SMU ECONOMICS & STATISTICS
WORKING PAPER SERIES

A Semi-parametric Two-component "Compound" Mixture Model and Its Application to Estimating Malaria Attributable Fractions

Jing Qin and Denis H. Y. Leung

August 2004

Paper No. 17-2004

A semi-parametric two-component "compound" mixture model and its application to estimating malaria attributable fractions

Jing Qin and Denis H. Y. Leung

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center

1275 York Avenue, New York, New York 10021

e-mail: qinj@biosta.mskcc.org

School of Economics and Social Sciences, Singapore Management University, 469 Bukit Timah Road, Singapore

e-mail: denisleung@smu.edu.sg

Summary

Malaria remains a major epidemiological problem in many developing countries. Malaria is defined as the presence of parasites and symptoms (usually fever) due to the parasites. In endemic areas, an individual may have symptoms attributable either to malaria or to other causes. From a clinical point of view, it is important to correctly diagnose an individual who has developed symptoms so that the appropriate treatments can be given. From an epidemiologic and economic point of view, it is important to determine the proportion of malaria affected cases in individuals who have symptoms so that policies on intervention programmes can be developed. Once symptoms have developed in an individual, the diagnosis of malaria can be based on analysis of the parasite levels in blood samples. However, even a blood test is not conclusive as in endemic areas, many healthy individuals can have parasites in their blood slides. Therefore, data from this type of studies can be viewed as coming from a mixture distribution, with the components corresponding to malaria and non-malaria cases. A unique feature in this type of data, however, is the fact that a proportion of the non-malaria cases have zero parasite levels. Therefore, one of the component distribu-

tions is itself a mixture distribution. In this article, we propose a semi-parametric likelihood approach for estimating the proportion of clinical malaria using parasite level data from a group of individuals with symptoms. Our approach assumes the density ratio for the parasite levels in clinical malaria and non-clinical malaria cases can be modeled using a logistic model. We use empirical likelihood to combine the zero and non-zero data. The maximum semi-parametric likelihood estimate is more efficient than existing non-parametric estimates using only the frequencies of zero and non-zero data. On the other hand, it is more robust than a fully parametric maximum likelihood estimate that assumes a parametric model for the non-zero data. Simulation results show that the performance of the proposed method is satisfactory. The proposed method is used to analyze data from a malaria survey carried out in Tanzania.

KEY WORDS: Attributable fraction; Density ratio model; Empirical likelihood; Malaria; Mixture methods.

Introduction

Recent reviews ("World Health Organization, Practical chemotherapy of malaria: report of a WHO scientific group", WHO Technical Report Series 805, Geneva, 1990) suggest that malaria causes around 110 million sickness episodes and one million deaths each year throughout the world. One of the symptoms of malaria is fever. In an endemicity, a person who has developed fever will be tested for parasite levels in his/her blood. However, the test is often not conclusive as healthy individuals living in endemic areas often tolerate malaria parasites. Furthermore, fever can be due to causes other than malaria. In other words, in individuals who have developed fever, there are some with low parasite levels but are truly malaria cases while there are some with high parasite levels but are non-malaria cases. Therefore, in analyzing parasite level data from individuals who have developed fever, the data can be viewed as coming from a two component mixture distribution, with the

components corresponding to the malaria and non-malaria population. A unique feature of this type of data is that, within the non-malaria population, there are some who have zero-parasite level. Therefore, the distribution of parasite level in the non-malaria population is itself a mixture distribution. More specifically, suppose a sample of parasite levels from n febrile individuals is collected from an endemicity. We let $x_1, x_2, ..., x_n$ be independent and identically distributed (i.i.d.) random variables representing the parasite levels. Then, $x_1, x_2, ..., x_n$ follow a two-component mixture distribution with density

$$f(x) = (1-\lambda)f_1^*(x) + \lambda f_2(x), \tag{1}$$

where f_1^* and f_2 are the densities of parasite levels in the non-malaria and malaria populations, respectively. The mixing parameter λ is the proportion of individuals with clinical malaria in the endemicity. It is also called the malaria attributable fraction in epidemiologic terminology. Furthermore, f_1^* can be decomposed as

$$f_1^*(x) = pI(x=0) + (1-p)f_1(x)I(x>0),$$

where p is the proportion in the non-malaria population with zero parasite level, f_1 is a density on $(0, \infty)$ and I is an indicator function. As a result, f can be written as

$$\begin{aligned} f(x) &= p(1-\lambda)I(x=0) + \{(1-\lambda)(1-p)f_1(x) + \lambda f_2(x)\}I(x>0) \\ &= p(1-\lambda)I(x=0) + [1-p(1-\lambda)]\{(1-\lambda^*)f_1(x) + \lambda^* f_2(x)\}I(x>0), \end{aligned} \tag{2}$$

where $\lambda^* = \lambda/\{1-p(1-\lambda)\}$ can be interpreted as the probability of an individual carrying malaria given he/she has positive parasite level from the endemicity. Based on (2), we can consider the distribution as a kind of "compound" mixture distribution. A partitioning of a typical set of data in an endemicity is given in Table 1.

In general, without specifying the forms of f_1^* and f_2, the mixture model (1) is not identifiable. However, the identifiability problem can be solved if additional information about these distributions can be obtained. Vounatsou, Smith and Smith (1998) described

a cross-sectional survey study of parasitaemia and fever among children up to one year old in a village in the Kilombero district in Tanzania (Kiua et al. 1996) where this is the case. In that study, in addition to data from the mixture distribution that were obtained in the endemicity, a secondary set of data $z_1, ..., z_m$ from f_1^* was obtained from the community. The secondary data can thus be considered as a training sample. Define the parasite prevalence probabilities in the endemicity and the community, respectively, as $p_f = 1 - p(1 - \lambda)$ and $p_a = 1 - p$, then Vounatsou et al (1998) and Smith, Schellenberg and Hayes (1994) showed that

$$\lambda = (p_f - p_a)/(1 - p_a). \tag{3}$$

Based on (3), a natural estimator of λ is to replace p_f and p_a by sample proportions. However, as Vounatsou et al (1998) pointed out, in general, p_a is very high and the proportion of community children without parasitaemia is low. As a result, the estimator of λ using the sample proportions can be either negative or imprecise when p_a is close to one. Also the estimator does not utilize the quantitative nature of the parasite level data. Another method to estimate λ is to use the binomial counts of zero and non-zero parasite level data. Finally, Vounatsou et al. (1998) suggested a multinomial likelihood by grouping the observations from the mixture and the training samples into a number of ordered categories. In this paper, we explore a method that makes use of the quantitative nature of the data and also does not require grouping of the data.

Let m_0 and m_1, respectively, be the numbers of observations with zero and non-zero parasite level in the training sample, $z_1, z_2, ..., z_m$. Similar definitions of n_0 and n_1 are applied to the mixture sample, $x_1, x_2, ..., x_n$. Without loss of generality, we assume the non-zero observations from the training sample and the mixture sample are $z_1, ..., z_{m_1}$ and $x_1, .., x_{n_1}$, respectively. Therefore,

$$z_1, z_2, ..., z_{m_1} \sim f_1(z),$$

$$x_1, ..., x_{n_1} \sim g(x) = (1 - \lambda^*) f_1(x) + \lambda^* f_2(x),$$

so that the density f_1 is the same in the endemicity and the training sample. The log-likelihood is

$$\ell = \ell_1 + \ell_2, \tag{4}$$

where

$$\ell_1 = m_0 \log p + m_1 \log(1-p) + n_0 \log\{(1-\lambda)p\} + n_1 \log\{1-(1-\lambda)p\} \tag{5}$$

and

$$\ell_2 = \sum_{i=1}^{m_1} \log f_1(z_i) + \sum_{j=1}^{n_1} \log\{(1-\lambda^*)f_1(x_j) + \lambda^* f_2(x_j)\}. \tag{6}$$

In (5) and (6), ℓ_1 is the marginal log-likelihood of the number of zeros in the data and ℓ_2 is the conditional likelihood given that the data are greater than zero. Furthermore, the parameter λ appears in both ℓ_1 and ℓ_2.

If inference is based on ℓ_1 alone, the method is that of making use of the binomial data of presence/absence of parasites. On the other hand, if inference is based only on ℓ_2, Lancaster and Imbens (1996) called this problem a case-control problem with contaminated controls. Applications can be found in econometrics literature, where, for example, the training sample is a random sample of female labor force participants and the mixture sample is a random sample of working age women whose labor force participating statuses are unknown. One can expect that the conditional log-likelihood ℓ_2 may contain information on λ^* (or λ). Unfortunately, if the forms of f_1 and f_2 are un-specified and λ^* is unknown, then the mixture model is not identifiable based on ℓ_2 alone. If there is an additional sample from f_2 beside the mixture and the training samples, then it is possible to estimate λ^* non-parametrically (Hall, 1981, Murray and Titterington, 1978, Hall and Titterington, 1984). Alternatively, if parametric models are assumed for f_1 and f_2, then maximum likelihood method for standard mixture model can be employed to find the underlying parameters (Titterington, Smith and Makov, 1985, Lindsay, 1995, McLachlan and Krishnan, 1997, McLachlan and Peel, 2001).

In exploring robust and efficient estimation methods, Smith et al (1994) considered a model-based approach, in which the relationship between parasite level and malaria fever

is modeled as a smooth function using a logistic regression. This model has been used for different applications of mixture models in fisheries, econometrics, clinical and genetics studies (Anderson, 1979, Imben and Lancaster, 1996, Qin, 1999, Nagelkerke, Borgdorff and Kim, 2001, Zou, Fine and Yandell, 2002). The logistic regression method is equivalent to a two sample semi-parametric modeling assumption, where the log density ratio is linearly related to the observed data,

$$\log \frac{f_2(x)}{f_1(x)} = \alpha + x\beta, \quad \text{or} \quad f_2(x) = \exp(\alpha + \beta x) f_1(x), \quad (\alpha, \beta) \neq (0, 0), \tag{7}$$

and the form of $f_1(x)$ is not specified. Using model (7) in the setting described here, we have a two-sample problem:

$$z_1, z_2, ..., z_{m_1} \sim f_1(x),$$

$$x_1, x_2, ..., x_{n_1} \sim g(x) = [(1 - \lambda^*) + \lambda^* \exp(\alpha + \beta x)] f_1(x),$$

This may be considered as a biased sample problem with weights $w_1(x) = 1, w_2(x) = (1 - \lambda^*) + \lambda^* \exp(\alpha + \beta x)$, which depend on parameters (α, β) and λ. In this paper, we propose using (7) to model the density ratio of the malaria and non-malaria populations.

A nice property of model (7) is that it is semi-invariant in the sense that if the data is transformed using a monotone increasing function $h(\cdot)$, the density ratio of the transformed data becomes $\exp[\alpha + \beta^{\mathrm{T}}\{h^{-1}(x)\}]$ with $h^{-1}(\cdot)$ being the inverse function of $h(\cdot)$. In other words, the new density ratio has the same form as the original one except for $h^{-1}(x)$ in place of x. Kay and Little (1987) discussed various choices of density ratio model for some well known distributions. For example, if f_1 and f_2 are normal densities with different means and variances, then the model (7) includes a quadratic term. White and Thompson (2003) have used model (7) to compare treatment effects in clinical trials. In a number of medical studies, it was found that many well known distributions did not fit the observed data well (Qin and Zhang, 1997, Qin et al., 2002, Zhang, 2001), whereas model (7) provided good fits. Therefore, we expect that a semi-parametric approach based on (7) to be more robust than

a parametric approach. On the other hand, a semi-parametric approach should be more efficient than a nonparametric approach.

The rest of this paper is organized as follows. In section 2, we consider estimation in a mixture model. Based on the assumption that the component densities are related by (7), we propose a semi-parametric method using empirical likelihood (Owen, 1988) and biased sampling estimating technique (Vardi, 1982, 1985). In section 3, we apply the proposed method to the malaria survey data. Simulation results are given in section 4. Concluding remarks are given in section 5. Proofs are relegated to the Appendix.

2. Main results

In this section we consider estimating the parameters $(p, \lambda^*, \alpha, \beta)$ in the mixture model when the component densities are related by model (7). Note that we have suppressed the parameter λ because it is a function of p, λ^*.

As defined in Section 1, $m_0 = \sum_{i=1}^{m} I(z_i = 0), m_1 = m - m_0$ and $n_0 = \sum_{i=1}^{n} I(x_i = 0), n_1 = n - n_0$. Under (7), the log-likelihood (4) becomes

$$\ell = \ell_1 + \ell_2, \tag{8}$$

where

$$\ell_1 = m_0 \log p + m_1 \log(1-p) + n_0 \log\{(1-\lambda)p\} + n_1 \log\{1-(1-\lambda)p\}$$

and

$$\ell_2 = \sum_{i=1}^{m_1} \log f_1(z_i) + \sum_{j=1}^{n_1} [\log\{(1-\lambda^*) + \lambda^* \exp(\alpha + \beta x_j)\} + \log f_1(x_j)].$$

As suggested in Section 1, a naive method is to estimate λ and p by using ℓ_1 alone. This is essentially using the binomial counts of the zero and non-zero data. We can easily

derive the maximum binomial likelihood estimators:

$$\hat{p}_B = \frac{m_0}{m}, \qquad \hat{\lambda}_B = 1 - \frac{n_0}{n\hat{p}_B}. \tag{9}$$

Clearly these estimators do not use the information provided by the quantitative part of the non-zero data. Next we will develop a method that utilizes the non-zero data in the conditional log-likelihood, ℓ_2.

In order to maximize ℓ_2, we only need to concentrate on those distribution functions with jumps at the observed data values. Let $(t_1, ..., t_{N_1}) = (z_1, .., z_{m_1}, x_1, ..., x_{n_1})$, $N_1 = m_1 + n_1$ and $q_i = dF_1(t_i)$, $i = 1, 2, ..., N_1$, be the non-negative jump sizes at the N_1 data values so that the total of all the jump sizes is unity. The semi-parametric likelihood of the data can be written as

$$\begin{aligned} & \prod_{i=1}^{m_1} dF_1(z_i) \prod_{k=1}^{n} [(1-\lambda^*) + \lambda^* \exp\{\alpha + \beta x_k\}] dF_1(x_k) \qquad (10) \\ = \; & \{\prod_{i=1}^{N_1} q_i\} \{\prod_{k=1}^{n_1} [(1-\lambda^*) + \lambda^* \exp\{\alpha + \beta x_k\}]\}. \end{aligned}$$

We will maximize the likelihood in two steps, as follows:

Step 1. For fixed $(\lambda^*, \alpha, \beta)$, maximize

$$\prod_{i=1}^{N_1} q_i$$

subject to the constraints

$$\sum_{i=1}^{N_1} q_i = 1, \qquad \sum_{k=1}^{n_1} q_k \{\exp(\alpha + \beta t_k) - 1\} = 0, \qquad q_i \geq 0, i = 1, ..., N_1.$$

Note that the second constraint comes from the fact that $F_2(t) = \int_{-\infty}^{t} \exp(\alpha + \beta x) dF_1(x)$ is a cumulative distribution function. Therefore $E_{F_1}\{\exp(\alpha + \beta x)\} = 1$. After maximizing over the q_i's, we have (Qin and Lawless, 1994)

$$\hat{q}_i = \frac{1}{N_1} \frac{1}{1 + \nu[\exp(\alpha + \beta t_i) - 1]}, \qquad i = 1, 2, ..., N_1,$$

where ν is a Lagrange multiplier determined by

$$\sum_{i=1}^{N_1} \frac{1}{N_1} \frac{\exp(\alpha + \beta t_i) - 1}{1 + \nu[\exp(\alpha + \beta t_i) - 1]} = 0. \tag{11}$$

It can be proved that in an $O(N_1^{-1/3})$ neighborhood of (α, β), $\nu = \nu(\alpha, \beta)$ is an implicit function of (α, β). Therefore the conditional log-likelihood is

$$\ell_2(\lambda^*, \alpha, \beta, \nu) \;=\; -\sum_{i=1}^{N_1} \log\{1 + \nu[\exp(\alpha + \beta t_i) - 1]\} + \sum_{k=1}^{n_1} \log\{(1 - \lambda^*) + \lambda^* \exp(\alpha + \beta x_k)\}.$$

Since, $\lambda = (\lambda^* - \lambda^* p)/(1 - \lambda^* p)$, we can change the variable from λ to λ^*, and the full semi-parametric log-likelihood becomes

$$\ell(p, \lambda^*, \alpha, \beta, \nu) = \ell_1(p, \lambda^*) + \ell_2(\lambda^*, \alpha, \beta, \nu), \tag{12}$$

where

$$\begin{aligned}
&\ell_1(p, \lambda^*) \\
=\;& m_0 \log p + m_1 \log(1 - p) + n_0 \log\{p(1 - \lambda^*)/(1 - \lambda^* p)\} + n_1 \log\{1 - p(1 - \lambda^*)/(1 - \lambda^* p)\} \\
=\;& (m_0 + n_0) \log p + (m_1 + n_1) \log(1 - p) + n_0 \log(1 - \lambda^*) - n \log(1 - \lambda^* p).
\end{aligned}$$

Step 2. Maximize the semi-parametric log-likelihood $\ell(p, \lambda^*, \alpha, \beta, \nu)$ with respect to $(p, \lambda^*, \alpha, \beta, \nu)$.

Differentiating ℓ with respect to $(p, \lambda^*, \alpha, \beta, \nu)$, we have

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n_1} \frac{\lambda^* \exp(\alpha + \beta x_i)}{(1 - \lambda^*) + \lambda^* \exp(\alpha + \beta x_i)} - \sum_{i=1}^{N_1} \frac{\nu \exp(\alpha + \beta t_i)}{1 + \nu\{\exp(\alpha + \beta t_i) - 1\}} = 0,$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{k=1}^{n_1} \frac{\lambda^* x_k \exp(\alpha + \beta x_k)}{(1 - \lambda^*) + \lambda^* \exp(\alpha + \beta x_k)} - \sum_{i=1}^{N_1} \frac{\nu t_i \exp(\alpha + \beta t_i)}{1 + \nu[\exp(\alpha + \beta t_i) - 1]} = 0,$$

$$\frac{\partial \ell}{\partial \lambda^*} = -\frac{n_0}{1 - \lambda^*} + \frac{np}{1 - \lambda^* p} + \sum_{k=1}^{n_1} \frac{-1 + \exp(\alpha + \beta x_k)}{(1 - \lambda^*) + \lambda^* \exp(\alpha + \beta x_k)} = 0,$$

$$\frac{\partial \ell}{\partial p} = \frac{m_0 + n_0}{p} - \frac{m_1 + n_1}{1 - p} + \frac{n\lambda^*}{1 - \lambda^* p} = 0.$$

Also, applying (11) to $\partial \ell / \partial \alpha = 0$, we have:

$$\nu = \lambda^* \frac{1}{N_1} \sum_{i=1}^{n_1} \frac{\exp(\alpha + \beta x_i)}{(1 - \lambda^*) + \lambda^* \exp(\alpha + \beta x_i)}. \tag{13}$$

Denote $\eta = (\alpha, \beta, \lambda^*, p, \nu)$, $N = m + n$, the true value of η as $\eta_0 = (\alpha_0, \beta_0, \lambda_0^*, p_0, \nu_0)$, the maximum semi-parametric likelihood estimate of η as $\hat{\eta} = (\hat{\alpha}, \hat{\beta}, \hat{\lambda}^*, \hat{p}, \hat{\nu})$ and assuming $m/N \to \rho$, $0 < \rho < 1$.

Theorem 1 *Suppose that:*

(1). The distribution function F_1 is non-degenerate, and $|\partial \ell_2 / \partial \eta_i \partial \eta_j \partial \eta_k|$, $i, j, k = 1, 2.., 5$ are bounded by some integrable functions in the neighbour of η_0.

(2). $E_{F_1}\{exp(3\beta x)\} < \infty$ in a neighbourhood of the true value of β_0.

(3). $0 < \lambda_0 < 1$ and $0 < \lambda_0^* < 1$.

(4). $(\alpha, \beta) \neq (0, 0)$.

Under regularity conditions (1)-(4), with probability 1, $\ell(\eta)$ has a local maximum in an $O(N^{-1/3})$ neighborhood of η_0. Moreover, the maximizer $\hat{\eta}$ satisfies the score equations $\partial \ell(\hat{\eta})/\partial \eta = 0$, and

$$\sqrt{N}(\hat{\eta} - \eta_0) \to N(0, \Sigma), \quad \Sigma = V^{-1} U V^{-1}, \tag{14}$$

where U and V are defined in (A.3) and (A.2) in the Appendix. As a result, by delta method, we can easily prove that

$$\sqrt{N}(\hat{\lambda} - \lambda_0) \to N(0, \sigma^2), \qquad \sigma^2 = \left(\frac{\partial \lambda(\eta_0)}{\partial \eta} \right) \Sigma \left(\frac{\partial \lambda(\eta_0)}{\partial \eta} \right)^T.$$

Next we consider the semi-parametric generalized likelihood ratio test statistic. As pointed out by Hall and La Scala (1990), the empirical likelihood method has many advantages over normal approximation methods and the usual bootstrap approximation approaches for constructing confidence intervals. For example, empirical likelihood confidence intervals do not have pre-defined shapes and are range and transformation respecting.

We now give a large sample likelihood ratio test for the parameter λ.

Theorem 2 *Denote* $\lambda^*(\lambda, p) = \lambda/\{1 - p(1 - \lambda)\}$ *and let*

$$R(\lambda) = 2\{\sup_{\alpha,\beta,\lambda,p} \ell(\alpha, \beta, \lambda^*(\lambda, p), p) - \sup_{\alpha,\beta,p} \ell(\alpha, \beta, \lambda^*(\lambda, p), p)\}. \tag{15}$$

Under the regularity conditions specified in Theorem 1, if $H_0 : \lambda = \lambda_0 \neq 0$ *is true, then*

$$R(\lambda_0) \to \chi^2_{(1)}.$$

If parametric models for $f_1(x, \theta_1)$ and $f_2(x, \theta_2)$ are postulated, we can consider a parametric approach. Let

$$\ell_P(\theta_1, \theta_2, \lambda^*, p) = \ell_1(p, \lambda^*) + \ell_{2P}(\lambda^*, \theta_1, \theta_2)$$

be the parametric log-likelihood, where

$$\ell_{2P}(\theta_1, \theta_2, \lambda^*) = \sum_{i=1}^{m_1} \log f_1(z_i, \theta_1) + \sum_{j=1}^{n_1} \log\{(1 - \lambda^*) f_1(x_j, \theta_1) + \lambda^* f_2(x_j, \theta_2)\}.$$

Denote the maximum parametric likelihood estimate as $(\hat{\theta}_{1P}, \hat{\theta}_{2P}, \hat{\lambda}^*_P, \hat{p}_P)$. For comparison, it can be shown that

Theorem 3 *Under some regularity conditions, the parametric likelihood ratio statistic:*

$$R_P(\lambda) = 2\{\max_{(\theta_1,\theta_2,\lambda,p)} \ell_P(\theta_1, \theta_2, \lambda^*(\lambda, p), p) - \max_{(\theta_1,\theta_2,p)} \ell_P(\theta_1, \theta_2, \lambda^*(\lambda, p), p)\} \tag{16}$$

converges to a $\chi^2_{(1)}$ *distribution if* $\lambda = \lambda_0$*, the true value of* λ*. Also the naive likelihood ratio based on binomial counts of zeros and non-zero observations is:*

$$R_B(\lambda) = 2\{\max_{(p,\lambda)} \ell_1(p, \lambda) - \max_{p} \ell_1(p, \lambda)\} \tag{17}$$

converges to a $\chi^2_{(1)}$ *distribution if* $\lambda = \lambda_0$*.*

An advantage of using the proposed semi-parametric likelihood proposed is that the distributions, F_1 and G (where G is the distribution of the community (training sample)

defined in (5) and (6)), can be estimated using the $\hat{q}_i$'s, i.e.:

$$\begin{aligned}\hat{F}_1(t) &= \frac{1}{N_1}\sum_{i=1}^{N_1}\frac{I(t_i \le t)}{1+\hat{\nu}[\exp(\hat{\alpha}+\hat{\beta}t_i)-1]},\\ \hat{G}(t) &= \frac{1}{N_1}\sum_{i=1}^{N_1}\frac{I(t_i \le t)[(1-\hat{\lambda}^*)+\hat{\lambda}^*\exp(\hat{\alpha}+\hat{\beta}t_i)]}{1+\hat{\nu}[\exp(\hat{\alpha}+\hat{\beta}t_i)-1]}.\end{aligned} \tag{18}$$

As Qin and Zhang (1997) suggested, the discrepancy between the distribution functions given in (18) and the empirical distribution functions:

$$\tilde{F}_1(t) = \sum_{i=1}^{m_1} I(z_i \le t)/m_1, \qquad \tilde{G}(t) = \sum_{i=1}^{n_1} I(x_i \le t)/n_1$$

can be used to form a goodness of fit statistic:

$$\Delta = \max_{-\infty<t<\infty} \sqrt{N}|\hat{F}_1(t) - \tilde{F}_1(t)|, \tag{19}$$

for the model (7). We do not give details of the theoretical results here. However, we will use (19) to assess the fit of the proposed method to the malaria data in the next section.

3. The malaria example

In this section we analyze the malaria dataset collected by Kitua et al (1996). The data were obtained from repeated cross-sectional surveys of parasitaemia and fever among children up to one year old in a village in the Kilombero district in Tanzania. Vounatsou et al (1998) used a subset of the data from children aged between 6 and 9 months that was collected in two seasons: the wet season (January-June) during which malaria prevalence is high and the dry season (July-December) during which the mosquito population, and also malaria prevalence, decreases. We use one of the datasets that can be obtained from

$$http://www.blackwellpublishers.co.uk/rss.$$

In this dataset, there are $n = 264$ observations in the mixture sample and $m = 144$ observations in the training sample. Among these, there are $n_0 = 53$ and $m_0 = 63$ observations with

zero parasite level in the mixture and training sample, respectively. Therefore, $m_1 = 81$ and $n_1 = 211$. The parasite level ranges from 0 to 399952.1. The data, after log-transformation for the non-zero data values are shown in Figure 1.

Three estimators were used to analyze the data: the binomial estimator based on maximizing ℓ_1, the semi-parametric estimator based on maximizing ℓ and the parametric estimator based on maximizing ℓ_P.

Using the binomial estimator, only (p, λ, λ^*) are relevant parameters. The binomial estimates for this dataset are

$$(\hat{p}_B, \hat{\lambda}_B, \hat{\lambda}_B^*) = (0.437, 0.541, 0.677).$$

The maximum semi-parametric likelihood estimates are

$$(\hat{\alpha}, \hat{\beta}, \hat{\lambda}, \hat{\lambda}^*, \hat{p}) = (-19.62, 2.038, 0.507, 0.641, 0.423).$$

To assess the goodness of fit of the semi-parametric method, the distribution function estimates of F_1 and F_2 using (18) are calculated and plotted against the corresponding empirical distribution functions (Figure 2). As seen in Figure 2, the semi-parametric distribution function estimates are extremely close to the empirical distribution functions. We also used 1000 bootstrap samples to calculate the significance of the statistic (19) and found the p-value to be 0.340, indicating no evidence of model lack of fit.

The maximum parametric likelihood estimation assumed normal models for the component distributions, $f_1 \sim N(\mu_1, \sigma^2)$ and $f_2 \sim N(\mu_2, \sigma^2)$. Note that $f_2(x)/f_1(x)$ satisfies (7) with

$$\alpha = \frac{\mu_1^2 - \mu_2^2}{2\sigma^2}, \qquad \beta = \frac{\mu_2 - \mu_1}{\sigma^2}$$

The estimated parameters are

$$(\hat{\alpha}_P, \hat{\beta}_P, \hat{\lambda}_P, \hat{\lambda}_P^*, \hat{p}_P) = (-9.427, 1.059, 0.627, 0.763, 0.478).$$

Clearly the choice of normal models for f_1 and f_2 is not good, λ is overestimated by an amount of 0.1, which is a large deviation considering that the range of λ is between 0 and 1.

The 95% semi-parametric likelihood ratio based confidence intervals for λ and λ^* are $(0.406, 0.615)$ and $(0.529, 0.748)$, respectively. Also the 95% binomial likelihood ratio based confidence intervals for λ and λ^* are $(0.380, 0.663)$ and $(0.497, 0.795)$, respectively. Note that the semi-parametric confidence intervals are much shorter that the binomial confidence intervals. We do not report the confidence intervals for the parametric method since its estimates are biased.

Another important problem in the Tanzania malaria survey data is to predict the malaria status in a child with a given non-zero parasite level. A popular approach is to diagnose the child with malaria if and only if the parasite level exceeds a given cutoff value. This approach is based on the observation that high parasite levels are less common among children without malaria. However there has been no clear criteria with which to select a suitable cutoff. Moreover, not all children from endemic areas have malaria. The conventional receive operational characteristic (ROC) analysis is biased if we do not take this fact into account since the malaria group contaminates the non-malaria group. Denote $D = 1$ or $D = 2$ as clinical non-malaria and malaria, respectively, for an individual from an endemicity. The conditional probability of $D = 2$ for a given parasite level, x, is

$$P(D = 2|x) = \frac{P(D = 2)f(x|D = 2)}{P(D = 1)f(x|D = 1) + P(D = 2)f(x|D = 2)} = \frac{\lambda^* f_2(x)}{\lambda^* f_2(x) + (1 - \lambda^*) f_1^*(x)}.$$

Under model (7), we have

$$P(D = 2|x) = \begin{cases} 0 & \text{if } x = 0 \\ \dfrac{\exp(\alpha^* + \beta x)}{1 + \exp(\alpha^* + \beta x)} & \text{if } x > 0 \end{cases}$$

where $\alpha^* = \alpha + \log \lambda^* - \log(1 - \lambda^*)$.

In Figure 3, we plotted the estimated conditional probability using the malaria data:

$$\hat{P}(D = 2|x) = \begin{cases} 0 & \text{if } x = 0 \\ \dfrac{\exp(-19.2 + 2.04x)}{1 + \exp(-19.62 + 2.04x)} & \text{if } x > 0 \end{cases}, \tag{20}$$

which can be used to predict the probability of malaria for observed parasite level of x in an endemicity. For example, if a case is to be diagnosed as malaria only if the probability is at least 80%, then the observed log-parasite level should be at least 9.

4. Simulation study

In this section, we present the results of a simulation study designed to evaluate the performance of the proposed estimator. In the study, we tried to mimic the malaria example by fixing $n = 264$ and $m = 144$. Data in the mixture sample were generated from a normal mixture model $(1 - \lambda^*)N(0, 1) + \lambda^* N(\mu, 1)$ and data in the training sample followed a standard normal distribution. One thousand simulations each were carried out under different combinations of λ^* and μ. For each combination, the means and standard deviations of the semi-parametric estimator are reported in Table 2. For comparison, we also report the corresponding values using the binomial estimator and the parametric estimator. For estimation of (λ, λ^*), $(\hat{\lambda}, \hat{\lambda}^*)$ and $(\hat{\lambda}_P, \hat{\lambda}^*_P)$ have better overall performance and smaller standard deviations than the binomial estimates $(\hat{\lambda}_B, \hat{\lambda}^*_B)$. This is expected since the binomial estimation only uses information from the binomial counts of zero and non-zero data. The advantages of the semi-parametric and the parametric methods over the binomial method are more significant when the two components (f_1 and f_2) in the mixture are well separated from each other and when the prevalence probability, $(1 - p)$, is high. On the other hand, when there is much overlap in the two components, the improvements are only moderate. These results are not surprising since in the latter case, not much information on λ (and λ^*) is contained in the mixture sample.

Comparing the semi-parametric and the parametric methods, the latter is more efficient in estimating the parameters α, β. However, for the more important parameters λ, λ^*, p, the semi-parametric method is nearly as efficient as the parametric method, in all the cases we studied. As demonstrated in the previous section, the semi-parametric method is more

robust than the parametric method under model mis-specification.

In Table 3, we report the empirical coverages of the 90% and 95% nominal confidence intervals for λ based on the semi-parametric likelihood ratio statistic (15), the binomial likelihood ratio statistic (17) and the parametric likelihood ratio statistic (16). From this table we can observe that the performances of all three likelihood ratio confidences are satisfactory. The empirical coverage levels are close to the nominal levels.

5. Conclusion

In this paper, we proposed a semi-parametric method for analyzing a "compound" mixture distribution problem with a training sample. The proposed method assumes the component densities are related by a density ratio model (or equivalently a logistic regression model). Based on this assumption, we used empirical likelihood to estimate the unknown parameters in the model. Unlike previous methods, which grouped data into distinct categories, the method discussed in this paper uses the original quantitative scale of the data. Therefore, the method avoids the arbitrariness in grouping and also gives more precise estimates. As demonstrated in the malaria example, the proposed method provided excellent fit to the data whereas the fully parametric method gave biased estimates.

The method described in this article depends on the existence of a training sample, as do other semi-parametric methods, for identifying the model parameters.

The method developed in this paper can also be applied to outputs of biomedical assays that classify samples into groups according to whether some outputs, such as parasite density or optical density, exceeds a given cut-off. The proposed method can also be generalized to cases where there are covariates.

APPENDIX: Proof of Theorem 1

First we establish some simple facts. Note that

$$E(n_1) = n[1 - p(1-\lambda)], \qquad E(m_1) = m(1-p), \qquad \lim(1-\lambda) = \frac{1-\lambda^*}{1-\lambda^* p}.$$

By the assumption $m/N \to \rho$, $(0 < \rho < 1)$ and the Weak Law of Large Number and (13), we have in probability

$$\nu_0 = \lambda_0^* \lim \frac{1}{1 + m_1/n_1} = \lambda_0^* \frac{1}{1 + E(m_1)/E(n_1)}.$$

Denote

$$Q_1 = \sum_{i=1}^{n_1} \frac{\lambda^* \exp(\alpha + \beta x_i)}{(1-\lambda^*) + \lambda^* \exp(\alpha + \beta x_i)} - N_1 \nu,$$

$$Q_2 = \sum_{k=1}^{n_1} \frac{\lambda^* x_k \exp(\alpha + \beta x_k)}{(1-\lambda^*) + \lambda^* \exp(\alpha + \beta x_k)} - \sum_{i=1}^{N_1} \frac{\nu t_i \exp(\alpha + \beta t_i)}{1 + \nu[\exp(\alpha + \beta t_i) - 1]},$$

$$Q_3 = -\frac{n_0}{1-\lambda^*} + \frac{np}{1-\lambda^* p} + \sum_{k=1}^{n_1} \frac{-1 + \exp(\alpha + \beta x_k)}{(1-\lambda^*) + \lambda^* \exp(\alpha + \beta x_k)},$$

$$Q_4 = \frac{m_0 + n_0}{p} - \frac{m_1 + n_1}{1-p} + \frac{n\lambda^*}{1-\lambda^* p},$$

and

$$Q_5 = -\sum_{i=1}^{N_1} \frac{\exp(\alpha + \beta t_i) - 1}{1 + \nu[\exp(\alpha + \beta t_i) - 1]}.$$

Then the maximum semi-parametric likelihood estimate, $\hat{\eta}$, is the solution of the equations $Q(\eta) = (Q_1, Q_2, Q_3, Q_4, Q_5) = 0$.

Expanding $Q(\hat{\eta})$ at the true value of η_0, we have

$$0 = Q(\hat{\eta}) = Q(\eta_0) + \frac{\partial Q(\eta_0)}{\partial \eta}(\hat{\eta} - \eta_0) + o_p(\|\hat{\eta} - \eta_0\|),$$

or

$$\sqrt{N}(\hat{\eta} - \eta_0) = \left(\frac{1}{N}\frac{\partial Q(\eta_0)}{\partial \eta}\right)^{-1} \frac{1}{\sqrt{N}} Q(\eta_0) + o_p(1).$$

By simple calculus, we have

$$\frac{\partial Q_1}{\partial \alpha} = \lambda^*(1-\lambda^*) \sum_{i=1}^{n_1} \frac{\exp(\alpha + \beta x_i)}{[(1-\lambda^*) + \lambda^* \exp(\alpha + \beta x_i)]^2},$$

$$\frac{\partial Q_1}{\partial \beta} = \lambda^*(1-\lambda^*)\sum_{i=1}^{n_1}\frac{x_i \exp(\alpha+\beta x_i)}{[(1-\lambda^*)+\lambda^*\exp(\alpha+\beta x_i)]^2},$$

$$\frac{\partial Q_1}{\partial \lambda^*} = \sum_{i=1}^{n_1}\frac{\exp(\alpha+\beta x_i)}{\{(1-\lambda^*)+\lambda^*\exp(\alpha+\beta x_i)\}^2}$$

$$\frac{\partial Q_1}{\partial p} = 0, \qquad \frac{\partial Q_1}{\partial \nu} = -N_1$$

$$\frac{\partial Q_2}{\partial \alpha} = \lambda^*(1-\lambda^*)\sum_{i=1}^{n_1}\frac{x_i \exp(\alpha+\beta x_i)}{[(1-\lambda^*)+\lambda^*\exp(\alpha+\beta x_i)]^2} - \nu(1-\nu)\sum_{j=1}^{N_1}\frac{t_j \exp(\alpha+\beta t_j)}{[1+\nu\{\exp(\alpha+\beta t_j)-1\}]^2}$$

$$\frac{\partial Q_2}{\partial \beta} = \lambda^*(1-\lambda^*)\sum_{i=1}^{n_1}\frac{x_i^2 \exp(\alpha+\beta x_i)}{[(1-\lambda^*)+\lambda^*\exp(\alpha+\beta x_i)]^2} - \nu(1-\nu)\sum_{j=1}^{N_1}\frac{t_j^2 \exp(\alpha+\beta t_j)}{[1+\nu\{\exp(\alpha+\beta t_j)-1\}]^2}$$

$$\frac{\partial Q_2}{\partial \lambda^*} = \sum_{i=1}^{n_1}\frac{x_i \exp(\alpha+\beta x_i)}{[(1-\lambda^*)+\lambda^*\exp(\alpha+\beta x_i)]^2}, \quad \frac{\partial Q_2}{\partial p} = 0, \quad \frac{\partial Q_2}{\partial \nu} = \sum_{j=1}^{N_1}\frac{t_j \exp(\alpha+\beta t_j)}{[1+\nu\{\exp(\alpha+\beta t_j)-1\}]^2}$$

$$\frac{\partial Q_3}{\partial \alpha} = \sum_{i=1}^{n_1}\frac{\exp(\alpha+\beta x_i)}{[(1-\lambda^*)+\lambda^*\exp(\alpha+\beta x_i)]^2}$$

$$\frac{\partial Q_3}{\partial \beta} = \sum_{i=1}^{n_1}\frac{x_i \exp(\alpha+\beta x_i)}{[(1-\lambda^*)+\lambda^*\exp(\alpha+\beta x_i)]^2}$$

$$\frac{\partial Q_3}{\partial \lambda^*} = -\frac{n_0}{(1-\lambda^*)^2} + n\frac{p^2}{(1-\lambda^* p)^2} - \sum_{i=1}^{n_1}\frac{[-1+\exp(\alpha+\beta x_i)]^2}{[(1-\lambda^*)+\lambda^*\exp(\alpha+\beta x_i)]^2}$$

$$\frac{\partial Q_3}{\partial p} = \frac{n}{(1-\lambda^* p)^2}, \qquad \frac{\partial Q_3}{\partial \nu} = 0$$

$$\frac{\partial Q_4}{\partial \alpha} = 0, \qquad \frac{\partial Q_4}{\partial \beta} = 0$$

$$\frac{\partial Q_4}{\partial \lambda^*} = \frac{n}{(1-\lambda^* p)^2}, \qquad \frac{\partial Q_4}{\partial p} = -\frac{m_0+n_0}{p^2} - \frac{m_1+n_1}{(1-p)^2} + \frac{n(\lambda^*)^2}{(1-\lambda^* p)^2}, \qquad \frac{\partial Q_4}{\partial \nu} = 0$$

$$\frac{\partial Q_5}{\partial \alpha} = -\sum_{i=1}^{N_1}\frac{\exp(\alpha+\beta t_i)}{\{1+\nu[\exp(\alpha+\beta t_i)-1]\}^2}$$

$$\frac{\partial Q_5}{\partial \beta} = -\sum_{i=1}^{N_1} \frac{t_i \exp(\alpha+\beta t_i)}{\{1+\nu[\exp(\alpha+\beta t_i)-1]\}^2}$$

$$\frac{\partial Q_5}{\partial \lambda^*} = 0, \qquad \frac{\partial Q_5}{\partial p} = 0,$$

$$\frac{\partial Q_5}{\partial \nu} = \sum_{i=1}^{N_1} \frac{[\exp(\alpha+\beta t_i)-1]^2}{\{1+\nu[\exp(\alpha+\beta t_i)-1]\}^2}$$

For simplicity, denote

$$a(x) = \frac{\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)}, \quad b(x) = \frac{\exp(\alpha_0+\beta_0 x)}{(1-\nu_0)+\nu_0\exp(\alpha_0+\beta_0 x)}, \quad \kappa = \frac{E(n_1/n)}{\rho+1}. \tag{A.1}$$

It can be proved that, in probability,

$$\begin{aligned}\frac{1}{N}\frac{\partial Q_1(\eta_0)}{\partial \alpha} &\to \lambda_0^*(1-\lambda_0^*)\kappa \int \frac{\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)} dF_1(x) \\ &= \lambda_0^*(1-\lambda_0^*)\kappa \int a(x) dF_1(x) = v_{11}\end{aligned}$$

$$\begin{aligned}\frac{1}{N}\frac{\partial Q_1(\eta_0)}{\partial \beta} &\to \lambda_0^*(1-\lambda_0^*)\kappa \int \frac{x\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)} dF_1(x) \\ &= \lambda_0^*(1-\lambda_0^*)\kappa \int x a(x) dF_1(x) = v_{12}\end{aligned}$$

$$\frac{1}{N}\frac{\partial Q_1(\eta_0)}{\partial \lambda^*} \to \kappa \int \frac{\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)} dF_1(x) = \kappa \int a(x) dF_1(x) = v_{13}$$

$$\frac{1}{N}\frac{\partial Q_1(\eta_0)}{\partial p} \to 0 = v_{14}, \qquad \frac{1}{N}\frac{\partial Q_1(\eta_0)}{\partial \nu} \to -\kappa\frac{\lambda_0^*}{\nu_0} = v_{15},$$

$$\begin{aligned}&\frac{1}{N}\frac{\partial Q_2(\eta_0)}{\partial \alpha} \\ \to\ & \kappa\lambda_0^*(1-\lambda_0^*) \int \frac{x\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)} dF_1(x) \\ -\ & \kappa\nu_0(1-\nu_0)\frac{\lambda_0^*}{\nu_0} \int \frac{x\exp(\alpha_0+\beta_0 x)}{(1-\nu_0)+\nu_0\exp(\alpha_0+\beta_0 x)} dF_1(x) \\ =\ & \kappa\lambda_0^*(1-\lambda_0^*) \int x a(x) dF_1(x) - \kappa(1-\nu_0)\lambda_0^* \int x b(x) dF_1(x) = v_{21}\end{aligned}$$

$$\frac{1}{N}\frac{\partial Q_2(\eta_0)}{\partial \beta}$$

$$
\begin{aligned}
\rightarrow \quad & \kappa\lambda_0^*(1-\lambda_0^*)\int \frac{x^2\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)}dF_1(x) \\
- \quad & \kappa\nu_0(1-\nu_0)\frac{\lambda_0^*}{\nu_0}\int \frac{x^2\exp(\alpha_0+\beta_0 x)}{(1-\nu_0)+\nu_0\exp(\alpha_0+\beta_0 x)}dF_1(x) \\
= \quad & \kappa\lambda_0^*(1-\lambda_0^*)\int x^2 a(x)dF_1(x) - \kappa(1-\nu_0)\lambda_0^*\int x^2 b(x)dF_1(x) = v_{22}
\end{aligned}
$$

$$
\frac{1}{N}\frac{\partial Q_2(\eta_0)}{\partial\lambda^*} \rightarrow \kappa\int \frac{x\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)}dF_1(x) = \int xa(x)dF_1(x) = v_{23},
$$

$$
\frac{1}{N}\frac{\partial Q_2(\eta_0)}{\partial p} \rightarrow 0 = v_{24}
$$

$$
\frac{1}{N}\frac{\partial Q_2(\eta_0)}{\partial\nu} \rightarrow \kappa\frac{\lambda_0^*}{\nu_0}\int \frac{x\exp(\alpha_0+\beta_0 x)}{(1-\nu)+\nu\exp(\alpha+\beta x)}dF_1(x), = \kappa\frac{\lambda_0^*}{\nu_0}\int xb(x)dF_1(x) = v_{25}
$$

$$
\frac{1}{N}\frac{\partial Q_3(\eta_0)}{\partial\alpha} \rightarrow \kappa\int \frac{\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)}dF_1(x) = \kappa\int a(x)dF_1(x) = v_{31}
$$

$$
\frac{1}{N}\frac{\partial Q_3(\eta_0)}{\partial\beta} \rightarrow \kappa\int \frac{x\exp(\alpha_0+\beta_0 x)}{(1-\lambda_0^*)+\lambda_0^*\exp(\alpha_0+\beta_0 x)}dF_1(x) = \kappa\int xa(x)dF_1(x) = v_{32}
$$

$$
\frac{1}{N}\frac{\partial Q_3(\eta_0)}{\partial\lambda^*} \rightarrow -\frac{\kappa p_0}{(1-\lambda_0^* p_0)(1-\lambda_0^*)} + \frac{\kappa p_0^2}{(1-p_0)(1-\lambda_0^* p_0)} - \frac{\kappa}{1-\lambda^*}\int \frac{[1-a(x)]^2}{1-\lambda^* a(x)}dF_1(x) = v_{33}
$$

$$
\frac{1}{N}\frac{\partial Q_3(\eta_0)}{\partial p} \rightarrow \frac{\kappa}{(1-\lambda_0^* p_0)(1-p_0)} = v_{34}, \qquad \frac{1}{N}\frac{\partial Q_3}{\partial\nu} \rightarrow 0 = v_{34}
$$

$$
\frac{1}{N}\frac{\partial Q_4(\eta_0)}{\partial\alpha} \rightarrow 0 = v_{41}, \qquad \frac{1}{N}\frac{\partial Q_4}{\partial\alpha} \rightarrow 0 = v_{42}
$$

$$
\frac{1}{N}\frac{\partial Q_4(\eta_0)}{\partial\lambda^*} \rightarrow \frac{\kappa}{(1-p_0)(1-\lambda_0^* p_0)} = v_{43},
$$

$$
\frac{1}{N}\frac{\partial Q_4(\eta_0)}{\partial p} \rightarrow -\frac{1}{p_0^2} + \frac{\rho E(m_1)+E(n_1)}{1+\rho}\left(\frac{1}{p_0^2} - \frac{1}{(1-p_0)^2}\right) = v_{44}
$$

$$
\frac{1}{N}\frac{\partial Q_4(\eta_0)}{\partial\nu} \rightarrow 0 = v_{45}
$$

$$
\frac{1}{N}\frac{\partial Q_5(\eta_0)}{\partial\alpha} \rightarrow -\frac{\kappa\lambda_0^*}{\nu_0}\int \frac{\exp(\alpha_0+\beta_0 x)}{1-\nu_0+\nu_0\exp(\alpha_0+\beta_0 x)}dF_1(x) = -\frac{\kappa\lambda_0^*}{\nu_0}\int b(x)dF_1(x) = v_{51}
$$

$$\frac{1}{N}\frac{\partial Q_5(\eta_0)}{\partial \beta} \to -\frac{\kappa\lambda_0^*}{\nu_0}\int \frac{x\exp(\alpha_0+\beta_0 x)}{1-\nu_0+\nu_0\exp(\alpha_0+\beta_0 x)}dF_1(x) = -\frac{\kappa\lambda_0^*}{\nu_0}\int xb(x)dF_1(x) = v_{52}$$

$$\frac{1}{N}\frac{\partial Q_5(\eta_0)}{\partial \lambda^*} \to 0 = v_{53}, \qquad \frac{1}{N}\frac{\partial Q_5(\eta_0)}{\partial p} \to 0 = v_{54}$$

$$\begin{aligned}\frac{1}{N}\frac{\partial Q_5(\eta_0)}{\partial \nu} &\to -\frac{\kappa\lambda^*}{\nu_0}\int \frac{[\exp(\alpha_0+\beta_0 x)-1]^2}{1-\nu_0+\nu_0\exp(\alpha_0+\beta_0 x)}dF_1(x)\\ &= -\frac{\kappa\lambda^*}{\nu_0(1-\nu_0)}\int \frac{[1-b(x)]^2}{1-\nu_0 b(x)}dF_1(x) = v_{55}\end{aligned}$$

Therefore, in probability, we have proved that

$$\frac{1}{N}\frac{\partial Q}{\partial \eta}|_{\eta=\eta_0} \to V = (v_{ij})_{1\le i,j\le 5}. \tag{A.2}$$

We can rewrite $Q_k(\eta_0), k = 1, 2, ..., 5$ as

$$Q_k(\eta_0) = \sum_{i=1}^{n} q_k(x_i) + \sum_{j=1}^{m} r_k(z_j), \qquad k = 1, 2, ..., 5$$

where

$$q_1(x) = [\lambda_0^* a(x_i) - \nu_0]I(x > 0), \qquad r_1(z) = -\nu_0 I(z_i > 0)$$

$$q_2(x) = \lambda_0^* x a(x) I(x > 0), \qquad r_2(z) = -\nu_0 z b(z) I(z > 0)$$

$$q_3(x) = -\frac{I(x=0)}{1-\lambda_0^*} + \frac{p_0}{1-\lambda_0^* p_0} + \frac{a(x)-1}{1-\lambda_0^*}I(x>0), \qquad r_3(z) = 0$$

$$q_4(x) = \frac{I(x=0)}{p_0} - \frac{I(x>0)}{1-p_0} + \frac{\lambda_0^*}{1-\lambda_0^* p_0}, \qquad r_4(z) = \frac{I(z=0)}{p_0} - \frac{I(z>0)}{1-p_0}$$

$$q_5(x) = -\frac{b(x)-1}{1-\nu_0}I(x>0), \qquad r_5(z) = -\frac{b(z)-1}{1-\nu_0}I(z>0).$$

Easily we have

$$Var(Q_k) = nVar(q_k(X)) + mVar(r_k(Z)),$$

and

$$Cov(Q_k, Q_l) = nCov(q_k(X), q_l(X)) + mCov(r_k(Z), r_l(Z)), \quad k \ne l$$

$$u_{ij} = \frac{1}{1+\rho}Cov(q_i(X), q_j(X)) + \frac{\rho}{1+\rho}Cov(r_i(Z), r_j(Z)), \qquad 1 \le i, j \le 5.$$

By the Central Limit Theorem, we can prove, in distribution,

$$\frac{1}{\sqrt{N}}Q(\eta_0) \to N(0, U), \qquad U = (u_{ij})_{1\leq i,j\leq 5}. \tag{A.3}$$

Finally by Slutsky's Theorem, we can show that, in distribution,

$$\sqrt{N}(\hat{\eta} - \eta_0) \to N(0, \Sigma), \quad \Sigma = V^{-1}UV^{-1}. \tag{A.4}$$

The proof of Theorems 2 and 3 are tedious but straightforward, therefore, we omit them. Similar proofs can be found in the first author's University of Waterloo Ph.D. dessertation.

References

Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika* **66** 17-26.

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

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *I. S. Review*, **58**, 109-127.

Hall, P. and Titterington, D. M. (1984). Efficient nonparametric estimation of mixture proportions. *Journal of the Royal Statistical Society, Ser.* B **46** 465-473.

Kitua, A. Y., Smith, T., Alonso, P. L., Masanja, H., Urassa, H., Menendez, C., Kimario, J. and Tanner, M. (1996). Plasmodium falciparum malaria in the first year of life in an area of intense and perennial transmission. *Trop. Med. Int. Hlth*, **1**, 475-484.

Lancaster, T. and Imbens, G. (1996). Case-control studies with contaminated controls. *J. Econometrics* **71** 145-60.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*, Haywood, CA: Institute of Mathematical Statistics.

McLachlan, G., J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, John Wiley & Sons, New York.

McLachlan, G. J. an Peel, D. (2001). *Finite Mixture Models*. John Wiley and Sons, New York.

Murray, G.D. and Titterington, D.M. (1978). Estimation problems with data from a mixture. *Applied Statistics* **27** 325-334.

Nagelkerke, N. J. D., Borgdorff, M. W. and Kim, S. J. (2001). Logistic discrimination of mixtures of M. tuberculosis and non-specific tuberculin reactions. *Statistics in Medicine* **20**, 1113-1124.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann. of Statist.*, **27**, 1368-84.

Qin, J., Berwick, M., Ashbolt, R. and Dwyer, T. (2002). Quantifying the change of melanoma incidence by Breslow thickness. *Biometrics*, **58**, 665-670.

Qin, J. and Lawless, J.F. (1994). Empirical likelihood and general estimating equations. *Ann. of Statist.*, **22**, 300-325.

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, **84**, 609-618.

Smith, T., Schellenberg, J. A. and Hayes, R. (1994). Attributable fraction estimates and case definitions for malaria in endemic areas. *Statist. Med.*, **13**, 2345-2358.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616-620.

Vardi, Y. (1985). Empirical distribution in selection bias models. *Ann. of Statist.* **13**, 178-203.

Vounatsou, P., Smith, T. and Smith, A. F. M. (1998). Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions. *Appl. Statist.*, **47**, 575-587.

White, I. R. and Thompson, S. G. (2003). Choice of test for comparing two groups, with particular application to skewed outcomes. *Statistics in Medicine*.

Zhang, B. (2001). An information matrix test for logistic regression models based on case-control data. *Biometrika*, **88**, 921-932.

Zou, F., Fine, J.P. and Yandell, B. S. (2002). On empirical likelihood for a semi-parametric mixture model. *Biometrika*. **89**, 61-75.

Table 1. Partitioning of a set of data in an endemicity

Parasite level	No Malaria	Malaria	Total
$X = 0$	$p(1-\lambda)$	0	$p(1-\lambda)$
$X > 0$	$(1-p)(1-\lambda)$	λ	$1-p(1-\lambda)$
Total	$1-\lambda$	λ	1

Table 2. Mean (standard deviation) of different estimators. Based on 1000 simulations

μ	Estimators	$p = 0.2, \lambda = 0.5$	$p = 0.3, \lambda = 0.6$	$p = 0.4, \lambda = 0.5$
$\mu = 2.0$	$\hat{\lambda}$	0.496 (0.069)	0.598 (0.059)	0.498 (0.062)
	$\hat{\lambda}^*$	0.551 (0.072)	0.679 (0.059)	0.622 (0.066)
	$\hat{\alpha}$	-2.424 (1.348)	-2.226 (0.840)	-2.359 (1.343)
	$\hat{\beta}$	2.272 (0.860)	2.170 (0.598)	2.257 (0.933)
	$\hat{p}$	0.201 (0.028)	0.299 (0.034)	0.401 (0.035)
	$\hat{\lambda}_B$	0.483 (0.136)	0.593 (0.089)	0.493 (0.084)
	$\hat{\lambda}_B^*$	0.535 (0.145)	0.672 (0.090)	0.615 (0.092)
	$\hat{p}_B$	0.197 (0.032)	0.302 (0.037)	0.401 (0.040)
	$\hat{\lambda}_P$	0.500 (0.052)	0.599 (0.048)	0.500 (0.048)
	$\hat{\lambda}_P^*$	0.555 (0.053)	0.680 (0.047)	0.624 (0.051)
	$\hat{\alpha}_P$	-2.065 (0.424)	-2.053 (0.379)	-2.059 (0.434)
	$\hat{\beta}_P$	2.053 (0.310)	2.051 (0.296)	2.047 (0.333)
	$\hat{p}_P$	0.201 (0.026)	0.299 (0.032)	0.401 (0.033)
$\mu = 1.5$	$\hat{\lambda}$	0.498 (0.085)	0.602 (0.066)	0.503 (0.066)
	$\hat{\lambda}^*$	0.552 (0.089)	0.683 (0.065)	0.626 (0.070)
	$\hat{\alpha}$	-1.371 (1.201)	-1.222 (0.459)	-1.244 (0.525)
	$\hat{\beta}$	1.678 (0.795)	1.586 (0.407)	1.604 (0.466)
	$\hat{p}$	0.200 (0.027)	0.303 (0.033)	0.402 (0.036)
	$\hat{\lambda}_B$	0.476 (0.135)	0.594 (0.089)	0.497 (0.085)
	$\hat{\lambda}_B^*$	0.527 (0.144)	0.674 (0.090)	0.619 (0.093)
	$\hat{p}_B$	0.197 (0.032)	0.302 (0.037)	0.401 (0.040)
	$\hat{\lambda}_P$	0.503 (0.072)	0.603 (0.059)	0.503 (0.061)
	$\hat{\lambda}_P^*$	0.558 (0.075)	0.685 (0.058)	0.619 (0.093)
	$\hat{\alpha}_P$	-1.188 (0.357)	-1.161 (0.296)	-1.185 (0.351)
	$\hat{\beta}_P$	1.551 (0.319)	1.538 (0.285)	1.556 (0.333)
	$\hat{p}_P$	0.201 (0.026)	0.303 (0.032)	0.402 (0.035)
$\mu = 1.0$	$\hat{\lambda}$	0.491 (0.107)	0.599 (0.076)	0.501 (0.075)
	$\hat{\lambda}^*$	0.544 (0.112)	0.680 (0.076)	0.624 (0.080)
	$\hat{\alpha}$	-0.673 (0.826)	-0.552 (0.258)	-0.561 (0.284)
	$\hat{\beta}$	1.146 (0.582)	1.060 (0.307)	1.068 (0.350)
	$\hat{p}$	0.199 (0.029)	0.303 (0.035)	0.402 (0.038)
	$\hat{\lambda}_B$	0.476 (0.135)	0.594 (0.089)	0.497 (0.085)
	$\hat{\lambda}_B^*$	0.527 (0.144)	0.674 (0.090)	0.619 (0.093)
	$\hat{p}_B$	0.197 (0.032)	0.302 (0.037)	0.401 (0.040)
	$\hat{\lambda}_P$	0.497 (0.099)	0.601 (0.074)	0.501 (0.074)
	$\hat{\lambda}_P^*$	0.550 (0.104)	0.681 (0.074)	0.624 (0.079)
	$\hat{\alpha}_P$	-0.580 (0.283)	-0.536 (0.195)	-0.550 (0.236)
	$\hat{\beta}_P$	1.074 (0.326)	1.044 (0.262)	1.058 (0.311)
	$\hat{p}_P$	0.200 (0.029)	0.302 (0.037)	0.402 (0.038)

Table 3. Empirical coverages of 90% (95%) likelihood ratio confidence intervals using three different methods. Based on 1000 simulations.

μ	Methods	$p = 0.2, \lambda = 0.5$	$p = 0.3, \lambda = 0.6$	$p = 0.4, \lambda = 0.5$
1.0	Semi-parametric	89.8% (94.6%)	90.1% (96.2%)	89.1% (95.0%)
	Binomial	88.7% (94.5%)	89.5% (96.1%)	89.9% (93.8%)
	Parametric	89.3% (94.5%)	88.7% (95.3%)	89.2% (94.7%)
1.5	Semi-parametric	89.3% (94.4%)	90.0% (96.1%)	90.5% (95.1%)
	Binomial	89.2% (94.6%)	91.3% (95.4%)	88.9% (94.1%)
	Parametric	89.3% (94.5%)	88.7% (95.3%)	89.2% (94.7%)
2.0	Semi-parametric	88.7% (93.8%)	91.1% (95.5%)	90.2% (94.8%)
	Binomial	88.9% (94.7%)	91.2% (96.1%)	88.4% (93.8%)
	Parametric	89.3% (94.5%)	88.7% (95.3%)	89.2% (94.7%)

Figure 1: Histograms showing parasite levels in the endemicity and community data (non-zero data only)

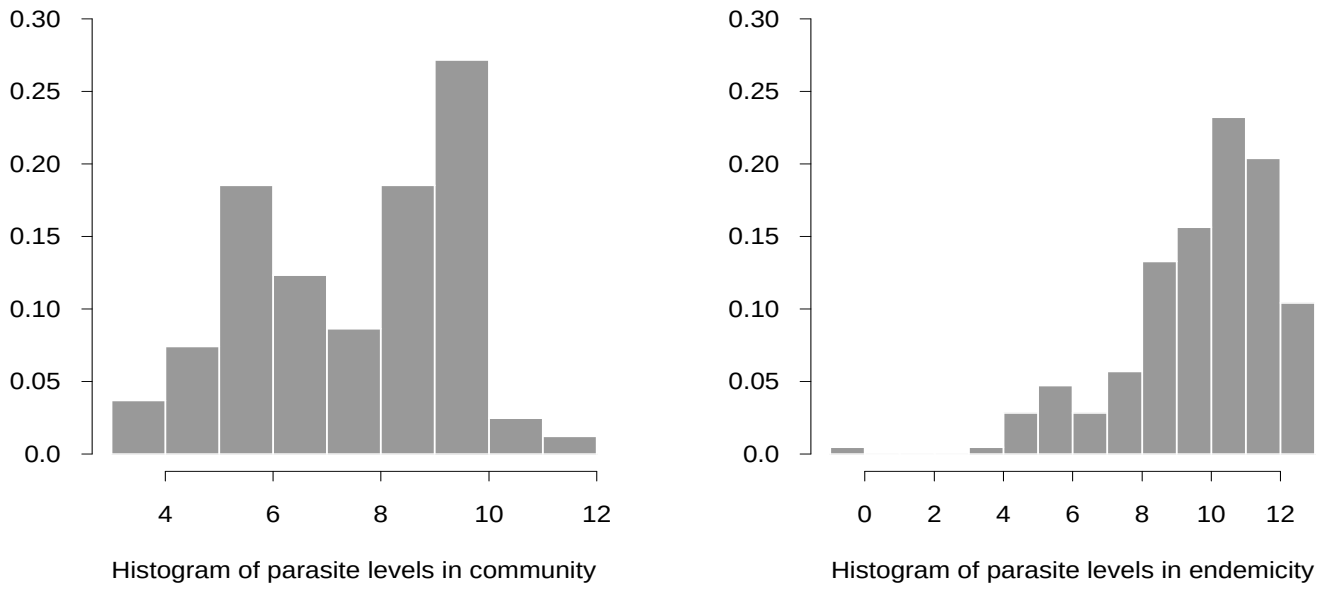

Figure 2: Estimated cumulative probability function of parasite levels using empirical and semi-parametric methods

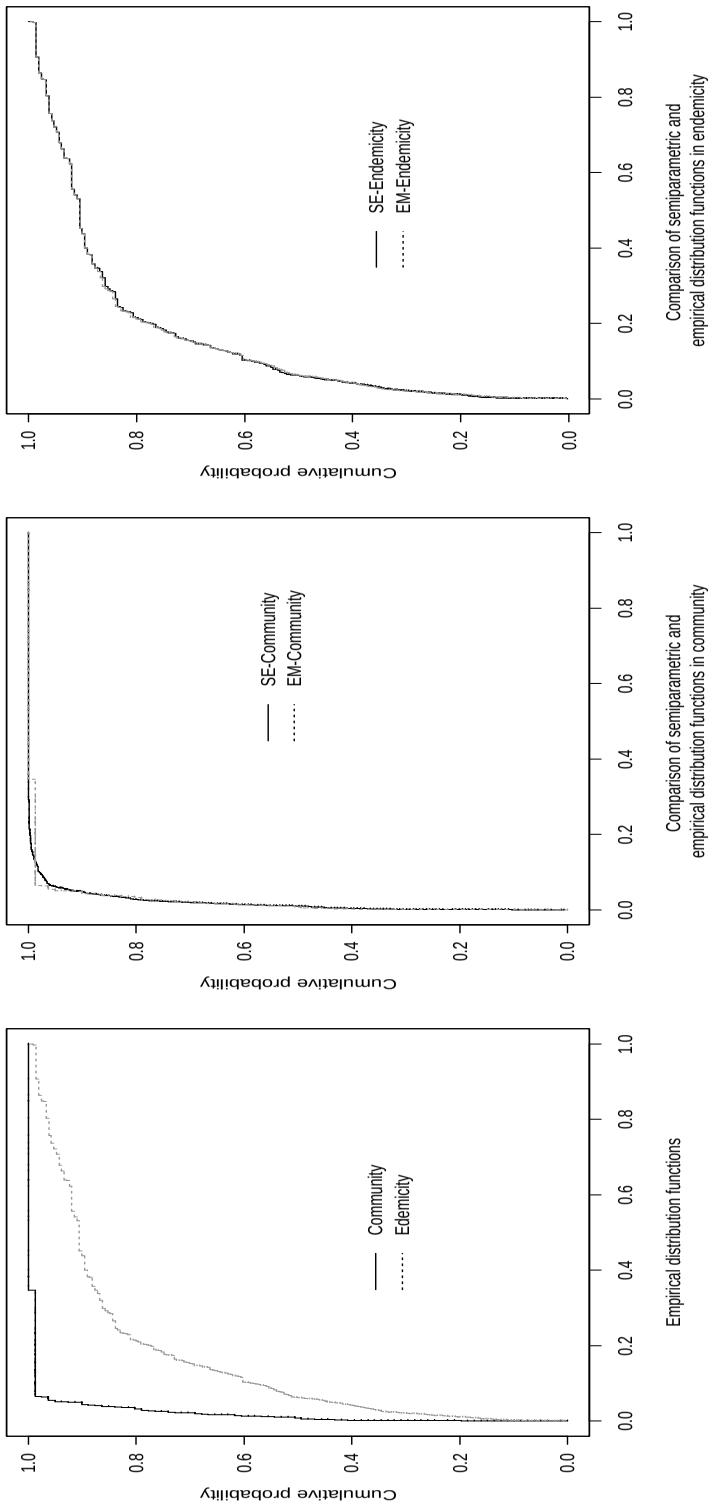

Figure 3: Semi-parametric estimate of probability of malaria using (20)