A hidden Markov model to predict early mastitis from test-day somatic cell scores.

J. C. Detilleux

Department of Quantitative Genetics, Veterinary Faculty, University of Liège, 4000 Liège, Belgium.

Corresponding author: Johann C. Detilleux. E-mail: jdetilleux@ulg.ac.be

Running head: Hidden Markov model in mastitis

Abstract

The absence of on farm recording systems in most countries precludes the identification of clinical mastitis cases after its occurrence. Therefore, in many countries high somatic cell scores (SCS) in milk are used as indicator for mastitis because they are collected on a routine basis. However, individual test day SCS are not very accurate in identifying infected cows. Mathematical models may improve the accuracy of the biological marker by making better use of the information contained in the available data. Here, a simple hidden Markov model (HMM) was applied on SCS recorded monthly on cows with or without clinical mastitis to evaluate its accuracy in estimating parameters (mean, variance and transition probabilities) under health or disease states. The SCS means were estimated at 1.96 (SD = 0.16) and 4.73 (SD = 0.71) for the hidden healthy and infected states, and the common variance at 0.83 (SD = 0.11). The probabilities to remain uninfected, to recover from infection, to get newly infected and to remain infected between consecutive test-days were estimated at 78.84%, 60.49%, 11.70% and 15%, respectively. Three different health related states were compared: clinical stages observed by farmers, subclinical cases defined for somatic cell counts below or above 250,000 cells/mL and infected stages obtained from the HMM. The results showed that HMM identifies infected cows before the apparition of clinical and subclinical signs which may critically improve the power of studies on the genetic determinants of SCS and reduce biases in predicting breeding values for SCS.

Key words: **mastitis; hidden Markov model; somatic cell counts.**

Implications

In most countries, somatic cell counts (SCC) are routinely used as indicators of mammary infection before milk is exploited for consumption before or after its transformation. However, SCC are not very sensitive in classifying cows as infected or healthy which leads to unnecessary costs and missed profits. Here, a simple hidden Markov model is proposed that improved the diagnostic accuracy of SCC by uncovering the hidden health status of the cows before the apparition of clinical signs and before SCC exceed the threshold of 250,000 cells/mL. This will critically improve the power of genetic studies of mastitis determinants and reduce biases in predicting breeding values.

Introduction

The absence of on farm recording systems in most countries precludes the identification of clinical mastitis cases after its occurrence. Therefore, in many countries high somatic cell counts (SCC) in milk are used as indicator for sub-clinical and clinical mastitis, especially for genetic evaluation to improve resistance to mammary infections that necessitate large amount of data (Shook and Schutz, 1994). However, the problem of identifying infected cows based on their SCC is still not satisfactorily solved as individual SCC are not very sensitive in diagnosing mammary infection, either at the quarter or cow levels (Djabri *et al.*, 2002; Sargeant *et al.*, 2001). This has relevant impact in animal selection because imperfect accuracy in the diagnosis of infectious diseases results in a reduction of heritability estimates (Bishop and Woolliams, 2010). It is also a source of misclassification as uninfected animals may have high SCC (and reversely). This may bias prediction of breeding values and decrease the power to detect association between a disease locus and a marker locus (Buyske et al., 2009). Selection for very low SCC might even not be the good objective because low initial SCC has been associated with increased susceptibility and severity of subsequent mastitis (Suriyasathaporn *et al.*, 2000).

Mathematical models improve the accuracy of SCC measures used to identify infected cows by making better use of the information contained in SCC data. For example, models developed by de Haas *et al.* (2004) lead to the identification of different SCC patterns according to the mammary pathogen: Clinical *E. coli*

mastitis is significantly associated with the presence of a short peak in SCC whereas *S. aureus* is associated with long increased SCC. Others have used the finite mixture model (FMM) methodology on SCC to infer the cow's individual probability of being infected (Detilleux and Leroy, 2000; Gianola, 2005). A simple FMM will assign SCC to one of two components hopefully representing SCC from cows with (IMI+) and without (IMI-) intra-mammary infection (IMI), respectively. Then, the identification of animals at risk is computed as the posterior probability of putative IMI, given SCC, rather than on crude SCC. However, after bacteriological examinations of goat milk samples, Boettcher *et al.* (2005) observed their FMM was able to classify correctly only 60% and 48% of the healthy and infected records, respectively. If these results are not encouraging, it should be noted the accuracy for detecting an IMI from bacteriological cultures of single composite or quarter milk samples in subclinically infected cows is known to be low (Lam *et al.*, 1996; Sears *et al.*, 1990). This is because pathogens such as *S. aureus* are often shed in an intermittent or cyclical pattern and in numbers too low to be detected by conventional culturing methods (Godden *et al.*, 2002). The *S. aureus* and coagulase negative staphylococci were the most prevalent pathogens in the above mentioned goat study (Moroni *et al.*, 2005).

Hidden Markov models (HMM) could be an alternative to FMM. A HMM is defined as a finite set of states, each of which is associated with a probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. The joint distribution over all states is a Markov

chain. An observation is associated to each state, according to a linked probability distribution, called the emission probability. Only observations are recordable, not the states that are ``hidden''; hence the name (Rabiner, 1989). In the mastitis context, the health status of the mammary gland could be considered as two hidden states and the SCC as the associated observations.

In a simulated data set (Detilleux, unpublished results), the accuracy of estimates obtained with a FMM was increased by incorporating information from previous SCC, as is done in HMM. In another study (Detilleux, 2008), estimates obtained with a mixed HMM were close to the true values unless the prevalence of the disease is low.

The objective of this study is to present the mathematical formalism behind the HMM methodology, to apply the model on SCC collected on first parity cows with known clinical status and to compare results on clinical (observed by farmers), subclinical (defined for SCC or above 250,000 cells/mL) and hidden (infected or not as obtained from the HMM) states**.**

Materials and methods

Animals and data collection

Data from the field study of Barkema *et al.* (1998) were used. Briefly, bacteriological samples were collected by the farmers from cows with signs of clinical mastitis. For the present analyses, only the first cases of clinical mastitis per lactation (CM_t) were considered and bacteriological results were consolidated

into negative ($B_t = 0$) and positive results ($B_t = 1$), with t representing the month in milk (MIM) at which the case was recorded. Records on clinical case were retrieved between December 1992 and August 1995. Conjointly, a total of 526,867 test days with SCC were recorded by the National Milk Recording System (NRS, Arnhem, The Netherlands). The somatic cell scores (SCS) were computed as $\log_2(\text{SCC}/100{,}000) + 3$ and averaged per MIM. After editing (birth year>1960, SCC<9,999,000, MIM $\leq$ 10, calving date $\leq$ test-day date), the data set included 128,748 records on SCC for the first 10 MIM, on 21,829 1st parity cows. A total of 951 mastitis cases were reported of which 774 were bacteriologically positive. Thereafter, clinical cases without positive bacteriological findings were considered as healthy.

For each MIM, three (two observed and one hidden) different health states were considered. The records were classified as being from a heifer with (CM+) or without (CM-) a reported clinical case. The SCC may be below (SCM-) or above (SCM+) the threshold of 250,000 cells/ml. This threshold was chosen as an indicator of subclinical mastitis and is the one chosen by de Haas *et al.* (2002) in her previous analyses of the data. The last stage is the hidden infected (IMI+) or uninfected (IMI-) stages that were obtained by the HMM.

Statistical analyses

Throughout, k indexes the individual cow, t is the MIM, y_k^t is the SCS observed at t on animal k, and z_k^t is the unknown state with $z_k^t = 0$ if y_k^t is from a hidden IMI- sample and $z_k^t = 1$ if y_k^t is from a hidden IMI+ sample. On each cow, data

consists of a series of repeated SCS: $\mathbf{y_k} = \{y_k^1, y_k^2, \ldots, y_k^T\}$ and the unobserved vector is $\mathbf{z_k} = \{z_k^1, z_k^2, \ldots, z_k^T\}$, for t = 1, 2, …, T. For simplicity, T is assumed constant for all cows.

General formulation of the model. A simple first-order HMM was assumed with 2 transient states corresponding to the hidden IMI- and IMI+ categories with the following parameters:

- probabilities of transition between hidden states:

$$a_k^{00} = \text{pr}(z_k^{t+1} = 0 \mid z_k^t = 0), a_k^{01} = \text{pr}(z_k^{t+1} = 1 \mid z_k^t = 0),$$
$$a_k^{10} = \text{pr}(z_k^{t+1} = 0 \mid z_k^t = 1), a_k^{11} = \text{pr}(z_k^{t+1} = 1 \mid z_k^t = 1),$$

- probability of being IMI- as an initial hidden state $\lambda_k = \text{pr}(z_k^1 = 0)$, and
- probabilities of SCS emission:

$$(y_k^t \mid z_k^t = 0) \sim N(\mu_0^t, \sigma^2) \text{ and } (y_k^t \mid z_k^t = 1) \sim N(\mu_1^t, \sigma^2).$$

The probabilities of transition represent the probabilities of observing a particular hidden (unknown) IMI state at time t + 1, given the hidden IMI state at time t. The probabilities of emission represent the probabilities of observing SCS (at time t) given the hidden IMI state (at time t). It is assumed that correlation between successive SCS is fully accounted for by the underlying Markov process structure so that each SCS are independent given the unknown IMI state (output independence assumption). It is also assumed that state transition probabilities are independent of the actual time at which the transition takes place and do not change across time (stationary assumption). Finally, it is assumed that values in

any hidden state are only influenced by the values of the state that directly preceded it (first-order Markov assumption). The suitability of these assumptions for analyzing repeated SCS are discussed afterward.

To obtain the maximum likelihood estimates (MLE) of the parameter set θ_k^t, where $\theta_k^t = (\lambda_k, a_k^{00}, a_k^{01}, a_k^{10}, a_k^{11}\ \mu_0^t, \mu_1^t, \sigma^2)$, the likelihood of the data must be maximized over all possible values of θ_k^t and this can be done through the expectation maximization (EM) algorithm.

Likelihood of the data. For one cow, the likelihood of one particular sequence of repeated SCS scores is given by:

$$\text{pr}(\mathbf{y_k}|\,\theta_k^t) = \alpha_{0,k}^t\,\beta_{0,k}^t + \alpha_{1,k}^t\,\beta_{1,k}^t,$$

with $\alpha_{i,k}^t = \text{pr}(y_k^1, y_k^2, ..., y_k^t, z_k^t = i \,|\, \theta_k^t)$ and

$$\beta_{i,k}^t = \text{pr}(y_k^{t+1}, y_k^{t+2}, ..., y_k^T \,|\, z_k^t = i, \theta_k^t),$$

for i = 0 and 1. The $\alpha_{i,k}^t$ represents the probability of a partial sequence and ending up in state i at time t and $\beta_{i,k}^t$ represents the probability of a partial sequence starting from t + 1to T given that the sequence started at state i at time t. This likelihood must be computed, for each cow, over all possible sequences of hidden states ($\mathbf{z_k}$). To do so, the naive way would be to sum, for each cow, the probabilities over all possible state sequences but their number can be huge (= 2^T) and the more efficient forward-backward algorithm is used in practice. This algorithm takes advantages of the sequential nature of the data, going forward (t =

1, 2, .., T) and backward (for t = T, T-1, …, 1) in time, knowing it must end in some particular state. For a practical description, see Eisner (2002) and its interactive spreadsheet for teaching the algorithm. After the likelihood is computed for one cow, the likelihood for all sequences of all cows is computed as the product of all individual likelihoods (assumption of independence between cows).

The EM algorithm. For a detailed derivation of the algorithm for HMM, please refer to Bilmes (1998) and Rabiner (1989). In short, the EM algorithm consists of a series of repeated E and M steps. In the E-step, one finds the expected value of the complete-data log-likelihood with respect to the unknown parameters, given the observed data ($\mathbf{y_k}$) and the current parameter estimates ($\theta_k^{(p)}$ at iteration p). To form the complete data, one assumes both observed ($\mathbf{y_k}$) and hidden ($\mathbf{z_k}$) vectors are known. Then, the expected complete-data log-likelihood is written as:

$$\sum_{k=1,N} \{ E[z_k^1 = 0 | \mathbf{y}_k, \theta_k^{(p)}] \log(\lambda_k) + E[z_k^1 = 1 | \mathbf{y}_k, \theta_k^{(p)}] \log(1 - \lambda_k)$$

$$+ \sum_{t=1,(T-1)} \sum_{i,j=0,1} E[z_k^t = i, z_k^{t+1} = j | \mathbf{y}_k, \theta_k^{(p)}] \log(a_k^{ij})$$

$$+ \sum_{t=1,(T-1)} \sum_{i,j=0,1} E[z_k^t = i | y_{\mathbf{k}}, \theta_k^{(p)}] \log p(y_k^t | z_k^t = i) \},$$

where the first two terms of the summation involve observations at the start of the sequence (t = 1), the third term counts how many times each i to j transition occurred in the sequence and the fourth includes all observations generated from state i.

In the M step, one maximizes each term by setting the derivative equal to zero and by using the constraint that $\sum_{i,j=0,1} a_k^{ij} = 1$ to obtain the following MLE (i = 0, 1 and j = 0,1):

$$\hat{\lambda}_k = \gamma_{0,k}^1, \quad \hat{a}_k^{ij} = \frac{\sum_{t=1,(T-1)} \xi_{ij,k}^t}{\sum_{t=1,T} \gamma_{i,k}^t}, \quad \hat{\mu}_i^t = \frac{\sum_{k=1,N} \gamma_{i,k}^t \, y_k^t}{\sum_{k=1,N} \gamma_{i,k}^t}, \quad \hat{\sigma}^2 = \frac{\sum_{k=1,N} \sum_{t=1,Ti=0,1} \gamma_{i,k}^t \, (y_k^t - \mu_i)^2}{\sum_{k=1,N} \sum_{t=1,Ti=0,1} \gamma_{i,k}^t},$$

$$\text{with } \gamma_{i,k}^t = E[z_k^t = i \mid y_k, \theta_k^{(p)}] = \frac{\alpha_{i,k}^t \, \beta_{i,k}^t}{\alpha_{0,k}^t \, \beta_{0,k}^t + \alpha_{1,k}^t \, \beta_{1,k}^t},$$

.

$$\text{and } \xi_{ij,k}^t = E[z_k^t = i, z_k^{t+1} = j \mid y_k, \theta_k^{(p)}] = \frac{\alpha_{i,k}^t \; a_k^{ij} \; \beta_{j,k}^{t+1} \; pr(y_k^{t+1} \mid z_k^{t+1} = j)}{\alpha_{0,k}^t \, \beta_{0,k}^t + \alpha_{1,k}^t \, \beta_{1,k}^t}.$$

Note $\gamma_{0,k}^t$ is the individual posterior probability of an IMI- sample, given the whole SCS sequence. Correspondingly, $\xi_{01,k}^t$ is the posterior probability, for the k^{th} cow, that a hidden state sequence that had to generate the SCS sequence went through IMI- at time t and transitioned into IMI+ at time t + 1.

Evaluation of the MLE. The HMM described in the preceding sections were used to analyze the SCS records. Missing SCS were restored through a multiple imputation procedure with the MCMC method (proc MI of SAS®) in an attempt to avoid loss of statistical power and selection bias associated with loss to follow-up, and to be able to use standard matrix algebra. In this method (refer to Horton and Kleinman, 2007 for a thorough discussion), each missing value is replaced by a

set of plausible values that represent the uncertainty about the right value to impute. After imputation, the set of imputed values were averaged for the subsequent analyses. Note the SCC were transformed in SCS to ensure normality which is an assumption of the MCMC method for imputing missing data.

Different priors for μ_0 (2 to 5), μ_1 (4 to 8) and σ^2 (1 or 2) were used to start the EM algorithms. After the MLE of the parameters were computed, the estimated number of transitions between successive MIM were obtained as

$$\hat{n}_{ij,k} = \sum_{t=1,T} \xi^{t}_{ij,k}$$

with i = j = 0 if the transition is from IMI- to IMI-, i = 0 and j = 1 if the transition is from IMI- to IMI+, i = 1 and j = 0 if the transition is from IMI+ to IMI-, and i = 1 and j = 1 if the transition is from IMI+ to IMI+. These numbers were compared to the observed numbers of transitions between successive MIM with SCM- or SCM+, and to the observed numbers of transitions between MIM with or without a clinical case (CM+ and CM-). The comparisons were done for lactations with or without reported clinical case associated to a positive bacteriological result. In lactations without reported clinical case, only transitions from CM- to CM- were achievable. The numbers of transitions from CM+ to CM+ were not computed because only the first clinical cases were considered.

Results

Before imputation, 6.01% of the 128,748 monthly records were bacteriologically positive. The average SCS over all lactations (first parity cows with or without

case) was at 2.65 (SD = 1.62) in the first MIM, decreased to a minimum at 2.08 (SD = 1.45) during the second MIM before increasing slowly to 2.70 (SD = 1.37) at the end of the lactation. A similar pattern was found for lactations without any case of mastitis (Figure 1), but here, SCS was slightly lower throughout the lactation. In lactations with mastitis, cases were detected on average on the 128th DIM and 27.6% of those occurred during the first MIM. The percentage decreased thereafter, from 12.6% in the second MIM to 5.5% in the last 3 MIM. The complete sequence (n = 10) of SCS records was available on 10.48% of the cows and 75.56% of lactations had information on 5 MIM or more. The missing pattern was not monotone, i.e., a missing SCS was not necessarily followed by missing SCS. After imputation, the complete data set included 218,290 monthly SCS records of which 3.23% were considered as CM+. The SCS trend was similar before and after imputation and the highest difference (about 0.08) was found when SCS were the smallest, in the second MIM. After imputation, the SCS averaged 2.64 (SD = 1.37) in the first MIM, decreased to a minimum at 2.00 (SD = 1.27) during the second MIM and increased to reach 2.69 (SD = 1.23) at the end of the lactation (Figure 1).

Figure 1 is about here

The estimated means and variance obtained with the HMM were: $\hat{\mu}_0$ = 1.96 (SD = 0.16), $\hat{\mu}_1$ = 4.73 (SD = 0.71), and $\hat{\sigma}^2$ = 0.83 (SD = 0.11). As comparison, the observed SCS means for CM- and CM+ lactations were 2.35 (SD = 0.99) and 3.18 (SD = 1.28), respectively. For SCM- and SCM+ lactations, the observed means were 1.97 (SD = 0.64) and 3.48 (SD = 1.01), respectively.

The average number of transitions between hidden (IMI- and IMI+) and observed states (SCM+ and SCM-, CM+ and CM-) states are shown in Figure 2 for lactations with or without at least one reported clinical case. The null hypothesis of no differences between these numbers was tested by a t student test ($p < 0.01$): The numbers of transitions from IMI- to IMI- and from SCM- to SCM- were lower than the observed number of transitions from CM- to CM-. For example, when no cases were reported during the entire lactation (Figure 2a), there were 9 transitions from CM- to CM- but the number of transitions from SCM- to SCM- was 8.14 (SD = 2.0) and the number of transitions from IMI- to IMI- was 7.27 (SD = 2.7).

Figures 2 is about here

The average probabilities of transition between hidden states are given in Figure 3 for lactations with or without a reported case of clinical mastitis. Overall, the probability to remain uninfected was $\hat{a}^{00} = 78.84\%$, to recover from infection was $\hat{a}^{10} = 60.49\%$, to get newly infected was $\hat{a}^{01} = 11.70\%$ and to remain infected was $\hat{a}^{11} = 15\%$. No significant differences were found in these probabilities between lactations with or without at least one reported clinical case.

Figure 3 is about here

Discussion

A naïve HMM is proposed to analyze sequences of monthly SCS as they are collected by the milk recording agencies with the intention of identifying cows

with or without mastitis. The data were previously analyzed by de Haas *et al.* (2004) to identify pathogen-specific SCC patterns. The SCS patterns in Figure 1 are similar to patterns from the previous study (Figure 1a in Haas *et al.*, 2004), with slight differences mainly due to different editing procedures and considering that SCS were averaged over each MIM.

The model provides useful features for genetic and genomic selections. Firstly, results from Figure 2 suggested that analyzing SCS with a HMM lead to the identification of infected cows before the apparition of clinical signs and before SCC gets higher than 250,000 cells/mL. Indeed, among cows for which at least a mastitis case was reported (Figure 2b), the model assigned the state IMI+ on three occurrences while the stage SCM+ was observed on two occasions. In heifers without any reported clinical mastitis case (Figure 2a), there were two IMI+ and one SCM+. The likely sequences for the IMI, SCM and CM stages are shown in Figure 4, considering that most clinical stages were reported in early lactation. It should however be noted that an experimental infection in a well-designed clinical trial is necessary to confirm this findings

Figure 4 is about here

Although these results should be confirmed in a well-designed clinical trial with experimental infection, this ability will lead to more accurate estimates of breeding values and an earlier and more accurate selection. It will also facilitate the identification of the genetic determinants of mastitis because hidden IMI states may be considered as intermediate phenotypes with stronger genetic determinants than SCM or CM. Secondly, HMM may be used to predict the future health

status of a cow, based on its previous sequence of SCS. Mathematically, this prediction is given by combining forward (α) and backward (β) probabilities used in the algorithm as:

$$\text{pr}(z_k^t = i \mid y_k^1, y_k^2, ..., y_k^{t-1}, \theta_k) = \frac{\alpha_{0,k}^{t-1}\, \hat{a}_k^{0i} + \alpha_{1,k}^{t-1}\, \hat{a}_k^{1i}}{\alpha_{0,k}^{t-1} + \alpha_{1,k}^{t-1}}$$

In these probabilities, the uncertainty about the time of exposure to infection, if it has occurred, is reduced because data on the entire available sequence of SCS is exploited. Therefore, it may lower the biases due to incomplete exposure on estimable heritabilities (Bishop and Wooliams, 2010). Thirdly, the model provides estimates of the probability of recovery (IMI+ to IMI- = a^{10}) and of new infection (IMI- to IMI+ = a^{01}) for each animal. These parameters are directly related to well-established selection objectives for better udder health and epidemiological concepts. For example, the force of infection (ω = the rate at which susceptible individuals become infected) and the recovery rate (δ = the rate at which infected individuals recover) may be obtained from a^{10} and a^{01} as:

$$a^{01} = \frac{\omega}{\varpi + \delta}(1 - e^{-(\omega + \delta)t}) \text{ and } a^{10} = \frac{\delta}{\varpi + \delta}(1 - e^{-(\omega + \delta)t}),$$

assuming a SI model (Anderson and May, 1992; Detilleux *et al*., 2006). Then, data from genetic and epidemiological studies could be combined to analyze the impact of selecting for a better ability to recover from disease on the spread of the disease at the population level. Finally, the model can be extended by adding genetic random effects to obtain breeding values for SCS (Detilleux, 2008) or even for the hidden IMI variable (Altman, 2007), considering the total genetic

effects on SCS is a combination of the effects of genes responsible for presence or not of infection and for the magnitude of the SCS response after infection.

The model is very flexible and allows the inclusion of prior knowledge (e.g., clinical or laboratory records) to the SCS information. The effects of covariates (e.g., treatment or culling, breed, parity) on the progression of the IMI could also be studied by comparing transition rates.

The HMM methodology presents also some limitations. The HMM, as proposed here, necessitated that the sequence of SCS was complete. One possibility was to discard lactations with incomplete information but this would have decreased the amount of available data. Missing data were instead imputed and a multiple imputation procedure was chosen as it increases robustness to departures from the true imputation model considerably compared to single imputation approaches that do not reflect uncertainty about the imputed values. The MCMC method was chosen because SCS were distributed normally and because the missing pattern was not monotone. After imputation, the SCS curves were slightly lower than before imputation (Figure 1). This may be explained by the fact that, in the MCMC method, missing SCS were replaced by randomly selecting a value (at any MIM) and that SCS at different MIM are correlated with the SCS being imputed. Another drawback was the assumption that probability of staying in a given state was independent of the duration of the state. It could have been modeled explicitly as a_{11}^{d-1} $(1 - a_{11})$ which is the probability of staying d times in state IMI+. The transition probabilities were assumed constant across time although it is known that susceptibility to IMI vary across lactation stages (Paape

et al., 2002). This stationary assumption is very strong but it could be relaxed by parameterizing the mean of the IMI+ distribution to account for various trend or seasonality in the data (Le Strat and Carrat, 1999). Another assumption of the HMM, the independence between successive SCS, could be released in its autoregressive form by allowing previous SCS to assist in predicting the current SCS (Laverty *et al.*, 2002; Ephraim and Roberts, 2005). Finally, the assumption of homoscedasticity can be relaxed by modeling different variances for the IMI+ and IMI- samples (Detilleux, 2008)

The maximum likelihood estimation via the EM algorithm has also some disadvantages. For example, it does not provide an estimated covariance matrix for the parameters. Bootstrap methods can be used but they are computationally intensive for this type of model. Other alternatives are to estimate parameters via the Gibbs sampler or Bayesian variational methods (Jaakkola and Jordan, 2000). Collinearity between parameter estimates can lead to identifiability problems (Brookhart *et al.*, 2002) and the EM may converge toward singular estimates at the boundary of the parameter space. It may also fail to converge. The problem becomes particularly severe when time series are short and data sparse (Cooper and Lipstich, 2004).

Conclusions

A simple hidden Markov model (HMM) was applied on SCS recorded monthly on cows with or without clinical mastitis to evaluate its accuracy in estimating

parameters under health or disease states. The SCS means were estimated at 1.96 (SD = 0.16) and 4.73 (SD = 0.71) for the hidden healthy and infected states, and the common variance at 0.83 (SD = 0.11). The probabilities to remain uninfected, to recover from infection, to get newly infected and to remain infected between consecutive test-days were estimated at 78.84%, 60.49%, 11.70% and 15%, respectively. Three different health related states were compared: clinical stages observed by farmers, subclinical cases defined for somatic cell counts below or above 250,000 cells/mL and infected stages obtained from the HMM. The results showed that HMM identifies infected cows before the apparition of clinical and subclinical signs which may critically improve the power of studies on the genetic determinants of SCS and reduce biases in predicting breeding values for SCS. The HMM provides also epidemiological parameters that describe the spread of mastitis at the population level.

Acknowledgements

My special thanks go to CRV BV and Prof. H. Barkema for providing the data and to EADGENE (European Animal Disease Genomics Network of Excellence for Animal Health and Food Safety) for financial support.

R eferences

Anderson RM and May RM, Infectious diseases of Humans. Oxford Science Publications, Oxford, 1992.

Altman RM 2007. Mixed hidden Markov model: An extension of the hidden Markov model to the longitudinal data setting. Journal Animal Science Association 102: 201-210.

Barkema, HW, Schukken YH, Lam TJGM, Beiboer ML, Wilmink H, Benedictus G, and Brand A 1998. Incidence of clinical mastitis in dairy herds grouped in three categories by bulk milk somatic cell counts. Journal Dairy Science 81:411-419.

Bilmes JA 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov models. Technical report, University of Berkeley.

Bishop SC, and Wooliams JA 2010. On the genetic interpretation of disease data. PLoS ONE 5(1): e8940. doi:10.1371/journal.pone.0008940

Boettcher PJ, Moroni P, Pisoni G, and Gianola D 2005. Application of finite mixture model to somatic cell scores of Italian goats. Journal Dairy Science 88, 2209-2216.

Brookhart MA, Hubbard AE, van der Laan MJ, Colford JM Jr, Eisenberg JN 2002. Statistical estimation of parameters in a disease transmission model: analysis of a Cryptosporidium outbreak. Statistics in Medicine 21, 3627–3638.

Buyske S, Yang G, Matise TC, Gordon D 2009. When a case is not a case: Effects of phenotype misclassification on power and sample size requirements for the transmission disequilibrium test with affected child trios. Human Heredity 67, 287-292.

Cooper B and Lipstich M 2004. The analysis of hospital infection data using hidden Markov models. Biostatistics 5, 223–237.

de Haas Y, Barkema HW, and Veerkamp RF 2002. The effect of pathogen-specific clinical mastitis on the lactation curve for somatic cell count. Journal Dairy Science 85, 1314-1323.

de Haas Y, Veerkamp RF, Barkema HW, Gröhn YT, and Schukken YH 2004. Associations between pathogen-specific cases of clinical mastitis and somatic cell count patterns. Journal Dairy Science 87, 95–105.

Detilleux JC, and Leroy P 2000. Application of a mixed normal mixture model for the estimation of mastitis-related parameters. Journal Dairy Science 83, 2341–2349.

Detilleux JC, Vangroenweghe F, and Burvenich C 2006. Mathematical model of the acute inflammatory response to *Escherichia coli* intramammary challenge. Journal Dairy Science 89,3455-65.

Detilleux JC 2008. The analysis of disease biomarker data using a mixed hidden Markov model. Genetics Selection Evolution 40, 491-509.

Djabri B, Bareille N, Beaudeau F, and Seegers H 2002. Quarter milk somatic cell count in infected dairy cows: a meta-analysis. Veterinary Research 33, 335-357.

Ephraim Y, and Roberts W 2005. Revisiting autoregressive hidden Markov modeling of speech signals. IEEE Signal Processing Letters 12, 166-169.

Eisner J 2002. An interactive spreadsheet for teaching the forward-backward algorithm. Conference at the ACL-02 Workshop on effective tools and methodologies for teaching natural language processing and computational linguistics, Philadelphia, Pennsylvania, pp 10-18.

Gianola D 2005. Prediction of random effects in finite mixture models with Gaussian components. Journal of Animal Breeding and Genetics 122, 145-159.

Godden SM, Jansen JT, Leslie KE, Smart NL, and Kelton DF 2002. The effect of sampling time and sample handling on the detection of *Staphylococcus aureus* in milk from quarters with subclinical mastitis. Canadian Veterinary Journal 43, 38-42.

Horton NJ and Kleinman KP 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician, 61, 79-90.

Jaakkola T, and Jordan MI 2000. Bayesian parameter estimation via variational methods. Statistics and Computing, 10: 25-37.

Lam T, van Wuijckhuise LA, Franken P, Morselt ML, Hartman EG, and Schukken YH 1996. Use of composite milk samples for diagnosis of *Staphylococcus aureus* mastitis in dairy cattle. Journal America Veterinary Medical Association 208, 1705–1708.

Laverty WH, Miket MJ, and Kelly IW 2002. Simulation of hidden Markov models with EXCEL. The statistician 51, 31-40.

Le Strat Y, and Carrat F 1999. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in Medicine 18, 3463-3478.

Moroni P, Pisoni G, Vimercati C, Rinaldi M, Castiglioni B, Cremonesi P, and Boettcher P 2005. Characterization of *Staphylococcus aureus* isolated from chronically infected dairy goats. Journal Dairy Science 88, 3500 – 3509.

Paape M, Mehrzad J, Zhao X, Detilleux J, and Burvenich C 2002. Defense of the bovine mammary gland by polymorphonuclear neutrophil leukocytes. Journal Mammary Gland Biology Neoplasia 7, 109–121.

Rabiner LR 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77:257-268, 1989.

Sargeant JM, Leslie KE, Shirley JE, Pulkrabek BJ, and Lim GH 2001. Sensitivity and specificity of somatic cell count and California mastitis test for identifying intramammary infection in early lactation. Journal Dairy Science 84, 2018-2024.

SAS OnlineDoc™:Version 8. Statistical Analysis System.

Sears PM, Smith BS, English PB, Herer PS, and Gonzales RN 1990. Shedding pattern of *Staphylococcus aureus* from bovine intramammary infections. Journal Dairy Science 73, 2785–2789.

Shook GE and Schutz MM 1994. Selection on somatic cell score to improve resistance to mastitis in the United States. Journal of Dairy Science 77, 648-658.

Suriyasathaporn W, Schukken YT, Nielen M, and Brand A. 2000. Low somatic cell count: a risk factor for subsequent clinical mastitis in a dairy herd. Journal Dairy Science 83, 1248-1255.

Figure 1 Monthly average somatic cell scores for all lactations (square) and for lactations without clinical mastitis (cross), before (straight line) and after (broken lines) imputation.

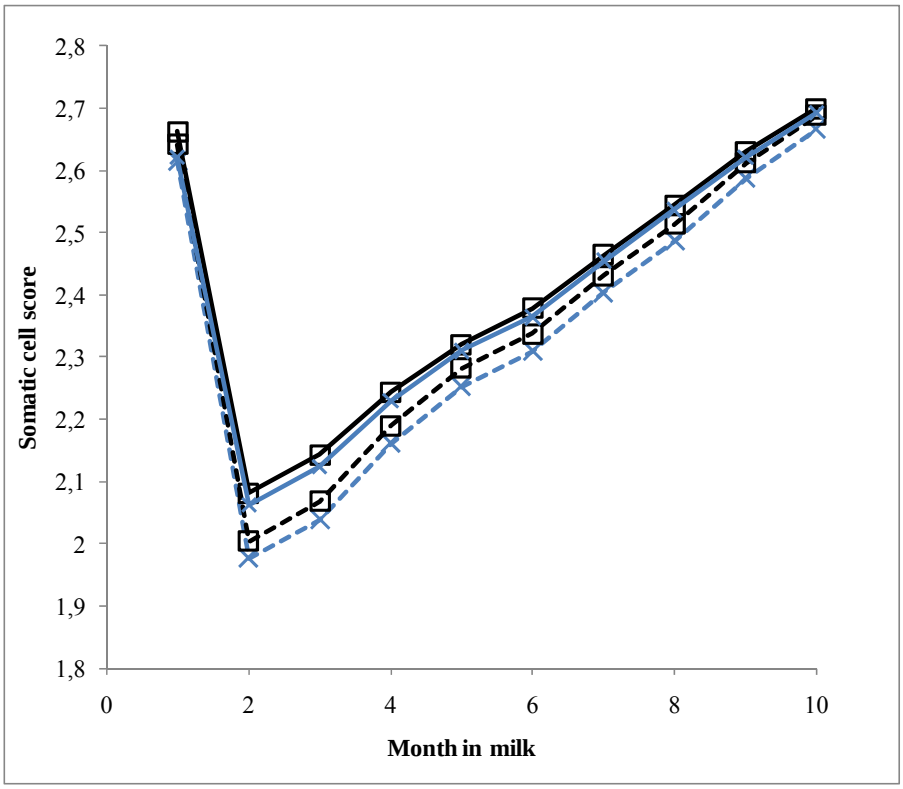
Figure 2 Mean number of transitions from one mastitis state to another across the 10 test-days for cows without (Figure 2a) and with (Figure 2b) at least one reported clinical case of mastitis. States are designed as CM+ or CM- (plain bar) when a clinical case is reported or not, as SCM+ or SCM- (spotted bar) when SCC are above or below 250,000 cells/ml, and IMI+ or IMI- (stripped bar) when the sample is classified as infected or not by the model, respectively.

Figure 3 Average probabilities of transition between IMI states for lactations with (stripped bar) and without (plain bar) a reported clinical case. The hidden states are IMI+ and IMI- when the sample is classified as infected or not by the model, respectively.

Figure 4 Examples of sequences for the IMI, SCM and CM stages across the 10 test-days based on the results shown in Figure 2, for cows without (Figure 4a) and with (Figure 4b) at least one reported clinical case of mastitis. The sign is + when the sample is positive for the stage at the test day. The sign is - when the sample is negative for the stage at the test day.

1 Figure 1.

2



3

4

Figure 2.

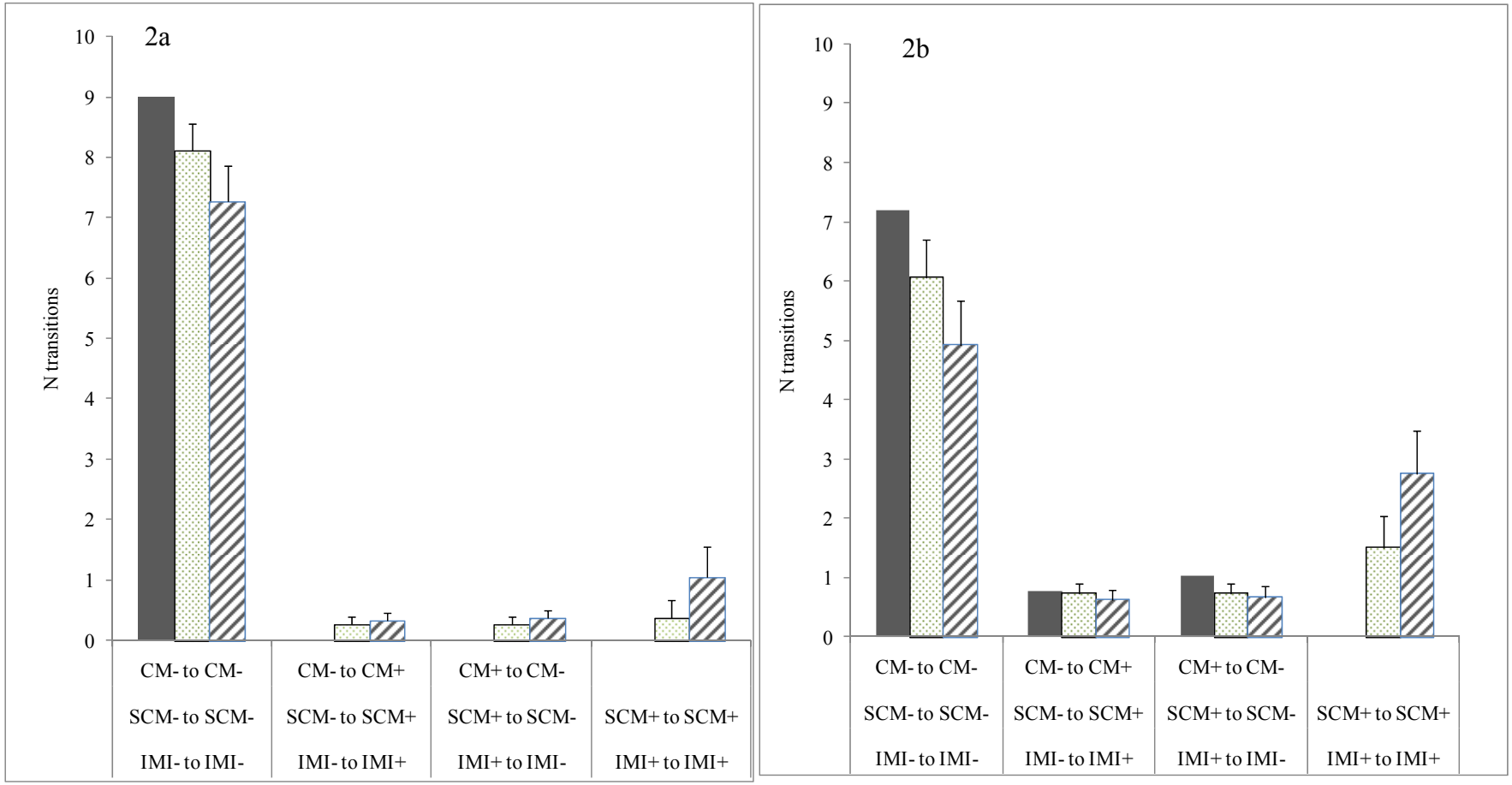
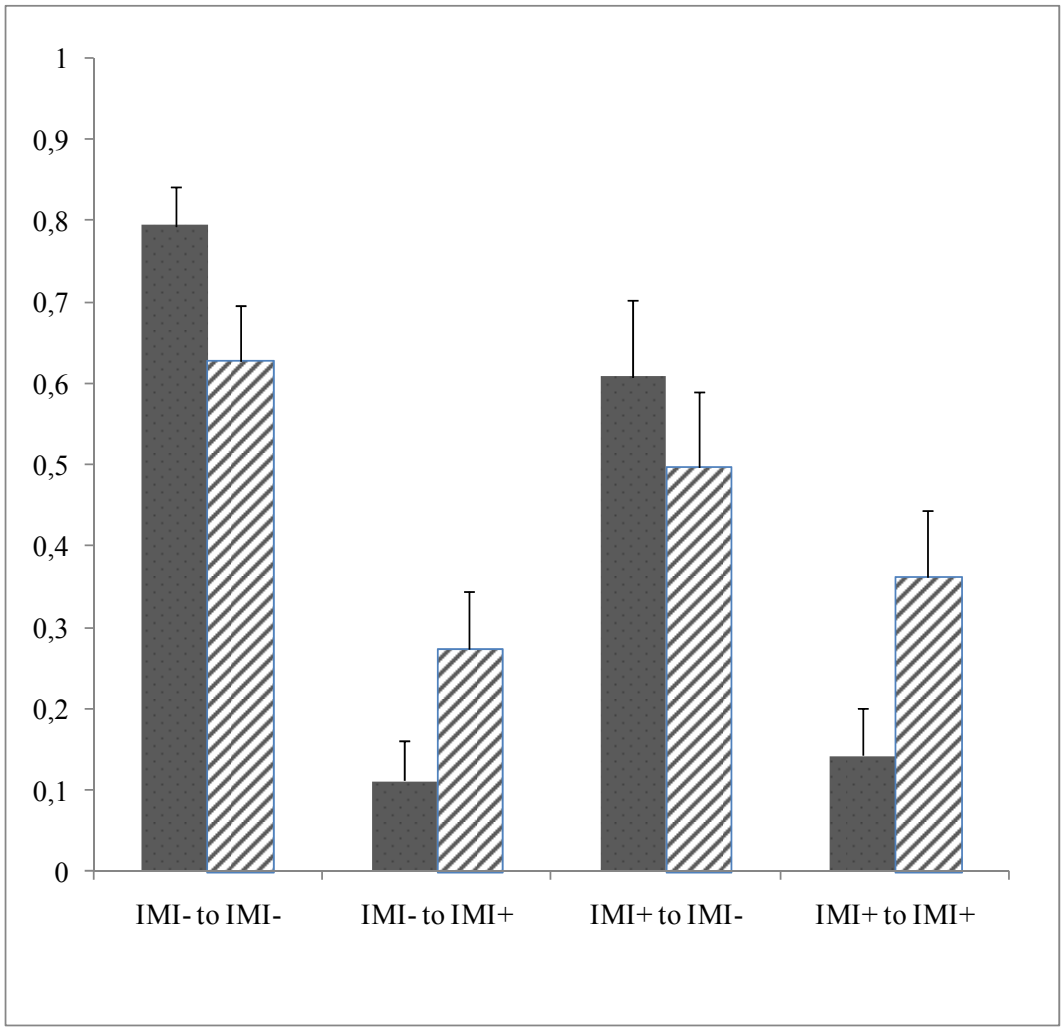
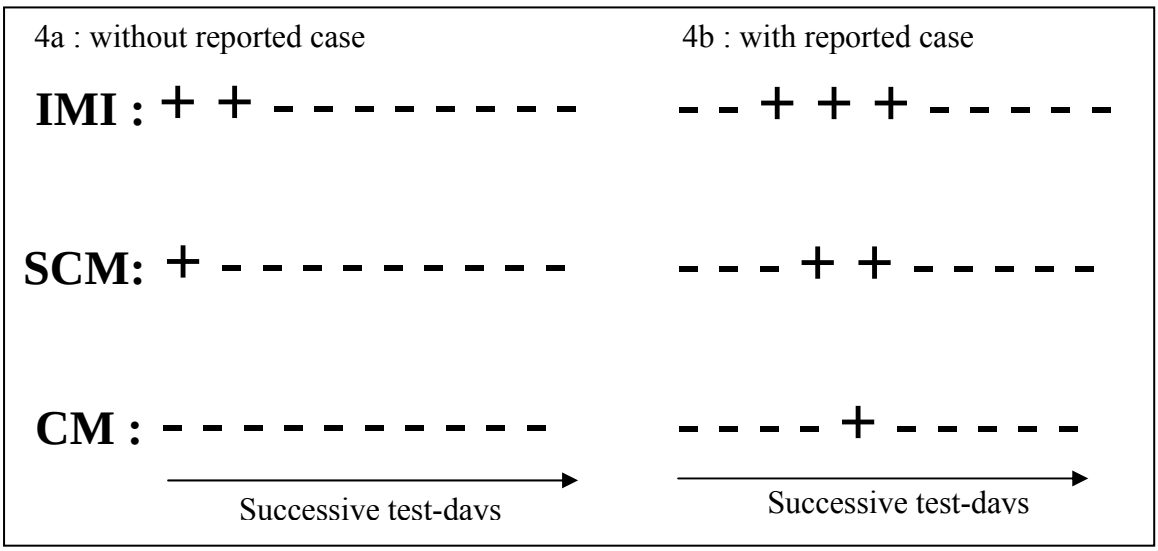

2a
N transitions
10
9
8
7
6
5
4
3
2
1
0
CM- to CM-
SCM- to SCM-
IMI- to IMI-
CM- to CM+
SCM- to SCM+
IMI- to IMI+
CM+ to CM-
SCM+ to SCM-
IMI+ to IMI-
SCM+ to SCM+
IMI+ to IMI+
2b
N transitions
10
9
8
7
6
5
4
3
2
1
0
CM- to CM-
SCM- to SCM-
IMI- to IMI-
CM- to CM+
SCM- to SCM+
IMI- to IMI+
CM+ to CM-
SCM+ to SCM-
IMI+ to IMI-
SCM+ to SCM+
IMI+ to IMI+

Figure 3


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
IMI- to IMI-
IMI- to IMI+
IMI+ to IMI-
IMI+ to IMI+

1 Figure 4.

2


4a : without reported case
4b : with reported case
IMI : + + - - - - - - - -
- - + + + - - - - -
SCM: + - - - - - - - - -
- - - + + - - - - -
CM : - - - - - - - - - -
- - - - + - - - - -
Successive test-davs
Successive test-davs