers have a noticeably larger impact on the shape of the LOD score curve on chromosome 3. (C) LOD score curves without mice No. 23 and No. 60. (D) Comparison of the LOD score curves obtained from the original dataset and the changed dataset.

To assess its significance, we conduct a parametric bootstrap according to the procedure in Section 4.3. We use Huber's robust method to estimate the regression coefficients and the error variance in (19). In our study, the `R` function `rlm()` with the option `psi=psi.huber` is used. Under model (19) with $J = \{3, 4, 5\}$, based on 1000 repeated simulations, the $p$-value for mouse No. 60 is estimated as 0.39. Model (19) with all loci on chromosome 3 (i.e., $J = \{1, 2, \ldots, 6\}$) gives similar result. Accordingly, No. 60 is not an influential mouse.

Using the PCA approach, we again search influential individuals for the shape of the LOD score curve on chromosome 3. In this method, projection matrix $\boldsymbol{H} = \boldsymbol{H}_1$ in (14) is used to remove the parallel shift of the LOD score curve. The scatter plot of the two influence score vectors is shown in Figure 3B.

The correlations between the eigenvectors $\boldsymbol{h}_l = (h_{li})_{1\le i\le n}$ and $\mathsf{EIFC}(\boldsymbol{c}_l) = \big(\mathsf{EIFC}_i(\boldsymbol{c}_l)\big)_{1\le i\le n}$ for $l = 1, 2$, obtained in the previous example are

$$\mathrm{Corr}\big((\boldsymbol{h}_1, \boldsymbol{h}_2), (\mathsf{EIFC}(\boldsymbol{c}_1), \mathsf{EIFC}(\boldsymbol{c}_2))\big) = \begin{pmatrix} 0.977 & 0.366 \\ -0.215 & 0.931 \end{pmatrix}.$$

This result suggests that the first eigenvector indicates the inclination and the second eigenvector indicates the curvature of the LOD score curve, even though they do not work in exactly the same way as the vectors $\boldsymbol{c}_l$.

The influence from mice No. 23 and No. 60 is shown in Figure 3C by comparing the original LOD score curve with those obtained from the dataset without these mice. The solid line connects the original LOD scores to form the curve for chromosome 3. The dotted line is obtained by removing mouse No. 23, which has the largest absolute value at the first eigenvector. We now observe that the first peak and the valley have gone down but the second peak has gone up. The dashed line indicates the LOD score curve without mouse No. 60, which has the largest value at the second eigenvector. In this case, the valley of this curve disappeared almost completely.

In our dataset, mouse No. 60 is not approved as influential. However, when we change its phenotype (4.776) to 3.842, the minimum phenotype in the dataset, its SEIF becomes −0.647, and the fictional mouse is detected as influential according to its small $p$-value 0.02 (see the LOD score curves in Figure 3D). This numerical experiment shows that the proposed method for $p$-value in Section 4.3 can accomplish its goal.

## 6 Simulation studies

### 6.1 Assessing the method in Section 4.3

In this subsection, we examine the validity of our method for $p$-values from Section 4.3. We confirm that in the case of the single marker analysis, this method controls the false positive rate when no outlier exists, and has statistical power when outliers exist.

The outline of our simulation study consists of the following steps:

(i) Assume a true QTL model. Repeat the steps (ii)–(v) $N$ times.
(ii) Generate a dataset $\mathcal{D}_0$ from the assumed true model. Calculate $T_0 = \max|\mathsf{SEIF}_i|$ from $\mathcal{D}_0$, where $\mathsf{SEIF}_i = \mathsf{EIFC}_i/\sqrt{\sum_{i=1}^{n} \mathsf{EIFC}_i^2}$ with $\mathsf{EIFC}_i$ given in (10).
(iii) From the assumed model, generate another dataset $\mathcal{D}_1$ with one outlier included. Calculate $T_1 = \max|\mathsf{SEIF}_i|$ from $\mathcal{D}_1$ as in (ii).
(iv) As stated in Section 4.3, fit the full model (19) to the dataset $\mathcal{D}_0$ by Huber's method as in Sections 4.3 and 5.3, and generate a new dataset $\widetilde{\mathcal{D}}_0$ using the estimated robust regression parameters. Calculate $\widetilde{T}_0 = \max|\mathsf{SEIF}_i|$ from $\widetilde{\mathcal{D}}_0$.
(v) Apply the same procedure as in (iv) to the data $\mathcal{D}_1$ to obtain the dataset $\widetilde{\mathcal{D}}_1$ and $\widetilde{T}_1$.
(vi) Compare empirical distributions of $T_0$, $T_1$, $\widetilde{T}_0$, and $\widetilde{T}_1$ based on the $N$ iterations.

Because $\widetilde{T}_0$ and $\widetilde{T}_1$ are samples from the model with robust estimators, the distributions of both $\widetilde{T}_0$ and $\widetilde{T}_1$ are expected to approximate the distribution of $T_0$ (i.e., null distribution). As stated below, our simulation study shows that this expectation is correct. That is, it is appropriate to use $\widetilde{T}$ (either $\widetilde{T}_0$ or $\widetilde{T}_1$, in practice) instead of $T_0$. This means that our proposal for estimating *p*-values is validated.

As the true model in step (i), we use model (19) with four plausible settings: $M_0 : J = \emptyset$ (no QTL), $M_1 : J = \{3\}$ (one QTL at $\gamma_3$), $M_2 : J = \{3, 5\}$ (two QTLs at $\gamma_3$ and $\gamma_5$), and $M_3 : J = \{3, 4, 5\}$ (the full model). The last model is also used as the full model in steps (iv) and (v). In steps (ii)–(v), genotypes $z_i^{(j)}$ are generated with Haldane's model, sexes $u_i$ are produced as a Bernoulli sequence independently of $z_i^{(j)}$, phenotypes in $\mathcal{D}_0$ and $\mathcal{D}_1$ are generated form the true model, and phenotypes in $\widetilde{\mathcal{D}}_0$ and $\widetilde{\mathcal{D}}_1$are generated from the full model. In step (iii), the outlier is generated as follows: genotypes $(z_i^{(3)}, z_i^{(4)}, z_i^{(5)})$ at the third, fourth, and fifth loci are set as the relatively rare genotype sequence $(0, -1, 0)$, which seemed to have the largest effect in the dataset, and is the same as that of mouse No. 60; error $\varepsilon_i$ of the phenotype is generated from $N(0, (3\sigma)^2)$.

Figure 4 shows the simulation results for the four models. The empirical distribution functions of $T_k$ and $\widetilde{T}_k$ $(k = 0, 1)$ are depicted as black and gray lines, respectively. We find that the distribution of $\widetilde{T}_0$ and $\widetilde{T}_1$ are close to each other, which means that the distribution of $\widetilde{T}$ ($\widetilde{T}_0$ or $\widetilde{T}_1$) is stable for the existence of outliers. We also find that $\widetilde{T}_0$ and $\widetilde{T}_1$ are close to $T_0$ and distinct from $T_1$. This indicates that when no outlier exists the false positive rate is appropriately controlled (i.e., unbiased) and when outliers exist it has statistical power in detecting influential individuals. We also tried simulations with two outliers. The results are similar and omitted.

### 6.2 Power comparisons by ROC analysis

In this subsection, we confirm the statistical power of our proposals by comparing them with existing diagnostics in regression analysis. We assess four diagnostics, the EIFC (10), the QEIF (16), Cook's *D* (Cook, 1977), and the standardized residual *r*, by a receiver operating characteristic (ROC) analysis (Fawcett, 2006). In the context of QTL analysis, Hayat et al. (2008) studied the detection power of a modified version of Cook's *D* (Zewotir and Galpin, 2005) in a QTL model with random effects. Since our QTL model is a fixed-effect model, we use the original Cook's *D* in the comparisons. We restrict our attention again to detecting individuals that influence the two-peak shape of the LOD score curve on chromosome 3.

Cook's *D* and the standardized residual *r* are defined through the regression model (19) with $J = \{3, 4, 5\}$. Note that Cook's *D* is the quadratic form of the EIF vector for the parameter vector $((\alpha_j, \beta_j)_{j\in\{3,4,5\}}, \mu, \nu)$.

In this simulation, the two-QTL model (19) with $J = \{3, 5\}$ is used as the true model. Each dataset contains two outliers and 168 normal individuals. The outliers are designed to have specified genotypes $(0, -1, 0)$ at the third, fourth, and fifth loci. These genotypes are the same as those used in Section 6.1. The error $\varepsilon_i$ of each outlier's phenotype is simulated in the following different ways: (a) Normal distribution $N(0, (2\sigma)^2)$, (b) $N(0, (3\sigma)^2)$, (c) *t*-distribution with 3 df and scale parameter $2\sigma$, and (d) Cauchy distribution with scale parameter $2\sigma$. Note that $\sigma$ is the scale parameter used in generating $\varepsilon_i$ for the 168 normal individuals. In each dataset, the genotypes and sexes of the normal individuals are generated from Haldane's model and Bernoulli distribution, respectively. Their phenotypes are simulated from the two-QTL model with the parameters estimated from the adiponectin dataset.

The ROC curves of the four indicators are shown in Figure 5 based on 1000 replicates. For the varying thresholds, the average rates of correctly classifying the true influential cases (detection rates), and the average rates of misclassifying the normal cases as influential cases (false positive rates) are plotted. As expected, outliers with larger $\sigma^2$ have larger absolute EIFs, and are thus more easily detected. In all panels, the EIFC has the largest area under its ROC curve and hence the best average performance. As the second-best method, QEIF is shown useful when the target shape cannot be fully specified.

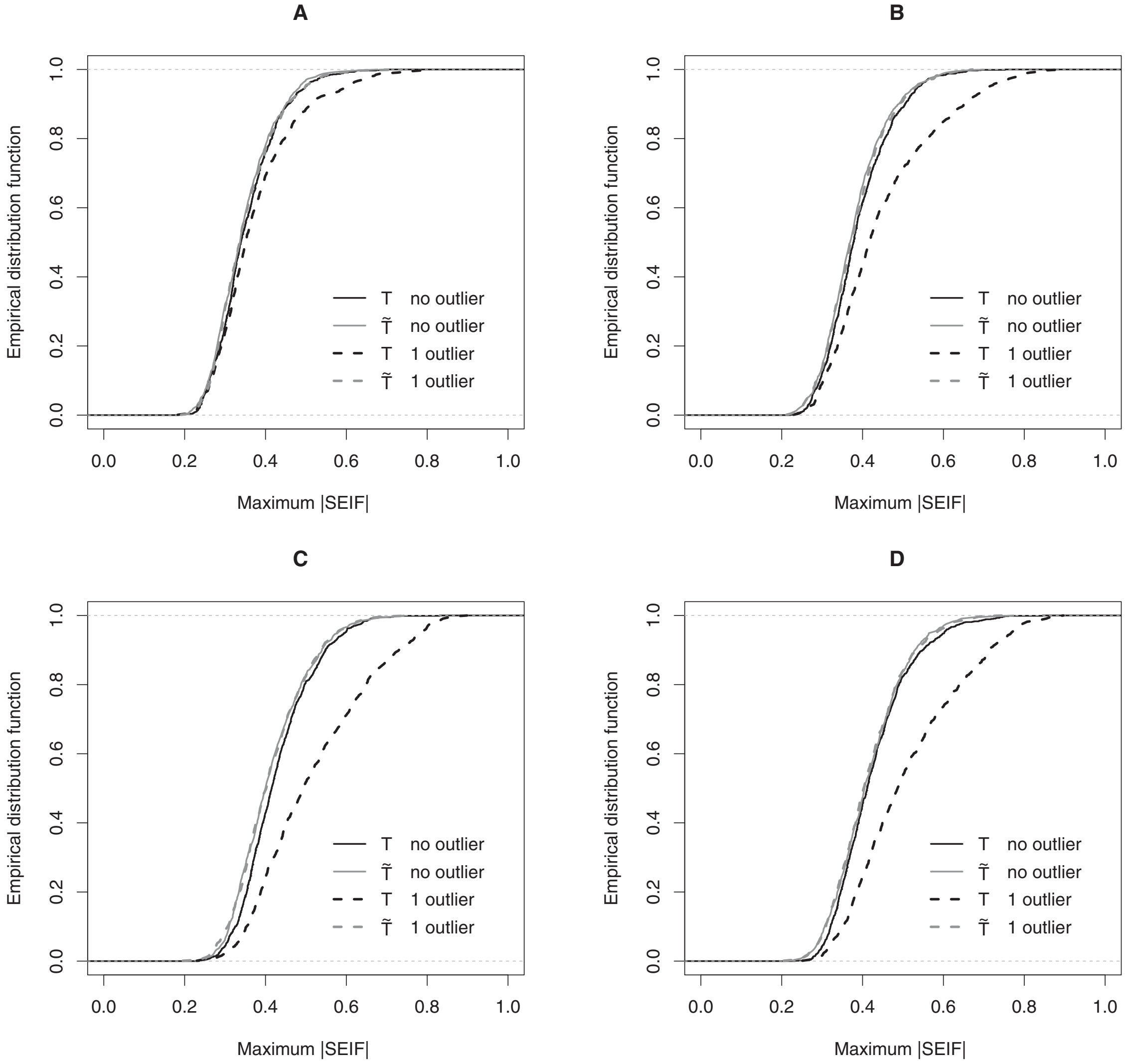

**Figure 4** Simulation results for Section 6.1. Distributions of the maximum |SEIF| and their approximations based on 1000 replicates. (A), (B), (C), and (D) are the simulation results under the models $M_0$, $M_1$, $M_2$, and $M_3$, respectively. Solid lines are the empirical distribution functions of simulated $T_0$ and their approximations $\widetilde{T}_0$ when there are no outliers. Dashed lines are empirical distribution functions of simulated $T_1$ and $\widetilde{T}_1$ when one outlier exists. ($T_0$, $\widetilde{T}_0$, $T_1$, $\widetilde{T}_1$ are referred to as "$T$ no outlier," "$\widetilde{T}$ no outlier," "$T$ 1 outlier," "$\widetilde{T}$ 1 outlier" in the legends, respectively.)

## 7 Discussion and guidelines

### 7.1 Summary and discussion

In this paper, we developed a general theory of profile likelihood function, and apply it to the linear functional of LOD score function. We also proposed methods to detect influential individuals to the shape of LOD score curves.

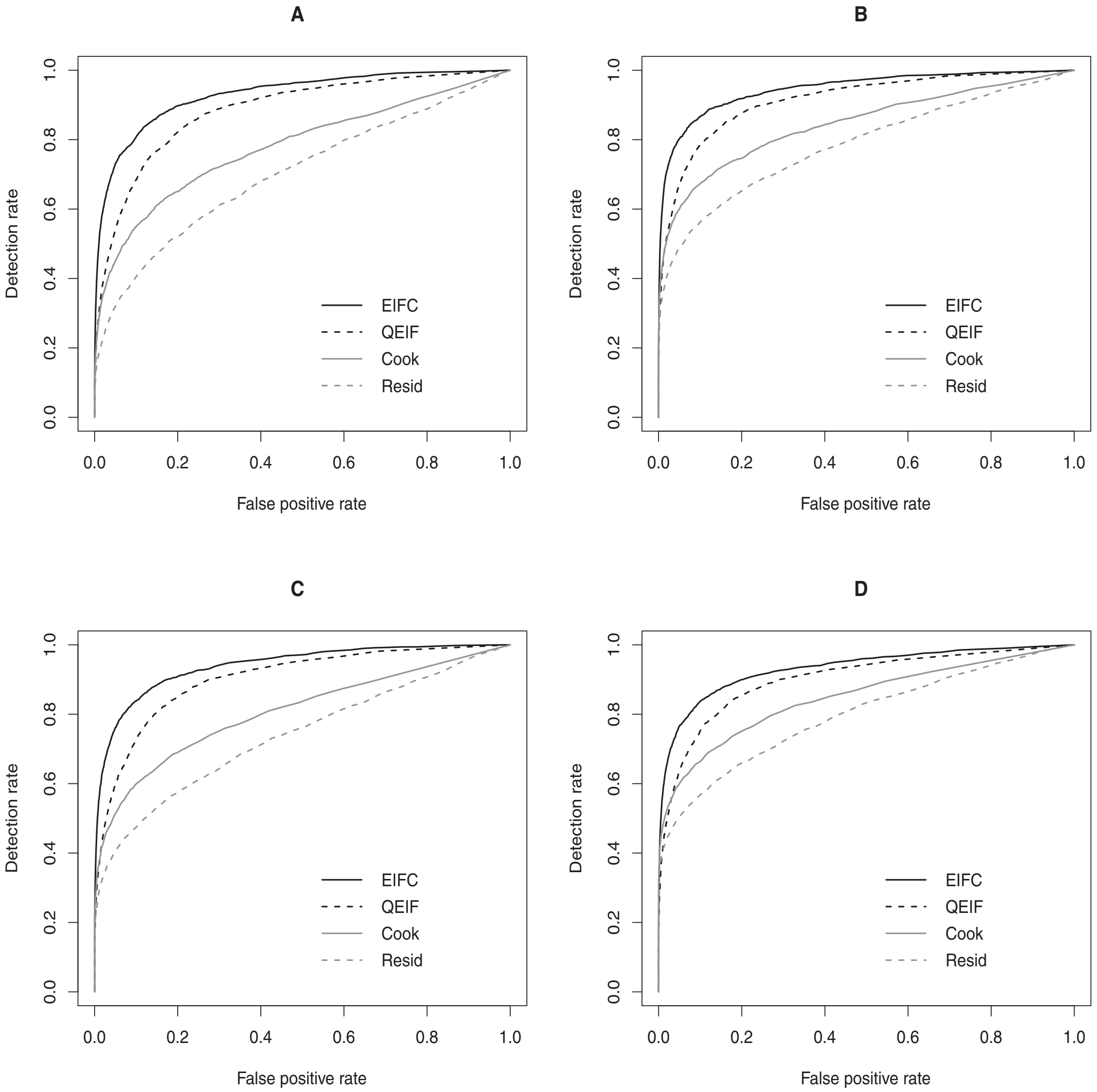

**Figure 5** Simulation results for Section 6.2. Comparison of ROC curves. EIFC: linear combination of EIFs, QEIF: quadratic EIF, Cook: Cook's $D$, Resid: standardized residual $r$. Distribution of outliers, $\varepsilon_i$, is (A) $N(0, (2\sigma)^2)$, (B) $N(0, (3\sigma)^2)$, (C) $t_3$ with scale $2\sigma$, and (D) Cauchy with scale $2\sigma$.

The proposed methods have the following four remarkable features.

(i) These methods focus on interactive effects of genotype and phenotype—Phenotype and genotypes are incorporated in influence analysis on LOD scores. For example, in our dataset, mouse No. 13 has genotypes $(-1, -1, -1)$ and the minimum phenotype 3.842, but its EIF on the curvature of the LOD score curve is 0.077, which is much less than that of 0.447 for mouse No. 60. However, as shown in Section 5.3, when the phenotype works with genotypes $(0, -1, 0)$, its influence becomes significant. The probability of genotypes $(0, -1, 0)$ under Haldane's model is 0.014. There are also some mice with rarer genotypes, such as mouse No. 9 $(-1, 0, -1)$

**Table 1** Three methods for designing coefficients.

| Method | Advantage [A] and drawback [D] |
|---|---|
| Orthogonal polynomial | [A] Coefficients are easy to interpret. Detection power is high.<br>[D] Need to choose the degree of polynomial, or need to try various degrees (e.g., linear, quadratic, cubic, . . .). |
| PCA-based method | [A] Useful as an exploratory data analysis.<br>[D] Results are not necessarily clear (not easy to interpret). |
| Quadratic form (QEIF) | [A] Omnibus test, and easy to use.<br>[D] Detection power is lower than for the orthogonal polynomial. |

and mouse No. 128 $(1, 0, -1)$, both with probability 0.007. However, neither of these mice has significant influence on the shape of the LOD score curve. Hence, the proposed methods do not separately detect outlier phenotype or rare genotypes. Note that, using these methods, individuals with outlier phenotypes, or rare multilocus genotypes including epistasis or concordant genotypes may also be identified as influential if they change the LOD score curve significantly.

(ii) The proposed approach based on the EIF can be applied to other QTL models —In this paper, two simple models are dealt with as examples. However, the proposed approach can be applied to more complicated multiple QTL models based on LOD score, such as multiple interval mapping (Kao et al., 1999).

(iii) Three methods are proposed to design coefficients of linear combination of LOD scores—They can be used when we have no clear idea for choosing the coefficients. Here we give a summary on the feature of these methods in Table 1.

(iv) A method to assess the significance of each detected individual is proposed—We proposed a simulation based method to assess the significance of detecting influential individuals, and confirmed that this method is approximately valid (i.e., controlling false positives) in the case of the single marker analysis.

### 7.2 Practical guidelines

The proposed influence analysis methods are designed to identify individuals that change significantly the LOD score curve. The data of the identified individuals may contain observation errors. Therefore, we should reexamine them at the first. The data of detected individuals that are confirmed to be accurate and reliable are potentially informative for the gene mapping process.

Influence analysis is a model-based method. It is desirable to do QTL detection and influence analysis using models that are well fitted to the data. As stated in Section 3, our proposed EIF approaches can be applied to any QTL detection models based on LOD score. When the assumed model is incorrect, it is not easy to interpret the results.

On the other hand, the one-QTL model (single marker analysis or the interval mapping method) is often used in initial scan even though it may not be the true model. Even in this case, the detected individuals provide important information. They may be true influential individuals in the assumed model, or they may suggest other possibilities of QTL model such as multiple-QTL or epistasis model. In any of these cases, the detected individuals are informative and should be investigated carefully in subsequent QTL analysis.

In the process of QTL analysis, aside from influence analysis, the significance of LOD scores should be checked by standard methods such as permutation test (Churchill and Doerge, 1994).

The dataset and software used in this article are available from the authors or at NIG Mouse Phenotype Database `http://molossinus.lab.nig.ac.jp/phenotype/index.html` .

**Acknowledgments** The authors would like to thank the Editor, the Associate Editor, and two referees for valuable and constructive comments and suggestions. The authors also thank Shigeharu Wakana and Ikuo Miura of RIKEN Bioresource Center for genotyping of SNP markers, Yutaka Tanaka and Chen-Hung Kao for their valuable advice, and Constantin Siriteanu and Xiaoqin Dou for careful proofreading. This work was supported by the Systems Genetics Project of the Research Organization of Information and Systems.

**Conflict of interest**

*The authors have declared no conflict of interest.*

# Appendix

## A.1 Regularity conditions and proof of Theorem 2.1

The conditions required in Theorem 2.1 are listed below. $N_\theta$ denotes a neighborhood of $\boldsymbol{\theta}$.

A1. $\ell(\gamma, \boldsymbol{\theta}; y)$ is a $C^2$-function of $\boldsymbol{\theta}$, where $C^2$ means twice continuously differentiable.
A2. $\int |\ell(\gamma, \boldsymbol{\theta}; y)| dF(y) < \infty$, $\int \|\ell_\theta(\gamma, \boldsymbol{\theta}; y)\| dF(y) < \infty$.
A3. For each $\boldsymbol{\theta}$, a function $g(y; \gamma, \boldsymbol{\theta})$ exists such that $\left\|\ell_{\theta\theta}(\gamma, \boldsymbol{\theta}'; y)\right\|_F \leq g(y; \gamma, \boldsymbol{\theta})$ $(\forall \boldsymbol{\theta}' \in N_\theta)$ and $\int g(y; \gamma, \boldsymbol{\theta}) dF(y) < \infty$, where $\| \ \|_F$ is the Frobenius norm of the matrix.
A4. The Hessian matrix $L_{\theta\theta}(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F)$ is negative definite.

B1. $\int\int |\ell(\gamma, \boldsymbol{\theta}; y)| dF(y) dV_C(\gamma) < \infty$, $\int\int \|\ell_\theta(\gamma, \boldsymbol{\theta}; y)\| dF(y) dV_C(\gamma) < \infty$, where $dV_C(\gamma)$ is the variation of $dC(\gamma)$, and $\int |\ell(\gamma, \boldsymbol{\theta}; x)| dV_C(\gamma) < \infty$, $\int \|\ell_\theta(\gamma, \boldsymbol{\theta}; x)\| dV_C(\gamma) < \infty$.
B2. For each $\boldsymbol{\theta}$, a function $g(y; \gamma, \boldsymbol{\theta})$ exists such that $\left\|\ell_\theta(\gamma, \boldsymbol{\theta}'; y)\right\| \leq g(y; \gamma, \boldsymbol{\theta})$ $(\forall \boldsymbol{\theta}' \in N_\theta)$ and $\int\int g(y; \gamma, \boldsymbol{\theta}) dF(y) dV_C(\gamma) < \infty$, $\int g(x; \gamma, \boldsymbol{\theta}) dV_C(\gamma) < \infty$.

C1. $\ell(\gamma, \boldsymbol{\theta}; y)$ is a $C^2$-function of $\gamma$ and $\boldsymbol{\theta}$.
C2. $\int |\ell(\gamma, \boldsymbol{\theta}; y)| dF(y) < \infty$, $\int \{|\ell_\gamma(\gamma, \boldsymbol{\theta}; y)| + \left\|\ell_\theta(\gamma, \boldsymbol{\theta}; y)\right\|\} dF(y) < \infty$.
C3. For each $\boldsymbol{\theta}$, a function $g(y; \gamma, \boldsymbol{\theta})$ exists such that $|\ell_{\gamma\gamma}(\gamma, \boldsymbol{\theta}'; y)| + \left\|\ell_{\gamma\theta}(\gamma, \boldsymbol{\theta}'; y)\right\| + \left\|\ell_{\theta\theta}(\gamma, \boldsymbol{\theta}'; y)\right\|_F \leq g(y; \gamma, \boldsymbol{\theta})$ $(\forall \boldsymbol{\theta}' \in N_\theta)$ and $\int g(y; \gamma, \boldsymbol{\theta}) dF(y) < \infty$.
C4. The Hessian matrix below is negative definite:

$$\begin{pmatrix} L_{\gamma\gamma}(\widehat{\gamma}(F), \widehat{\boldsymbol{\theta}}(F); F) & L_{\gamma\theta}(\widehat{\gamma}(F), \widehat{\boldsymbol{\theta}}(F); F) \\ L_{\theta\gamma}(\widehat{\gamma}(F), \widehat{\boldsymbol{\theta}}(F); F) & L_{\theta\theta}(\widehat{\gamma}(F), \widehat{\boldsymbol{\theta}}(F); F) \end{pmatrix}.$$

**Lemma A.1.** *Assume A1–A4. Then,*

*(a)* $\widehat{\boldsymbol{\theta}}(\gamma, F_x^\epsilon)$ *with* $F_x^\epsilon = (1-\epsilon)F + \epsilon\delta_x$ *is* $C^1$ *in* $\epsilon$ *when* $|\epsilon|$ *is small. The influence function of* $\widehat{\boldsymbol{\theta}}(\gamma, \cdot)$ *is given as*

$$\mathsf{IF}(x, \widehat{\boldsymbol{\theta}}(\gamma, \cdot), F) = -L_{\theta\theta}(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F)^{-1} \ell_\theta(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); x).$$

*(b) The influence function of the profile likelihood at* $\gamma$, $M(\gamma; \cdot)$, *is given as*

$$\mathsf{IF}(x, M(\gamma; \cdot), F) = \ell(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); x) - L(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F).$$

**Proof of Lemma A.1.** Define a vector-valued function $\boldsymbol{H}(\epsilon, \boldsymbol{\theta}) = \Big(H_j(\epsilon, \boldsymbol{\theta})\Big)_{1\le j\le p}$ with $H_j(\epsilon, \boldsymbol{\theta}) = \int \ell_{\theta_j}(\gamma, \boldsymbol{\theta}; y) dF_x^{\epsilon}(y)$. Then, under conditions A1–A3, $\boldsymbol{H}(\epsilon, \boldsymbol{\theta})$ is $C^1$ in $(\epsilon, \boldsymbol{\theta})$. Because of the assumption of the unique existence of the maximizer $\widehat{\boldsymbol{\theta}}(\gamma, F) \in \Theta$, $0 = L_{\theta_j}(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F) = \int \ell_{\theta_j}(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); y) dF(y) = H_j(0, \widehat{\boldsymbol{\theta}}(\gamma, F))$. Here, the exchangeability of differential and integral signs is assured by A3. Noting that $\frac{\partial}{\partial\epsilon} H_j(\epsilon, \boldsymbol{\theta})|_{(\epsilon,\theta)=(0,\widehat{\theta}(\gamma,F))} = \ell_{\theta_j}(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F), x)$, and the Jacobian matrix

$$\left(\frac{\partial}{\partial\theta_k} H_j(\epsilon, \boldsymbol{\theta})\Big|_{(\epsilon,\theta)=(0,\widehat{\theta}(\gamma,F))}\right)_{1\le j,k\le p} = L_{\theta\theta}(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F)$$

is nonsingular by A4, we obtain the result for part (a) by the implicit function theorem and the definition that $(\frac{d}{d\epsilon})_0 \widehat{\boldsymbol{\theta}}(\gamma, F_x^{\epsilon}) = \mathsf{IF}(x, \widehat{\boldsymbol{\theta}}(\gamma, \cdot), F)$.

For part (b), since $L(\gamma, \boldsymbol{\theta}; F)$ is linear in $F$,

$$\begin{aligned}\mathsf{IF}(x; M(\gamma; \cdot), F) &= (\tfrac{d}{d\epsilon})_0 M(\gamma, F_x^{\epsilon}) = (\tfrac{d}{d\epsilon})_0 L(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F_x^{\epsilon}); F_x^{\epsilon}) \\ &= (\tfrac{d}{d\epsilon})_0 L(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F_x^{\epsilon}); F) - L(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F) + \ell(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); x).\end{aligned}$$

Here, the first term becomes $\sum_j L_{\theta_j}(\gamma, \widehat{\boldsymbol{\theta}}(\gamma, F); F)\mathsf{IF}(x, (\widehat{\boldsymbol{\theta}}(\gamma))_j; F) = 0$. □

**Regularity conditions and proof of (i) of Theorem 2.1.** Assume A1–A4. Under these regularity conditions, $\int L(\gamma, \boldsymbol{\theta}; F) dC(\gamma)$ and $\int \ell(\gamma, \boldsymbol{\theta}; x) dC(\gamma)$ are $C^1$ in $\boldsymbol{\theta}$, and differentiation $(\frac{d}{d\epsilon})_0$ and the integration with respect to $dF(y)$ and $dC(\gamma)$ are exchangeable. Equation (3) is proved in the same way as part (b) of Lemma A.1. □

**Regularity conditions and proof of (ii) and (iii).** Assume A1–A4 (if $\Gamma$ is discrete) or C1–C4 (if $\Gamma$ is continuous). Here, we only consider the case where $\Gamma$ is continuous. Note that the maximizer $\widehat{\gamma}$ of the profile likelihood $M(\gamma, F)$ is simply the MLE based on the ordinary likelihood $L(\gamma, \boldsymbol{\theta}, F)$, and hence $\max_{\gamma} M(\gamma; F) = \max_{(\gamma,\boldsymbol{\theta})} L(\gamma, \boldsymbol{\theta}, F)$. Then, from Lemma A.1, (b), the IF of the maximum profile likelihood is given as (4). Moreover, from Lemma A.1, (a), the IF of $(\widehat{\gamma}(\cdot), \widehat{\boldsymbol{\theta}}(\cdot))$ is given as

$$\begin{pmatrix}\mathsf{IF}(x, \widehat{\gamma}(\cdot), F)\\ \mathsf{IF}(x, \widehat{\boldsymbol{\theta}}(\cdot), F)\end{pmatrix} = -\begin{pmatrix}L_{\gamma\gamma}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F) & L_{\gamma\theta}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F)\\ L_{\theta\gamma}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F) & L_{\theta\theta}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; F)\end{pmatrix}^{-1} \begin{pmatrix}\ell_{\gamma}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; x)\\ \ell_{\theta}(\widehat{\gamma}, \widehat{\boldsymbol{\theta}}; x)\end{pmatrix}\Bigg|_{\widehat{\gamma}=\widehat{\gamma}(F),\, \widehat{\theta}=\widehat{\theta}(F)}.$$

By means of the inversion formula of the partitioned matrix, we obtain (5). □

## A.2 Conditional distribution of the putative QTL genotype

The conditional probability $\mathrm{P}(z_i^*|\boldsymbol{z}_i; \gamma)$ can be obtained as follows: Let $\boldsymbol{\epsilon}_i = \big(\epsilon_i^{(1)}, \ldots, \epsilon_i^{(m)}\big)$ and $\boldsymbol{\delta}_i = \big(\delta_i^{(1)}, \ldots, \delta_i^{(m)}\big)$ denote the genotypes originating from the mother and the father, respectively, of each $\mathrm{F}_2$ individual. Each element of $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\delta}_i$ takes 1 or $-1$, because they are from the $\mathrm{F}_1$ population. Thus, the genotype of the $\mathrm{F}_2$ individual can be written as $\boldsymbol{z}_i = \frac{1}{2}(\boldsymbol{\epsilon}_i + \boldsymbol{\delta}_i)$. Although the elements of $\boldsymbol{\epsilon}_i = \big(\epsilon_i^{(1)}, \ldots, \epsilon_i^{(m)}\big)$ take the values $\pm 1$ with the same probability $\frac{1}{2}$, they are strongly correlated by the linkage. For example, under the assumption of the most basic linkage model, Haldane's map function, the vector $\big(\epsilon_i^{(1)}, \ldots, \epsilon_i^{(m)}\big)$ can be considered as a Markov sequence with probability $\mathrm{P}\big(\epsilon_i^{(1)} = \pm 1\big) = \frac{1}{2}$, $\mathrm{P}\big(\epsilon_i^{(j+1)} = \pm\epsilon_i^{(j)}\big|\epsilon_i^{(j)}\big) = \frac{1}{2}(1 \pm \rho_{j,j+1})$, where $\rho_{jk} = e^{-2|d_k - d_j|/100}$ ($j$ and $k$ are on the same chromosome), 0 (otherwise), and $d_j$ denotes the location of the marker locus $j$ measured in cM.

Similarly, $\boldsymbol{\delta}_i = (\delta_i^{(1)}, \ldots, \delta_i^{(m)})$ has the same Markov property. Because $\boldsymbol{\delta}_i$ and $\boldsymbol{\epsilon}_i$ are independent, the $m$-vector $\boldsymbol{z}_i$ has the probability distribution

$$\mathrm{P}(\boldsymbol{z}_i) = \sum_{\boldsymbol{z}_i} \frac{1}{2^{2m}} \prod_{j=1}^{m-1} \left(1 + \epsilon_i^{(j)} \epsilon_i^{(j+1)} \rho_{j+1,j}\right) \left(1 + \delta_i^{(j)} \delta_i^{(j+1)} \rho_{j+1,j}\right),$$

where the summation $\sum_{\boldsymbol{z}_i}$ is taken over all possible $\boldsymbol{\delta}_i, \boldsymbol{\epsilon}_i \in \{1, -1\}^m$ such that $\boldsymbol{z}_i = (\boldsymbol{\epsilon}_i + \boldsymbol{\delta}_i)/2$.

Similarly, when genotype $z_i^* = (\epsilon_i^* + \delta_i^*)/2$ of the putative QTL at location $\gamma$ is taken into consideration, the joint probability of the $(m+1)$-vector $(\boldsymbol{z}_i, z_i^*)$ can be obtained as

$$\mathrm{P}(\boldsymbol{z}_i, z_i^*; \gamma) = \sum_{\boldsymbol{z}_i, z_i^*} \frac{1}{2^{2(m+1)}} \prod_{k=1, k \neq j}^{m-1} \left(1 + \epsilon_i^{(k)} \epsilon_i^{(k+1)} \rho_{k+1,k}\right)\left(1 + \delta_i^{(k)} \delta_i^{(k+1)} \rho_{k+1,k}\right)$$

$$\times\left(1 + \epsilon_i^{(j)} \epsilon_i^* e^{-2(\gamma - d_j)}\right)\left(1 + \epsilon_i^* \epsilon_i^{(j+1)} e^{-2(d_{j+1} - \gamma)}\right)\left(1 + \delta_i^{(j)} \delta_i^* e^{-2(\gamma - d_j)}\right)\left(1 + \delta_i^* \delta_i^{(j+1)} e^{-2(d_{j+1} - \gamma)}\right)$$

for $d_j \le \gamma \le d_{j+1}$, where the summation $\sum_{\boldsymbol{z}_i, z_i^*}$ is taken over all possible $\boldsymbol{\delta}_i, \boldsymbol{\epsilon}_i \in \{1, -1\}^m$, $\delta_i^*, \epsilon_i^* \in \{1, -1\}$ such that $\boldsymbol{z}_i = (\boldsymbol{\epsilon}_i + \boldsymbol{\delta}_i)/2$ and $z_i^* = (\epsilon_i^* + \delta_i^*)/2$.

Then, the conditional probability is obtained as (6).

### A.3 Regularity conditions and proof of Theorem 4.1

We derive the asymptotic null covariance (17) of $\mathsf{LOD}(\gamma) = n\{L_n(\gamma, \widehat{\boldsymbol{\theta}}(\gamma)) - L_n(\widetilde{\boldsymbol{\theta}})\}/\log 10$ in (9), where $\widehat{\boldsymbol{\theta}}(\gamma)$ and $\widetilde{\boldsymbol{\theta}}$ are the MLEs for $\boldsymbol{\theta} = (\alpha, \beta, \mu, \nu, \sigma^2)$ under the $H_1$ and $H_0$, respectively.

$L_n$ is the quasi-likelihood in the sense that the $y_i$'s are independent but not identically distributed. For the quasi-likelihood, the asymptotic properties of the MLE and the likelihood ratio test still hold under regularity conditions (White, 1996). We use the asymptotic equivalence of the LR and Rao's score statistic under $H_0$ in Lemma A.2. Write $f_i(\gamma, \boldsymbol{\theta}) = f(y_i | \boldsymbol{z}_i, u_i; \gamma, \boldsymbol{\theta})$, $g_i(z^*, \boldsymbol{\theta}) = g(y_i | z^*, u_i; \boldsymbol{\theta})$, and $P_i(z^*; \gamma) = \mathrm{P}(z^* | \boldsymbol{z}_i; \gamma)$ for simplicity. Let $\boldsymbol{s}_i(\gamma, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta} \log f_i(\gamma, \boldsymbol{\theta})$ be the efficient score vector. Let $\boldsymbol{\theta}_0 = (0, 0, \mu_0, \nu_0, \sigma_0^2)$ be the true parameter in $H_0$.

We assume the regularity conditions:

D1. The covariates $u_i$'s are bounded.

D2. As $n \to \infty$, $\frac{1}{n}\begin{pmatrix} \boldsymbol{Z}(\gamma)' \\ \boldsymbol{U}' \end{pmatrix}\begin{pmatrix} \boldsymbol{Z}(\gamma) & \boldsymbol{U} \end{pmatrix}$ converges to a positive definite matrix.

**Lemma A.2.** *Assume D1 and D2 above. Then, under* $H_0$,

$$2n\{L_n(\gamma, \widehat{\boldsymbol{\theta}}(\gamma)) - L_n(\widetilde{\boldsymbol{\theta}})\} = \boldsymbol{S}_n(\gamma, \widetilde{\boldsymbol{\theta}})' \boldsymbol{I}_n^{-1}(\gamma, \widetilde{\boldsymbol{\theta}}) \boldsymbol{S}_n(\gamma, \widetilde{\boldsymbol{\theta}}) + o_p(1), \tag{A.1}$$

where

$$\boldsymbol{S}_n(\gamma, \boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{s}_i(\gamma, \boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{I}_n(\gamma, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathrm{Cov}_{(\gamma, \theta)}(\boldsymbol{s}_i(\gamma, \boldsymbol{\theta})).$$

**Proof of Lemma A.2.** First, it is straightforward to show that under assumptions D1 and D2, $\lim_{n \to \infty} \boldsymbol{I}_n(\gamma, \boldsymbol{\theta}) = \boldsymbol{I}(\gamma, \boldsymbol{\theta})$ exists, and the central limit theorem $\boldsymbol{S}_n(\gamma, \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{I}(\gamma, \boldsymbol{\theta}_0))$ works. Moreover, under assumptions D1 and D2, it is easy to construct functions $h_i^{(j)}(y)$ $(j = 0, 1, 2)$ such that $\left\|(\frac{\partial}{\partial\theta})^j \log f_i(\gamma,\boldsymbol{\theta})\right\| \le h_i^{(j)}(y_i)$ and $\mathrm{E}_{(\gamma,\boldsymbol{\theta})}\left[h_i^{(j)}(y_i)\right] \le M$ for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$. Using these facts, we can confirm all of the regularity conditions of Theorems 8.9 and 8.10 of White (1996), from which the stochastic equivalence (A.1) follows. □

Under $H_0$, the efficient score vector and the Fisher information matrix have simple forms. Since under $H_0$, $g_i(z^*,\boldsymbol{\theta}_0)$ does not depend on $z^*$, the efficient score vector is

$$\begin{aligned}
\boldsymbol{s}_i(\gamma,\boldsymbol{\theta}_0) &= \frac{\frac{\partial}{\partial\theta} f_i(\gamma,\boldsymbol{\theta}_0)}{f_i(\gamma,\boldsymbol{\theta}_0)} = \frac{\sum_{z^*=-1}^{1}\left\{\frac{\partial}{\partial\theta}\log g_i(z^*,\boldsymbol{\theta}_0)\right\} g_i(z^*,\boldsymbol{\theta}_0)P_i(z^*;\gamma)}{\sum_{z^*=-1}^{1} g_i(z^*,\boldsymbol{\theta}_0)P_i(z^*;\gamma)} \\
&= \frac{g_i(*,\boldsymbol{\theta}_0)\sum_{z^*=-1}^{1}\left\{\frac{\partial}{\partial\theta}\log g_i(z^*,\boldsymbol{\theta}_0)\right\}P_i(z^*;\gamma)}{g_i(*,\boldsymbol{\theta}_0)\sum_{z^*=-1}^{1} P_i(z^*;\gamma)} \\
&= \sum_{z^*=-1}^{1}\left\{\frac{\partial}{\partial\theta}\log g_i(z^*,\boldsymbol{\theta}_0)\right\}P_i(z^*;\gamma) \\
&= \sum_{z^*=-1}^{1}\frac{1}{\sigma_0^2}\left(z^* e_i(\xi_0), w(z^*)e_i(\xi_0), e_i(\xi_0), u_i e_i(\xi_0), \frac{1}{2}\left(\frac{e_i(\xi_0)^2}{\sigma_0^2}-1\right)\right)' P_i(z^*;\gamma) \\
&= \frac{1}{\sigma_0^2}\left(\left(\bar{z}_i(\gamma), \bar{w}_i(\gamma), 1, u_i\right)e_i(\xi_0), \frac{1}{2}\left(\frac{e_i(\xi_0)^2}{\sigma_0^2}-1\right)\right)',
\end{aligned}$$

where $\xi_0 = (\mu_0, \nu_0)$ and $e_i(\xi_0) = y_i - \mu_0 - \nu_0 u_i$. Because $\boldsymbol{e}(\xi_0) = (e_i(\xi_0))_{1\le i\le n} \sim N_n(\mathbf{0}, \sigma_0^2\boldsymbol{I}_n)$,

$$\boldsymbol{S}_n(\gamma,\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}\sigma_0^2}\begin{pmatrix} \begin{pmatrix}\boldsymbol{Z}(\gamma)' \\ \boldsymbol{U}'\end{pmatrix}\boldsymbol{e}(\xi_0) \\ \frac{1}{2}\left(\frac{\boldsymbol{e}(\xi_0)'\boldsymbol{e}(\xi_0)}{\sigma_0^2} - n\right)\end{pmatrix}, \quad \boldsymbol{I}_n(\gamma,\boldsymbol{\theta}_0) = \frac{1}{\sigma_0^2}\begin{pmatrix} \frac{1}{n}\begin{pmatrix}\boldsymbol{Z}(\gamma)' \\ \boldsymbol{U}'\end{pmatrix}(\boldsymbol{Z}(\gamma), \boldsymbol{U}) & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2}\end{pmatrix}.$$

Moreover, substituting $\widetilde{\boldsymbol{\theta}} = (0, 0, \widetilde{\mu}, \widetilde{\nu}, \widetilde{\sigma}^2)$ into $\boldsymbol{\theta}_0$ in $\boldsymbol{S}_n(\gamma,\boldsymbol{\theta}_0)$ and $\boldsymbol{I}_n(\gamma,\theta_0)$, and taking into account that $\widetilde{\sigma}^2 = \boldsymbol{e}(\widetilde{\xi})'\boldsymbol{e}(\widetilde{\xi})/n$ with $\widetilde{\xi} = (\widetilde{\mu}, \widetilde{\nu})$, (A.1) can be rewritten as

$$\boldsymbol{S}_n(\gamma,\widetilde{\boldsymbol{\theta}})'\boldsymbol{I}_n^{-1}(\gamma,\widetilde{\boldsymbol{\theta}})\boldsymbol{S}_n(\gamma,\widetilde{\boldsymbol{\theta}}) = \frac{1}{\widetilde{\sigma}^2}\boldsymbol{e}(\widetilde{\xi})'\{\boldsymbol{I}_n - \boldsymbol{Q}(\gamma)\}\boldsymbol{e}(\widetilde{\xi}) \tag{A.2}$$

with

$$\boldsymbol{Q}(\gamma) = \boldsymbol{I}_n - \begin{pmatrix}\boldsymbol{Z}(\gamma) & \boldsymbol{U}\end{pmatrix}\left\{\begin{pmatrix}\boldsymbol{Z}(\gamma)' \\ \boldsymbol{U}'\end{pmatrix}\begin{pmatrix}\boldsymbol{Z}(\gamma) & \boldsymbol{U}\end{pmatrix}\right\}^{-1}\begin{pmatrix}\boldsymbol{Z}(\gamma)' \\ \boldsymbol{U}'\end{pmatrix}.$$

Under $H_0$, $e_i(\widetilde{\xi})$ is the residual error in the regression model $y_i = \mu + \nu u_i + e_i$, $e_i \sim N(0,\sigma^2)$. Hence, $\boldsymbol{e}(\widetilde{\xi}) \sim N_n(\mathbf{0}, \sigma_0^2\boldsymbol{Q})$ and $\boldsymbol{Q} = \boldsymbol{I}_n - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'$ under $H_0$. Representing the residuals as $\boldsymbol{e}(\widetilde{\xi}) = \sigma_0\boldsymbol{Q}\boldsymbol{\varepsilon}$ with an $n$-vector $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{I}_n)$ and noting that $\widetilde{\sigma}^2/\sigma_0^2 = 1 + o_p(1)$, it is seen from (A.2) that

$$\boldsymbol{S}_n(\gamma,\widetilde{\boldsymbol{\theta}})'\boldsymbol{I}_n^{-1}(\gamma,\widetilde{\boldsymbol{\theta}})\boldsymbol{S}_n(\gamma,\widetilde{\boldsymbol{\theta}}) = \boldsymbol{\varepsilon}'\boldsymbol{Q}\{\boldsymbol{I}_n - \boldsymbol{Q}(\gamma)\}\boldsymbol{Q}\boldsymbol{\varepsilon} + o_p(1) = \boldsymbol{\varepsilon}'\boldsymbol{R}(\gamma)\boldsymbol{\varepsilon} + o_p(1), \tag{A.3}$$

where $\boldsymbol{R}(\gamma) = \boldsymbol{Q}\{\boldsymbol{I}_n - \boldsymbol{Q}(\gamma)\}\boldsymbol{Q} = \boldsymbol{Q} - \boldsymbol{Q}(\gamma)$ is given in (18). Combining (A.1) and (A.3) with Lemma A.3 below, we get the result

$$\mathrm{Cov}\left(2n\left\{L_n(\gamma_j, \widehat{\boldsymbol{\theta}}_{\gamma_j}) - L_n(\widetilde{\boldsymbol{\theta}})\right\}, 2n\left\{L_n(\gamma_k, \widehat{\boldsymbol{\theta}}_{\gamma_k}) - L_n(\widetilde{\boldsymbol{\theta}})\right\}\right) = 2\mathrm{tr}\left(\boldsymbol{R}(\gamma_j)\boldsymbol{R}(\gamma_k)\right) + o(1),$$

and (17) follows. The proof of Theorem 4.1 is completed.

**Lemma A.3.** *Let* $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$ *be a zero-mean Gaussian vector with a covariance matrix* $\boldsymbol{\Sigma}$. *Let* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be symmetric matrices. Then,* $\mathrm{Cov}(\boldsymbol{\varepsilon}'\boldsymbol{A}\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'\boldsymbol{B}\boldsymbol{\varepsilon}) = 2\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma})$.

**Proof of Lemma A.3.** Write $\boldsymbol{\varepsilon} = (\varepsilon_i)$, $\boldsymbol{\Sigma} = (\sigma_{ij})$, $\boldsymbol{A} = (a_{ij})$, and $\boldsymbol{B} = (b_{ij})$. For $1 \leq i, j, k, l \leq n$, since $\varepsilon_i$'s have mean 0, we have

$$\begin{aligned}\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l] &= \mathrm{E}[\varepsilon_i\varepsilon_j]\mathrm{E}[\varepsilon_k\varepsilon_l] + \mathrm{E}[\varepsilon_i\varepsilon_k]\mathrm{E}[\varepsilon_j\varepsilon_l] + \mathrm{E}[\varepsilon_i\varepsilon_l]\mathrm{E}[\varepsilon_j\varepsilon_k]\\ &= \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk},\end{aligned}$$

and hence

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\varepsilon}'\boldsymbol{A}\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'\boldsymbol{B}\boldsymbol{\varepsilon}) &= \mathrm{E}[\boldsymbol{\varepsilon}'\boldsymbol{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{B}\boldsymbol{\varepsilon}] - \mathrm{E}[\boldsymbol{\varepsilon}'\boldsymbol{A}\boldsymbol{\varepsilon}]\mathrm{E}[\boldsymbol{\varepsilon}'\boldsymbol{B}\boldsymbol{\varepsilon}]\\ &= \sum a_{ij}b_{kl}\big(\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l] - \mathrm{E}[\varepsilon_i\varepsilon_j]\mathrm{E}[\varepsilon_k\varepsilon_l]\big)\\ &= 2\sum a_{ij}b_{kl}\sigma_{jk}\sigma_{li} = 2\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma}).\end{aligned}$$

□

## References

Bollen, K. A. and Jackman, R. (1990). Regression diagnostics: an expository treatment of outliers and influential cases. In: Fox, J. and Scott Long, J. (Eds.), *Modern Methods of Data Analysis*. Sage, Newbury Park, CA, pp. 257–291.

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Cook, D. (1977). Detection of influential observation in linear regression. *Technometrics* **19**, 15–18.

Cook, D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B* **2**, 133–169.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874.

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

Hayat, Y., Yang, J., Xu, H.-M., and Zhu, J. (2008). Influence of outliers on QTL mapping for complex traits. *Journal of Zhejiang University, Science B* **9**, 931–937.

Kao, C.-H. and Zeng, Z.-B. (1997). General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653–665.

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

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

Lu, J., Ko, D., and Chang, T. (1997). The standardized influence matrix and its applications. *Journal of the American Statistical Association* **92**, 1572–1580.

Manly, K. F. and Olson, J. M. (1999). Overview of QTL mapping software and introduction to Map Manager QT. *Mammalian Genome* **10**, 327–334.

Murphy, A. A. and van der Vaart, A. A. (2000). On profile likelihood. *Journal of the American Statistical Association* **95**, 449–465.

Sen, Ś. and Churchill, G. (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**, 371–387.

Siegmund, D. and Yakir, B. (2007). *The Statistics of Gene Mapping*, Springer, New York, NY.

Takada, T., Mita, A., Maeno, A., Sakai, T., Shitara, H., Kikkawa, Y., Moriwaki, K., Yonekawa, H., and Shiroishi, T. (2008). Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits. *Genome Research* **18**, 500–508. NIG Mouse phenotype database http://molossinus.lab.nig.ac.jp/phenotype/index.html

Tanaka, Y. (1994). Recent advance in sensitivity analysis in multivariate statistical methods. *Journal of the Japanese Society of Computational Statistics* **7**, 1–25.

White, H. (1996). Estimation, *Inference and Specification Analysis*. Cambridge University Press, Cambridge, UK.

Wright, F. A. and Kong, A. (1997). Linkage mapping in experimental crosses: the robustness of single gene models. *Genetics* **146**, 417–425.

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

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences USA* **90**, 10972–10976.

Zewotir, T. and Galpin, J. S. (2005). Influence diagnostics for linear mixed models. *Journal of Data Science* **3**, 153–177.