The analysis of disease biomarker data using mixed hidden Markov model.

Johann C. Detilleux

Quantitative Genetics Group, Department of Animal Production, Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

Corresponding author:

e-mail: jdetilleux@ulg.ac.be

Tel: 32 4 366 4215

Fax: 32 4 366 4122

Short title : Mixed hidden Markov model

Abstract - A mixed hidden Markov model (HMM) is developed for predicting breeding values of a biomarker (here, somatic cell score) and the individual probabilities of health and disease (here, mastitis) based upon measurements of the biomarker. At a first level, the unobserved disease process (Markov model) is introduced and, at a second level, the measurement process is modelled, making the link between the unobserved disease states and the observed biomarker values. This hierarchical formulation allows joint estimation of the parameters of both processes. The flexibility of this approach is illustrated on simulated data. Firstly, lactation curves for the biomarker are generated based upon published parameters (mean, variance and probabilities of infection) for cows with known clinical condition (health or mastitis due to *E. coli* or *S. aureus*). Next, estimation of the parameters is performed via Gibbs sampling, assuming the health status is unknown. Results from the simulations and the mathematics showed that the mixed HMM is appropriate to estimate the quantities of interest although the accuracy of the estimates is moderate when the prevalence of the disease is low. The paper ends with some indications for further developments of the methodology.

hidden Markov model/ mixed model/ mastitis/ somatic cell score

1. INTRODUCTION

Studies have shown variability among cows for natural resistance to intra-mammary infection (IMI). Selection is therefore possible but direct measures of IMI are not readily available. Usually, information on IMI is based upon biomarkers such as somatic cell scores (SCS), electrical conductivity, immunoglobulin or acute phase proteins (reviewed in Detilleux, accepted). One important difficulty in using these biomarkers to find the most resistant animals is that factors known to influence their expression may be different in healthy (IMI-) from infected (IMI+) cows. As these are usually unidentified, breeding values tend to be biased. To reduce this bias and to infer more precisely cows' individual probabilities to be IMI- or IMI+, several authors have used the mixture model methodology on SCS [2], [8], [12], [17]. A generalization of the mixture model is the hidden Markov model (HMM) that presents the advantages to not only estimate individual probabilities of being infected but also to predict individual probabilities of new infection and of recovery, both useful to compute epidemiological measures of IMI spread within a population and to assist mastitis control programs.

The objective of this study is to present the mathematical formalism behind the HMM methodology as it may apply to the analysis of infectious disease biomarkers assumed to be dependent upon the genetic make-up of the cows. The fit of the HMM will be assessed on data simulated based on parameters obtained in a survey of clinical mastitis cases. Bayesian estimates of the parameters will be obtained using the Gibbs sampler. Finally, limitations and possible extensions of the current approach will be discussed.

2. MATERIALS AND METHODS

Throughout, k indexes the individual cow, t (t = 1 to T) is the follow-up time point during the lactation (e.g., month in milk), y_k^t is the value of the biomarker observed at t on animal k, and z_k^t is the corresponding unknown health status (IMI- or IMI+). Let $z_k^t = 0$ if y_k^t is from an unknown IMI- sample and $z_k^t = 1$ if y_k^t is from an unknown IMI+ sample. For simplicity, T is assumed constant for all cows. We use the notation Ødegård *et al.* [17] utilized for their finite mixture model, with slight modifications.

2. 1. General formulation of the model

Conditionally on the unknown vector **z**, it was assumed that the vector of observations **y** could be described by the linear model:

$$\mathbf{y} = \mathbf{M_0}\ \boldsymbol{\mu_0} + \mathbf{M_1}\ \boldsymbol{\mu_1} + \mathbf{Z\ a} + \mathbf{e}$$

where **y** is the (NT X 1) data vector of y_k^t, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are (T X 1) vectors of fixed effects for data on a IMI- or IMI+ cow, respectively, **a** is the (Na X 1) vector of random additive genetic effects; $\mathbf{M}_0$ is the (NT X T) matrix with elements = 1 if $z_k^t = 0$ and = 0 otherwise; $\mathbf{M}_1$ is the (NT X T) matrix with elements = 1 if $z_k^t = 1$ and = 0 otherwise; **e** is the (NT X 1) vector of residuals; **Z** is the (NT X Na) incidence matrix relating **a** to **y**, N is the number of animals with data and Na is the number of animals with pedigree records.

The conditional distribution of **y,** given the vector **z**, the location and scale parameters, was assumed to be:

$$(\mathbf{y}|\, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \sigma_0^2, \sigma_1^2, \mathbf{a}, \mathbf{z}) \sim \mathrm{N}\,[(\mathbf{M}_0\, \boldsymbol{\mu}_0 + \mathbf{M}_1\, \boldsymbol{\mu}_1 + \mathbf{Z}\, \mathbf{a}), \mathbf{R}]$$

with $\mathbf{R} = \mathbf{F}_0 \sigma_0^2 + \mathbf{F}_1 \sigma_1^2$, where $\mathbf{F}_\mathbf{i}$ is the (NT X NT) diagonal matrix with elements = 1 if $z_k^t = i$ and = 0 otherwise. The parameters σ_0^2 and σ_1^2 are the residual variances associated to a record on an IMI- and IMI+ cow, respectively. For the additive effects, it was assumed that $(\mathbf{a}\, |\sigma_a^2) \sim \mathrm{N}\ [0, \mathbf{A}\sigma_a^2]$ where σ_a^2 is the additive genetic variance and **A** is the matrix of additive genetic relationship between animals.

2. 2. Sampling distribution of the observations given group status

The density of the vector **y** for the subset of the N_i observations with z_k^t = i, i.e. {**z** = i}, given the location parameters and the residual variances can be written as:

pr ($\mathbf{y}|\,\boldsymbol{\mu}_\mathbf{i}, \sigma_i^2$, {**z** = i})

$$\propto (\sigma_i^2)^{N_i/2} \exp\{(\frac{-1}{2\sigma_i^2})\ (\mathbf{y} - \mathbf{M_i}\ \boldsymbol{\mu}_\mathbf{i} - \mathbf{Z}\ \mathbf{a})'\ \mathbf{F_i}\,(\mathbf{y} - \mathbf{M_i}\ \boldsymbol{\mu}_\mathbf{i} - \mathbf{Z}\ \mathbf{a})\}.$$

2.3. Prior distributions of parameters and of the unknown status vector

For i = 0 or 1, normal prior densities were assumed for the location parameters:

$$\mathrm{pr}(\mu_i) \propto (s_i^2)^{-T/2} \exp\ \{(-\frac{1}{2s_i^2})(\mu_i - \mathbf{1}\ m_i)'\ (\mu_i - \mathbf{1}\ m_i)\},$$

where **1** is the (TX1) vector of 1. The prior density for the additive effects, conditionally on the additive variance, was:

$$\mathrm{pr}(\mathbf{a}\,|\sigma_a^2) \propto (\sigma_a^2)^{-N/2} \exp\{(-\frac{1}{2\sigma_a^2})\ \mathbf{a}'\ \mathbf{A}^{-1}\ \mathbf{a}\ \}.$$

Under simple mixture models, the individual elements of the classification vector **z** are assumed to be independent a priori and to follow the same Bernoulli distribution with the mixing proportion as parameter. Here, under a equally simple mixed HMM, the variables z_k^t do not follow the same distribution. The first element of the series (z_k^1) follows a Bernoulli distribution with λ_k as parameter while the other elements follow Bernoulli distributions with state transition probabilities from z_k^{t-1} to z_k^t as parameters. Formally, the unknown state at time t may be decomposed in:

$$\mathrm{pr}(z_k^t = i) = p(z_k^t = i | z_k^{t-1} = 0)\, p(z_k^{t-1} = 0) + p(z_k^t = i | z_k^{t-1} = 1)\, p(z_k^{t-1} = 1)$$

where $p(z_k^t = i | z_k^{t-1} = j)$ are the state transition probabilities with i, j = 0 or 1. The state transition probabilities are assumed to possess the first-order Markov property namely that, given the present state, the future and past states are independent or that the current value (z_k^t) depends solely on the most recent past value (z_k^{t-1}). Transition probabilities are also independent of the actual time at which the transition takes place (stationarity assumption). Then, we have $\mathrm{pr}(z_k^t = i | z_k^{t-1} = j) = \pi_k^{ij}$ for all t and $(z_k^t = i | z_k^{t-1} = 0) \sim \mathrm{Ber}(\pi_k^{00})$ and $(z_k^t = i | z_k^{t-1} = 1) \sim \mathrm{Ber}(\pi_k^{01})$.

2. 4. Priors for variance components and probabilities

Scale-inverse chi-square distributions with ν degrees of freedom and scale parameters (s^2_a, s^2_0, and s^2_1) were used for the variance components:

$$\mathrm{pr}(\sigma_a^2) \propto (\sigma_a^2)^{-(\nu+2)/2} \exp(-\nu\, s_a^2/2\,\sigma_a^2),$$

$$\mathrm{pr}(\sigma_0^2) \propto (\sigma_0^2)^{-(\nu+2)/2} \exp(-\nu\, s_0^2/2\sigma_0^2), \text{ and}$$

$$\mathrm{pr}(\sigma_1^2) \propto (\sigma_1^2)^{-(\nu+2)/2} \exp(-\,\nu\, s_1^2/2\,\sigma_1^2)\,.$$

Finally, λ_k, π_k^{00} and π_k^{01} were assigned uniform (i.e. Beta(1,1)) prior distributions.

2. 5. Joint posterior distributions

For all cows, the joint posterior density of all unknown parameters is given by:

$$\mathrm{pr}(\mu_0, \mu_1, \sigma_a^2, \sigma_0^2, \sigma_1^2, \mathbf{z}, \mathbf{a}, \boldsymbol{\pi}^{\mathbf{00}}, \boldsymbol{\pi}^{\mathbf{01}}, \boldsymbol{\lambda}|\mathbf{y})$$

$$\propto \quad \mathrm{pr}(\mathbf{y}|\ \mu_0, \mu_1, \sigma_a^2, \sigma_0^2, \sigma_1^2, \mathbf{z}, \mathbf{a}, \boldsymbol{\pi}^{\mathbf{00}}, \boldsymbol{\pi}^{\mathbf{01}}, \boldsymbol{\lambda})$$

$$\mathrm{pr}\ (\mathbf{z}|\, \mu_0, \mu_1, \sigma_a^2, \sigma_0^2, \sigma_1^2, \mathbf{a}, \boldsymbol{\pi}^{\mathbf{00}}, \boldsymbol{\pi}^{\mathbf{01}}, \boldsymbol{\lambda})$$

$$\mathrm{pr}(\mathbf{a}|\, \mu_0, \mu_1, \sigma_a^2, \sigma_0^2, \sigma_1^2, \boldsymbol{\pi}^{\mathbf{00}}, \boldsymbol{\pi}^{\mathbf{01}}, \boldsymbol{\lambda})$$

$$\mathrm{pr}(\mu_0)\, \mathrm{pr}(\mu_1)\, \mathrm{pr}(\sigma_0^2)\, \mathrm{pr}(\sigma_1^2)\, \mathrm{pr}(\sigma_a^2)\ \ \mathrm{pr}(\boldsymbol{\pi}^{\mathbf{00}})\ \mathrm{pr}(\boldsymbol{\pi}^{\mathbf{01}})\ \mathrm{pr}(\ \boldsymbol{\lambda})$$

where $\boldsymbol{\lambda} = [\lambda_1, .., \lambda_N]'$, $\boldsymbol{\pi}^{\mathbf{00}} = [\pi_1^{00}, ..., \pi_N^{00}]$ and $\boldsymbol{\pi}^{\mathbf{01}} = [\pi_1^{01}, ..., \pi_N^{01}]\,'$.

Explicitly, the joint posterior is:

$$(\sigma_0^2)^{-(N_0+\nu+2)/2} \exp - \frac{1}{2\sigma_0^2}\{\nu\, s_0^2 + (\mathbf{y} - \mathbf{M_0}\ \boldsymbol{\mu_0} - \mathbf{Z\ a})'\ \mathbf{F_0}\,(\mathbf{y} - \mathbf{M_0}\ \boldsymbol{\mu_0} - \mathbf{Z\ a})\}$$

$$(\sigma_1^2)^{-(N_1+\nu+2)/2} \exp - \frac{1}{2\sigma_1^2}\{\nu\, s_1^2 + (\mathbf{y} - \mathbf{M_1}\ \boldsymbol{\mu_1} - \mathbf{Z\ a})'\ \mathbf{F_1}\,(\mathbf{y} - \mathbf{M_1}\ \boldsymbol{\mu_1} - \mathbf{Z\ a})\}$$

$$(s_0^2)^{-T/2} \exp \{(-\frac{1}{2s_0^2})(\boldsymbol{\mu_0} - \mathbf{1}\, m_0)' (\boldsymbol{\mu_0} - \mathbf{1}\, m_0)\}$$

$$(s_1^2)^{-T/2} \exp \{(-\frac{1}{2s_1^2})(\boldsymbol{\mu_1} - \mathbf{1}\, m_1)' (\boldsymbol{\mu_1} - \mathbf{1}\, m_1)\}$$

$$(\sigma_a^2)^{-(N+\nu+2)/2} \quad \exp - \frac{1}{2\sigma_a^2} \{\nu\, s_a^2 + \mathbf{a}' \, \mathbf{A}^{-1}\, \mathbf{a}\}$$

$$\prod_{k=1}^{N} (\lambda_k)^{K_k^{0,1}+1} \, (1-\lambda_k)^{K_k^{1,1}+1}$$
$$\prod_{k=1}^{N} (\pi_k^{00})^{n_k^{00}+1} \, (1 - \pi_k^{00})^{n_k^{10}+1} \;\; x \;\prod_{k=1}^{N} (\pi_k^{01})^{n_k^{01}+1} (1 - \pi_k^{01})^{n_k^{11}+1}$$

where $K_k^{i,1}$ is an indicator function which takes the value 1 if $z_k^1 = i$ and 0 otherwise and n_k^{ij} = number of transitions from $z_k^t = j$ to $z_k^{t+1} = i$.

2. 6. Fully conditional posterior distributions.

The conditional posterior distributions of each parameter (or block of parameters) are required for implementing a Gibbs sampler. Conditional on **y** and **z**, these conditional posterior densities are analytical because they only involve one of the possible realizations in the space of all possible sequences of **z**. For the location parameters, we have:

$$(\mu_i^t \mid \Theta, \mathbf{y}, \mathbf{z}) \sim \mathrm{N}\left(\frac{s_i^2 \sum_k^N (y_k^t - a_k)\, K_k^{i,t} + m_i\, \sigma_i^2}{(s_i^2 \sum_k^N \eta_k^{i,t}) + \sigma_i^2}, \frac{s_i^2\, \sigma_i^2}{(s_i^2 \sum_k^N \eta_k^{i,t}) + \sigma_i^2} \right),$$

where Θ refers to values of all parameters that the conditional distributions depend upon (i.e., all parameters except the one under consideration), $\eta_k^{i,t}$ is the number of cows with IMI- (i = 0) or IMI+ (i = 1) unknown state at the tth time.

Let $\mathbf{W} = [\mathbf{Z}\ \mathbf{M_0}\ \mathbf{M_1}]$ and the vector of parameters $\boldsymbol{\theta} = [\mathbf{a}\ \boldsymbol{\mu_0}\ \boldsymbol{\mu_1}]'$. Hence, one can write the model as: $\mathbf{y} = \mathbf{Z}\ \mathbf{a} + \mathbf{M_0}\ \boldsymbol{\mu_0} + \mathbf{M_1}\ \boldsymbol{\mu_1} + \mathbf{e} = \mathbf{W}\ \boldsymbol{\theta} + \mathbf{e}$. By partitioning the parameter vector $\boldsymbol{\theta}$ as $\boldsymbol{\theta_1} = \mathbf{a}$ and $\boldsymbol{\theta_2} = [\boldsymbol{\mu_0}\ \boldsymbol{\mu_1}]'$, we can compute the conditional posterior distribution of the vector of additive genetic values as $(\mathbf{a}|\Theta, \mathbf{y}, \mathbf{z}) \sim \mathrm{N}\ (\hat{\mathbf{a}}, \mathbf{C}_{11}^{-1})$ with $\hat{\mathbf{a}} = C_{11}^{-1}[r_1 - C_{12}\theta_2]$and $\mathbf{r}_1$, $\mathbf{C_{11}}$, $\mathbf{C_{12}}$ = the corresponding partition of $\mathbf{C} = [\mathbf{W'}\ \mathbf{R^{-1}}\ \mathbf{W} + \mathbf{A^{-1}}/\sigma_a^2]$ and $\mathbf{r} = \mathbf{W'}\ \mathbf{R^{-1}}\ \mathbf{y}$.

The fully conditional posterior density of the genetic variance is

$$\mathrm{pr}\,(\sigma_a^2 \mid \Theta, \mathbf{y}, \mathbf{z}) \propto (\sigma_a^2)^{-(N+\nu+2)/2}\ \ \exp - \frac{1}{2\sigma_a^2}\{\nu\, s_a^2 + \mathbf{a'}\ \mathbf{A^{-1}}\ \mathbf{a}\}$$

which is in the form of a scale-inverse chi-square density, with [N + ν] degrees of freedom and scale parameter $[\mathbf{a'}\ \mathbf{A^{-1}}\ \mathbf{a} + \nu\ s^2_a]$. Likewise, the fully conditional densities of the residual variances for IMI- and IMI+ observations are:

$$\mathrm{pr}\,(\sigma_i^2 \mid \Theta, \mathbf{y}, \mathbf{z})$$

$$\propto (\sigma_i^2)^{-(N_i+\nu+2)/2} \exp - \frac{1}{2\sigma_i^2}\{\nu\, s_i^2 + (\mathbf{y} - \mathbf{M_i}\ \boldsymbol{\mu}_\mathbf{i} - \mathbf{Z\ a})'\ \mathbf{F_i}\,(\mathbf{y} - \mathbf{M_i}\ \boldsymbol{\mu}_\mathbf{i} - \mathbf{Z\ a})\}$$

which are in the form of scale-inverse chi-square densities, with $[N_i + \nu]$ degrees of freedom, and with scale parameter = $\{\nu\, s_i^2 + (\mathbf{y} - \mathbf{M_i}\ \boldsymbol{\mu}_\mathbf{i} - \mathbf{Z\ a})'\ \mathbf{F_i}$ $(\mathbf{y} - \mathbf{M_i}\ \boldsymbol{\mu}_\mathbf{i} - \mathbf{Z\ a})\}$ for i = 0 and 1.

For the k^th cow, the fully conditional posterior densities of the parameters λ_k, π_k^{00} and π_k^{01} are:

$$\mathrm{pr}\ (\lambda_k \mid \Theta, \mathbf{y}, \mathbf{z})$$

$$\propto \lambda^{K_k^{0,1}+1}(1-\lambda)^{K_k^{1,1}+1}, \mathrm{pr}\,(\pi_k^{00} \mid \Theta) \propto (\pi_k^{00})^{n_k^{00}+1}(1-\pi_k^{00})^{n_k^{10}+1},$$

and $\mathrm{pr}\,(\pi_k^{01} \mid \Theta, \mathbf{y}, \mathbf{z}) \propto (\pi_k^{01})^{n_k^{01}+1}(1-\pi_k^{01})^{n_k^{11}+1}$ which are in the form of beta distributions.

Finally, one must compute the fully conditional distribution for each individual z_k^t. These may be obtained either from the $\mathrm{pr}(\mathbf{z}|\Theta, \mathbf{y})$ or by considering $\mathrm{pr}(z_k^t\, |\mathbf{z}(\text{-}z_k^t), \Theta, \mathbf{y})$ where $\mathbf{z}(\text{-}z_k^t)$ represent the hidden vector **z** without z_k^t, as suggested by one referee. Under the first alternative, pr($\mathbf{z}|\Theta$) can be decomposed as :

$$\mathrm{pr}(\mathbf{z}|\Theta, \mathbf{y}) = \mathrm{pr}(z_k^1|\Theta, \mathbf{y}) \prod_{t=2}^{T} \mathrm{pr}(z_k^t|z_k^{t-1}, \Theta, \mathbf{y}),$$

which leads to a stochastic version of the forward-backward algorithm in which z_k^1 is sampled from a Bernoulli distribution with parameter $\mathrm{pr}(z_k^1 = 0 \cap \mathbf{y})$ and each z_k^t is sampled successively (for t = 2 to T) from Bernoulli distributions with parameter $\xi_k^{ij,t} = \mathrm{pr}(z_k^t = i \mid z_k^{t-1} = j, \mathbf{y})$. The computations are reduced as components of $\xi_k^{ij,t} = \dfrac{\alpha_k^{j,t-1}\,\pi_k^{ij}\,b_k^{i,t}\,\beta_k^{i,t}}{\alpha_k^{j,t-1}\,\beta_k^{j,t-1}}$ may be stored gradually as t increases from 1 to T:

$$\begin{aligned}
&\alpha_k^{j,t} = \mathrm{pr}([y_k^1, y_k^2, \ldots y_k^t] \cap z_k^t = j),\\
&\beta_k^{i,t} = \mathrm{pr}([y_k^{t+1}, \ldots, y_k^T] \mid z_k^t = i),\\
&\pi_k^{ij} = \mathrm{pr}(z_k^t = i \mid z_k^{t-1} = j) \text{ and } b_k^{i,t} = \mathrm{pr}(y_k^t \mid z_k^t = i).
\end{aligned}$$

The forward and backward probabilities can be efficiently calculated by the following recursion formulae [10]:

$$\begin{aligned}
&\alpha_k^{j,t} = [\alpha_k^{0,t-1}\,\pi_k^{j0} + \alpha_k^{1,t-1}\,\pi_k^{j1}]\; b_k^{j,t}\\
&\beta_k^{i,t} = \left[\beta_k^{0,t+1}\,\pi_k^{0i}\;\; b_k^{0,t+1}\right] + \left[\beta_k^{1,t+1}\,\pi_k^{1i}\;\; b_k^{1,t+1}\right]
\end{aligned}$$

with initial conditions given by: $\alpha_k^{0,1} = \lambda_k \ b_k^{0,1}$, $\alpha_k^{1,1} = (1-\lambda_k) \ b_k^{1,1}$ and $\beta_k^{i,T} = 1$ for i = 0 and 1.

In the second alternative, pr(z_k^t |**z**(-z_k^t), Θ, **y**) is reduced to pr(z_k^t | z_k^{t-1}, z_k^{t+1}, Θ, **y**) because of the first-order Markov property on **z**. Then, pr(z_k^t = i| z_k^{t-1} = j, z_k^{t+1} = r, Θ ,**y**) $\propto$ $\mathrm{pr}(y_k^1 \mid z_k^1 = i)\,\mathrm{pr}(z_k^1 = i)$ if t = 1. It is proportional to $\mathrm{pr}(z_k^t = i \mid z_k^{t-1} = j)\,\mathrm{pr}(y_k^t \mid z_k^t = i, \Theta)\,\mathrm{pr}(z_k^{t+1} = r \mid z_k^t = i)$ for t = 2 to T-1 and to $\mathrm{pr}(y_k^T \mid z_k^T = i)\,\mathrm{pr}(z_k^T = i \mid z_k^{T-1} = j)$ if t = T. Note this alternative uses T different components while the first alternative generates a realization of **z** directly from its conditional p(**z**|**y**, Θ). It presents also a more complicated correlation structure (since each z_k^t depends on both z_k^{t-1} and z_k^{t+1}) than the first alternative which may lead to a slower mixing chain.

2. 7. Implementation of a Gibbs Sampler

The following steps describe how a Gibbs sampling can be implemented for our model, using the stochastic version of the forward-backward algorithm to sample **z**:

1. Set initial values for parameters as needed.

2. Select the block ($\boldsymbol{\theta}_1$) of the vector $\boldsymbol{\theta}$, compute $\tilde{\theta}_1 = C_{11}^{-1}[r_1 - C_{12}\theta_2]$ and replace **a** with $[\tilde{\theta}_1 + C_{11}^{-0.5}$ rannor(0)] where rannor(0) is a random draw from a standard normal distribution.

3. Replace μ_i (i = 0 and 1) with

$$\left[\frac{s_i^2 \sum_k^N (y_k^t - a_k) K_k^{1,t} + m_i \sigma_i^2}{(s_i^2 \sum_k^N n_{i,k}) + \sigma_i^2}\right] + \left[\left(\frac{s_i^2 \sigma_i^2}{(s_i^2 \sum_k^N n_{i,k}) + \sigma_i^2}\right)^{0.5} \text{rannor}(0)\right].$$

4. Replace σ^2_a with $(\mathbf{a}' \mathbf{A}^{-1} \mathbf{a} + \nu s^2_a)/\chi^2_{N+\nu}$, where $\chi^2_{N+\nu}$ is a random draw from a central chi-square distribution with [ν + *N*] degrees of freedom.

5. Replace σ^2_i with $\{\nu s_i^2 + (\mathbf{y} - \mathbf{M}_i \boldsymbol{\mu}_i - \mathbf{Z}\,\mathbf{a})' \mathbf{F}_i (\mathbf{y} - \mathbf{M}_i \boldsymbol{\mu}_i - \mathbf{Za})\}/\chi^2_{N_i+\nu}$ for i = 0 or 1, where $\chi^2_{N_i+\nu}$ is a random draw from a central χ-square distribution with [N_i+ν] degrees of freedom.

6. Compute $\zeta_k^{0,1} = \alpha_k^{0,1}\, \beta_k^{0,1} = \text{pr}(z_k^1 = 0 \cap \mathbf{y})$ and sample z_k^1 from Ber($\zeta_k^{0,1}$).

7. Compute and store $\zeta_k^{0j,t}$ for t = 2, ... , T and j = 0 or 1. Then, sample z_k^t from Ber($\zeta_k^{0j,t}$) if $z_k^{t-1} = j$ for t = 2, .., T.

8. Sample λ_k and π_k^{ij}, from their corresponding beta distributions with parameters $K_k^{i,1} + 1$ and $n_k^{ij} + 1$, for i ,j = 0 and 1, respectively.

9. Repeat (2 through 8) ρ times for burn-in as needed. Then, sample all parameters δ times. The total number of cycles is $\rho + \delta$.

In this study, values for the hyperparameters are: $s^2_0 = 0.5$, $s^2_1 = 1$, $m_0 =$ overall average computed from the data, $m_1 = m_0 + 3$, $\nu = 2$, $s^2_a = h^2 s^2_p$ (s^2_p = variance computed from the data) and $h^2 = 0.1$.

2. 8. Simulations

The model was evaluated using simulated values for the biomarker (here, SCS) with genetic effects considered as having the same distributions for cows with IMI+ and IMI- samples. Each simulation was replicated 10 times. Simulated rather than real data were used because a negative diagnosis, even based on absence of bacteria in cell culture, is not a guarantee of health and the reverse has also been observed [22].

2. 8. 1. Simulated data

Results from the field study of de Haas *et al.* [6], [7] on pathogen-specific SCC curves among multiparous cows were used to simulate the means of monthly samples from IMI- and IMI+ cows. In the Figure 3b of de Haas's paper [6], it is shown that, in cows clinically infected with *E. coli*, SCC increase rapidly after infection occurring around the 2nd month-in-milk, peak at 2,000 cells per μL above pre-infection values and return to pre-infection levels one month later. On the other hand, presence of a long increased SCC, without recovery within 4 consecutive months, was common in lactations with clinical *Staph. aureus* mastitis. In the cows without clinical mastitis, SCC followed an approximate inverse lactation curve. The SCC values were $\log_2$-transformed in SCS and used to simulate the SCS means, as explained below. In the simulations, it was also considered that cows may be classified as high and moderate responders on the basis of the extent of their immune response to a particular infection [14]. Therefore, SCS were considered at higher values and of longer duration in high than moderate responders (Figure 1).

In the simulations, 3 discrete generations were considered with 400 cows per generation. No selection was applied, sires were selected from 30 different bulls, each cow was replaced by a daughter, and mating was at random. Breeding values for base animals were sampled from a normal distribution with null mean and additive variance of 0.15 or 0.25. The values for the additive variance was found in the literature [4]. Breeding values for non-base animals were sampled from a normal distribution with the mid-parent value as mean and variance = 0.15/2 or 0.25/2. Inbreeding was ignored.

Somatic cell scores under healthy ($\mathbf{SCS}_0$) and infected ($\mathbf{SCS}_1$) states were simulated as follows:

$$\mathbf{SCS_0} = \mathbf{M_0}\ \boldsymbol{\mu_0} + \mathbf{a} + \mathbf{e_0} \text{ and } \mathbf{SCS_1} = \mathbf{M_1}\ \boldsymbol{\mu_1} + \mathbf{a} + \mathbf{e_1},$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are the (T X 1) vector means of both distributions, **a** is the (N X 1) vector of breeding values (computed as above), and $\mathbf{M_0}$ and $\mathbf{M_1}$ are the incidence matrices relating $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ to $\mathbf{SCS_0}$ and $\mathbf{SCS_1}$, respectively. The number of observations per cow was set at T = 10 or 20 observations. The vectors $\mathbf{e}_0$ and $\mathbf{e}_1$ were sampled from 2 normal distributions with null means and residual variances set at 1.0 and 1.4. The values for the residual variances were found in the literature [13]. Each element of $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ was taken from the curves observed in cows without and with mastitis, and for high and low responders (Figure 1). The cows were assigned to a group (IMI+ or IMI-) at random using appropriate membership probabilities: The proportion of cows with at least one IMI+ sample was set at P_{cow} = 20% and 50% and, among IMI+ cows, the proportion infected with *E. coli* was set at P_{coli} = 0%, 50% and 100% (the other IMI+ cows were considered infected with *S. aureus*). If a cow was assigned to the IMI+ group, the time at which the clinical episode starts (= t*) was sampled from an exponential distribution with scale parameter 3 which is in agreement with the reported median time of first occurrence of mastitis, i.e., 2 to 3 months [6].

2.8.2. Evaluation of the accuracy of the estimates

The estimates $(\hat{\mu}_i^t, \hat{\sigma}_0^2, \hat{\sigma}_1^2, \hat{\sigma}_a^2, \hat{\mathbf{a}})$ of the parameters $(\mu_i^t, \sigma_0^2, \sigma_1^2, \sigma_a^2, \mathbf{a})$ were computed, after burn-in, as the means of the posterior distributions. Their accuracies were assessed over the range of parameter values (sensitivity analysis) as follows. For the predicted breeding values, the Spearman correlation coefficient (corr_{BV}) with the true breeding values was computed for each replicate and averaged over the 10 replicates. For residual and additive variances, the differences ($\mathrm{bias}_{\sigma 0}$, $\mathrm{bias}_{\sigma 1}$, and $\mathrm{bias}_{\sigma a}$) between estimates and simulated values were computed for each replicate and averaged over the 10 replicates. For the location parameters, the biases ($\mathrm{bias}_{\mu 0}$ and $\mathrm{bias}_{\mu 1}$) were calculated between the estimates and $\bar{y}_i^t$, where $\bar{y}_i^t = \dfrac{\sum_{k=1,n_t^i} (y_k^t \mid z_k^t = i)}{n_t^i}$ computed with known values for z_k^t. Finally, sensitivity (SE), specificity (SP), and probability of correct classification (PCC), were computed at each iterative step as:

$$\mathrm{SE} = \sum_{k=1,N} \sum_{t=1,T} \mathrm{p}\left(\hat{z}_k^t = 1 \middle| z_k^t = 1\right), \mathrm{SP} = \sum_{k=1,N} \sum_{t=1,T} \mathrm{pr}\left(\hat{z}_k^t = 0 \middle| z_k^t = 0\right), \text{ and}$$

$$\mathrm{PCC} = \sum_{k=1,N} \sum_{t=1,T} \mathrm{pr}\left[(z_k^t = 1 \cap \hat{z}_k^t = 1) \cup (z_k^t = 0 \cap \hat{z}_k^t = 0)\right].$$

After burn-in, these were averaged over the δ Gibbs rounds and the 10 replicates.

3. RESULTS AND DISCUSSION

From visual inspection of the algorithmic convergence, it was found that a total of 1,000 cycles and a burn-in (ρ) of 200 runs were sufficient to remove the influence of the prior values and obtain stable estimates. All results presented thus correspond to the last (δ = 800) runs of the Gibbs algorithm. This may seem very few cycles but results were checked for 3 simulated data sets over a higher number of cycles of the Gibbs sampler. Convergence rates were also checked with an EM algorithm and the Gibb sampler on models similar to those used in the simulation of this study but without genetic covariance structure ($\mathbf{SCS_i = M_i \ \mu_i + \ e_i}$). Explanations may be linked to the simplicity of the pedigree structure, small number of cows and the fact that value for m_0 and $s^2{}_p$ were obtained from the data.

3. 1. Overall accuracy of the estimates

Overall, the sensitivity was high (SE ~ 90%) but the specificity low (SP ~ 60%). Because of this high sensitivity, we can be confident that a cow with $\hat{z}_k^t = 0$ is healthy and spare the costs of further testing (e.g., bacteriological cultures) or useless treatment. On the other end, the low specificity indicates that cows with $\hat{z}_k^t = 1$ should be further tested to confirm the clinical suspicion. These observations may suggest some economic interest in HMM.

Before any testing, the probability for a cow to be IMI+ can only be estimated from the prevalence of the disease in the population, while, after testing, this probability is estimated from the posterior probability of being IMI+ given a positive test (also called the positive predictive value). With SE = 90% and SP = 60%, the difference between prior and posterior probabilities is maximum at disease frequencies between 20% and 50%, with posterior probabilities 20% higher than the prior probabilities. These frequencies are within the range of prevalence typically reported for mastitis, as illustrated in the following few studies. In Finland, Pitkälä *et al.* [18] reported 31% of cows with SCC>300,000/mL (mastitis) in 2001. In Switzerland, Roesch *et al.* [19] reported 40% of cows showing at least 1 positive California Mastitis Test in at least one quarter at 31 d and 102 d post partum. In a survey of clinical and subclinical mastitis in England and Wales, the mean incidence of clinical mastitis recorded by the farmer was 47 cases per 100 cows per year [3]. In Canada, Sargeant *et al.* [21] observed 19.8% of cows experienced one or more cases of clinical mastitis during a 2-year observational study. Therefore, HMM may also be of interest in field studies, when it is necessary to precisely identify infected cows.

Breeding values from the HMM seemed accurate in predicting the true additive genetic merit of the cows. Indeed, the correlation ($\mathrm{corr}_{\mathrm{BV}}$) between simulated and estimated breeding values varied from 65% to 79% over the whole data sets. This is close to the correlations of 70 to 75% computed as the square root of the coefficient of determination (CD), where $\mathrm{CD} = 1 - \frac{\mathrm{PEV}}{\mathrm{V}}$, PEV = prediction error variance = $[\mathbf{W'R^{-1}}\ \mathbf{W} + \mathbf{A}^{-1}/\sigma_a^2]^{-1}$ and V = true additive variance = $\mathbf{A}\ \sigma_a^2$ [11]. The PEV were computed with the values of the parameters used in the simulation and weighted by the true proportion of IMI- and IMI+ per cow.

On the other hand, the HMM was less efficient in estimating the parameters for the IMI+ group. Indeed, $\hat{\sigma}_1^2$ had a tendency to underestimate and $\hat{\mu}_1^t$ to overestimate the values used in the simulation. The biases varied from -1.33 to -0.13 (mean = -0.59) for $\hat{\sigma}_1^2$ and from -0.02 to 3.26 (mean = 1.14) for $\hat{\mu}_1^t$. The magnitude of the biases decreased when the amount of information available on IMI+ cows increased, as discussed in the sensitivity analyses below.

3. 2. Sensitivity analyses

The robustness of the HMM approach was assessed by computing the biases in the estimates over a wide range of values for the simulated parameters.

Over all, estimates of means and variances were rather insensitive to the values of the corresponding simulated values but they were sensitive to the proportion of cows with at least one IMI+ sample (P_{cow}) and to the proportion of *E. coli* among infected cows (P_{coli}). This suggests HMM estimates are sensitive to the amount of data available to compute them. For example, biases in the estimation of both location parameters $(\hat{\mu}_0^t, \hat{\mu}_1^t)$ were highest when P_{cow} was lowest (Figure 2), suggesting that it is necessary to have a sufficient number of observations per cow when the disease prevalence is low. Similarly, SE, SP and PCC decreased as the proportion of *E. coli* infection (P_{coli}) increased (Figure 3). This was not surprising because, in cows infected with *E. coli*, only few simulated SCS were higher than SCS for IMI- samples, as is observed in naturally occurring *E. coli* infections usually of short duration.

Level of response to infection influenced estimates of transition probabilities, in opposition to estimates of both location parameters and of breeding values. For example, SE and PCC were higher among high (SE = 92%; PCC = 64%) than moderate (SE = 80%; PCC = 60%) responders suggesting that HMM is more accurate when IMI- and IMI+ distributions are farther apart. Conversely, accuracy of $\hat{\sigma}_1^2$ worsened when the distance between IMI- and IMI+ distributions increased with $\text{bias}_{\sigma 1}$ = -0.51 for moderate and $\text{bias}_{\sigma 1}$= -0.80 for high responders.

Note that SE and SP were insensitive to change in disease frequency (P_{cow}), as they should by definition, conversely to PCC that is, by definition, a function of the disease frequency: PCC = [SE*pr(IMI+)] + [SP*pr(IMI-)].

Finally, note SE and SP reported here are different from SE and SP in Ødegård *et al.* [17] in which

$$\mathrm{SE} = \frac{\sum_{i=1,n} t_i \mathrm{PPM}_i}{\sum_{i=1,n} t_i} \text{ and } \mathrm{SPE} = \frac{\sum_{k=1,n} (1 - t_i)(1 - \mathrm{PPM}_i)}{n - \sum_{i=1,n} t_i}$$

where PPM_i is the posterior mean of the estimates of z_i averaged over Gibbs samples (after burn-in), t_i = 0 if IMI-, t_i = 1 if IMI+, and i = 1 to n cows.

3. 3. General discussion

The main advance of this paper is the presentation of a HMM in which genetic random effects were added to the conditional model for the observed data. In the subject-area literature, HMMs with random effects have been used in a very limited way. Only recently, Altman [1] introduced a mixed HMM to study lesion counts in multiple sclerosis patients. In her model, parameters for the observed and hidden data are allowed to vary randomly among patients, although they are assumed independent from each other (no genetic relationship). This suggests a natural extension of the present HMM, i.e., to allow the parameters of the hidden Markov chain to vary randomly among cows. However, interpretation of the results of such extended model will be delicate because sets of identical genes may be associated to both IMI and SCS (confounding effects). Stated otherwise, the total genetic effects on SCS would be a combination of the effects of genes responsible for presence or not of IMI (resistance to infection) and for the magnitude of the SCS response after IMI (tolerance after infection).

Structural equation modeling is a technique to evaluate models with different hypothesized relationships among variables. In this context, it would be interesting to evaluate the different models proposed in Figure 4 to determine the amount of relationships between genes insuring tolerance or resistance to infection. In the model proposed here, biomarker value at one specific time is independently influenced by the IMI status and by some genes. But, both the IMI status and the biomarker values could also be under the influence of this same set of genes (model b of Figure 4). The relationship between genes, biomarker and IMI status can become even more complicated with different sets of correlated genes influencing the expression of both traits (model e). This is important for the long-term because some epidemiological models predict that selection for resistant cows (no infection) may not be as durable as selection for tolerant (infection but no disease) cows [16], [20]. Increased resistance would reduce disease transmission, reducing the fitness advantage of carrying the resistant genes, and possibly impose pressure upon the pathogen to evade the control strategy. By contrast, as genes conferring disease tolerance spread within a population, the disease incidence rises, increasing the evolutionary advantage of carrying the tolerance genes, without leading to genetic changes in the parasite population.

Other extensions of the HMM are possible. Trends and seasonality in SCS can be readily accommodated to relax the assumption of time-independence between transition probabilities [15]. Prior information on the parameters can be included to increase accuracy and speed up convergence. Location parameters can be made more realistic by considering effects affecting SCS values, such as age, herd, or season, as a few examples. Elements of the M matrices could take different values than zeroes or ones to reflect the different effects on SCS for different parts of the lactation.

The genetic variance could also be different for IMI- and IMI+ samples and would allow for genetic difference in the response in SCS to IMI.

The first-order Markov assumption is also a limiting feature of the HMM and mechanisms of transmission of the IMI between cows could also be considered more precisely in deriving the transition probabilities. Indeed, transmission of infection is a complex process that involves the mixing structure of the population (as it determines the probability of contact between animals), the infectiousness of the contagious animal (or infective dose) and the susceptibility of a healthy cow (i.e., its probability of getting infected after contact with a contagious animal). To solve these issues, Cooper and Lipsitch [5] proposed to model the transition probabilities of the hidden Markov chain in terms of the parameters of epidemiologic models used to describe the transmission of an infectious disease at the population level.

3. 4. Conclusions

In summary, it is shown that the mixed HMM provides a good fit to the data sets simulated in this study. The advantages of the HMM over other approaches are the prediction of health or disease status, the reduction of confirmatory diagnosis costs, and the increased accuracy in breeding values. However, future work needs to be done to extend the HMM proposed here, the most important piece of which is to quantify the level of resistance and tolerance to infection while considering the mechanisms of transmission between healthy and sick cows.

4. ACKNOWLEDGMENTS

This study was supported by EADGENE (European Animal Disease Genomics Network of Excellence for Animal Health and Food Safety).

5. REFERENCES

[1] Altman R. M., Mixed hidden Markov model: An extension of the hidden Markov model to the longitudinal data setting, J. Am. Stat. Assoc. 102 (2007) 201-210.

[2] Boettcher P.J., Moroni P., Pisoni G., Gianola D., Application of finite mixture model to somatic cell scores of Italian goats, J. Dairy Sci. 88 (2005) 2209-2216.

[3] Bradley A. J., Leach K. A., Breen J. E., Green L. E., Green M. J., Survey of the incidence and aetiology of mastitis on dairy farms in England and Wales, Vet. Rec. 160 (2007) 253-257.

[4] Carlén E., Strandberg E., Roth A., Genetic parameters for clinical mastitis, somatic cell score, and production in the first three lactations of Swedish Holstein cows, J. Dairy Sci. 87 (2004) 3062-3070.

[5] Cooper B., Lipsitch M., The analysis of hospital infection data using hidden Markov model, Biostat. 5 (2004) 223-237.

[6] de Haas Y., Barkema H. W., Veerkamp R. F., The effect of pathogen-specific clinical mastitis on the lactation curve for somatic cell count, J. Dairy Sci. 85 (2002) 1314-1323.

[7] de Haas Y., Veerkamp R. F., Barkema H. W., Gröhn Y. T., Schukken Y. H., Associations between pathogen-specific cases of clinical mastitis and somatic cell count patterns, J. Dairy Sci. 87 (2004) 95 – 105.

[8] Detilleux J.C., Leroy P., Application of a mixed normal mixture model for the estimation of mastitis-related parameters, J. Dairy Sci. 83 (2000) 2341–2349.

[9] Detilleux J., Genetic factors affecting susceptibility to udder pathogens, Vet. Microbiol. (accepted)

[10] Eisner J., An interactive spreadsheet for teaching the Forward-Backward algorithm, in: Proceedings of the ACL workshop on effective tools and methodologies for teaching NLP and CL, July 2002, Philadelphia, pp.10-18.

[11] Fouilloux M-N., Laloë D., A sampling method for estimating the accuracy of predicted breeding values in genetic evaluation, Genet. Sel. Evol. 33 (2001) 473-486.

[12] Gianola D., Prediction of random effects in finite mixture models with Gaussian components, J. Anim. Breed. 122 (2005) 145-159.

[13] Heringstad B., Gianola D., Chang Y. M., Ødegård J., Klemetsdal G., Genetic associations between clinical mastitis and somatic cell score in early first-lactation cows, J. Dairy Sci., 89 (2006) 2236 – 2244.

[14] Hernández A., Karrow N., Mallard B. A., Evaluation of immune responses of cattle as a means to identify high and low responders and use of a human microarray to differentiate gene expression, Genet. Sel. Evol. 35 (2003) 67-81.

[15] Le Strat Y., Carrat F., Monitoring epidemiologic surveillance data using hidden Markov models, Stat. Medicine 18 (1999) 3463-3478.

[16] Miller M. R., White A., Boots M., The evolution of host resistance: Tolerance and control as distinct strategies, J. Theor. Biol. 236 (2005) 198-207.

[17] Ødegård J., Jensen J., Madsen P., Gianola D., Klemetsdal G., Heringstad B., Detection of mastitis in dairy cattle by use of mixture models for repeated somatic cell scores: A Bayesian approach via Gibbs sampling, J. Dairy Sci. 86 (2003) 3694–3703.

[18] Pitkälä A., Haveri M., Pyörälä S., Myllys V., Honkanen-Buzalski T., Bovine mastitis in Finland 2001 – Prevalence, Distribution of bacteria, and antimicrobial resistance, J. Dairy Sci., 87 (2004) 2433-2441.

[19] Roesch M., Doherr M. G., Schären W., Schällibaum M., Blum J. W., Subclinical mastitis in dairy cows in Swiss organic and conventional production systems, J. Dairy Res. 74 (2007) 86-92.

[20] Roy B. A., Kirchner J. W., Evolutionary dynamics of pathogen resistance and tolerance, Evolution 54 (2000) 51-63.

[21] Sargeant J.M., Scott H. M., Leslie K. E., Ireland M. J., Bashiri A., Clinical mastitis in dairy cattle in Ontario: frequency of occurrence and bacteriological isolates, Can. Vet J. 39 (1998) 33-38.

[22] Wenz J. R., Barrington G. M., Garry F. B., McSweeney K. D., Dinsmore P., Goodell G., Callan R. J., Bacteremia associated with naturally occurring coliform mastitis in dairy cows, J. Am. Vet. Med. Assoc., 219 (2001) 976 – 981.

FIGURE CAPTIONS

Figure 1. Means of SCS for lactations without clinical mastitis (plain line) and lactations with clinical mastitis associated with *Staphylococcus aureus* (square) or *Escherichia coli* (triangle) occurring on the median MIM for multiparous cows (Adapted from de Haas *et al.*, [6]) .

Figure 2. Differences between simulated and estimated values for the means of the distributions for healthy (dot bar) and infected (slash bar) cows as a function of the proportion of infected cows.

Figure 3. Sensitivity (plain bar), specificity (open bar) and probability of correct classification (circled bar) as a function of the proportion of *E. coli* among infected cows.

Figure 4. Five different hypothetical models of the relationship between genetic background (G), intra-mammary infection (IMI) and biomarker (Bio). The first model (a) is the model of this study (The dependent variables are the targets of one-headed arrows).

Figure 1

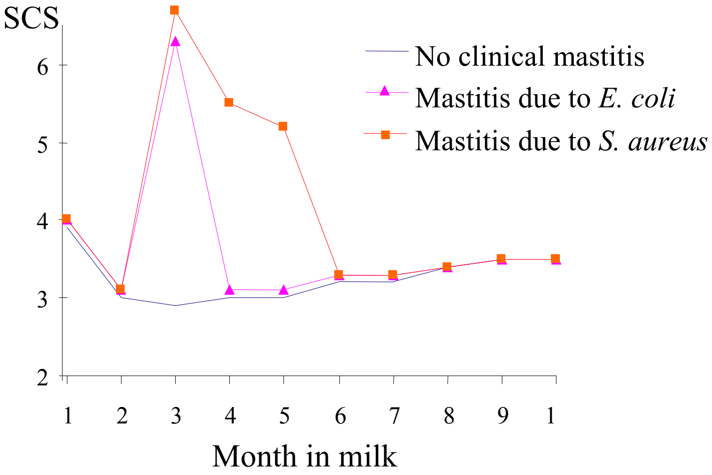

SCS
6
5
4
3
2
1
2
3
4
5
6
7
8
9
1
Month in milk
No clinical mastitis
Mastitis due to *E. coli*
Mastitis due to *S. aureus*

Figure 2.

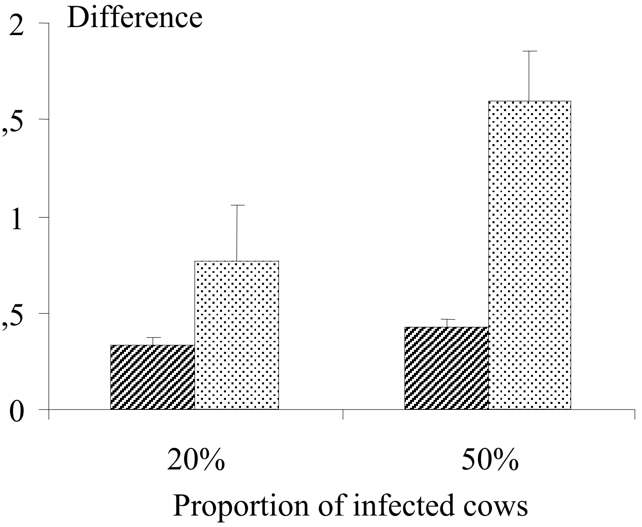


Figure 3.

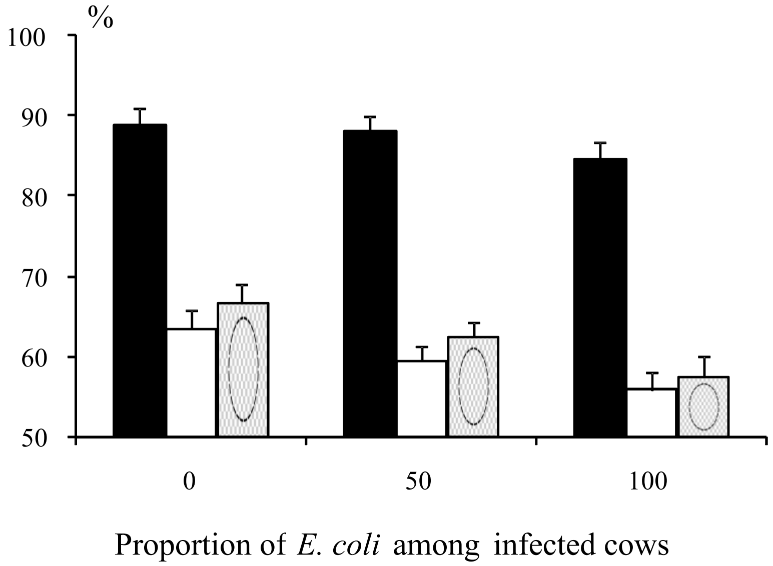
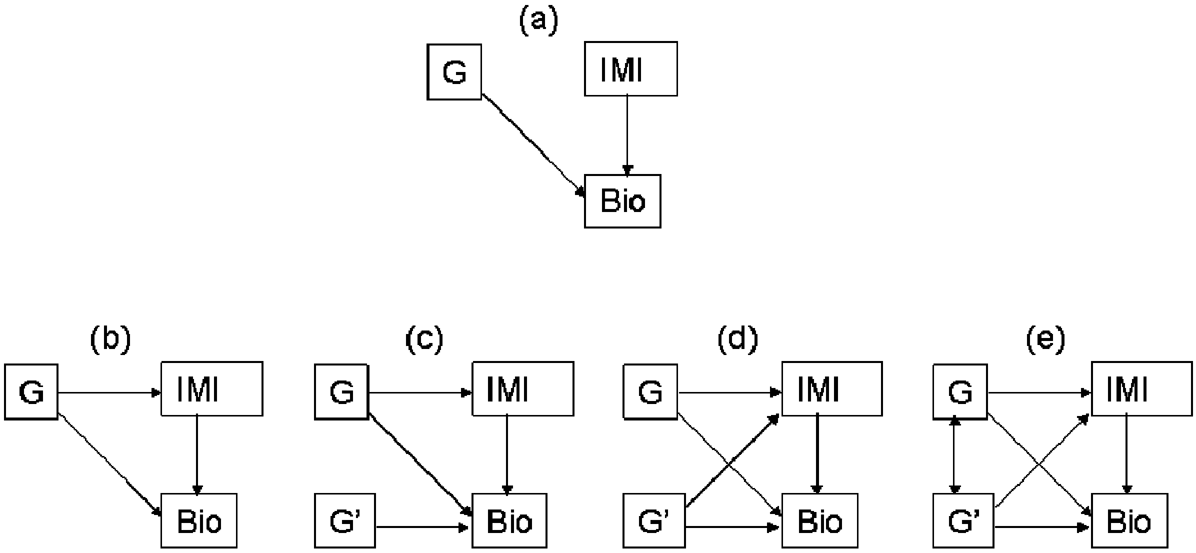

%
100
90
80
70
60
50
0
50
100
Proportion of *E. coli* among infected cows

Figure 4.



ANNEX

Table I. Sensitivity (SE), specificity (SP), and probability of correct classification (PCC) as a function of the level of response to infection, high (H) or moderate (M) responders, number of samples per cow (T), percentage of cows with at least one IMI+ sample (P_{cow}), percentage infected with *E. coli* (P_{coli}), and residual and additive genetic variances ($\sigma_0^2, \sigma_1^2, \sigma_a^2$). Data sorted by SE.

SE	SP	PCC	T	P_{cow}	P_{coli}	σ_0^2	σ_1^2	σ_a^2
High responders								
95.03	59.65	63.70	10	50	50	1.0	1.0	0.15
94.50	58.19	60.64	10	20	0	1.4	1.4	0.15
94.25	49.59	56.73	10	20	50	1.4	1.4	0.15
94.03	58.05	59.90	20	20	50	1.0	1.0	0.25
93.92	62.71	65.98	20	50	0	1.0	1.0	0.25
93.79	58.88	60.63	20	20	50	1.4	1.4	0.25
93.20	57.51	59.31	20	20	50	1.4	1.4	0.25
93.08	55.15	56.95	10	20	50	1.4	1.4	0.25
92.64	58.23	62.16	10	50	50	1.4	1.4	0.15
92.64	65.99	68.16	20	20	0	1.4	1.4	0.25
92.63	57.49	58.34	20	20	50	1.4	1.4	0.25
92.03	59.91	61.49	20	20	50	1.4	1.4	0.25
90.41	50.89	51.65	10	20	100	1.4	1.4	0.15
89.58	50.60	51.34	10	20	100	1.4	1.4	0.15
89.05	69.75	73.53	20	50	0	1.0	1.0	0.15
88.81	68.09	72.19	20	50	0	1.4	1.4	0.25
88.19	66.02	70.42	20	50	0	1.4	1.4	0.25
88.14	68.43	72.38	20	50	0	1.0	1.4	0.15
85.06	68.53	71.84	20	50	0	1.0	1.4	0.25

84.27	55.36	55.94	20	20	100	1.4	1.4	0.25
Moderate responders								
94.24	57.41	59.28	20	20	50	1.0	1.0	0.25
79.74	52.41	52.95	20	20	50	1.0	1.0	0.25
79.09	54.89	56.74	20	20	0	1.4	1.4	0.25
77.95	53.64	54.81	20	20	50	1.4	1.4	0.25
77.67	64.32	67.03	20	50	0	1.0	1.4	0.15
77.06	63.14	65.90	20	50	0	1.0	1.4	0.25
75.77	51.78	52.24	20	20	100	1.4	1.4	0.25
73.04	58.81	61.60	20	50	0	1.0	1.4	0.25

Table II. Accuracy of the estimates of the mixed hidden Markov model as a function of the level of response to infection, high (H) or moderate (M) , number of samples per cow (T), percentage of cows with at least one IMI+ sample (P_{cow}), percentage infected with *E. coli* (P_{coli}), and residual and additive genetic variances ($\sigma_0^2, \sigma_1^2, \sigma_a^2$). The accuracy is determined as the differences between values used in the simulations and estimates of means ($bias_{\mu 0}$, $bias_{\mu 1}$) and residual variances ($bias_{\sigma 0}$, $bias_{\sigma 1}$) in IMI- and IMI+ cows, respectively; the differences between values used in the simulations and estimates of additive genetic variance ($bias_{\sigma a}$); and the correlation between predicted and simulated breeding values ($corr_{BV}$). Data sorted by $corr_{BV.}$

$corr_{BV}$	$bias_{\sigma 0}$	$bias_{\sigma 1}$	$bias_{\sigma a}$	$bias_{\mu 0}$	$bias_{\mu 1}$	T	P_{cow}	P_{coli}	σ_0^2	σ_1^2	σ_a^2
High responders											
0.79	0.00	-0.66	-0.08	0.24	0.47	20	50	0	1.0	1.4	0.15
0.79	0.02	-0.65	-0.02	0.21	0.28	20	50	0	1.0	1.0	0.15
0.78	-0.02	-0.78	0.00	0.22	0.43	20	50	0	1.0	1.4	0.25
0.77	0.01	-0.70	0.01	0.28	0.51	20	50	0	1.4	1.4	0.25
0.77	0.02	-0.63	0.04	0.23	0.52	20	50	0	1.4	1.4	0.25
0.74	-0.01	-0.29	0.05	0.41	2.16	20	20	100	1.4	1.4	0.25
0.74	0.06	-0.46	-0.01	0.50	2.93	10	20	100	1.4	1.4	0.15
0.73	0.04	-0.57	0.02	0.31	0.80	20	20	0	1.4	1.4	0.25
0.73	0.09	-0.48	-0.03	0.55	3.26	10	20	100	1.4	1.4	0.15
0.72	0.03	-0.42	0.04	0.52	1.26	20	20	50	1.4	1.4	0.25
0.71	0.02	-0.46	0.04	0.42	1.22	20	20	50	1.4	1.4	0.25
0.71	0.03	-0.48	0.05	0.40	1.13	20	20	50	1.4	1.4	0.25
0.71	0.09	-0.65	-0.02	0.44	1.86	10	20	50	1.4	1.4	0.15
0.70	0.02	-0.44	0.04	0.38	1.17	20	20	50	1.4	1.4	0.25
0.70	0.09	-0.60	0.06	0.51	1.73	10	20	50	1.4	1.4	0.25
0.69	0.03	-0.57	0.04	0.36	0.87	20	50	0	1.0	1.0	0.25

0.69	0.11	-0.74	-0.03	0.40	1.69	10	20	0	1.4	1.4	0.15
0.68	0.08	-1.25	-0.02	0.38	1.48	10	50	50	1.0	1.0	0.15
0.67	0.03	-0.44	0.06	0.43	1.06	20	20	50	1.0	1.0	0.25
0.67	0.07	-1.21	-0.03	0.39	1.46	10	50	50	1.4	1.4	0.15
Moderate responders											
0.76	-0.02	-0.46	-0.02	0.24	0.00	20	50	0	1.0	1.4	0.15
0.75	-0.01	-0.13	0.05	0.48	1.61	20	20	100	1.4	1.4	0.25
0.75	-0.01	-0.14	0.07	0.47	1.30	20	20	50	1.0	1.0	0.25
0.75	-0.03	-0.21	0.04	0.32	0.70	20	20	0	1.4	1.4	0.25
0.74	-0.02	-0.18	0.06	0.32	0.82	20	20	50	1.4	1.4	0.25
0.73	-0.03	-0.46	0.04	0.32	0.19	20	50	0	1.0	1.4	0.25
0.72	-0.04	-0.36	0.05	0.39	-0.02	20	50	0	1.0	1.4	0.25
0.66	0.03	-0.45	0.06	0.44	1.22	20	20	50	1.0	1.0	0.25