Biostatistics (2004), **5**, 4, *pp.* 501–513
doi: 10.1093/biostatistics/kxh004

Nonparametric estimation of the effects of quantitative trait loci

JASON P. FINE*
Department of Statistics, University of Wisconsin, Madison, WI 53706, USA
fine@biostat.wisc.edu

FEI ZOU
Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

BRIAN S. YANDELL
Department of Statistics, University of Wisconsin, Madison, WI 53706, USA

SUMMARY

Interval mapping of quantitative trait loci from breeding experiments plays an important role in understanding the mechanisms of disease, both in humans and other organisms. Standard approaches to estimation involve parametric assumptions for the component distributions and may be sensitive to model misspecification. Some nonparametric tests have been studied. However, nonparametric estimation of the phenotypic distributions has not been considered in the genetics literature, even though such methods might provide essential nonparametric summaries for comparing different loci. We develop a sufficient condition for identifiability of the phenotypic distributions. Simple nonparametric estimators for the distributions are proposed for uncensored and right censored data. They have a closed form and their small and large sample properties are readily established. Their practical utility as numerical summaries which complement nonparametric tests is demonstrated on two recent genetics examples.

Keywords: Discrete mixture model; Empirical distribution; Genetic linkage; Least squares; Molecular marker; Nonparametric identifiability.

1. INTRODUCTION

1.1 *Motivation*

The motivating application for this paper is the genetic mapping of quantitative traits in breeding populations. In these settings, inbred lines are mated in a predetermined manner and the progeny are genotyped at known markers. The goals are to determine the genomic regions influencing the traits and to quantify genotypic effects. These studies have proven invaluable in understanding the mechanisms of disease, both in humans and other organisms. Controlled crosses of rats and mice have been widely used for complex human disorders, like diabetes, cardiovascular disease, and cancer, where the variability in natural populations precludes in-depth biological investigations. Such experiments are also critical to commercial agriculture and animal breeding, where the ultimate objective is the identification of genes influencing profitable phenotypes. Examples include grain yield in rice and milk production in cows.

*To whom correspondence should be addressed.

The statistical issues in quantitative trait mapping are complicated. A typical application which we use to motivate these issues is a rat study of the genetics of tumor development (Lan *et al.*, 2001). The animals were derived from a backcross design with the Wistar–Kyoto (WKy) strain, which is resistant to cancer, and the Wistar–Furth (WF) strain, which is not resistant. These strains have genotypes WKy/Wky and WF/WF at each locus, respectively, while the study animals are either WKy/WF or WF/WF at each locus. When searching for genes, one evaluates the tumor count distributions at all loci in the genome, not only the markers. The difficulty is that while the genotypes are known at the markers, they are unknown between the markers. This leads to mixture models where the probability distribution of the genotypes at putative loci between markers may be calculated with genotypes observed at the markers and known distances between the loci and the markers. Note that the observed genotypes at the markers may vary across animals, and hence the mixture models may also vary across animals. The goal is to evaluate the tumor count distribution for each genotype at the putative loci.

Formally, the observations arise from K ($\geqslant 2$) discrete mixtures, that is, mixtures consisting of a finite number of components, L ($\geqslant 2$). In the genetics set-up, K is the number of possible genotypes at the flanking markers and L is the number of possible genotypes at the locus between the markers. Each observation originates in one of the K mixtures and the label of the mixture generating the datum is known. The mixing probabilities are also known and may vary amongst the K mixtures, while the L component distributions are common to all mixtures. The data consist of $n = \sum_{k=1}^{K} n_i$ independent observations, denoted $(X_{kj}, k = 1, \ldots, K, j = 1, \ldots, n_k)$, where n_k is the number of observations from mixture k $(= 1, \ldots, K)$. For given k, X_{kj}, $j = 1, \ldots, n_k$, are i.i.d. with

$$\Pr(X_{kj} \leqslant t) = H_k(t) = \sum_{l=1}^{L} \lambda_{kl} F_l(t). \tag{1.1}$$

The unknown mixture distributions, $H_k(t), k = 1, \ldots, K$, and the unknown component distributions, $F_l(t), l = 1, \ldots, L$, are nondegenerate. The known mixing proportions $0 \leqslant \lambda_{kl} \leqslant 1, l = 1, \ldots, L$, satisfy $\sum_{l=1}^{L} \lambda_{kl} = 1$, for $k = 1, \ldots, K$. The determination of K and L and the computation of λ_{kl} is discussed for the rat study and other experimental designs in Section 5.1.

The standard analyses for this data posit parametric models for F_l (Doerge *et al.*, 1997). The normal mixture is the default in the widely used software Mapmaker/QTL (Lander and Botstein, 1989). In practice, the traits may be non-normal. In Lan *et al.* (2001), the outcome is a tumor count, which is discrete and has an asymmetric distribution. Model misspecification may lead to reduced power to detect genes effecting a trait or to invalid estimates of the locations of the genes. It may also result in biased estimates of the genetic effects. Nonparametric tests can be obtained with Wilcoxon rank statistics (Kruglyak and Lander, 1995). Unfortunately, the parametric analyses are often used because of their availability in the mapping packages and model checking has been largely overlooked in practice.

Identifying the locations of genes is the primary objective in a mapping study. This is accomplished by taking as point estimates those locations where the largest test statistics are obtained. However, estimating the phenotypic distributions at these peak locations is important in quantifying the magnitude of the genetic effects. These effects are essential for comparing and ranking different genes, as well as for assessing the validity of parametric assumptions. The former point is not well understood by geneticists, who tend to rank the locations according to the magnitude of the test statistics, with larger test statistics assumed to correspond to larger genetic effects. However, the magnitudes of the effects are determined by the model parameters, not the test statistics. For parametric models, these estimates are a by-product of the maximum likelihood analysis. However, with rank tests, no such estimates are available. With this in mind, we investigate model (1.1) with the component distributions completely unspecified. Our objective is nonparametric estimates of genetic effects which complement the existing nonparametric tests (Kruglyak and Lander, 1995).

1.2 *Overview of paper*

In many applications, estimation of discrete mixture models may involve both unknown mixing proportions and component distributions (Titterington *et al.*, 1985). Parametric analyses may be based on maximum likelihood or the method of moments (Hosmer, 1973). Robust alternatives use kernel estimates of the component densities or the empirical distribution functions (Murray and Titterington, 1978; Hall, 1981; Hall and Titterington, 1984, 1985). These nonparametric approaches involve a single mixture and require data from the component distributions. Without the categorized data, nonparametric identification of the component distributions is impossible. In the multiple mixture formulation (1.1) for quantitative trait mapping, the genotypes at the loci of interest may not be observed and there may not be direct information from the components. Thus, the previous results are not appropriate. In Section 2, we provide a sufficient condition for nonparametric identifiability of the component distributions with $K \geqslant L$ and known λ_{kl}. The variability of the weights is critical to establishing the result. Of course, if $K = 1$, then $F_l, l = 1, \ldots, L$, are nonidentifiable, just as with unknown λ_{kl}. To our knowledge, our identifiability result for the phenotypic distributions has not appeared in the genetics literature.

In Section 3, we introduce intuitive least squares estimators for the components in the general mixture set-up (1.1). The estimators are linear combinations of the empirical distributions for $H_k, k = 1, \ldots, K$. They have a closed form, are unbiased and their finite sample variance may be evaluated explicitly and estimated unbiasedly. Large sample behavior is derived from the theoretical properties of the empirical distributions and functional mapping concepts. It is established that the estimators are uniformly consistent and converge weakly to Gaussian processes with variances which may be consistently estimated. In Section 4, the estimators are extended to right censored data. The asymptotic distribution is given, with the technical details in the Appendix. A competing approach to estimation is nonparametric maximum likelihood, which can be implemented using the EM algorithm (Dempster*et al.*, 1977). A difficulty is that the algorithm may be quite slow to converge, especially in the nonparametric setting. Furthermore, because the model is infinite dimensional, the usual asymptotic results for parametric maximum likelihood are not applicable. The theoretical properties of the estimators are unclear and variance estimation may not be straightforward.

Our simple nonparametric estimators complement the rank-based tests of Kruglyak and Lander (1995). We view their role in quantitative trait analyses as somewhat analogous to that of the Kaplan and Meier (1958) estimator in clinical research, where the logrank statistic (Mantel, 1966) is first used to test for survival differences. The usefulness of this strategy is illustrated in Section 5.2 in a reanalysis of the rat data (Lan *et al.*, 2001) on mammary carcinoma. The standard assumptions, e.g. normality, are inappropriate for the tumor count phenotype and the nonparametric estimates of the component distributions are helpful in understanding the nature of the genetic effects. In Section 5.3, the censored data methodology is demonstrated on flowering times from a breeding experiment of the plant species *Brassica napus* (Ferreira *et al.*, 1995). Existing mapping software cannot handle the censored data and may give misleading results if the censoring is naively disregarded. At the least, the estimates of mean genotypic effects based on parametric analyses are invalid, owing to the improper use of information in the tail of the phenotypic distributions. The nonparametric analyses enable robust inferences about descriptive measures, i.e. the median, which are nonparametrically identifiable with such censoring. In Section 6, the paper concludes with general remarks on the practicability of nonparametric methods in gene mapping studies.

2. Identifiability of phenotypic distributions

Let $\tilde{F}(t) = \{F_1(t), \ldots, F_L(t)\}^T$, $\tilde{H}(t) = \{H_1(t), \ldots, H_K(t)\}^T$ and let Λ be a known $K \times L$ matrix with kth row $(\lambda_{k1}, \lambda_{k2}, \ldots, \lambda_{kL})$. For fixed t, the relationship of the mixture distributions and

the component distributions is described by a system of K linear equations

$$\Lambda \tilde{F}(t) = \tilde{H}(t). \tag{2.1}$$

Because data are observed directly from $H_k, k = 1, \ldots, K$, these mixture distributions are nonparametrically identifiable via the corresponding empirical distributions. The only quantities not yet identified are $F_l(t), l = 1, \ldots, L$. Evidently, the component distributions are nonparametrically identifiable if there is a unique $\tilde{F}$ solving equations (2.1) with Λ and $\tilde{H}$ fixed. The reason for requiring $K \geqslant L$ is that if $K < L$, then this cannot be guaranteed without stronger conditions on the components and the mixtures.

For $K \geqslant L$, identifying the components uses the variability of the mixing proportions. A sufficient condition for nonparametric identification is that the $L \times L$ matrix $\Lambda^T \Lambda$ is nonsingular, as occurs when Λ is full rank. The invertibility of $\Lambda^T \Lambda$ gives that

$$\tilde{F}(t) = (\Lambda^T \Lambda)^{-1} \Lambda^T \tilde{H}(t). \tag{2.2}$$

Since the weights in Λ are known, the rank of the matrix is easily verified in applications. In quantitative trait studies, $K > L$ and the proportions from different mixtures are linearly independent, giving the nonsingularity of $\Lambda^T \Lambda$.

3. ESTIMATING PHENOTYPIC DISTRIBUTIONS WITH UNCENSORED DATA

It seems natural to estimate $\tilde{F}(t)$ by replacing $\tilde{H}(t)$ in (2.2) with its empirical counterpart $\hat{H}(t) = \{\hat{H}_1(t), \ldots, \hat{H}_K(t)\}^T$, where $\hat{H}_k(t) = n_k^{-1} \sum_{j=1}^{n_k} I(X_{kj} \leqslant t)$ is the empirical distribution based on the sample from mixture k $(= 1, \ldots, K)$ and $I(\cdot)$ is the indicator function. The resulting plug-in estimator is $\hat{F}^*(t) = \{\hat{F}_1^*(t), \ldots, \hat{F}_L^*(t)\}^T = (\Lambda^T \Lambda)^{-1} \Lambda^T \hat{H}(t)$. Since $\hat{H}$ is unbiased for $\tilde{H}$, $\hat{F}^*(t)$ is unbiased for $\tilde{F}$. In finite samples, each component in $\hat{H}(t)$ is a left continuous step function. This means that $\hat{F}_l^*(t)$ $(l = 1, \ldots, L)$ has potential jumps (in t) at the distinct values in $(X_{kj}, k = 1, \ldots, K, j = 1, \ldots, n_k)$.

The large sample properties of $\hat{F}^*$ follow from those of $\hat{H}$. For $k = 1, \ldots, K$, $\hat{H}_k(t)$ is uniformly consistent for $H_k(t)$ for $t \in [0, \tau]$, where $F_l(\tau) < 1$ $(l = 1, \ldots, L)$. The uniform consistency of $\hat{F}^*(t)$ for $\tilde{F}(t)$ for $t \in [0, \tau]$ follows from a continuous mapping theorem on functional spaces (Andersen *et al.*, 1993). The empirical distributions $\hat{H}_i$ and $\hat{H}_j$ are computed from independent samples for $i \neq j$. Thus, it is easy to show that $n^{1/2}\{\hat{H}(t) - \tilde{H}(t)\}$ converges in distribution to a K-variate Gaussian process for $t \in [0, \tau]$ as $n \to \infty$ and $n_k n^{-1} \to \rho_k > 0$. The value ρ_k is the limiting proportion of observations from mixture k. Under these asymptotic conditions, the functional delta method (Andersen *et al.*, 1993) gives that $n^{1/2}(\hat{F}^* - \tilde{F})$ behaves like an L-variate Gaussian process.

When $K = L$ and Λ is full rank, there is a one-to-one transformation between $\tilde{H}$ and $\tilde{F}$. One might expect that the invariance property of maximum likelihood estimation would hold in this set-up. That is, since $\hat{H}$ is the nonparametric maximum likelihood estimator (NPMLE) for $\tilde{H}$ (Wellner, 1982), $\hat{F}^*$ is the NPMLE for $\tilde{F}$ and is efficient. However, as we discuss later, transforming $\hat{H}$ using the weight matrix may not yield $\hat{F}$ whose components satisfy the definition of distribution functions. This means that the estimator may be inadmissible. For the case where $\hat{F}$ is admissible, that is, each component is monotone increasing on [0, 1], it is the NPMLE, but this cannot be guaranteed in a given sample. An exception is when $\lambda_{kk} = 1, k = 1, \ldots, K$, in which case the data are from the component distributions and the estimators reduce to the usual empirical distributions.

When $K > L$, $\hat{F}^*$ may give large weight to $\hat{H}_k$ with small ρ_k, leading to suboptimal estimation. It is desirable to improve the efficiency of the plug-in estimator without compromising its ease of implementation and its theoretical tractability. This may be achieved by using the linear model $E\{N_{kj}(t)\} = \sum_{k=1}^{L} \lambda_{kl} F_l(t)$ to construct a least squares criterion for each $t \in [0, \tau]$, where $N_{kj}(t) =$

$I(X_{kj} \leqslant t)$. The component distributions can be estimated separately at different time points, which may not be possible with maximum likelihood. Observe that the mixture model for X_{kj} induces a linear model for the expectation of $N_{kj}(t)$, not a logistic model, which is popular for binary data. In the regression analogy, λ_{kl} and $F_l(t), l = 1, \ldots, L$, are the 'covariates' and the 'coefficients', respectively.

Now, consider $\hat{F}^{**}(t) = \{\hat{F}_1^{**}(t), \ldots, \hat{F}_L^{**}(t)\}^T$ minimizing

$$Q\{\tilde{F}(t)\} = \sum_{k=1}^{K} \sum_{j=1}^{n_k} \{N_{kj}(t) - \sum_{l=1}^{L} \lambda_{kl} F_l(t)\}^2. \tag{3.1}$$

The objective function (3.1) gives equal weight to each observation. If Λ is full rank and $n_k > 0, k = 1, \ldots, K$, then the estimator exists and has a closed form. Let $N(t)$ be the $n \times 1$ vector $\{N_{11}(t), N_{12}(t), \ldots, N_{1n_1}(t), \ldots, N_{Kn_K}(t)\}^T$ and let $\tilde{\Lambda}$ be the $n \times L$ matrix with kth row equal to the vector of mixture proportions for the kth element of $N(t), k = 1, \ldots, n$. The formula for $\hat{F}^{**}(t)$ is $(\tilde{\Lambda}^T\tilde{\Lambda})^{-1}\tilde{\Lambda}^T N(t)$. As with $\hat{F}^*$, the estimators in $\hat{F}^{**}$ are unbiased and are left continuous step functions with possible jumps at the distinct values in the observed data. That the estimators are piecewise constant in t is computationally convenient. One only minimizes Q at the change points.

Since the observations from each mixture are exchangeable, the estimator can be written as linear combinations of the empirical distributions. Some matrix algebra gives that $\hat{F}^{**} = \sum_{k=1}^{K} c_k \hat{H}_k$, where c_k is a random $L \times 1$ vector defined to be the summation of those n_k columns of $(\tilde{\Lambda}^T\tilde{\Lambda})^{-1}\tilde{\Lambda}^T$ with corresponding elements $N_{kj}(t)$ in $N(t), k = 1, \ldots, K$. If $K = L$, then one may establish that c_k equals the kth column of $(\Lambda^T\Lambda)^{-1}\Lambda^T, k = 1, \ldots, K$. For $K > L$, the two estimators are connected by showing that if $\rho_k = K^{-1}, k = 1, \ldots, K$, then, as $n \to \infty$, c_k has deterministic limit $\tilde{c}_k$ which is the kth column of $(\Lambda^T\Lambda)^{-1}\Lambda^T, k = 1, \ldots, K$, and $n^{1/2}(\hat{F}^{**} - \tilde{F})$ and $n^{1/2}(\hat{F}^* - \tilde{F})$ are asymptotically equivalent.

The linear representation for $\hat{F}^{**}(t)$ makes its uniform consistency and its weak convergence transparent. The derivations use similar arguments to those for $\hat{F}^*$. For inferences about $\tilde{F}$ using either $\hat{F}^*$ or $\hat{F}^{**}$, variance estimation is needed. We obtain results for a general estimator $\hat{F}^g = (\hat{F}_1^g, \ldots, \hat{F}_L^g)^T$ of the form $\sum_{k=1}^{K} w_k \hat{H}_k$, where $(w_k, k = 1, \ldots, K)$ are bounded $L \times 1$ vectors independent of $\hat{H}$. Assume the estimator is unbiased, that is $E(\hat{F}^g) = \tilde{F}$, and uniformly consistent, and has a limiting Gaussian distribution. Then, the exact (fixed n) covariance function, $\mathrm{cov}[n^{1/2}\{\hat{F}^g(s) - \tilde{F}(s)\}, n^{1/2}\{\hat{F}^g(t) - \tilde{F}(t)\}]$ is

$$\Sigma_n(s, t) = n \sum_{k=1}^{K} n_k^{-1} w_k w_k^T H_k(s)\{1 - H_k(t)\}$$

for $s \leqslant t$. The single summation occurs because $\mathrm{cov}(\hat{H}_i, \hat{H}_j) = 0, i \neq j$. A consistent estimator $\hat{\Sigma}(s, t)$ may be calculated with H_k in $\Sigma_n(s, t)$ replaced by $\hat{H}_k, k = 1, \ldots, K$. A more complicated but unbiased estimator is

$$\hat{\Sigma}^u(s, t) = n \sum_{k=1}^{K} n_k^{-1} w_k w_k^T \left[\hat{H}_k(s) - \left\{ \frac{1}{n_k(n_k - 1)} \sum_{j \neq j'} N_{jk}(s) N_{j'k}(t) \right\} \right].$$

A $(1 - 2\alpha)$ confidence interval for $F_i(t)\,(i = 1, \ldots, L)$ may be constructed with $\hat{\Sigma}$ (or $\hat{\Sigma}^u$) and the normal approximation. Let m be an invertible and differentiable function chosen to stabilize the variance of $\hat{F}_i^g$ and to bound the interval in [0, 1]. The endpoints of the interval are

$$m^{-1}\left[m\{\hat{F}_i^g(t)\} \pm n^{-1/2} \dot{m}\{\hat{F}_i^g(t)\} \sqrt{\hat{\Sigma}_i(t, t)}\, \psi_\alpha \right], \tag{3.2}$$

where $\dot{m}(x) = \mathrm{d}\{m(x)\}/\mathrm{d}x$, $\hat{\Sigma}_i(s,t)$ is the ith diagonal element of $\hat{\Sigma}(s,t)$, and ψ_α is the α quantile of the standard normal distribution.

The rank tests of Kruglyak and Lander (1995) may have low power to detect differences between the phenotypic distributions under certain alternatives. A test for homogeneity of the components may also be conducted using the proposed nonparametric estimators. For given t, the null hypothesis is $H_0 : A\tilde{F}(t) = 0$, where A is an $(L-1)\times L$ matrix containing $(L-1)$ linearly independent contrasts of $F_1(t), \ldots, F_L(t)$. Under H_0, the statistic

$$\mathcal{L}(t) = \{A\hat{F}^g(t)\}\{A\hat{\Sigma}(t,t)A^T\}^{-1}\{A\hat{F}(t)^g\}^T$$

has a chi-squared distribution with $L-1$ degrees of freedom. Evaluating the distribution of $\mathcal{L}$ as a process in $t \in [0,\tau]$ would enable omnibus testing procedures which are sensitive to differences amongst the component distributions at all time points. For example, using $\sup_t \mathcal{L}(t)$ would provide a statistic which is sensitive to all alternatives, unlike the test of Kruglyak and Lander (1995). Unfortunately, the theoretical developments are beyond the scope of the current paper and appear to be rather challenging. In practice, permutation methods (Churchill and Doerge, 1994) can be used to generate the reference distribution of the sup test under H_0.

A minor issue is that in small samples $\hat{F}_i^g(t)$ $(i = 1, \ldots, L)$ are not constrained to be monotone increasing in t and to be in [0, 1]. The problem occurs mainly in the right tail of the distribution, which is difficult to estimate well. Note that it applies to $\hat{F}^*$ and $\hat{F}^{**}$. We propose a straightforward modification of $\hat{F}_i^g$ which has the properties of a distribution function. The estimator is $\hat{F}_i^{gm}(t) = \min[\max_{s\leqslant t}\{\hat{F}_i^g(s)\}, 1]$. It accepts $\hat{F}_i^g(t)$ if it satisfies the constraints. Otherwise, it equals either the largest value of $\hat{F}_i^g(s)$ for $s \leqslant t$ or 1. If $\hat{F}_i^g$ is monotone and in [0, 1], then $\hat{F}_i^{gm} = \hat{F}_i^g$. Since $\hat{F}_i^g$ is uniformly consistent, the same holds for $\hat{F}_i^{gm}$. Under mild conditions, $n^{1/2}(\hat{F}^g - \tilde{F})$ and $n^{1/2}(\hat{F}^{gm} - \tilde{F})$ have the same limiting distribution, where $\hat{F}^{gm} = (\hat{F}_1^{gm}, \ldots, \hat{F}_L^{gm})^T$. Thus, inferences for $\hat{F}^{gm}$ may use $\hat{\Sigma}$.

4. ESTIMATING PHENOTYPIC DISTRIBUTIONS WITH CENSORED DATA

Next, we consider the case where the data are subject to right censoring. The potential censoring times $C_{kj}, k = 1, \ldots, K, j = 1, \ldots, n_k$, are independently distributed, where $\mathrm{pr}(C_{kj} > t) = G_k(t)$ is the survivor function for censoring in mixture k. It is assumed that C_{kj} and X_{kj} are independent. The observable data are $W_{kj} = \min(X_{kj}, C_{kj})$ and $\Delta_{kj} = I(X_{kj} \leqslant C_{kj})$, $k = 1, \ldots, K$, $j = 1, \ldots, n_k$.

It is not obvious that the least squares criterion can be adapted to the censored data. First, replace $N_{kj}(t)$ with $\tilde{N}_{kj}(t) = I(W_{kj} > t)$ in (3.1). Next, note that the binary response $\tilde{N}_{kj}(t)$ has expectation

$$\Pr(W_{kj} > t) = P_k(t) = G_k(t)\{1 - H_k(t)\}, \tag{4.1}$$

for $k = 1, \ldots, K$, $j = 1, \ldots, n_k$. The mixture model for H_k and equation (4.1) suggest substituting $G_k(t)\{\sum_{l=1}^L \lambda_{kl} S_l(t)\}$ for H_k in Q, where $S_l(t) = 1 - F_l(t)$, $l = 1, \ldots, L$. The modified criterion is

$$Q^*\{\tilde{S}(t)\} = \sum_{k=1}^{K}\sum_{j=1}^{n_k}\{\tilde{N}_{kj}(t) - G_k(t)\sum_{l=1}^{L}\lambda_{kl}S_l(t)\}^2,$$

where $\tilde{S}(t) = \{S_1(t), \ldots, S_L(t)\}^T$. Minimizing Q^* yields $(\tilde{\Lambda}^T\tilde{\Lambda})^{-1}\tilde{\Lambda}^T\tilde{N}^*(t)$, where $\tilde{N}^*(t)$ is $N(t)$ with $\tilde{N}_{kj}(t)\{G_k(t)\}^{-1}$ in place of $N_{kj}(t)$, $k = 1, \ldots, K$, $j = 1, \ldots, n_k$.

If censoring is fixed in advance, that is G_k is known, then the minimizer is unbiased and can be shown to be consistent and asymptotically normal using ideas from Section 3. In practice, the censoring distribution may be unknown. Assume $G_k = G, k = 1, \ldots, K$, and replace G_k in N^* with $\hat{G}$, the right

continuous version of the Kaplan–Meier estimator for G based on $(W_{kj}, 1 - \Delta_{kj}, k = 1, \ldots, K, j = 1, \ldots, n_k)$. Denote this estimator for $\tilde{S}$ by $\hat{S}^{**} = (\hat{S}_1^{**}, \ldots, \hat{S}_L^{**})^T$. It is easy to see that

$$\hat{S}^{**}(t) = \{\hat{G}(t)\}^{-1} \sum_{k=1}^{K} c_k \hat{P}_k(t),$$

where c_k is given in Section 3 and $\hat{P}_k(t) = n_k^{-1} \sum_{j=1}^{n_k} \tilde{N}_{kj}(t)$ is the empirical estimator for $P_k(t)$ in (4.1). Each term in $\hat{S}^{**}$ is a right continuous step function with possible jumps at the distinct values in $(W_{kj}, k = 1, \ldots, K, j = 1, \ldots, n_k)$. When the censoring distributions differ amongst mixtures, one may use $\hat{G}_k$ instead of $\hat{G}$, where $\hat{G}_k$ is the Kaplan–Meier estimator for G_k using only data from mixture k. The theoretical developments are slightly more difficult than the results for $G_k = G$ below.

Because of $\hat{G}$, $\hat{S}^{**}$ may not be unbiased in finite samples. However, $\hat{G}(t)$ converges uniformly to $G(t)$ and $\hat{P}_k(t)$, the empirical distribution for $\tilde{N}_{kj}(t)$, converges uniformly to $P_k(t)$ $(k = 1, \ldots, K)$, for $t \in [0, \tau]$. Therefore, assuming that $G(\tau) > 0$, a continuous mapping theorem gives the uniform consistency of $\hat{S}^{**}$ for $\tilde{S}$.

Obtaining the weak convergence of $n^{1/2}(\hat{S}^{**} - \tilde{S})$ is challenging compared to without censoring. The issues are that $\hat{G}$ and $\hat{P}_k$ are dependent for $k = 1, \ldots, K$, and the influence function for $\hat{G}$ involves counting process martingales. Deriving the result requires considering the joint distribution of $(\hat{G}, \hat{P}_1, \ldots, \hat{P}_K)$. In the Appendix, we demonstrate that

$$J(t) = n^{1/2}\{\hat{G}(t) - G(t), \hat{P}_1(t) - P_1(t), \ldots, \hat{P}_K(t) - P_K(t)\}^T$$

converges weakly to a $(K + 1)$-variate Gaussian process on $[0, \tau]$ with covariance function $\Psi(s, t)$. The functional delta method provides the weak convergence of $n^{1/2}\{\hat{S}^{**}(t) - \tilde{S}(t)\}$.

The asymptotic covariance function $\Sigma^*(s, t) = \text{cov}[n^{1/2}\{\hat{S}^{**}(s) - \tilde{S}(s)\}, n^{1/2}\{\hat{S}^{**}(t) - \tilde{S}(t)\}]$ is $\kappa(s)\Psi(s, t)\kappa(t)^T$ for $s \leqslant t$, where κ is defined in the Appendix. The theoretical quantities in Σ^* may be estimated with plug-in formulae using the observed data. The details are tedious and are sketched in the Appendix. Denote the estimators for κ and Ψ by $\hat{\kappa}$ and $\hat{\Psi}$, respectively. A consistent estimator for $\Sigma^*(s, t)$ is $\hat{\Sigma}^*(s, t) = \hat{\kappa}(s)\hat{\Psi}(s, t)\hat{\kappa}(t)^T$. As with uncensored data, inferences may be based on the asymptotic normality of the estimators and the variance estimates. For example, a $(1 - 2\alpha)$ confidence interval for $S_i(t)$ is

$$m^{-1}\left[m\{\hat{S}_i^{**}(t)\} \pm n^{-1/2}\dot{m}\{\hat{S}_i^{**}(t)\}\sqrt{\hat{\Sigma}_i^*(t, t)}\psi_\alpha\right], \tag{4.2}$$

where $\hat{\Sigma}_i^*(t, t)$ is the ith diagonal element of $\hat{\Sigma}^*(t, t)$.

For small n, the estimators in $\hat{S}^{**}$ may not be survivor functions. They can be modified similarly to $\hat{F}_g$ in Section 3 without affecting the validity of the large sample inferential procedures.

5. Examples

5.1 *Background*

In this section, we give background material for the examples in Sections 5.2 and 5.3, which involve the analysis of data from backcross and double haploid experiments.

In a backcross, there are two inbred parents which differ in a quantitative trait. The allele of parent P1 is labelled m and that of P2 is labelled M. Because of inbreeding, P1 has genotype m/m at each locus, while P2 has genotype M/M at each locus. Mating P1 and P2 yields an F1 generation which is m/M at each locus. Crossing F1 with a parent, P1 say, generates the backcross progeny, BC. By the laws of

Mendelian inheritance, the BC specimens have equal probability of genotypes m/M and m/m at each locus. Within each BC, the genotype may vary amongst loci depending on the pattern of crossover events during meiosis in the F1 parent.

In a double haploid experiment, the haploid gametes from an F1 individual are chemically treated to create double haploid, DH, individuals. Ordinarily, gametes only contain one of the F1 chromosomes, either that from P1 or that from P2. However, the treatment causes a duplication leading to individuals with two identical copies of the parental chromosome. These specimens are either m/m or M/M at each locus with probability 0.5. As in the backcross, recombination in F1 during meiosis means that the genotype for DH individuals may vary amongst loci.

While the constructions of the BC and DH lines are quite different, the rationale for analysing the data is identical. The idea is to test for differences in the distributions of the traits amongst genotypes across the genome. That is, one screens all loci for genetic effects. Genotypes are only observed at markers at known chromosomal positions. At these loci, the mixing weights are either 0 or 1, and each mixture corresponds to a component distribution for a particular genotype. However, for both BC and DH, for a locus in an interval between markers, there are two possible genotypes and four possible pairs of genotypes at the flanking markers. Hence, $L = 2$ and $K = 4$.

Using basic genetic principles, the distribution of genotypes between markers may be computed conditional on the genotypes at the markers. To illustrate, suppose a QTL is situated between a left marker $M1$ and a right marker $M2$. Let the recombination fractions between the QTL and $M1$ and between the QTL and $M2$ be θ_1 and θ_2, respectively. In BC, if the genotypes at $M1$ and $M2$ are m/m and m/m, then with probability $p_{mm} = (1-\theta_1)(1-\theta_2)\{(1-\theta_1)(1-\theta_2)+\theta_1\theta_2\}^{-1}$ the genotype at the QTL is m/m, and with probability $1 - p_{mm}$ the genotype is m/M. This calculation is based on the Haldane (1919) map function, which assumes recombination events in nonoverlapping intervals are independently distributed. In DH, the same logic gives that the probabilities of m/m and M/M at the QTL, when both flanking genotypes are m/m, are also p_{mm} and $1 - p_{mm}$, respectively.

Such reasoning generates the conditional probability distributions for the genotype at the QTL for all possible pairs of genotypes at the flanking markers. It is these distributions which define the mixing weights. For instance, if $k = 1$ denotes m/m at $M1$ and m/m at $M2$ and $l = 1$ denotes m/m at the QTL, then $\lambda_{11} = p_{mm}$ and $\lambda_{12} = 1 - p_{mm}$.

5.2 *Mammary tumors in rats*

Female rats from the WKy strain resistant to carcinogenesis were crossed with male rats from the WF strain (Lan *et al.*, 2001). The F1 progeny were mated to WF animals, producing 383 female rats in the BC generation. These backcross rats were scored for number of mammary carcinomas and were genotyped at 58 markers on chromosome 5. Using Mapmaker/QTL, (Lan *et al.*, 2001) found that marker D5Rat22 was strongly associated with lower tumor counts. That is, the female rats with a WKy allele at DFRat22 had fewer tumors than those rats with no WKy allele. The mean numbers of counts estimated from the normal mixture are 2.68 and 5.43 for the WKy/WF and WF/WF genotypes, respectively.

The data are now analysed with nonparametric methods. Lan *et al.* (2001) used rank tests (Kruglyak and Lander, 1995) to confirm the genomic region linked to the tumor counts. The maximal test statistic is at a site very near to D5Rat22, between markers at 42.88 cM and 45.88 cM on chromosome 5. We compute nonparametric estimates of the carcinoma distributions for the WKy/WF and WF/WF genotypes at this locus. The recombination fractions are $\theta_1 = 0.011$ and $\theta_2 = 0.018$, giving $\lambda_{11} = 0.9998$, $\lambda_{21} = 0.627$, $\lambda_{31} = 0.373$ and $\lambda_{41} = 0.0002$, with $\lambda_{k2} = 1 - \lambda_{k1}$, $k = 1, \ldots, 4$.

The estimated tumor count distributions for genotypes Wky/WF and WF/WF are displayed in Figure 1 along with 0.95 pointwise confidence intervals based on the untransformed estimators with $m(u) = u$ in (3.2). The plots exhibit that WF/WF rats have higher tumor counts. The estimated means in the WKy/WF

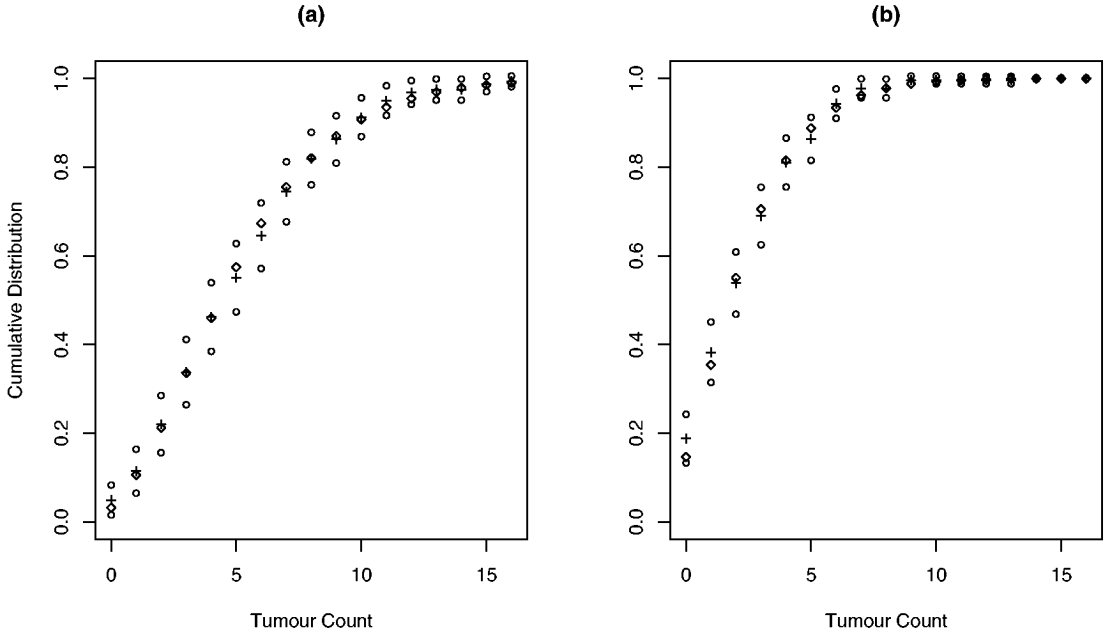

Fig. 1. Nonparametric point estimates (+) and 0.95 pointwise confidence intervals (○) for tumor count distributions from Lan *et al.* (2001), and point estimates (◇) from the negative binomial mixture model. (a) WF/WF, (b) WKy/WF.

and WF/WF groups are $\int x \mathrm{d}\hat{F}_1(x) = 2.64$ and $\int x \mathrm{d}\hat{F}_2(x) = 5.46$, respectively, which agrees with Mapmaker/QTL. However, the estimated distributions from the normal mixture are rather different from the nonparametric estimates. The estimates from the normal mixture are not shown here but their lack of fit is evidenced in Figure 2 of Zou *et al.* (2002). Instead, the estimated components from a model with F_1 and F_2 assumed to be negative binomial, which was fitted by Lan *et al.* (2001), are displayed. These fall entirely within the 0.95 limits, indicating that this model matches the data well.

Interestingly, while the normal model does not seem to fit very well, the estimates of the gene's location are rather similar to those from the rank and negative binomial mapping procedures. Research on lod score analyses of pedigree data in human genetics suggests that, in certain cases, estimates of genetic parameters may be robust to misspecification of the genetic model (Clerget-Darpoux, 1982; Clerget-Darpoux *et al.*, 1986). Preliminary numerical studies (Zou *et al.*, 2003) indicate that similar properties may be shared by likelihood based analyses of quantitative trait loci in plant and animal studies. However, under model misspecification, the parametric analyses may be less efficient than the rank based tests (Kruglyak and Lander, 1995). A rigorous theoretical analysis of the behavior of the statistical procedures under different types of genetic model misspecifications has not yet been undertaken.

5.3 *Flowering times in Brassica napus*

Flowering in *Brassica napus* is important in oilseed production. Understanding the genes controlling the trait may be valuable for commercial breeding programs. Some variants require a period of low temperature (vernalization) to flower (biennials), while others do not (annuals). In this experiment, a single plant of cv Major (a biennial rapeseed cultivar) was crossed as a female to a double haploid line

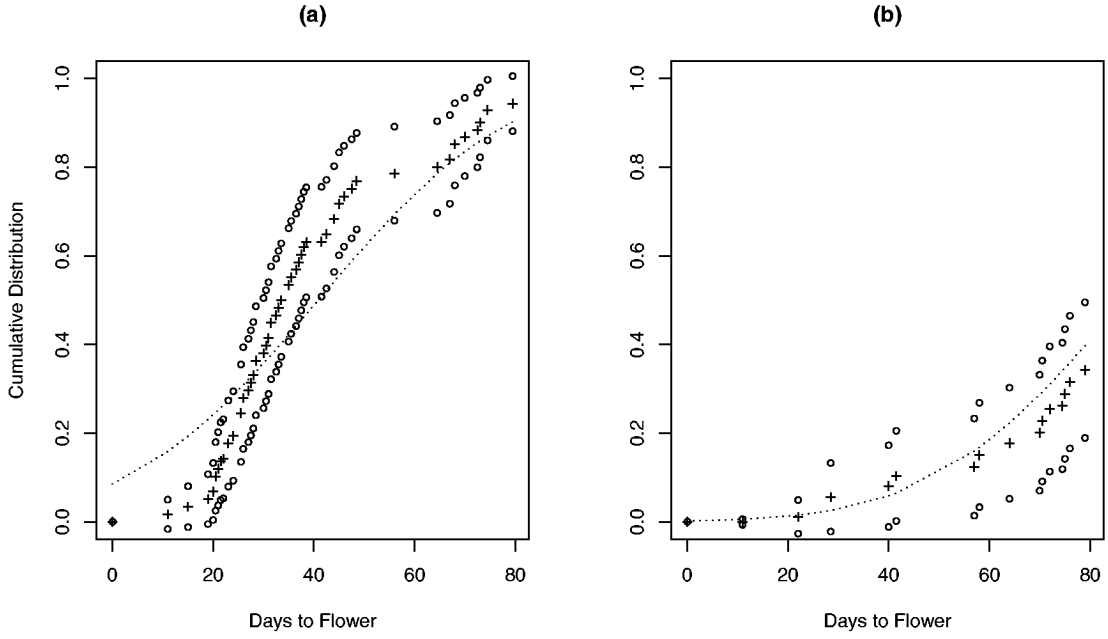

Fig. 2. Points estimates (+) and 0.95 pointwise confidence limits (o) for flowering time distributions from Ferreira *et al.* (1995). Dotted lines are point estimates from the normal mixture model. (a) S/S, (b) M/M.

from cv Stellar (an annual canola cultivar). Ninety-nine F1-derived DH lines were evaluated in the field for flowering time. Plants from each line were given no vernalization, 4 weeks or 8 weeks of vernalization.

We analyse the non-vernalization data. Out of 99 plants, 27 plants had not flowered after 83 days. In Ferreira *et al.* (1995), the flowering time was set to 100 days for the non-flowering plants and Mapmaker/QTL was used with F_l assumed normal. The genomic region N2 was strongly associated with flowering time. Censoring has not been formally addressed for quantitative trait loci in the genetics literature and a typical analysis is the one just described. Because the censoring time is fixed and hence independent of the flowering time in this example, the ad hoc approach is suitable for testing the null hypothesis of no genotypic effects, but is invalid for estimating the phenotypic distributions. It is worth emphasizing that with general random censoring patterns, this strategy is invalid and may lead to biased tests and biased estimates of a gene's location. Such patterns occur frequently in animal studies where dropout cannot be completely controlled by the experimenter.

The estimated means from the normal mixture are 41 days and 81 days for S/S and M/M, respectively, where S denotes the Stellar allele and M the Major allele. Since the data are heavily skewed and the censoring rate is high, the estimates may severely underestimate the true means. Nonparametric methods which avoid assumptions on the tails of the distributions seem more appropriate. In fact, since flowering is not observed beyond 83 days, the distributions are not identifiable beyond that time. This means that nonparametric estimation of the mean flowering times is not possible and brings into question the values in Ferreira *et al.* (1995). We will report medians, which are supported by the observed data.

Setting the censored times to 100, the nonparametric testing procedure in Kruglyak and Lander (1995) reaches a maximum at a site in N2 close to that in Ferreira *et al.* (1995). The locus is between markers at 63.4 cM and 70.6 cM in this region, with $\theta_1 = 0.044$ and $\theta_2 = 0.025$. The mixing weights are $\lambda_{11} = 0.999, \lambda_{21} = 0.361, \lambda_{31} = 0.639$ and $\lambda_{41} = 0.001$, with $\lambda_{k2} = 1 - \lambda_{k1}, k = 1, \ldots, 4$. Figure 2

exhibits $1 - \hat{S}^{**}$ and 0.95 pointwise confidence intervals based on the untransformed estimators for S/S and M/M with $m(u) = u$ in (4.2). The estimated distributions from Mapmaker/QTL are also shown and fit poorly, particulary in the S/S group. Based on the nonparametric estimates, the estimated median is ≈ 33 days for S/S and the estimate has a lower bound of 83 days for M/M. Without vernalization, substitution of M/M for S/S substantially increases the flowering time.

6. Remarks

Nonparametric methods provide robust analyses for genetic studies of quantitative traits. The distribution estimators complement existing rank tests and are useful in summarizing genetic effects for comparing different genes and in assessing parametric assumptions, e.g. normality. Diagnosing model misspecification is an area which has received little attention in the genetics literature. Formal tests of the assumptions in Mapmaker/QTL (Lander and Botstein, 1989) and other mapping software are not available. We plan future work on the construction of numerical goodness-of-fit tests using the nonparametric techniques.

If the normality assumption holds, then there may be a modest reduction in power to detect genes using the rank tests and a modest loss of efficiency in estimating the magnitude of the genetic effects (Zou *et al.*, 2003). These losses are comparable to those in the simple two-sample problem where the t-test and rank sum test have been thoroughly studied. While the likelihood analyses give unbiased estimates of the location of a gene under model misspecification when using permutation methods (Churchill and Doerge, 1994), empirical studies have shown large gains in power and efficiency with rank methods when the data are either heavy tailed or skewed (Zou *et al.*, 2003). In addition, maximum likelihood estimation of the genetic effects may exhibit large biases under model misspecification, which may impact the comparison of genes. For this reason, we recommend robust nonparametric methods, particularly when it is obvious that the normality assumption fails to hold, as in Sections 5.2 and 5.3.

An R function implementing the nonparametric procedure in Section 3 is available at http://www.bios.unc.edu/~fzou. It provides estimates of the phenotypic distributions, their standard errors, and their means and medians, at a fixed locus along with nonparametric hypothesis tests based on these estimates, including $\sup \mathcal{L}(t)$ and $\int \mathcal{L}(t)$. The function is compatible with R/qtl (Broman *et al.*, 2003), which should enable genome screens using permutation methods (Churchill and Doerge, 1994). It would be worthwhile to conduct simulations to evaluate the extent to which power to detect genes and efficiency in estimation of genetic effects is either gained or lost using our new nonparametric approach relative to existing rank methods and likelihood methods.

The single gene models for backcross and double haploid experiments in Section 5.1 have $K = 4$ and $L = 2$ in (1.1). An alternative to the backcross design is the intercross design, where an F1 is mated to another F1, instead of to a P1, which leads to $K = 9$ and $L = 3$. The nonparametric formulation is also applicable to genetic models with multiple loci (Zeng *et al.*, 1999). The estimators should prove helpful in evaluating complex genetic hypotheses, such as whether a gene has either an additive or a dominant effect and whether several genes have nonlinear interactions, e.g. epistasis. Typically, such hypotheses are explored in the context of parametric models, specifically, the linear model with normal errors. The model (1.1) allows these issues to be addressed nonparametrically.

Acknowledgments

The authors are grateful to the associate editor and referee for helpful comments.

APPENDIX A

The martingale representation for $n^{1/2}(\hat{G} - G)$ (Gill, 1980, p. 37) is

$$n^{1/2}\{\hat{G}(t) - G(t)\} = n^{-1/2} \sum_{k=1}^{K} \sum_{j=1}^{n_k} g_{kj}(t) + o_p(1), \tag{A.1}$$

where $g_{kj}(t) = -G(t) \int_0^t \pi(u)^{-1} \mathrm{d}M_{kj}(u)$, $\pi(u) = G(u) \sum_{k=1}^{K} \rho_k \{1 - H_k(u)\}$ is the deterministic limit of $\hat{\pi}(u) = n^{-1} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \tilde{N}_{kj}(u)$, $M_{kj}(u) = \{1 - \tilde{N}_{kj}(u)\}\Delta_{kj} - \int_0^u \tilde{N}_{kj}(s)\gamma_c(s)\mathrm{d}s$ and $\gamma_c(u) = -\mathrm{d}[\log\{1 - G(u)\}]/\mathrm{d}u$ is the common hazard function of $(C_{kj}, k = 1, \ldots, K, j = 1, \ldots, n_k)$. A martingale limit theorem provides the weak convergence of (A.1) for $t \in [0, \tau]$. The empirical process $n^{1/2}\{\hat{P}_k(t) - P_k(t)\}$ also converges weakly on $[0, \tau]$ and is asymptotically equivalent to

$$n^{-1/2} \sum_{j=1}^{n_k} h_{kj}(t) + o_p(1), \tag{A.2}$$

where $h_{kj}(t) = \rho_k^{-1/2}\{\tilde{N}_{kj}(t) - P_k(t)\}$.

Combining (A.1) and (A.2) gives $J(t) = n^{-1/2} \sum_{k=1}^{K} \sum_{j=1}^{n_k} r_{kj}(t) + o_p(1)$, where $r_{kj}(t) = \{g_{kj}(t), I(k = 1)h_{1j}(t), \ldots, I(k = K)h_{Kj}(t)\}^T$. The multivariate c.l.t. gives convergence to a $(K+1)f$-variate normal distribution for any finite collection of times $0 \leqslant t_1, \ldots, t_f \leqslant \tau$, $f < \infty$. The weak convergence of the terms in (A.1) and (A.2) gives tightness of these processes in t. Hence, the convergence in distribution of $J(t)$ is uniform on $[0, \tau]$. The limiting covariance function $\mathrm{cov}\{J(s), J(t)\} = \Psi(s, t)$, $s \leqslant t$, may be consistently estimated by replacing theoretical quantities with empirical estimates.

Since $\hat{S}^{**}(t)$ is compactly differentiable in $\{\hat{G}(t), \hat{P}_1(t), \ldots, \hat{P}_K(t)\}$, the functional delta method gives that $Z(t) = n^{1/2}\{\hat{S}^{**}(t) - \tilde{S}(t)\}$ converges weakly. The covariance function $\Sigma^*(s, t) = \mathrm{cov}\{Z(s), Z(t)\}$, $s \leqslant t$ is $\kappa(s)\Psi(s, t)\kappa(t)^T$, where the jth column of $\kappa(s)$ is the partial derivative of $G(s)^{-1} \sum_{k=1}^{K} \tilde{c}_k P_k(s)$ with respect to the jth element of $\{G(s), P_1(s), \ldots, P_K(s)\}$, $j = 1, \ldots, K + 1$.

The terms κ, Ψ are estimated by replacing theoretical quantities with empiricals. Let $\hat{\Psi}(s, t) = n^{-1} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \hat{r}_{kj}(t)\hat{r}_{kj}(s)^T$, where $\hat{r}_{kj}(t) = \{\hat{g}_{kj}(t), I(k = 1)\hat{h}_{1j}(t), \ldots, I(k = K)\hat{h}_{Kj}(t)\}^T$. The term $\hat{g}_{kj}(t) = -\hat{G}(t) \int_0^t \hat{\pi}(u)^{-1} \mathrm{d}\hat{M}_{kj}(u)$, where $\hat{\pi}(u) = n^{-1} \sum_{k=1}^{K} \sum_{j=1}^{n_k} I(W_{kj} > u)$, $\hat{M}_{kj}(u)$ equals $\{1 - \tilde{N}_{kj}(u)\}\Delta_{ij} - \int_0^u \tilde{N}_{kj}(s)\mathrm{d}\hat{\Gamma}_c(s)$ and $\hat{\Gamma}_c(u)$ is the Nelson–Aalen estimator of the cumulative hazard function for censoring distribution. The term $\hat{h}_{kj}(t) = (n_k/n)^{-1/2}\{\tilde{N}_{kj}(t) - \hat{P}_k(t)\}$. For fixed s, $\kappa(s)$ is a $L \times (K + 1)$ matrix with first column equal to $-G(s)^{-1}\tilde{S}(s)$ and $(j + 1)$th column equal to $G(s)^{-1}\tilde{c}_j$, $j = 1, \ldots, K$. Define $\hat{\kappa}$ to be κ with c_k, $\hat{G}$ and $\hat{S}^{**}$ in place of $\tilde{c}_k$, G and $\tilde{S}$.

REFERENCES

ANDERSEN, P. K., BORGAN, O., GILL, R. D. AND KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

BROMAN, K., WU, H., SEN, S. 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.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

CLERGET-DARPOUX, F. M. (1982). Bias of the estimated recombination fraction and lod score due to an association between a disease gene and a marker. *Annals of Human Genetics* **46**, 363–372.

CLERGET-DARPOUX, F. M., BONAITI-PELLIE, C. AND HOCHEZ, J. (1986). Effects of misspecifying genetic paramters in lod score analysis. *Biometrics* **42**, 393–399.

DOERGE, R. W., ZENG, Z.-B. AND WEIR, B. S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* **12**, 195–219.

FERREIRA, M. E., SATAGOPAN, J., YANDELL, B. S., WILLIAMS, P. H. AND OSBORN, T. C. (1995). Mapping loci controlling vernalization requirement and flower time in *Brassica napus*. *Theoretical and Applied Genetics* **90**, 727–732.

GILL, R. D. (1980). *Censoring and Stochastic Integrals: Mathematical Centre Tracts 124*. Amsterdam: Mathematisch Centrum.

HALDANE, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 299–309.

HALL, P. (1981). On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B* **43**, 147–156.

HALL, P. AND TITTERINGTON, D. M. (1984). Efficient nonparametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Series B* **46**, 465–473.

HALL, P. AND TITTERINGTON, D. M. (1985). The use of uncategorized data to improve the performance of a nonparametric estimator of a mixture density. *Journal of the Royal Statistical Society, Series B* **47**, 155–163.

HOSMER, D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics* **29**, 761–770.

KAPLAN, E. L. AND MEIER, P. (1958). Nonparametric estimation from incomplete observations, I, II. *Journal of the American Statistical Association* **53**, 457–481, 562–563.

KRUGLYAK, L. AND LANDER, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

LAN, H., KENDZIORSKI, C. M., HAAG, J. D., SHEPEL, L. A., NEWTON, M. A. AND GOULD, M. N. (2001). Genetic loci controlling breast cancer susceptibility in the Wistar–Kyoto rat. *Genetics* **157**, 331–339.

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

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Report* **50**, 163–170.

MURRAY, G. D. AND TITTERINGTON, D. M. (1978). Estimation problems with data from a mixture. *Journal of the Royal Statistical Society, Series C* **27**, 325–334.

TITTERINGTON, D. M., SMITH, A. F. M. AND MAKOV, U. E. (1985). *Analysis of Finite Mixture Distributions*. New York: Wiley.

WELLNER, J. (1982). Asymptotic optimality of the product limit estimator. *Annals of Statistics* **10**, 595–602.

ZENG, Z. B., KAO, C. H. AND BASTEN, C. J. (1999). Estimating the genetic architecture of quantitative traits. *Genetical Research* **74**, 279–289.

ZOU, F., FINE, J. P. AND YANDELL, B. S. (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika* **89**, 61–75.

ZOU, F., YANDELL, B. S. AND FINE, J. P. (2003). Rank-based statistical methodologies for quantitative trait locus mapping. *Genetics* **165**, 1599–1605.

[*Received 15 December 2003; accepted for publication 5 March 2004*]