

# **Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models**

Tom Smith & Penelope Vounatsou

*Swiss Tropical Institute, Socinstrasse 57, Postfach CH-4002, Basel, Switzerland.*

## **Summary**

A Bayesian hierarchical model is proposed for estimating parasitic infection dynamics for highly polymorphic parasites when detectability of the parasite using standard tests is imperfect. The parasite dynamics are modelled as a nonhomogeneous hidden two-state Markov process, where the observed process is the detection or failure to detect a parasitic genotype. This is assumed to be conditionally independent given the hidden process, i.e. the underlying “true” presence of the parasite, which evolves according to a first order Markov chain. The model allows the transition probabilities of the hidden states as well as the detectability parameter of the test to depend on a number of covariates.

## **1. Introduction**

The dynamics of parasitic infections, that is the rate by which the infections are acquired and lost is an important aspect of any disease control program. It is often convenient to estimate infection and recovery rates from repeated observations of the presence or absence of the infection in the same group of individuals (panel data). However, such studies are often complicated by the fact that laboratory tests used to detect infectious agents often have imperfect sensitivity, especially for light infections. Statistical modelling should take into account the occurrence of false negatives, otherwise naïve estimates will provide misleading information on the transition dynamics of the infection [1].

Markov models have been used to describe the dynamics of malaria infections in both discrete and continuous time frameworks. Bekessy *et al*[2] used a first-order Markov model (corresponding to the Reversible Catalytic model introduced into epidemiology by Muench[3]) to study the infection dynamics of malaria in Garki, Nigeria. The presence or absence of infection defined the states and transition probabilities were estimated graphically. Such models can now easily be fitted by maximum likelihood to give estimates of infection and recovery probabilities, but the pattern of transitions in the Garki data does not correspond to a homogeneous Markov process[4]. Moreover, the model ignores the fact that detectability of malaria parasites is imperfect. Multi-state Markov models have subsequently been used to allow for undetected parasites[5], but only limited relevant data were available because no laboratory techniques were then available for distinguishing parasites derived from different infection events.

Nagelkerke *et al* [6] adapted Bekessy’s model to study the dynamics of *Giardia lamblia* infection among children in Kenya when detectability is not 100%. They estimated simultaneously the transition probabilities and the sensitivity of the test by

numerically maximising the partial likelihood augmented by the ‘true’ status of the infection and obtained approximate standard errors based on asymptotic arguments. This model assumes that the true parasite dynamics correspond to a Markov chain, but that this is hidden as a result of the imperfect detection. Such hidden Markov Models (HMMs) have been applied in fields varying from automatic speech recognition[7], the analysis of DNA sequences[8] in econometrics [9] and in other areas of epidemiology[10;11] [12].

HMMs provide a natural approach for estimating infection and recovery rates of malaria parasites. However malaria transmission is generally extremely focal [13] and hence infection rates are expected to vary by individual. This focality would lead to severe bias in estimates of the ratio of recrudescences : new infections in an HMM model fitted to only presence/absence data. In principle, better estimates of the dynamics of acquisition and loss can be obtained from analyses of sub-populations (types), identified via molecular or serological techniques.

We now present a HMM in which the observed process is the detection or failure to detect a parasite genotype. This is assumed to be conditionally independent given the hidden process. This is the underlying true presence of that genotype, which is assumed to evolve according to a first order Markov chain. In a mixed population of a polymorphic pathogen, separate HMMs cannot be fitted for each type since there is usually substantial variation in type-specific prevalence, so data for rare parasite types are sparse and the corresponding estimates would be imprecise or non-identifiable. To overcome this difficulty we incorporate data for multiple parasite types in the same model, using random effects terms to allow for variation between types in incidence, but assuming common duration of infection and detectibilities. To estimate such a model we use Markov Chain Monte Carlo (MCMC) simulation for full Bayesian inference. We extend previous work on simulation-based Bayesian inference for HMMs[14] by allowing covariate dependence of both the transition probabilities and detectability parameter.

Existing approaches for modelling dynamics of infection under perfect and imperfect detectibility are given in Section 2. The formulation of the problem as a hidden Markov model and computational details are given in Section 3.

## 2. Existing approaches in modelling infection dynamics

Our aim is to develop a model that explicitly incorporates the infection and recovery processes as well as the detectability parameter. Consider  $n$  individuals, each observed at  $T_i$  times and let  $X_{i,t}$  be a binary response variable indicating whether an infection was observed ( $X_{i,t} = 1$ ) or not ( $X_{i,t} = 0$ ) and  $\xi_{it}$  be the true underlying state taking the values 1 or 0 depending on the presence or absence of the parasite in individual  $i$  and at time  $t$ .

### 2.1 Perfect detectability

Assuming that the test has perfect specificity, and detectability  $\nu$ , that is  $p(X_{i,t}^{(k)} = 0 | \xi_{i,t} = 0) = 1$  and  $p(X_{i,t} = 1 | \xi_{i,t} = 1) = \nu$ , [2] assume perfect detectability and model parasite infection dynamics using a first-order two-state Markov model with transition probabilities  $p_{0,1}$  and  $p_{1,0}$  estimated by the model:

$$\frac{dp}{dt} = -\mu p + \lambda(1 - p),$$

where  $p$  is the point prevalence (corresponding to the proportion of the population in the infected compartment);  $\lambda$  is the force of infection and  $\mu$  is the recovery rate. Solving (1), we obtain:

$$p(t) = \frac{\lambda}{\lambda + \mu} + c \exp[-(\lambda + \mu)t]$$

where  $c$  is a constant depending on the initial conditions.

Assuming a discrete time Markov model and that at  $t=0$  the infection was not present, that is,  $p_0 = 0$ , we can estimate  $c$  and therefore the infection probability for intervals of duration  $t$ ,  $p_{0,1}(t)$ ;

$$p_{0,1}(t) = \frac{\lambda}{\lambda + \mu} (1 - \exp(-(\lambda + \mu)t)),$$

Similarly, assuming that at  $t=0$ , the infection was present, that is  $p_0 = 1$ , we can calculate  $c$  and therefore the recovery probability,  $p_{1,0}(t)$ :

$$p_{1,0}(t) = \frac{\mu}{\lambda + \mu} (1 - \exp(-(\lambda + \mu)t))$$

Estimates of  $\lambda$  and  $\mu$  can be obtained from these equations via maximum likelihood since the transitions represent i.i.d. Bernoulli trials with probabilities  $p_{0,1}(t)$  and  $p_{1,0}(t)$  respectively. Indeed, the likelihood can be written as:

$$\left\{ \prod_{i=1}^n Pr(X_{i,0}) \right\} \prod_{i=1}^n \prod_{t=1}^{T_i} Pr(X_{i,t} | X_{i,t-1})$$

The second part of the likelihood, which in fact is the partial likelihood obtained by conditioning on the first measurement  $X_{i,0}$ , is the same as that of a set of two Binomial distributions, that is,

$$\prod_{i=1}^n \prod_{t=1}^{T_i} Pr(X_{i,t} | X_{i,t-1}) \propto p_{0,1}^{N_{0,1}} (1 - p_{0,1})^{N_{0,0}} p_{1,0}^{N_{1,0}} (1 - p_{1,0})^{N_{1,1}}$$

where  $N_{j,l}$  are the total number of transitions from state  $l$  to state  $j$ , and therefore explicit maximization is possible. Nagelkerke *et al.* [6] suggest estimation of the transition probabilities by maximizing this partial likelihood instead of the full likelihood, since the first measurement  $X_{i,0}$  contributes a limited amount of information only if some steady-state assumptions are made. The maximum likelihood estimates obtained this way are:

$$\tilde{p}_{0,1} = \frac{N_{0,1}}{N_{0,1} + N_{0,0}} \text{ and } \tilde{p}_{1,0} = \frac{N_{1,0}}{N_{1,0} + N_{1,1}}$$

By replacing the above estimates of the transition probabilities in (3) and (4) we obtain estimators of the transition rates  $\lambda$  and  $\mu$ .

## 2.2 Imperfect detectability

Nagelkerke *et al.* ([6]) extend the Bekessy *et al* [2] model to the case where the detectability of the test,  $\nu$  (in our notation,  $\pi_l$ ) is less than 1. They augment the partial likelihood by the ‘true’ status of the infection and obtain maximum likelihood estimates numerically. The likelihood in this case is more complicated because the Markov property for the observed states does not hold. It can be written as

$$\prod_{i=1}^n Pr(X_{i,0}, \dots, X_{i,T_i}) = \left\{ \prod_{i=1}^n Pr(X_{i,0}) \right\} \prod_{i=1}^n \prod_{t=1}^{T_i} Pr(X_{i,t} | X_i^{t-1})$$

where  $X_i^{t-1} = (X_{i,0}, \dots, X_{i,t-1})$ . By the law of total probability,

$$\begin{aligned} Pr(X_{i,0} = 1) &= Pr(X_{i,0} = 1 | \xi_{i,0} = 1) Pr(\xi_{i,0} = 1) + Pr(X_{i,0} = 1 | \xi_{i,0} = 0) Pr(\xi_{i,0} = 0) \\ &= \pi_l Pr(\xi_{i,0} = 1) \end{aligned}$$

Also,

$$Pr(X_{i,t} = 1 | X_i^{t-1}) = \pi_l \left\{ (1 - p_{1,0}) \rho_{i,t-1} + p_{0,1} \rho_{i,t-1} \right\}$$

where  $\rho_{i,t} = P(\xi_{i,t} = 1 | X_{i,t}, X_i^{t-1})$ . When  $X_{i,t} = 1$ , the  $\rho_{i,t} = 1, \forall t$  and when  $X_{i,t} = 0$ ,  $\rho_{i,t}$  can be expressed by the following recursive formula,

$$\rho_{i,t} = \begin{cases} \frac{(1-\pi_1)[(1-p_{0,1}-p_{1,0})\rho_{i,t-1}+p_{0,1}]}{1-\pi_1[(1-p_{0,1}-p_{1,0})\rho_{i,t-1}+p_{0,1}]}, & t > 0 \\ \frac{(1-\pi_1)p(\xi_{i,0}=1)}{(1-\pi_1)p(\xi_{i,0}=1)+p(\xi_{i,0}=0)}, & t = 0 \end{cases}$$

[6] provide maximum likelihood inference, by maximising the partial and full likelihood numerically and obtain approximate standard errors for  $v$ ,  $p_{1,0}$ , and  $p_{0,1}$  and via the Jacobian for  $\lambda$  and  $\mu$  based on asymptotic arguments.

Although, the above approach provides estimates of the parameters of interest, it is intractable when these parameters depend on a number of covariates. In the next section we formulate the problem as a hidden Markov model and develop a Bayesian hierarchical model to estimate effects of covariates on  $\lambda$ ,  $\mu$  and  $\pi_1$ .

### 3. A hidden Markov model for infection dynamics with imperfect detectability

#### 3.1 Likelihood computations

We assume that given the true state  $\xi_{i,t} = r$  the observations  $X_{it}$ , are conditionally independent of  $\xi_{i,t'}, t \neq t'$  and arise from a Bernoulli distribution, that is  $P(X_{i,t} = 1 | \xi_{i,t} = r) \equiv \text{Bernoulli}(\pi_{i,r}(t))$ , where  $r = 1, 0$  or more generally  $P(X_{i,t} | \xi_{i,t}) \equiv \text{Bernoulli}(\pi_i(t))$  where  $\pi_i(t) = \sum_{r=1,0} \pi_{i,r}(t) I_{i,r}(t)$  and  $I_{i,r}(t)$  is the indicator of the event  $\{\xi_{i,t} = r\}$ . The hidden process, due to the fact that the underlying ‘‘true’’ presence of the parasite  $\{\xi_{i,t}\}$  is not observed, is assumed to evolve according to a first order two-state Markov chain, with transition probabilities,  $p_{1,0}^{(i)} = P(\xi_{i,t} = 0 | \xi_{i,t-1} = 1)$ ,  $p_{0,0}^{(i)} = 1 - p_{1,0}^{(i)}$ ,  $p_{01}^{(i)} = P(\xi_{i,t} = 1 | \xi_{i,t-1} = 0)$ ,  $p_{11}^{(i)} = 1 - p_{01}^{(i)}$  and initial probability distribution,  $P(\xi_{i,1} = s_1) = q_{s_1}^{(i)}$ .

Let  $\tilde{\theta}$  be the parameter vector (transition probabilities and probabilities of the Bernoulli distribution) to be estimated. The likelihood function for  $\tilde{\theta}$  given the data, is defined by

$$\begin{aligned} L(Y; \tilde{\theta}) &= \prod_{i=1}^n P(X_{i,1} = r_1, X_{i,2} = r_2, \dots, X_{i,T} = r_T) \\ &= \prod_{i=1}^n \prod_{t=1}^T P(X_{i,t} = r_t | \xi_{i,t} = s_t) \left[ \prod_{t=2}^T P(\xi_{i,t} = s_t | \xi_{i,t-1} = s_{t-1}) \right] P(\xi_{i,1} = s_1) \\ &\propto \prod_{i=1}^n \prod_{t=1}^T \pi_{i,s_t}^{r_t}(t) (1 - \pi_{i,s_t}(t))^{1-r_t} \prod_{s_{t-1}, s_t} p_{s_{t-1}, s_t}^{(i)} q_{s_1}^{(i)} \end{aligned}$$

Estimation and maximisation of the above likelihood is intractable especially when the transition probabilities are covariate-dependent [15] developed an algorithm (later

shown to be equivalent to the EM algorithm [16] to obtain maximum likelihood estimates for similar models, by considering the hidden “true” states to be missing data. The EM algorithm has also been applied to non-homogeneous Markov models [17]. Another approach for maximising such likelihoods is the forward-backward recursive algorithm (e.g. [7]). Bayesian estimation of hidden Markov chains using Markov Chain Monte Carlo (MCMC) simulation has been proposed by [14;18]. We have extended this by allowing the transition probabilities to be covariate-dependent and have used Winbugs to implement MCMC simulation for full Bayesian inference.

### 3.2 Bayesian hierarchical model

Our model has the following structure,

$$\xi_i^{(k)} \square \Pr(\xi_i^{(k)} | p_i, q_i) \propto \left[ \prod_{t=2}^T P(\xi_{i,t}^{(k)} = s_t | \xi_{i,t-1}^{(k)} = s_{t-1}) \right] P(\xi_{i,1}^{(k)} = s_1) = \prod p_{s_{t-1}, s_t}^{(ik)} q_{s_{t,1}}^{(k)}$$

where k denotes the parasite genotype. We allow the detectability parameter, the infection and recovery rates to be covariate dependent by introducing the following parameterizations into the model specification:

$$\text{logit}(\pi_{i,1}^{(k)}(t)) = \sum_{p=0} \beta_{1,p} Z_{i,t}, \quad \text{logit}(P(\xi_{i,t}^{(k)} = 1 | \xi_{i,t-1}^{(k)} = 0)) = \beta_{2,0}^{(k)} + \sum_{c=1} \beta_{2,c} Z_{i,t},$$

$\text{logit}(P(\xi_{i,t}^{(k)} = 0 | \xi_{i,t-1}^{(k)} = 1)) = \sum_{l=0} \beta_{3,l} Z_{i,t}$  and  $\pi_{i,0}(t) = 0$  as we assume the specificity of

the PCR technique to be perfect. This defines a non-homogeneous Markov model since it allows the transition probabilities of the hidden states to depend on a set of observed covariates (such as age). Genotype dependence arises only in the overall mean term for the infection probabilities.

Within the Bayesian hierarchical framework we need to specify prior distributions for the initial state probabilities,  $q^{(k)}$ , and the regression coefficients  $\beta$ , including the random effects  $\beta_{2,0}^{(k)}$ . We can adopt non-informative normal priors for the regression coefficients, Uniform(0,1) priors for  $q^{(k)}$  and assume that  $\beta_{2,0}^{(k)}$  are i.i.d Normally distributed, that is,  $\beta_{2,0}^{(k)} \square N(\alpha, \sigma^2), k = 1, \dots, K$ , where the hyperparameters,  $\alpha$  and  $\sigma^2$ , have vague normal and inverse gamma priors respectively.

## Reference List

1. Walter, S.D. and Irwig, L.M. (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J.Clin.Epidemiol.*, **41**, 923-937.
2. Bekessy A, Molineaux L, and Storey J (1976) Estimation of incidence and recovery rates of *Plasmodium falciparum* parasitaemia from longitudinal data. *Bulletin WHO*, **54**, 685-691.
3. Muench H (1959) *Catalytic models in epidemiology*. Harvard University Press, Cambridge, Mass.

4. Cohen JE and Singer B (1979) Malaria in Nigeria: constrained continuous-time Markov models for discrete-time longitudinal data on human mixed-species infections. *Lectures on Mathematics in the Life Sciences*, **12**, 69-133.
5. Nedelman, J. (1984) Inoculation and recovery rates in the malaria model of Dietz, Molineaux and Thomas. *Mathematical Biosciences*, **69**, 209-233.
6. Nagelkerke, N.J., Chunge, R.N., and Kinoti, S.N. (1990) Estimation of parasitic infection dynamics when detectability is imperfect. *Stat. Med.*, **9**, 1211-1219.
7. Juang, B.H. and Rabiner, L.R. (1991) Hidden Markov models for speech recognition. *Technometrics*, **33**, 251-272.
8. Churchill GA and Lazareva B (1999) Bayesian restoration of a hidden Markov chain with applications to DNA sequencing. *J Comput. Biol.*, **6**, 261-277.
9. Chib S (1996) Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75**, 79-97.
10. MacDonald, I.L. and Lerer, L.B. (1994) A time-series analysis of trends in firearm-related homicide and suicide [published erratum appears in *Int J Epidemiol* 1994 Oct;23(5):1108]. *Int. J. Epidemiol.*, **23**, 66-72.
11. Le Strat, Y. and Carrat, F. (1999) Monitoring epidemiologic surveillance data using hidden Markov models. *Stat. Med.*, **18**, 3463-3478.
12. Guihenneuc-Jouyaux, C., Richardson, S., and Longini, I.M. (2000) Modelling markers of disease progression by a hidden Markov process: application to characterising CD4 cell decline. *Biometrics*.
13. Woolhouse, M.E., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., Hagan, P., Hii, J.L., Ndhlovu, P.D., Quinnell, R.J., Watts, C.H., Chandiwana, S.K., and Anderson, R.M. (1997) Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 338-342.
14. Robert, C.P., Celeux, G., and Diebolt, J. (1993) Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics and Probability Letters*, **16**, 77-83.
15. Baum LE, Petrie T, Soules G, and Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **37**, 1554-1563.
16. Dempster AN, Laird NM, and Rubin D (1977) Maximum Likelihood from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Association, Series B*, **39**, 1-38.
17. Hughes, J.P., Guttorp, P., and Charles, S.P. (1999) A Nonhomogeneous hidden Markov model for precipitation. *Applied Statistics*, **48**, 15-30.

18. Robert, C.P., Ryden, T., and Titterton, D.M. (2000) Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Applied Statistics*, 57-75.