Journal of Statistical Theory and Applications, Vol. 12, No. 4 (December 2013), 392-402

# Interval Mapping Using Nonparametric Accelerated Failure Time Cure Model

**Devrim Bilgili [1] *, Nader Ebrahimi [2]**

*[1] Department of Mathematics and Statistics, University of North Florida,*
*1 UNF Drive,*
*Jacksonville, FL 32224, USA*

*E-mail: devrim.bilgili@unf.edu*

*[2] Department of Mathematical Sciences, Statistics Divison, Northern Illinois University*
*1425 W. Lincoln Hwy,*
*Dekalb, IL 60115, USA*

*E-mail: nader@math.niu.edu*

Received 29 October 2012

Accepted 9 August 2013

**Abstract**

Many important problems in evolutionary biology begin with observations of phenotypic variation. Suppose time to an event-data are used to map quantitative trait loci (QTL) and underlying population is a mixture of susceptible and non-susceptible subjects. If the cured subjects are ignored we may fail to detect the responsible genetic factors or find false significant locations. In this article, we propose a nonparametric accelerated failure time cure model which takes cured subjects as well to model time to an event.

*Keywords:* Survival Function, Interval Mapping, AFT Cure Model, LOD score, EM, Imputation.

## 1. Introduction

Finding and mapping genetic loci(genes, markers are genetic loci) which is responsible for variation in a quantitative phenotype is a key step toward understanding the molecular bases of a disease. With the development of genetic markers and genetic linkage maps, using data from experimental crosses, it is now possible to detect and localize chromosomal regions of interest known as quantitative trait loci (QTL) by applying models relating observed phenotype values with genotypes. In standard interval mapping of QTL, the trait(or some transformation of trait) distribution is often modeled as a mixture of two (or more) normal components corresponding to two (or more) different genotypes at the putative QTL, see Lander and Botstein (1989), Zeng (1993) and many references cited there. Diao et al., (2004) proposed QTL mapping for censored observations. Non-parametric methods have been also developed to test the presence of QTL, see Krugylak and Lander (1995), Broman (2003), Poole and Drinkwater (1996), Basrak et al. (2004), Fine, Zou, Yandell (2004), Li, Boehnke, Abecasis, Song (2006), Zak et al. (2007), Manichaikul (2008) and many references cited there.

In some genetic studies, with survival end points, there are heterogenous populations of subjects which is divided into two groups. One group consists of subjects who become immune or insusceptible to a disease. They are said to be cured. The other group consists of susceptible subjects who would eventually experience the event in the absence of censoring. They are said to be uncured. Survival models with a cure rate referred to as “cure rate models” have received much attention

* 1 UNF Drive Jacksonville, FL 32224 USA.

Published by Atlantis Press

in recent years. These models are useful when a proportion of study subjects are cured. In fact, the cure rate models have been used for modeling time to an event data for various types of cancers, including breast cancer, leukemia, prostate cancer and hand and neck cancer, where a significant proportion of patients are cured, see Farewell (1986), Kuk and Chen (1992), Maller and Zhou (1992), Sposto et al. (1992), Lu and Ying (2004) and many references cited there for various cure models.

When the primary phenotype is time to event and the proportion of a population is cured, various cure models have been used to model phenotype distribution. See Broman (2003), Liu et al. (2006) and others. In this paper we focus on joint estimation of QTL location and its effects on survivals of both cured and non-cured subjects using nonparametric accelerated failure time (AFT) cure model. See Bilgili (2009) and Piao et al.(2011) for parametric AFT cure models. Recently, Xua and Zhan (2010) proposed a multiple imputation method to impute cure and not cured latent variable based on the rank estimation method and profile likelihood method for nonparametric AFT cure model. They assumed all the covariates are observable in AFT part. However, in our situation some of the covariates in AFT part are also not observable, i.e., some covariates are treated as latent variables. There are two ways to estimate QTL location as well as its effect on survivals of both cured and non-cured subjects for our situation. The first approach is to borrow strength from the marker information. Second approach is to use the imputation method. In this paper, we use both EM method as well as multiple imputation method where genotypes, unobservable covariates, are imputed randomly, but they are conditioned on the observed marker data. In other words, we simulate from the joint genotype distribution given the observed data for all subjects. We then follow Wei (1990,1992), Jin et.al (2003) and Zhang and Peng (2007) for estimation of unknown parameters.

The paper is organized as follows. Section 2 introduces notation and model. In Section 3, we obtain the estimates of unknown parameters in the model using the imputation method. Section 4 is devoted to the analysis of Listeria monocytogenes data. In Section 5 we illustrate simulation study. Some concluding remarks are given in Section 6. Throughout this paper we consider only single QTL. However, the methods described in this paper can easily be extended to multiple QTL and/or different designs.

## 2. Notation and Model Specifications

A cure model can be considered as a survival model where part of the population is not affected by the hazard of interest. More specifically, consider $n$ subjects in the study. Let $T_i$ be the potential time-to-event for the $i$-th subject. Then,

$$T_i = \eta_i T_i{}^* + (1-\eta_i)\infty, i = 1, \cdots, n, \tag{1}$$

where $\eta_i$ is the susceptibility indicator which attains a value 1 if the subject is susceptible otherwise 0 and $T_i{}^* < \infty$ is the failure time for the $i$-th subject when the subject is susceptible.

We consider the following model for $W_i{}^* = \ln T_i{}^*$.

$$W_i{}^* = \ln T_i{}^* = \beta' z_i + \beta_0 + \beta_g G_i + \varepsilon, \tag{2}$$

where $z_i$ is the covariates of interest, such as environmental factors which are assumed to be independent of $G_i$, $i = 1, \cdots, n$ and $\beta' = (\beta_1, \cdots, \beta_p)$ and $G_i$'s are genotypes of QTLs. Let the random error $\varepsilon$ in the equation (2) has the survival function $S_0$. Then, it is clear that

$$\begin{aligned} S_{W_i{}^*}(w_i{}^*) &= P(W_i{}^* > w_i{}^*) \\ &= P(\beta' z_i + \beta_0 + \beta_g G_i + \varepsilon > w_i{}^*) \\ &= P(\varepsilon > w_i{}^* - \beta' z_i - \beta_0 - \beta_g G_i) \\ &= S_0(w_i{}^* - \beta' z_i - \beta_0 - \beta_g G_i). \end{aligned}$$

In our setup, censoring indicator can be defined as $\delta_i = I(W_i{}^* \leqslant C_i) = \min(W_i{}^*, C_i)$ where $C_i$ is the random censoring time and the censoring time is assumed to be noninformative, $i = 1, \cdots, n$. That is, $\delta_i = 1$ if the actual failure time is observed and $\delta_i = 0$ if the censoring time is observed for the $i$-th individual.

A logistic model is assumed for the susceptibility indicator $\eta_i$. That is,

$$\pi(G_i, z_i) \equiv P(\eta_i = 1|G_i, z_i) = \frac{e^{\gamma' z_i + \gamma_0 + \gamma_g G_i}}{1 + e^{\gamma' z_i + \gamma_0 + \gamma_g G_i}}, \tag{3}$$

where $\gamma' = (\gamma_1, \cdots, \gamma_p)$ and $i = 1, \cdots, n$.

Several remarks are in order in regards to equations (2) and (3).

(i) In Zhang and Peng (2007), since all covariates are assumed to be observed, they combined $z_i$ and $G_i$. However, in our situation to distinguish between observable and unobservable covariates we separate the two terms.

(ii) In (2) we make no assumptions on $S_0$, distribution of $\varepsilon$. The model in (2) is often called the nonparametric AFT model, see Cheng and Tzheng (2009). Combination of (2) and (3) are referred to as the nonparametric AFT cure model. It should be noted that if we assume some distribution for $\varepsilon$ in (2), then we refer to it as the parametric AFT. For details see Bilgili (2009). To avoid confusion, throughout the paper we will assume $G_i$'s has one locus.

## 3. Estimation of Unknown Parameters

If $G_i$ is observable for $i = 1, \cdots, n$ in (2) and (3), then one can use the method proposed by Xu and Zhan (2010) to estimate all the unknown parameters in both models. In our situation, however, the problem is that $G_i$, genotypes of QTLs, are not observable for $i = 1, \cdots, n$. In this section first we propose methods to obtain the genotypes and then estimate unknown parameters in (2) and (3).

### *3.1. Obtaining $G_i$ using imputation and EM algorithm*

#### *3.1.1. Obtaining $G_i$ using imputation*

One way to obtain $G_i$'s is to use the hidden Markov models, see Baum(1970), and multiple imputation which imputes all the missing genotype data and then perform standard interval mapping methods for QTL mapping. Note that here we need to impute the missing genotype data multiple times and as a rule of thumb, usually 16 different imputations are recommended. In fact, the more the missing $G_i$ the more imputations are needed. Imputations were conducted with the help of Rqtl package using *sim.geno* function by specifying *n.draws* argument. It should be noted that the imputed genotypes at the markers match those observed assuming no genotyping errors have been made. For general case see Broman(2009).

Sen and Churchill (2001) use imputation method in their work and successfully apply this method on reanalysis of a hypertension cross described in Sugiyama et al. (2001). Li (2009) et al., showed that genotype imputation works very well for the susceptibility locus for age-related macular degeneration. In their imputation approach, they masked 5% of the genotypes at the locus and showed that masked genotypes could be imputed correctly >99% of the time.

#### *3.1.2. EM algorithm*

In order to obtain the unknown genotypes (maybe a putative QTL), QTL genotype probabilities need to be obtained conditioned on the marker data using the nearest flanking markers. Flanking marker is an identifiable region located close to a gene which is used in linkage studies to understand how gene under investigation is inherited. For example, in a backcross design, $G_i = 1/0$ stands for the genotypes $AA/AB$ respectively. We look for the probability that an individual attains genotype $AA$ or $AB$ at the locus based on the marker data. Table 1 displays these probabilities: $P(AA|M_1 = AA, M_2 = AA), P(AB|M_1 = AA, M_2 = AA)$, $P(AA|M_1 = AA, M_2 = AB), P(AB|M_1 = AA, M_2 = AB)$, $P(AA|M_1 =$

$AB, M_2 = AA), P(AB|M_1 = AB, M_2 = AA)$ and $P(AA|M_1 = AB, M_2 = AB), P(AB|M_1 = AB, M_2 = AB)$. In Table 1, $r_{12}$ denotes the recombination fraction between two nearest flanking markers in a backcross and $r_{sP}$ is the recombination fraction between a putative QTL and marker $s$.

Table 1. Conditional probabilities of a putative QTL given two flanking marker genotypes for a backcross population.

| $M_1$ | $M_2$ | AA | AB |
|---|---|---|---|
| AA | AA | $\frac{(1-r_{1P})(1-r_{2P})}{(1-r_{12})}$ | $\frac{r_{1P}r_{2P}}{(1-r_{12})}$ |
| AA | AB | $\frac{(1-r_{1P})r_{2P}}{r_{12}}$ | $\frac{r_{1P}(1-r_{2P})}{r_{12}}$ |
| AB | AA | $\frac{r_{1P}(1-r_{2P})}{r_{12}}$ | $\frac{(1-r_{1P})r_{2P}}{r_{12}}$ |
| AB | AB | $\frac{r_{1P}r_{2P}}{(1-r_{12})}$ | $\frac{(1-r_{1P})(1-r_{2P})}{(1-r_{12})}$ |

Since the above conditional probabilities depend on unknown recombinaton fractions, below we describe the derivation of the joint maximum likelihood estimates (MLE's) of the recombination fractions.

Let $r = (r_1, \cdots, r_{N-1})$ denote the set of recombination fractions where $N$ is the number of ordered loci for an individual. From now on, we use $r$ as a recombination fraction between two markers or marker and a QTL. Let $M_i$ denote the observed marker data for individual $i$, for $i = 1, \cdots, n$. To avoid complication, we first descibe the procedure when $N$=2.

Consider $AaBb \times aabb$, where possible genotypes are $AaBb$, $aabb$, $aaBb$ and $Aabb$. For $N$=2, the likelihood function can be written as $L(r) = \sum_{l=1}^{2}\sum_{q=1}^{2} g_{lq}\log(\text{ef})_{lq}$, where observed genotype counts $g_{lq}$ and expected genotype frequencies (ef) are shown in Table 2. From this case, the maximum likelihoood estimate of $r$ is $\widehat{r} = \frac{g_{12}+g_{21}}{n}$. In general, the likelihood can be written as $L(r) = \prod_{i=1}^{n} P(M_i|r)$ where $M_i$ is the observed marker data for individual $i$, $i = 1, \cdots, n$.

Table 2. Expected genotype frequencies for a backcross experiment: $g_{lq}$ is the observed genotype count for the $l$-th genotype of locus $A$ and $q$-th genotype of locus $B$. $r$ is the recombination fraction between locus $A$ and locus $B$.

| Genotypes | Observed Counts($g_{lq}$) | Expected Frequency ef$_{lq}$ |
|---|---|---|
| AaBb | $g_{11}$ | $0.5(1-r)$ |
| aabb | $g_{12}$ | $0.5(1-r)$ |
| aaBb | $g_{21}$ | $0.5r$ |
| Aabb | $g_{22}$ | 0.5r |

Now, to obtain maximum likelihood estimates of $r$ we use EM algorithm. Generally speaking, the EM algorithm consists of two steps: E-step and M-step. In the E-step we calculate the expected number of recombinants for every interval. In the M-step, we obtain the MLE's by replacing the unobserved quantities with their expected values. For more details see Broman (2009).

### 3.2. Estimation for parametric AFT cure model

Assume, the relationship between covariates and the log of failure time is

$$W_i^* = \beta' z_i + \beta_0 + \beta_g G_i + \sigma W. \tag{4}$$

where $\sigma$ is a scale parameter, $\beta_0$ is the intercept and $W$ is the error which has the unknown distribution $S_0$. Note that combining (3) and (4) gives the parametric AFT cure model.

Assuming that models (3) and (4) are correct, we use the likelihood procedure in order to estimate these parameters based on the observed data. The likelihood under the observed data $(W_i{}^*,\delta_i,z_i,M_i)$, is

$$\begin{aligned} L &= \prod_{i=1}^{n} P(W_i{}^*,\delta_i,z_i,M_i) \\ &= \prod_{i=1}^{n} \sum_{G_i,\eta_i} P(W_i{}^*,G_i,\delta_i,z_i,M_i,\eta_i) \\ &= \prod_{i=1}^{n} [P(W_i{}^*,1,\delta_i,z_i,M_i,1)+P(W_i{}^*,1,\delta_i,z_i,M_i,0)+P(W_i{}^*,0,\delta_i,z_i,M_i,1)+P(W_i{}^*,0,\delta_i,z_i,M_i,0)]. \end{aligned}$$

$$\begin{aligned} = & \prod_{i=1}^{n}\{[(f(W_i{}^*)^{\delta_i}(S(W_i{}^*))^{1-\delta_i})P(\eta_i=1|G_i=1,z_i)P(G_i=1|z_i,M_i)]+P(\eta_i=0|G_i=1,z_i)P(G_i=1|z_i,M_i)(1-\delta_i) \\ + & \ [(f(W_i{}^*)^{\delta_i}(S(W_i{}^*))^{1-\delta_i})P(\eta_i=1|G_i=0,z_i)P(G_i=0|z_i,M_i)]+[P(\eta_i=0|G_i=0,z_i)P(G_i=0|z_i,M_i)(1-\delta_i)]\}.(5) \end{aligned}$$

It should be noted here the parameters of interest are $(\beta_0,\beta_g,\gamma_0,\gamma_g,\beta',\gamma',\sigma)$. We refer to (5) as the incomplete (observed) likelihood function. Since it is difficult to get the maximum likelihood estimate of parameters by maximizing (5), one option is to use the EM algorithm.

For $\delta_i$=1 and given the fact that $\delta_i$=1 implies $\eta_i$=1,

$$P(W_i{}^*,\delta_i,\eta_i,G_i|M_i,z_i) \quad = \quad P(W_i{}^*|G_i,M_i,\eta_i=1,z_i)P(\eta_i=1|G_i,M_i,z_i)P(G_i|M_i,z_i).$$

For $\delta_i$=0, $\eta_i$ is unobservable and therefore,

$$\begin{aligned} P(W_i{}^*,\delta_i,\eta_i,G_i|M_i,z_i) &= [P(W_i{}^*|G_i,M_i,\eta_i=1,z_i)P(\eta_i=1|G_i,M_i,z_i)P(G_i|M_i,z_i)P(G_i|M_i,z_i)] \\ &+ [P(W_i{}^*|G_i,M_i,\eta_i=0,z_i)P(\eta_i=0|G_i,M_i,z_i)P(G_i|M_i,z_i)]. \end{aligned}$$

Thus, the likelihood function for the complete data $(W_i{}^*,\delta_i,\eta_i,G_i,M_i,z_i)$ is

$$\begin{aligned} L_{\text{Complete}} &= \prod_{i=1}^{n}\left(\frac{e^{\gamma' z_i+\gamma_0+\gamma_g G_i}}{1+e^{\gamma' z_i+\gamma_0+\gamma_g G_i}}\right)^{\eta_i}\left(\frac{1}{1+e^{\gamma' z_i+\gamma_0+\gamma_g G_i}}\right)^{(1-\eta_i)} \\ &\times \prod_{i=1}^{n}\left[(1/\sigma)f_0\left(\frac{W_i{}^*-\beta' z_i-\beta_g G_i-\beta_0}{\sigma}\right)\right]^{\delta_i\eta_i} \\ &\times S_0\left(\frac{W_i^*-\beta' z_i-\beta_g G_i-\beta_0}{\sigma}\right)^{(1-\delta_i)\eta_i} \\ &\times \prod_{i=1}^{n} P(G_i|M_i). \end{aligned} \qquad (6)$$

This complete data likelihood is composed of three parts: logistic model, the AFT cure model, and the conditional probability of $G_i$ given the flanking markers genotypes $M_i$. It should be noted here the parameters of interest are $(\beta_0,\beta_g,\gamma_0,\gamma_g,\beta',\gamma',\sigma)$.

This likelihood can also be expressed in terms of hazard function $h(W_i{}^*)$. That is,

$$\begin{aligned} L_{\text{Complete}} &= \prod_{i=1}^{n}\left(\frac{e^{\gamma' z_i+\gamma_0+\gamma_g G_i}}{1+e^{\gamma' z_i+\gamma_0+\gamma_g G_i}}\right)^{\eta_i}\left(\frac{1}{1+e^{\gamma' z_i+\gamma_0+\gamma_g G_i}}\right)^{(1-\eta_i)} \\ &\times \prod_{i=1}^{n}\left[(1/\sigma)h_0\left(\frac{W_i{}^*-\beta' z_i-\beta_g G_i-\beta_0}{\sigma}\right)\right]^{\delta_i\eta_i} \\ &\times S_0\left(\frac{W_i{}^*-\beta' z_i-\beta_g G_i-\beta_0}{\sigma}\right)^{\eta_i} \\ &\times \prod_{i=1}^{n}P(G_i|M_i). \end{aligned} \tag{7}$$

where $h_0$ is the hazard for $S_0$ and $P(G_i|M_i)$ can be obtained from Table 1. It is clear that $L_{Complete}$ is a function of unobserved quantities $\{\eta_i, G_i, \eta_i G_i\}$.

The EM algorithm consists of two steps: E-step and M-step. In the E-step we will calculate the conditional expectation of $l_{\text{complete}}$, with respect to unobserved quantities $\{\eta_i, G_i, \eta_i G_i\}$ given the current estimated parameter values and the observed data $O_i = \{W_i{}^*, \delta_i, M_i, z_i\}$. In the M-step we make our initial guesses for the parameters of interest and maximize and update them until we reach the convergence. Note that $l_{\text{complete}} = \log(L_{Complete})$. In order to calculate $E(l_{complete})$ we need to find conditional expectations of unobserved quantities namely $E(G_i)$, $E(\eta_i)$ and $E(\eta_i G_i)$. In this paper these conditional expectations are derived only for backcross design and are:

If $\delta_i = 1$ then $\eta_i = 1$, and $P_{\eta_i} = E(\eta_i|M_i, z_i, W_i{}^*)$ is 1.

If $\delta_i = 0$, then

$$E(\eta_i|M_i,z_i,W_i{}^*) = \frac{\pi_i(1)p_iS_0(\frac{W_i{}^*-\beta' z_i-\beta_g-\beta_0}{\sigma})}{D_{i0}} + \frac{\pi_i(0)(1-p_i)S_0(\frac{W_i{}^*-\beta' z_i-\beta_0}{\sigma})}{D_{i0}}$$

where,

$$\begin{aligned} D_{i0} &= \pi_i(1)p_iS_0(\frac{W_i{}^*-\beta' z_i-\beta_g-\beta_0}{\sigma})+(1-\pi_i(1))(p_i) \\ &+ \pi_i(0)(1-p_i)S_0(\frac{W_i{}^*-\beta' z_i-\beta_g-\beta_0}{\sigma})+(1-\pi_i(0))(1-p_i). \end{aligned}$$

If $\delta_i = 1$,

$$E(G_i|M_i,z_i,W_i{}^*) = \frac{f_0(\frac{W_i{}^*-\beta' z_i-\beta_g-\beta_0}{\sigma})\pi_i(1)p_i}{D_{i1}},$$

where,

$$D_{i1} = f_0(\frac{W_i{}^*-\beta' z_i-\beta_g-\beta_0}{\sigma})\pi_i(1)p_i + f_0(\frac{W_i{}^*-\beta' z_i-\beta_0}{\sigma})\pi_i(0)(1-p_i).$$

If $\delta_i = 0$,

$$E(G_i|M_i,z_i,W_i{}^*) = \frac{S_0(\frac{W_i{}^*-\beta' z_i-\beta_g-\beta_0}{\sigma})\pi_i(1)p_i}{D_{i0}} + \frac{(1-\pi_i(1))p_i}{D_{i0}}.$$

If $\delta_i = 1$, then

$$E(\eta_i G_i|M_i,z_i,W_i{}^*) = \frac{f_0(\frac{W_i{}^*-\beta' z_i-\beta_g-\beta_0}{\sigma})\pi_i(1)p_i}{D_{i1}}.$$

And if $\delta_i = 0$, then

$$E(\eta_i G_i|M_i, z_i, W_i{}^*) = \frac{S_0(\frac{W_i{}^* - \beta' z_i - \beta_g - \beta_0}{\sigma})\pi_i(1)p_i}{D_{i0}}.$$

It should be noted that the same procedure can be used for the other multiple crosses including intercross.

### *3.3. Nonparametric estimation*

Estimation method for the unknown parameters is based on EM algorithm and rank estimator. Similar methods used by Wei et.al (1990,1992) and Jin et.al (2003) in the AFT model. Zhang and Peng (2007) considered semiparametric AFT model with the cure part. Using Gehan type weight in the estimating equation, they obtained updated estimates in the M step through minimizing a convex function using a linear programming method. If there exists a solution for the estimating function, it will be unique and consistent. For details see Zhang and Peng (2007). Let $f_0$ be the density probability function for $\varepsilon$ and $S_0$ be the corresponding survival function. The conditional survival function of $W_i{}^*$, given the patient is not cured, is $S_0(W_i{}^* - \beta' z_i - \beta_0 - \beta_g G_i)$. Let $O_i = (W_i{}^*, \delta_i, M_i, z_i)$ denote the observed data for the $i$-th individual, $i = 1, \cdots, n$. It should be noted that $G_i$'s are not provided in the original data set. However, they are obtained through imputation method. The contribution to the likelihood from the $i$-th individual is given in section 3.2. The likelihood is the same as the parametric likelihood which coincides with Zhang and Peng (2007).

We can obtain the estimates through maximizing observed likelihood directly if the distribution of the error term is known to us. However, due to lack of information about $S_0$, this is difficult to do. We then follow Zhang and Peng(2007).

As mentioned earlier, the latent variable $\eta_i$ defines whether the subject is susceptible ($\eta_i = 1$) to a disease or not ($\eta_i = 0$). It can be easily seen that if $\delta = 1$ then $\eta_i = 1$. However, if $\delta = 0$, we have no information on $\eta_i$.

To get the unknown parameter estimates, we will make use of EM algorithm which requires $\eta_k$. Let $\eta_i{}^{(k)} = E(\eta_i|\alpha^{(k)}, O_i)$ where $\alpha = (\beta_0, \beta_g, \gamma_0, \gamma_g, \beta', \gamma', S_0)$. Note that $\eta_i{}^{(k)}$ can be interpreted as the conditional probability that the $i$-th individual is uncured at the $k$-th iteration of the algorithm. If $\delta_i$=0, then

$$E(\eta_i|O_i, \alpha^{(k)}) \quad = \quad \frac{P(W_i{}^*|\eta_i = 1, Q_i)P(\eta_i = 1|Q_i)}{D_{i2},} \tag{8}$$

where,

$$D_{i2} \quad = \quad P(W_i{}^*|\eta_i = 1, Q_i)P(\eta_i = 1|Q_i) + P(W_i{}^*|\eta_i = 0, Q_i)P(\eta_i = 0|Q_i)$$

and $Q_i = (O_i, \alpha^{(k)})$. Combining (8) and using the fact $\delta_i = 1$ implies $\eta_i = 1$, we get the following.

$$\eta_i^{(k)} = \delta_i + \frac{(1 - \delta_i)\pi(G_i, z_i)S_0(W_i^* - \beta' z_i - \beta_0 - \beta_g G_i)}{D_{i3}}, \tag{9}$$

where,

$$D_{i3} = [1 - \pi(G_i, z_i) + \pi(G_i, z_i)S_0(W_i{}^* - \beta' z_i - \beta_0 - \beta_g G_i)].$$

It is worth note that the difference between (9) and $E(\eta_i|M_i, z_i, W_i{}^*)$ for the parametric case. Once the genotypes are imputed, we make use of equation (9), otherwise we use $E(\eta_i|M_i, z_i, W_i{}^*)$ where $G_i$'s are not known.

Our interest is to estimate the unknown parameters $(\beta_0, \beta_g, \gamma_0, \gamma_g, \beta', \gamma', S_0)$. Following Zhang and Peng (2007) we can obtain the estimate of $S_0$ through the residuals of the model. Let $r_1 < r_2 < \cdots < r_m$ be the the failure residuals and $d_{r_j}$ denote the number of failures at time corresponding to $r_j$ and $R(r_j)$ denote the risk set at $r_j$, then the estimate of $S_0$ in the current maximization step is

$$\widehat{S_0}^{(k+1)}(\varepsilon) = \exp\left(-\sum_{j:r_j<\varepsilon} \frac{d_{r_j}}{\sum_{i\in R(r)} \eta_j{}^{(k)}}\right) \tag{10}$$

The estimate of survival function is stable. To see that, since stability of estimation of unknown parameters, guarantees the stability of estimate of survival function we showed the stability of parameters.

Table 3 shows the estimated parameters with different initial values of unknown parameters.

Table 3. Parameter estimates for listeria data for different inital values.

| Initial values for $\beta_g$,$\gamma_0$,$\gamma_g$ | Estimate |
|---|---|
| (-100,0,0) | (-0.436,0.265,1.181) |
| (-10,0,0) | (-0.437,0.223,1.130) |
| (-1,0,0) | (-0.433,0.223,1.310) |
| (0,0,0) | (-0.437,0.265,1.181) |
| (1,0,0) | (-0.437,0.265,1.181) |
| (10,0,0) | (-0.431,0.265,1.181) |
| (100,0,0) | (-0.434,0.265,1.181) |

Table 4. Parameter estimates and their standard errors.

| Parameters | Estimate | Std Err |
|---|---|---|
| $\beta_g$ | -0.484 | 0.005 |
| $\beta_0$ | 4.934 | 0.004 |
| $\gamma_g$ | 1.330 | 0.029 |
| $\gamma_0$ | 0.218 | 0.016 |

Because of the lack of having a complete log-likelihood function, it is not easy to obtain the variances of the estimated parameters in the proposed semiparametric AFT cure model. In order to handle these challenges in estimation of variances for the semiparametric AFT cure model, we used bootstrap method to estimate the standard error of the parameters. We generated 1000 data sets from the listeria data. Table 4 shows the estimates and their standard errors.

For convergence criterion, we looked at the absolute differences between the estimated values of parameters at the $k$-th iteration and the estimated values of parameters at $(k+1)$-th iteration until the difference between $k$-th and $(k+1)$-th iteration values is less than 0.05. We usually achieved convergence in three updates. During simulations the convergence is guaranteed and resulted in consistent estimates.

### 3.4. LOD score

Evidence in identifying QTL is summarized by a LOD score, defined as the log10-likelihood ratio comparing the alternative hypothesis of a QTL at the position of interest with the null hypothesis of no QTL. More specifically, for a given location $\xi$, suppose we want to test,

$$H_0 : \beta_g(\xi) = 0, \gamma_g(\xi) = 0. \tag{11}$$

Then it is clear that under $H_0$, the LOD score at a given putative location $\xi$ can be defined as

$$LOD(\xi) \quad = \quad \log_{10}\frac{L_1}{L_2} \tag{12}$$

where

$$L_1 = L_{\widehat{\beta_g}(\xi),\widehat{\gamma}_g(\xi),\widehat{\beta_0}(\xi),\widehat{\sigma}(\xi),\widehat{\boldsymbol{\beta}}(\xi),\widehat{\boldsymbol{\gamma}}(\xi)}$$

$$L_2 = L_{\widehat{\boldsymbol{\beta}}(\xi),\widehat{\boldsymbol{\gamma}}(\xi),\widehat{\beta_{\mathbf{0}}}(\xi),\widehat{\sigma}(\xi)|\beta_g(\xi)=0,\gamma_g(\xi)=0)}.$$

We maximize numerator under alternative hypothesis, $H_a$, while keeping all the parameters in the general model. Also, we maximize the denominator in the absence of $\widehat{\beta}_g(\xi)$ and $\widehat{\gamma}_g(\xi)$ corresponding to $H_0$ for every putative and marker locations for the whole genome. In other words, we scan the whole genome by calculating the LOD scores. We choose the location which gives the highest LOD score.

In parametric case, $W$ has an extreme value distribution then $T_i^*$ follows a Weibull distribution. We used this distribution because it can be used to model a variety of life behaviors and has wide applications in biomedical research. There are theoretical motivations such as time to appearance of tumor or until death in animals. See Pike (1966), Peto and Lee (1973). It is shown that Weibull model fits data well dealing with the time to appearance of tumors in animals. See Lee and O'Neill (1971) and Doll (1971).

## 4. Applying Methodology to the Listeria Data Set

In this section we illustrate the methods with an application to Listeria data and conduct simulations to evaluate the performance of our methods.

### *4.1. Illustration of our method using the listeria data*

We apply our methods described in Section 3 to the Listeria monocytogenes data set by Boyartchuk, Broman, Mosher, D'orazio, Starnbach (2001) and Broman (2003).

The data set consists of 120 age-matched (9 weeks of age) female $BALB/cByJ \times C57BL/6ByJ$ intercross (CB6F2/ByJ) mice that were infected by intravenous injection and were monitored to find their time of death within eight hours. For more details see Broman (2000). All animals at the point of death were recorded as dead. Animals surviving past 264 hours ($\delta_i = 0$) were considered recovered from the disease. Traditional interval mapping, which relies on normally distributed traits, is thus not appropriate. We checked all chromosomes and chromosome 13 showed promising results from other investigations as well as our investigation, we decided to analyze only chromosome 13 where $m = 29$, $n = 116$ and $k = 12$. Here $m$ is the number of putative QTL including observed markers, $n$ is the number of mice and $k$ is the number of observed markers. We used $R$ and *optim* function for the maximization. Now, we are ready to apply our methodology to the data set we described in Section 4.1. We provide the LOD scores of 29 locations on Chromosome 13 for the nonparametric method. They are: 0.19, 0.89, 0.83, 0.36, 1.01, 0.23, 0.72, 0.54, 0.88, 1.09, 1.12 , 0.73, 0.42, 0.82, 1.32, 1.58, 1.63, 2.02, 2.85, 4.97, 5.66, 6.18, 4.86, 4.69 ,4.46, 3.43, 3.26, 2.85, 2.56 from location 1 through 29 respectively. It is clear that location 22 has the highest LOD score. Large LOD score indicates evidence for the presence of a QTL. In other words, the larger the LOD score the greater the evidence.

## 5. Simulation Study

From Table 5, it can be seen that the results using parametric and and nonparametric methods are similar. When $n = 115$ it took approximately 9 hours to complete 1000 simulations. When $n = 50$, it took about 5.5 hours and it was around 3 hours for $n = 25$ with Intel Core 2 Duo CPU at 2.10GHz. We also counted how many times location 22 was picked (in terms of percentage) for the parameter values specified above and for different sample sizes. We obtained the bias and standard errors of the estimates for parameters. Here we assumed that there are no environmental covariates for AFT as well as logistic part.

Table 5 contains bias, MSE and proportion of times that correct location which is location 22 is picked. From Table 5 it is clear that both methods performs well in terms of bias, MSE and proportion of time that location 22 was picked. Parametric approach does slightly better. However, the difference is small. Based on Table 5 it can be seen that parametric and nonparametric approaches are comparable. Also, as *n* increases our results improve as one expected.

Table 5. Coverage, Biases and MSE of simulated data sets.

| Method | QTL coverage (%) | Size | Par | Bias | MSE |
|---|---|---|---|---|---|
| Parametric | 68 | 25 | $\widehat{\beta}_g$ | 0.524 | 0.000291 |
| | | | $\widehat{\gamma}_g$ | 0.228 | 0.010414 |
| | | | $\widehat{\beta}_0$ | 0.304 | 0.000138 |
| | | | $\widehat{\gamma}_0$ | 0.144 | 0.046140 |
| | 79 | 50 | $\widehat{\beta}_g$ | 0.236 | 0.000026 |
| | | | $\widehat{\gamma}_g$ | 0.063 | 0.000808 |
| | | | $\widehat{\beta}_0$ | 0.198 | 0.000022 |
| | | | $\widehat{\gamma}_0$ | 0.022 | 0.012139 |
| | 83 | 115 | $\widehat{\beta}_g$ | 0.178 | 0.000018 |
| | | | $\widehat{\gamma}_g$ | 0.038 | 0.000297 |
| | | | $\widehat{\beta}_0$ | 0.140 | 0.000014 |
| | | | $\widehat{\gamma}_0$ | 0.017 | 0.000126 |
| Nonparametric | 58 | 25 | $\widehat{\beta}_g$ | 0.476 | 0.001772 |
| | | | $\widehat{\gamma}_g$ | 0.277 | 0.018913 |
| | | | $\widehat{\beta}_0$ | 0.369 | 0.001364 |
| | | | $\widehat{\gamma}_0$ | 0.159 | 0.051305 |
| | 75 | 50 | $\widehat{\beta}_g$ | 0.279 | 0.000915 |
| | | | $\widehat{\gamma}_g$ | 0.062 | 0.010715 |
| | | | $\widehat{\beta}_0$ | 0.249 | 0.000868 |
| | | | $\widehat{\gamma}_0$ | 0.024 | 0.012407 |
| | 82 | 115 | $\widehat{\beta}_g$ | 0.224 | 0.000748 |
| | | | $\widehat{\gamma}_g$ | 0.019 | 0.003115 |
| | | | $\widehat{\beta}_0$ | 0.139 | 0.000510 |
| | | | $\widehat{\gamma}_0$ | 0.005 | 0.000160 |

## 6. Concluding Remarks

In this article, we described interval mapping using nonparametric accelerated failure time cure model. We proposed a nonparametric method using imputation for QTL location and effects on both cured and non-cured subjects. We applied our methodologies to a real data set. Both parametric and nonparametric approaches provided similar results of unknown parameters as well as the same QTL location. Imputation seems a good choice without the need of computing all expected values over an EM algorithm. We also conducted simulations to evaluate performance of our methodology. The results show that our methods perform well in terms of bias, MSE and detecting the QTL location.

## 7. Acknowledgments

We would like to thank Professor Karl W Broman for his kind help in answering our questions.

## References

1. E. S. Lander and D. Botstein, "Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps," *Genetics,* **121**, 185–199 (1989).

2. Z. B. Zeng, "Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci," *Proc Natl Acad Sci,* **90**, 10972–10976 (1993).
3. G. Diao, D. Y. Lin and F. Zou, "Mapping Quantitative Trait Loci With Censored Observations," *Genetics,* **168**, 1689–1698 (2004).
4. L. Krugylak and E. S. Lander, "A Nonparametric Approach for Mapping Quantitative Trait Loci," *Genetics,* **139**, 1421–1428 (1995).
5. K. W. Broman, "Mapping quantitative trait loci in the case of a spike in the phenotype distribution," *Genetics,* **163**, 1169–1175 (2003).
6. T. M. Poole and N. R. Drinkwater, "Two genes abrogate the inhibition of murine hepatocarcinogenesis by ovarian hormones," *Proc Natl Acad Sci,* **93 (12)**, 5848–5853 (1996).
7. B. Basrak, C.A.J. Klaassen, M. Beekman, N.G. Martin and D.I. Boomsma, "Copulas in QTL mapping," *Behav Genet*, **34**, 161–171 (2004).
8. J. P. Fine, F. Zou and B.S. Yandel, "Nonparametric estimation of mixture models, with application to quantitative trait loci," *Biostatistics,* **5**, 501–513 (2004).
9. M. Li, , M. Boehnke, G. R. Abecasis and P. X. Song, "Quantitative trait linkage analysis using Gaussian copulas," *Genetics,* **173**, 2317–2327. (2006).
10. M. Zak, A. Baierl, M. Bogdan and A. Futschik, "Locating multiple interacting quantitative trait loci using rank-based model selection," *Genetics,* **176(3)**, 1845–1854 (2007).
11. A. Manichaikul, "Statistical methods for mapping quantitative trait loci in experimental crosses," *Dissertation*, Johns Hopkins University. (2008).
12. V. T. Farewell, "Mixture models in survival analysis: Are they worth the risk?," *Can J Stat,* **14**, 257–262 (1986).
13. Kuk A. Y. C. and C. H. Chen, "A mixture model combining logistic regression with proportional hazards regression," *Biometrika,* **79**, 187–200 (1992).
14. R. A. Maller and S. Zhou, "Estimating the proportion of immunes in a censored sample," *Biometrika,* **79**, 731–739 (1992).
15. R.Sposto, D. L. Preston, Y. Shimizu and K. Mabuchi, "The effect of diagnostic misclassification on noncancer and cancer mortality dose," *Biometrics,* **48(2)**, 605–617 (1992).
16. W. Lu and Z. Ying, "On semiparametric transformation cure models," *Biometrika,* **91**, 331–343 (2004).
17. M. Liu, W. Lu and Y. Shao, "Mixture cure model with an application to interval mapping of quantitative trait loci," *Lifetime Data Anal,* **12**, 421–440 (2006).
18. M. Liu, W. Lu and Y. Shao, "Interval Mapping of Quantitative Trait Loci for Time-to-Event Data with the Proportional Hazards Mixture Cure Model," *Biometrics,* **62**, 1053–1061 (2006).
19. D. Bilgili, "Quantitative Trait Loci (QTL) detection using Accelerated Failure Time Cure Model," *Dissertation*, Northern Illinois University. (2009).
20. Z. Piao, X. Zhou, L. Yan, Y. Guo, R. Yang, , Z.Luo and D. R. Prows, "Statistical optimization of parametric accelerated failure time model for mapping survival trait loci," *Theor Appl Genet,* **122**, 855–863 (2011).
21. L. Xua and J. Zhang, "Multiple imputation method for the semiparametric accelerated failure time mixture cure model," *Comput Stat Data An,* **54**, 1808–1816 (2010).
22. L. J. Wei, "Linear regression analysis of censored survival data on rank tests," *Biometrika,* **77**, 845–851 (1990).
23. L. J. Wei, "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis," *Stat Med,* **11**, 1871–1879 (1992).
24. Z. Jin, D. Y. Lin, L. J. Wei and Z. Ying, "Rank-based inference for the accelerated failure time model," *Biometrika,* **90**, 341–353 (2003).
25. J. Zhang and Y. Peng, " A new estimation method for the semiparametric accelerated failure time mixture cure model," *Stat Med,* **26 (16)**, 3157–3171 (2007).
26. J.. Y. Cheng and S. J. Tzeng, "Parametric and semiparametric methods for mapping quantitative trait loci," *Comput Stat Data An*, **53**, 1843–1849 (2009).
27. L.E. Baum, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Ann Math Statist*, **41**, 164–171 (1970).
28. K. W. Broman and S. Sen, "A Guide to QTL Mapping with R/qtl," Springer, New York (2009).
29. S. Sen and G. A. Churchill, "A Statistical Framework for Quantitative Trait Mapping," *Genetics,* **159**, 371–387 (2001).
30. F. Sugiyama, G. A. Churchill, D. C. Higgins , C. Johns, K. P. Makaritsis, et al., "Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci," *Genomics,* **71**, 70–77 (2001).
31. Y. Li, C. Willer, S. Sanna and G. R. Abecasis,"Genotype imputation," *Annu Rev Genomics Hum Genet,* **10**, 387–406 (2009).
32. M. C. Pike, "A new method of analysis of certain class of experiments in carcinogenesis," *Biometrics,* **22**, 142–161 (1966).
33. R. Peto and P. Lee, "Weibull distributions for continuous carcinogenesis experiments," *Biometrics,* **29**, 457–470 (1973).
34. P. N. Lee and J. A. O'Neill, "The effect both of time and dose applied on tumour incidence rate in benzopyrene skin painting experiments," *Brit.J.Cancer,* **25**, 759–770. (1971).