

Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions.

By P. VOUNATSOU[†] T. SMITH

Swiss Tropical Institute, Basle, Switzerland

and A. F. M. SMITH

Department of Mathematics, Imperial College, London, UK

SUMMARY

Malaria illness can be diagnosed by the presence of fever and parasitaemia. However, in highly endemic areas the diagnosis of clinical malaria can be difficult since children may tolerate parasites without fever and may have fever due to other causes. We propose a novel, simulation-based Bayesian approach for obtaining precise estimates of the probabilities of children with different levels of parasitaemia having fever due to malaria, by formulating the problem as a mixture of distributions. The suggested methodology is a general one for decomposing any two-component mixture distribution nonparametrically, when an independent training sample is available from one of the components. It is based on the assumption that one of the component distributions lies on the left of the other but there is some overlap between the distributions.

Keywords: Attributable fractions; Bayesian methods; Malaria; MCMC simulation; Mixture decomposition; Nonparametric methods

[†]Address for correspondence: Swiss Tropical Institute, Socinstrasse 57, Basle, Switzerland.
Email: penelope@tropi.sti.unibas.ch

1 Introduction

Clinical malaria can be diagnosed by the presence of parasites and high temperature (fever). However, in endemic areas children can tolerate parasites without symptoms and may have fever due to other causes. It is important to identify the cause of morbidity because the implementation of intervention programmes first requires an assessment of the number of cases of malaria.

A simple statistical approach to estimating the frequency of malaria among fever cases is to compare the parasite prevalence p_f in febrile children, who can be malaria or non-malaria cases, with the parasite prevalence p_a , in afebrile controls from the community, by calculating the malaria attributable fraction

$$\lambda_{AF} = \frac{p_f - p_a}{1 - p_a} \quad (1)$$

(Greenwood, 1987). However in areas of high transmission, p_a is very high and the proportion of afebrile children without parasitaemia is low, resulting either in negative estimates of λ_{AF} (when p_f is less than p_a) or imprecise estimates when p_a is close to 1. Alternative methods were explored by Smith *et al.* (1994), and their limitations discussed by Smith and Vounatsou (1997).

We propose a novel approach for obtaining precise estimates of the frequency of clinical malaria by formulating the problem as a mixture of distributions. The mixture consists of parasite densities in children with fever either due to malaria or due to other causes. One component of the mixture corresponds to children without clinical malaria and the other to children with clinical malaria. Parasite levels in children from the community are available and are used as a training sample, that is a sample that comes from the component of the mixture corresponding to children without clinical malaria but who may have parasites. The mixing proportion estimates the proportion of children whose fever is attributable to malaria.

The data arose from repeated cross-sectional surveys of parasitaemia and fever among

426 children up to one year old in a village in the Kilombero district in Tanzania. The data are described by Kitua *et al.* (1996).

Standard statistical methods for decomposing mixture distributions assume known form for the component distributions (Everitt and Hand, 1981; Titterton, 1985). However, in the malaria problem the underlying mechanisms generating the observations are not known. Non-parametric approaches for estimating mixing proportions are available in the literature. Some are based on kernel-type density estimators (Murray and Titterton, 1978) and are sensitive to the choice of the smoothing parameter; others use the EM algorithm or a stochastic variant (e.g Diebolt and Robert, 1994). Smith and Vounatsou (1997) discussed four different estimation techniques including a latent class model which uses the EM algorithm for estimating the mixing proportion. However, practical implementation of the EM algorithm is not straightforward since it may converge to a local maximum or a saddle point, (Wu, 1983).

Recently, the development of simulation-based methods in Bayesian statistics (Smith and Roberts, 1993) have made possible the implementation of the Bayesian approach to mixture problems. Diebolt and Robert (1994) and Robert (1996) derived the posterior distribution using conjugate priors and introduced indicator variables to classify observations, according to the component of the mixture from which they come. They simulate indicator variables and parameters, using a simulation algorithm. To avoid non-identifiability, which can arise under non-informative priors, Robert (1996) proposed reparametrization of the posterior. It remains unclear how this can be applied in a completely nonparametric approach.

In this paper, we present a model for decomposing nonparametrically a two-component mixture distribution. The underlying assumptions are that a "training" sample is available from one of the components and that the one component of the mixture is on the left of the other. We group the ordered observations from the "training" sample and the mixture into a number of categories and derive the likelihood as a product of two multinomial dis-

tributions. Adopting a non-informative uniform prior, we estimate the parameters of the posterior using the Gibbs sampler (Smith and Roberts, 1993). Results on simulated data verify the validity of our methodology. Comparison of our estimates on real data with those obtained by earlier methods showed that our approach gives comparable results, without the drawbacks of these methods. In Section 2 the model is described, and in Section 3, we estimate the parameters of the model by applying the simulation-based approach. The performance of the approach is discussed in Section 4 where it is tested on a number of simulated data sets. In Section 5 we implement the methodology on real data from malaria epidemiology to estimate malaria attributable fractions during two seasons; the wet season during which the mosquito population, and hence exposure to malaria infection is high and the dry season during which the mosquito population is lower. Further applications and conclusions are discussed in Section 6.

2 Model

Let x_1, x_2, \dots, x_n be a set of independent, identically distributed random variables from the two-component mixture distribution

$$f(x) = (1 - \lambda)g_1(x) + \lambda g_2(x),$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are the unknown component distributions and λ is the mixing proportion. Observations from one component distribution, $g_1(\cdot)$, are assumed to be smaller than those from the other and there is some overlap between the two distributions. Also, assume that a "training" sample, (say from the first component distribution, $g_1(\cdot)$) is available. The objective is to estimate the proportion λ of observations that come from $g_2(\cdot)$.

First we divide the data into K ordered categories over the range of x . Let n_i be the number of observations from the mixture that belong to category i , $i = 1, 2, \dots, K$, and m_i be the number of observations from the training sample (say from $g_1(\cdot)$) that belong to category i , $i = 1, 2, \dots, K$. If x is not continuous, each of its discrete values can be considered

TABLE 1: Data summarized into the K ordered categories of the mixture f and of the g_1 component distribution.

Category i	1	2	3	...	K
Sample from g_1	m_1	m_2	m_3	...	m_K
Sample from f	n_1	n_2	n_3	...	n_K

a separate category. Thus the data can be summarized as above.

From Table 1, we can assume that the data come from two multinomial distributions, specified by the parameters

$$\begin{aligned}\theta_i &= P(x \in \text{category } i \mid P_1) \\ \phi_i &= P(x \in \text{category } i \mid P_2) \quad \text{and} \\ \lambda &= P(x \in P_2),\end{aligned}$$

where P_1 and P_2 are the distribution functions of the two components, $g_1(\cdot)$ and $g_2(\cdot)$, respectively. Then

$$(m_1, m_2, \dots, m_K) \sim Mn\left(\sum_{i=1}^K m_i, \theta_1, \theta_2, \dots, \theta_k\right)$$

and

$$(n_1, n_2, \dots, n_K) \sim Mn\left(\sum_{i=1}^K n_i, p_1, p_2, \dots, p_k\right),$$

where

$$\begin{aligned}p_i &= P(x \in \text{category } i) \\ &= P(x \in \text{category } i \mid P_1)p(x \text{ from } P_1) + P(x \in \text{category } i \mid P_2)p(x \text{ from } P_2) \\ &= (1 - \lambda)\theta_i + \lambda\phi_i\end{aligned}$$

The joint density of the data $\mathbf{m} = (m_1, \dots, m_k)^T$ and $\mathbf{n} = (n_1, \dots, n_k)^T$ is

$$\mathcal{L}(\mathbf{m}, \mathbf{n} \mid \boldsymbol{\theta}, \boldsymbol{\phi}, \lambda) \propto \prod_{i=1}^{K-1} \theta_i^{m_i} \left(1 - \sum_{i=1}^{K-1} \theta_i\right)^{m_K} \times$$

$$\prod_{i=1}^{K-1} ((1-\lambda)\theta_i + \lambda\phi_i)^{n_i} \left((1-\lambda)(1 - \sum_{i=1}^{K-1} \theta_i) + \lambda(1 - \sum_{i=1}^{K-1} \phi_i) \right)^{n_K}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^T$ and $\boldsymbol{\phi} = (\phi_1, \dots, \phi_K)^T$. As $\boldsymbol{\theta}, \boldsymbol{\phi}$ and λ are probabilities,

$$0 < \theta_i, \phi_i, \lambda < 1, \quad \sum_{i=1}^K \theta_i = 1 \text{ and } \sum_{i=1}^K \phi_i = 1. \quad (2)$$

Also, assume that observations from the first component, $g_1(\cdot)$ are smaller than the ones from $g_2(\cdot)$, that is

$$P(x \text{ from } P_2 \mid \text{category } 1) = 0 \text{ and}$$

$$P(x \text{ from } P_2 \mid \text{category } i-1) < P(x \text{ from } P_2 \mid \text{category } i) \text{ for } i = 2, \dots, K$$

By simple probability calculations, this constraint is equivalent to

$$0 = \frac{\phi_1}{\theta_1} < \frac{\phi_2}{\theta_2} < \dots < \frac{\phi_{K-1}}{\theta_{K-1}} < \frac{1 - \sum_{i=1}^{K-1} \phi_i}{1 - \sum_{i=1}^{K-1} \theta_i}. \quad (3)$$

It can be incorporated into the model by assuming a uniform prior distribution in $[0, 1]$

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}, \lambda) \propto 1$$

with parameters satisfying constraints (2) and (3). The above constraints make the problem identifiable. To minimise any possible bias coming from the first constraint (i.e $\phi_1 = 0$), we should try to choose the first category such that it has the smallest possible range of values of x . We further explain this point in Section 5. Then the posterior density, derived by combining the likelihood and the prior, is proportional to the likelihood, that is $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \lambda \mid \mathbf{m}, \mathbf{n}) \propto \mathcal{L}(\mathbf{m}, \mathbf{n} \mid \boldsymbol{\theta}, \boldsymbol{\phi}, \lambda)$ under constraints (2) and (3).

3 The simulation approach

The Gibbs sampling algorithm can be used to simulate from the posterior $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \lambda \mid \mathbf{m}, \mathbf{n})$ under the constraints (2) and (3). To facilitate simulation $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \lambda \mid \mathbf{m}, \mathbf{n})$ is reparametrized in terms of $\boldsymbol{\theta}, \mathbf{z} = (z_1, z_2, \dots, z_{k-1})^T$ and λ , where $z_i = \frac{\phi_i}{\theta_i}$ for $i = 1, \dots, K-1$. Adopting the

notation of Gelfand and Smith, (1990) for density functions, the full conditionals required for implementing the Gibbs sampler can be written as follows:

$$[\theta_i \mid \theta_j, z_i, z_j, j = 1, \dots, K-1, j \neq i, \mathbf{m}, \mathbf{n}] \propto \theta_i^{m_i+n_i+1} \left(1 - \sum_{i=1}^{K-1} \theta_i\right)^{m_K} \quad (4)$$

$$\times (1 - \lambda + \lambda z_i)^{n_i} \left((1 - \lambda) \left(1 - \sum_{i=1}^{K-1} \theta_i\right) + \lambda \left(1 - \sum_{i=1}^{K-1} \theta_i z_i\right) \right)^{n_k},$$

where $0 < \theta_i < \min \left\{ 1, 1 - \sum_{j \neq i}^{K-1} \theta_j, (1 - \sum_{j \neq i}^{K-1} \theta_j z_j) / z_j \right\}$ and $i = 1, \dots, K-1$,

$$[z_i \mid z_j, \theta_i, \theta_j, j = 1, \dots, K-1, j \neq i, \mathbf{m}, \mathbf{n}] \propto (1 - \lambda + \lambda z_i)^{n_i} \quad (5)$$

$$\times \left((1 - \lambda) \left(1 - \sum_{i=1}^{K-1} \theta_i\right) + \lambda \left(1 - \sum_{i=1}^{K-1} \theta_i z_i\right) \right)^{n_k},$$

where $0 = z_1 < z_2 < \dots < z_{K-1} < \frac{1 - \sum_{i=1}^{K-1} \theta_i z_i}{1 - \sum_{i=1}^{K-1} \theta_i}$ and $i = 1, \dots, K-1$,

$$[\lambda \mid \theta_j, z_j, j = 1, \dots, K, \mathbf{m}, \mathbf{n}] \propto (1 - \lambda + \lambda z_i)^{n_i} \left((1 - \lambda) \left(1 - \sum_{i=1}^{K-1} \theta_i\right) + \lambda \left(1 - \sum_{i=1}^{K-1} \theta_i z_i\right) \right)^{n_k} \quad (6)$$

where $0 < \lambda < 1$.

We simulate from the above conditionals using the generalised ratio-of-uniforms method (Wakefield *et al.*, 1991) which requires the conditionals to be defined on the real line. Therefore we apply the logit transformation to the conditionals in equations (4) - (6). To simulate from (5), we first reparametrize in terms of

$$q_i = \frac{z_i - z_{i-1}}{z_{i+1} - z_{i-1}}, \text{ where } 0 < q_i < 1,$$

and then apply the logit transformation.

4 Assessing the performance of the approach on simulated data

Before applying our approach to the malaria data we examine its performance on simulated data sets where the true value of λ and the degree of overlap between the two component distributions are known. The sizes of the data sets generated are chosen to reflect typical sample sizes for the real application.

4.1 Data simulated from a mixture of Poissons

A set of 100 observations was generated from each of 5 mixtures of a Poisson(2) and a Poisson(6) distribution, with mixing parameters $\lambda = \{0.05, 0.20, 0.50, 0.80, 0.95\}$, and a training sample of 100 observations was obtained from a Poisson(2) distribution. The Gibbs sampling approach was applied to each of the mixtures. The number of classes is defined by the number of discrete values from the mixture of the two Poissons. Ergodic averages of λ together with their standard errors are given in Table 2. For each value of λ , four different data sets are simulated. The maximum likelihood estimates obtained for a mixture of the Poisson distributions are given for comparative purposes. The estimates of the standard errors were calculated numerically, using the inverse of the information matrix. The results show that the Gibbs sampling approach gives estimates of λ with comparable precision to that of the parameter model found by Maximum Likelihood.

Similarly, data sets were generated from a mixture of a Poisson(4) and a Poisson(6) distribution and the above approach applied. Here the two mixing distributions have a high degree of overlap and maximum likelihood fails, for most values of λ , to give an estimate of λ close to the true value. However, our approach gives estimates close to the value of λ which generated the data (Table 3) and within valid ranges.

TABLE 2: Estimates of λ using the Gibbs sampling and the maximum likelihood approaches for data generated under a mixture of two Poisson distributions with means 2 and 6.

Value of λ	Gibbs sampling		Maximum Likelihood	
	Ergodic mean	Standard error	Estimate	Standard error
0.05	0.0614	0.0618	0.0376	0.1398
0.05	0.0817	0.0856	0.0769	0.1459
0.05	0.0624	0.0657	0.1041	0.1656
0.05	0.0577	0.0617	0.1636	0.2437
0.20	0.1845	0.1365	0.5159	0.2428
0.20	0.1451	0.0928	0.1559	0.1349
0.20	0.2330	0.1803	0.2605	0.1528
0.20	0.2255	0.1801	0.4507	0.3509
0.50	0.4194	0.1351	0.1777	0.1778
0.50	0.4919	0.1462	0.5488	0.1019
0.50	0.4702	0.1461	0.6299	0.1569
0.50	0.4732	0.1898	0.4135	0.1192
0.80	0.7861	0.1164	0.8576	0.1300
0.80	0.7459	0.1000	0.5675	0.0793
0.80	0.7164	0.1278	0.8431	0.1154
0.80	0.7195	0.1222	0.8080	0.1201
0.95	0.9560	0.0398	0.9999	0.0046
0.95	0.8893	0.0769	0.9396	0.0588
0.95	0.9161	0.0822	0.9999	0.1474
0.95	0.8910	0.0751	0.8707	0.0738

TABLE 3: Estimates of λ using the Gibbs sampling and the maximum likelihood approaches for data generated under a mixture of two Poisson distributions with means 4 and 6.

Value of λ	Gibbs sampling		Maximum Likelihood	
	Ergodic mean	Standard error	Estimate	Standard error
0.05	0.0987	0.1231	0.9999	7.7087
0.05	0.0634	0.0718	0.7414	1.1372
0.05	0.0771	0.0924	0.9998	12.3452
0.05	0.0701	0.0806	0.9170	6.8504
0.20	0.2167	0.1796	0.2033	0.4712
0.20	0.2021	0.1766	0.1609	0.1835
0.20	0.1975	0.1687	0.0356	0.1388
0.20	0.1963	0.2075	0.12E-08	0.1478
0.50	0.4859	0.2436	0.7640	0.2518
0.50	0.2971	0.2524	0.3738	0.2689
0.50	0.5366	0.2417	0.4628	0.2006
0.50	0.3841	0.2848	0.8137	0.6657
0.80	0.7543	0.2136	0.9999	0.4101
0.80	0.7702	0.1988	0.7860	0.2061
0.80	0.7388	0.2125	0.9999	0.5205
0.80	0.7155	0.1953	0.6942	0.3196
0.95	0.7514	0.2162	0.9999	0.4203
0.95	0.7442	0.2308	0.7644	0.1816
0.95	0.7930	0.1581	0.9999	0.2916
0.95	0.6179	0.2663	0.7456	0.1097

TABLE 4: Estimates of the mixing proportion, λ using the Gibbs sampling approach for data generated from a mixture of two Uniform distributions; $g_1(\cdot) \equiv U(0, 4)$ and $g_2(\cdot) \equiv U(2, 6)$.¹

Value of λ	Ergodic mean	Standard error
0.05	0.1108	0.0904
0.05	0.0630	0.0619
0.05	0.0758	0.0662
0.05	0.0683	0.0687
0.20	0.1510	0.0965
0.20	0.2005	0.1230
0.20	0.2289	0.1243
0.20	0.2742	0.1299
0.50	0.4720	0.1188
0.50	0.4468	0.1394
0.50	0.5730	0.1159
0.50	0.5992	0.1170
0.80	0.8470	0.0871
0.80	0.9037	0.0603
0.80	0.7807	0.0869
0.80	0.8488	0.0652
0.95	0.9632	0.0361
0.95	0.9298	0.0455
0.95	0.9607	0.0359
0.95	0.9528	0.0445

¹Estimates are based on $K = 10$ categories. Standard errors were calculated from the outputs of a multiple-chain Gibbs-sampling algorithm using the formula: $m^{-1} \sum_{j=1}^m (\hat{\lambda}_j^2 - m\bar{\lambda}^2)/(m-1)$ (Schmeiser, 1990), where m is the number of replicate chains (here $m=30$), $\bar{\lambda}$ is the ergodic mean of λ and $\hat{\lambda}_j$ is the mean of all sampled values in chain j .

4.2 Data simulated from a mixture of continuous Uniform distributions

Next, we present the performance of the approach to continuous mixtures. To demonstrate the robustness of the approach, we have chosen to show results on a mixture of Uniforms. Table 4 gives the results when the data are generated from continuous uniform distributions on the intervals $[0, 4]$ and $[2, 6]$, for $g_1(\cdot)$ and $g_2(\cdot)$, respectively. The classes are defined by splitting the range of sample values from $g_1(\cdot)$ into $K - 1$ categories. The K th class consists of observations from $g_2(\cdot)$ outside the range of sample values from $g_1(\cdot)$. For this last group m_K is 0. This way of grouping ensures categories with enough observations to capture the degree of overlap and also a smooth increase in $\lambda_i = P(x \in P_2 \mid x \in \text{category } i)$. The estimates in Table 4 are based on $K = 10$ categories.

The effect of the choice of K on the estimate of λ was investigated empirically. Table 5 shows the estimates of λ when K is equal to 6, 8, 10 and when the data are generated from a mixture of the two Uniform distributions with $\lambda = 0.20$. Taking into account the standard error of $\hat{\lambda}$, it seems that the number of groups does not seriously bias the estimates of λ . We have chosen K to vary from 6 to 10 because there is no point in having few or many groups, since the discriminating power of the approach in classifying observations into the one or the other component of the mixture is reduced. Using a large number of categories will introduce many unnecessary parameters into the model, whilst using only a few categories may not be able to capture the degree of overlap between the two component distributions.

5 Application

We now apply the method of Section 2 to the clinical malaria data. We analyse a subset of the data corresponding to children aged between 6 and 9 months old and to two seasons; the wet season (January - June) during which malaria prevalence is high and the dry season

TABLE 5: Estimates of the mixing proportion, λ when $\lambda = 0.20$ for data generated from a mixture of two Uniform distributions; $g_1(\cdot) \equiv U(0, 4)$ and $g_2(\cdot) \equiv U(2, 6)$

No. of categories	Ergodic mean	Standard error ²
6	0.2008	0.1041
8	0.2359	0.1139
10	0.2529	0.1137

² Standard errors were calculated as described in Table 3.

(July - December) during which the mosquito population, and also malaria prevalence, decreases.

We group the data into 10 categories (Table 6). The first category includes children with zero parasitaemia. This assumes that the constraint, $\phi_1 = 0$ makes sense since it is expected that aparasitaemic children cannot have fevers due to malaria. The next $k - 2$ categories are taken by dividing the range of sample values from $g_1(\cdot)$ for $x > 0$ into $k - 2$ intervals. The last category includes all children with parasite levels x outside the range of sample values from $g_1(\cdot)$.

We ran the Gibbs sampler for 15000 iterations using a multiple-chain scheme of 30 replicate chains. Plots of ergodic averages of the parameters obtained against iteration number, showed that 15 000 iterations were more than enough for convergence and it took less than two hours to run on a HP 712/80 workstation. Estimates of the posterior means and standard errors are calculated using ergodic first and second moments (Schmeiser, 1990) over the last 12 000 iterations. In particular, for the wet period $\hat{\lambda}_{\text{wet}} = 0.444(0.054)$, and for the dry season, $\hat{\lambda}_{\text{dry}} = 0.305(0.118)$. These estimates indicate that around 44% of fever cases during the wet season are attributable to malaria and only 30% during the dry season. Marginal posterior densities (Fig. 1) of λ during the two seasons are estimated using

TABLE 6: Distribution of malaria parasite densities in blood slides taken from children between 6 and 9 months old and during the wet and dry seasons. The parasite level refers to the midpoint of the category; m_i and n_i are as in Table 1.

Category	Wet season			Dry season		
	Parasite level	m_i	n_i	Parasite level	m_i	n_i
1	0	43	60	0	43	42
2	3251	40	58	11370	68	116
3	9673	3	14	34029	8	30
4	16095	3	13	56689	2	16
5	22518	2	10	79348	0	7
6	28940	1	8	102008	0	7
7	35362	0	7	124668	0	6
8	41785	1	6	147327	0	2
9	48207	1	6	169987	1	3
10	225685	0	69	290634	0	16
Total		94	251		122	245

the Rao-Blackwell density estimator (Smith and Roberts, 1993) and are based on samples $\left\{ \theta_1^{(s)}, \dots, \theta_{10}^{(s)}, \phi_1^{(s)}, \dots, \phi_{10}^{(s)}, \lambda^{(s)} \right\}_{s=1}^{600}$ drawn from the posterior density. To collect an independent sample of size 600 we increase the replicate chains to 600, at the end of the 15 000 iterations and run the algorithm for a further 100 iterations. Then, we select an independent sample from the final states of the chains.

Fig. 2 shows the probabilities that children with a certain level of parasite densities have clinical malaria during the wet and dry period. These probabilities are calculated as

$$\lambda_i = P(x \in P_2 \mid x \in \text{category } i) = \frac{\lambda \phi_i}{(1 - \lambda)\theta_i + \lambda \phi_i}.$$

The error bars represent ± 1 standard error of the estimate of the corresponding λ_i 's and are calculated from the final independent sample from the posterior.

To see the effect of different groupings on the estimates of λ we ran the algorithm for $K = 8$ and $K = 6$. Tables 7 and 8 show the distributions of malaria parasite densities when $K = 8$ and $K = 6$ respectively. Estimates of λ in Table 9 together with their standard errors verify the robustness of the approach to the number of groups. The variation in the estimates of posterior means during the dry season on K may be explained by the large standard error and the shape of the marginal $p(\lambda \mid \mathbf{m}, \mathbf{n})$ in Fig. 1.

We further compare these estimates with the ones obtained by applying the methods described in Smith, (1994). They discuss the usual estimate of population attributable fraction λ_{AF} given in (1) as well as logistic regression models which model fever risk as a continuous function (linear or polynomial) of parasite density. The estimates of λ are given in Table 10. Estimates of the standard error of λ_{AF} were obtained analytically (Bruzzi *et al.*, 1985) and estimates of the standard error of λ in the logistic models were obtained by using the bootstrap method and taking 1000 bootstrap samples. Table 10 shows that our approach gives comparable results with the current methods, without suffering from their disadvantages, such as negative estimates of probabilities and imprecise standard errors.

TABLE 7: Distribution of malaria parasite densities in blood slides taken from children between 6 and 9 months old during the wet and dry seasons. The parasite level refers to the midpoint of the category; m_i and n_i are as in Table 1

Category	Wet season			Dry season		
	Parasite level	m_i	n_i	Parasite level	m_i	n_i
1	0	43	60	0	43	42
2	4322	41	63	15164	72	128
3	12884	4	19	45359	5	32
4	21447	3	13	75572	1	9
5	30010	1	11	105785	0	8
6	38574	0	7	135997	0	6
7	47137	2	9	166210	1	4
8	225685	0	69	290634	0	16
Total		94	251		122	245

TABLE 8: Distribution of malaria parasite densities in blood slides taken from children between 6 and 9 months old during the wet and dry seasons. The parasite level refers to the midpoint of the category; m_i and n_i are as in Table 1

Category	Wet season			Dry season		
	Parasite level	m_i	n_i	Parasite level	m_i	n_i
1	0	43	60	0	43	42
2	6462	43	72	22699	76	146
3	19307	5	23	68019	2	23
4	32151	1	15	113338	0	13
5	44996	2	12	158657	1	5
6	225685	0	69	290634	0	16
Total		94	251		122	245

TABLE 9: Estimates of the mixing proportion λ for 6, 8 and 10 categories for the clinical malaria data and the wet and dry seasons.

Number of categories	Wet season		Dry season	
	Mean	Standard error	Mean	Standard error
6	0.452	0.062	0.412	0.111
8	0.448	0.058	0.359	0.119
10	0.444	0.054	0.305	0.118

TABLE 10: Estimates of λ using the earlier methods described in Smith (1994).

Method	Wet season		Dry season	
	Estimate	Standard error	Estimate	Standard error
λ_{AF}	0.477	0.106	0.513	0.144
Logistic model	0.490	0.045	0.365	0.058
Logistic power model	0.548	0.054	0.433	0.081

6 Discussion

In this paper we have successfully developed a model for decomposing nonparametrically a two-component mixture distribution. The underlying assumption is that the distribution of the one population is on the left of the other but there is a degree of overlap between them. Also, a set of outputs known to come from the one component is available. Results on simulated data verify the validity of the approach even when there is a high degree of overlap between the components and maximum likelihood estimation fails to converge or provides out of range estimates.

The type of mixtures considered arises frequently in biology and epidemiology. Established statistical methods, such as the classical estimate of the population attributable fraction (λ_{AF}) in (1), and the logistic model approach of Smith, (1994), are not able to provide accurate estimates. The estimate of λ_{AF} can be negative or imprecise. The main

disadvantages of the logistic models are that these models do not always fit well and that estimates of the standard errors of λ are not straightforward to obtain. Our approach does not suffer from these disadvantages. In addition it can provide estimates of the posterior distribution of λ and samples from the parameters which can be used to estimate other quantities of interest.

The implementation of the approach to malaria data showed that even though parasite densities among fever cases were higher in the dry season, reflecting differences in their exposure history and immunological status, the proportion of fevers attributable to malaria was indeed lower in the dry season than the wet season. Often covariates should be taken into account; for example the effects of age or season in the proportion of fevers attributable to malaria. These can be incorporated in an hierarchical structure model and are currently being investigated from the authors.

This methodology can be further applied to outputs of many biomedical assays which are expected to classify samples into two groups, according to whether some output (eg. parasite density, optical density) exceeds a given cutoff. Examples include many immunological assays, and diagnostic tests where a control group of negative samples is available.

The software for implementing the methodology is written in Fortran and can be obtained on request from the authors who are currently working on a user friendly version of the program.

Acknowledgements

The authors would like to thank Dr Eduardo Gutiérrez-Peña for very helpful discussions and the support and encouragement of Professor Marcel Tanner. Special thanks as well to A. Kitua for permission to analyse the data of the example. Penelope Vounatsou is supported by Swiss National Science Foundation grant 32-43527.95.

References

- Bruzzi, P., Green, S., Byar, D. P., Brinton, L. A., and Schairer, C. (1985). Estimating the population attributable risk for multiple risk factors using case-control data. *Am. J. Epid.*, **122**, 904-914.
- Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. B*, **56**, 363-375.
- Everitt, B. and Hand, D. (1981) *Finite Mixture Distributions*. New York: Chapman and Hall.
- Gelfand A., and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398-409.
- Greenwood, B. M. (1987) Asymptomatic malaria infections -Do they matter? *Parasitol. Today*, **3(7)**, 206-214.
- 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. Health*, **1 (4)**, 475-484.
- Murray, G. D. and Titterton, D. M. (1978) Estimation problems with data from a mixture. *Appl. Statist.*, **27**, 325-334.
- Robert, C. P. (1996) Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 441-464. London: Chapman and Hall.
- Schmeiser, B. (1990) Simulation experiments. In *Handbooks in Operations Research and*

Management Science, Vol 2: Stochastic models (eds. D.P. Heyman and M. J. Sobel). Amsterdam: North-Holland, 295-330.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc. B*, **55**, 3-23.

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.

Smith, T. and Vounatsou P. (1997) Logistic regression and latent class models for estimating positivities in diagnostic assays with poor resolution. *Commun. Statist.*, **26**, 1677-1700.

Titterton, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical analysis of finite mixture distributions*. Chichester: Wiley.

Wakefield, J. C., Gelfand, A. E. and Smith, A. F. M. (1991) Efficient Computation of random variates via the ratio-of-uniforms method. *Statist. Comp.*, **1**, 129-133.

Wu, C. F. (1983) On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95-103.

List of Figures

Fig. 1 Marginal posterior density of $p(\lambda \mid \text{data})$ for the malaria data.

Fig. 2 Probabilities (with error bars of ± 1 standard errors) of malaria attributed fevers in the 10 groups of parasite densities.