A mixed hidden Markov model for analyzing somatic cell scores

J. Detilleux
University of Liège
Faculty of veterinary medicine

jdetilleux@ulg.ac.be
Outline

- Introduction
- Mixed Hidden Markov model
- Simulation
- Conclusions
1. Introduction

Biological marker
Accuracy of available marker
Resistance/tolerance
1a. Biological marker

- Reduction of bovine mastitis prevalence (eg. breeding for disease resistance)

- Early detection of mammary infection (IMI)
  - Biomarker = objective indicator of disease state (eg., SCC, M-SAA3, Hp, LDH, NAGase, ...)
  - Surrogate endpoint = substitute for disease endpoint (eg. early predictor of infection, survival, or clinical signs)

→ mathematical models to estimate \( \text{pr}(\text{IMI}) \)
1b. Accuracy

Frequency distribution of SCC for IMI- or IMI+ quarters

\[ \hat{a} = G Z' V^{-1} (y - \mu) \]

Imperfect detectability

→ misleading info on transmission dynamics
  (prevalence, incidence, association with disease, ...)
1c. Resistance/ tolerance

- Resistance = low proba of infection
- Tolerance = little fitness loss after infection

Breeding goals?

+ herd immunity
- natural selection

+ disease spread
+ natural selection
2. Mixed hidden Markov model

General formulation
Bayes estimation
2a. Formulation

\[
Pr(\text{SCC})_t = \text{pr}(\text{IMI}^-)_t \times \text{pr}(\text{SCC}|\text{IMI}^-)_t \\
+ \text{pr}(\text{IMI}^+)_t \times \text{pr}(\text{SCC}|\text{IMI}^+)_t
\]

\[
\text{Tolerance: } \text{IMI}^+_t \rightarrow \text{IMI}^-_t
\]

\[
\text{Resistance: } \text{IMI}^+_t \rightarrow \text{IMI}^+_t
\]

\[
\text{Recovery: } \text{IMI}^+_t \rightarrow \text{IMI}^-_t
\]

\[
\text{Recurrence: } \text{IMI}^-_{t-1} \rightarrow \text{IMI}^+_t
\]

\[
\text{Persistence}^+: \text{IMI}^+_t \rightarrow \text{IMI}^+_t
\]

\[
\text{Persistence}^-: \text{IMI}^-_{t-1} \rightarrow \text{IMI}^-_t
\]
Hidden states:

- IMI+
- IMI-
- IMI-
- IMI-

Observed sequences:

- SCC
- SCC
- SCC
- SCC

Time:

- t=1
- t=2
- t=3
- t=4

Emission probability

Transition probability

Genetic values
Time:

\begin{align*}
t &= 1 \\
t &= 2 \\
t &= 3 \\
t &= 4
\end{align*}

\begin{align*}
Z_k^0 &\rightarrow Z_k^1 \\
Z_k^1 &\rightarrow Z_k^2 \\
Z_k^2 &\rightarrow Z_k^3 \\
Z_k^3 &\rightarrow Z_k^4
\end{align*}

\begin{align*}
y_k^1 &= \text{value of biomarker at } t^{th} \text{ time on } k^{th} \text{ cow} \\
Z_k^t &= 0 \text{ if IMI-} \\
Z_k^t &= 1 \text{ if IMI+}
\end{align*}
- **Output independence**: observations are independent given the unknown IMI state
  \[ p(y_k^t \mid z_k^t, y_{k-1}^t, y_{k-2}^t, \ldots) = p(y_k^t \mid z_k^t) \]

- **Time is discrete**

- **Markov property**: the next state depends only on the current state
  \[ p(z_{k+1}^t \mid z_k^t, z_{k-1}^t, \ldots, z_{k}^1) = p(z_{k+1}^t \mid z_k^t) \]
  
  "Conditioned on the present, the past & future are independent"

- **Stationarity**: transition probabilities are time invariant
  \[ p(z_{k+1}^t = i \mid z_k^t = j) = a_{ij}^t \]
\[ p(y | \mu_0, \mu_1, \sigma_0^2, \sigma_1^2, a, z) \sim N(M_0\mu_0 + M_1\mu_1 + Za, R) \]

- \( y \) = (NT X 1) vector of data
- \( z \) = (NT X 1) vector of hidden states
- \( \mu_0 \) = (T X 1) vector of fixed effects for data on a IMI- cow
- \( \mu_1 \) = (T X 1) vector of fixed effects for data on a IMI+ cow
- \( a \) = (N X 1) vector of random additive genetic effects

- \( M_0 \) = (NT X T) matrix with elements = 1 if \( z_k^t = 0 \)
- \( M_1 \) = (NT X T) matrix with elements = 1 if \( z_k^t = 1 \)
- \( Z \) = (NT X N) incidence matrix relating \( a \) to \( y \).

\[ R \sim J_0\sigma_0^2 + J_1\sigma_1^2 \quad J_i = (NT \times NT) \text{ matrix with elements } = 1 \text{ if } z_k^t = i \]
2b. Estimation

Prior distributions

\[
\begin{align*}
\mu_{0} & \sim N (1 m_0, I s_0^2) \\
\mu_{1} & \sim N (1 m_1, I s_1^2) \\
\alpha & \sim N (0, A \sigma^2_a) \\
\sigma^2_a & \sim \chi^{-2} (s^2_a) \\
\sigma^2_0 & \sim \chi^{-2} (s^2_0) \\
\sigma^2_1 & \sim \chi^{-2} (s^2_1) \\
\end{align*}
\]

\[
\begin{align*}
z_k^0 & \sim Br(\lambda_k) \\
(z_k^t = 0 \mid z_{k-1}^t = 0) & \sim Br(\pi_{k}^{00}) \\
(z_k^t = 1 \mid z_{k-1}^t = 0) & \sim Br(\pi_{k}^{01}) \\
\lambda_k, \pi_{k}^{00}, \pi_{k}^{01} & \sim Un[0,1]
\end{align*}
\]
Joint posterior distribution

\[ \theta = (\mu_0, \mu_1, \sigma_a^2, \sigma_0^2, \sigma_1^2, \pi^{00}, \pi^{01}, z, a, \lambda) \]

\[ p(\theta \mid y) = \sum_{z} p(y \mid \theta) p(\theta) \]

nb hidden sequences per cow = 2^T
\[ \Rightarrow \text{total number of operations} \sim 4N^T \]

Use of the trellis structure of HMM
\[ \Rightarrow \text{Forward-backward algorithm} \]
The algorithm takes on the order of $4T$ computations.
**Forward probabilities**: proba that, given $H$, at time $t$, the state is $i$ and the sequence of partial observation ($y_1 \ldots y_t$) has been generated.

$$\alpha^t_k(i) = p(y^1 \ldots y^t, z^t_k = i)$$

**3 steps**
- induction ($t=0$)
- recursion (increasing $t$)
- termination ($t=T$)
Backward probabilities: proba that, given H and given the state $i$ at time $t$, a sequence of partial observation $(y_{t+1} \ldots y_T)$ has been generated

$$\beta_k^t(i) = p(y_k^T \ldots y_{k+1}^t | z_k^t = i)$$

3 steps
- induction ($t=T$)
- recursion (decreasing $t$)
- termination ($t=1$)
Fully conditional distributions

\[
\mu_i^t \sim N \left( \frac{s_i^2 \sum_{k}^{N} (y_k^t - a_k) I_{k_i}^t + m_i \sigma_i^2}{s_i^2 \sum_{k}^{N} n_{i,k}^t + \sigma_i^2} \right), \quad \frac{s_i^2 \sigma_i^2}{s_i^2 \sum_{k}^{N} n_{i,k}^t + \sigma_i^2}
\]

\[
a \sim N(\widehat{\theta}_1, C_{11}^{-1}) \quad \tilde{\theta}_1 = C_{11}^{-1} [r_1 - C_{12} \theta_2]
\]

\[
(\sigma_a^2) \sim \chi_{N+v}^{-2} (a' A^{-1} a + \nu s_a^2)
\]

\[
(\sigma_i^2) \sim \chi_{N_i+v}^{-2} [\nu s_i^2 + (y - M_i \underline{\mu}_i - Z a)' J_i (y - M_i \underline{\mu}_i - Z a)]
\]

\[
\lambda_k \sim \text{beta}(I_{k_i}^{0,1} + 1; I_{k_i}^{1,1} + 1)
\]

\[
\pi_{k_0}^{00} \sim \text{beta}(n_{k_0}^{00} + 1, n_{k_0}^{10} + 1) \quad \pi_{k_0}^{01} \sim \text{beta}(n_{k_0}^{01} + 1, n_{k_0}^{11} + 1)
\]
\[ z_{k_1}^t \sim \text{Br}(\zeta_{0,k}^1) \quad \zeta_{0,k}^0 = p[z_k^0 = 0 \cap y_{\underline{k}_1}] = \alpha_k^1(0) \beta_k^1(0) \]

It is the probability of being healthy at the start with the observation sequence \( y_1 y_2 \ldots y_T \).

\[ z_{k_1}^t \sim \text{Br}(\zeta_{ij,k}^t) \quad \xi_{ij,k}^t = p[z_k^t = i | z_k^{t-1} = j, y_{\underline{k}_1}] = \frac{\alpha_{k_1}^{t-1}(j) \pi_{k_1}^{ij} \beta_k^t(i) \text{pr}(y_k^t | z_k^t = i)}{\alpha_{k_1}^{t-1}(j) \beta_k^{t-1}(j)} \]

It is the probability of being in state \( i \) at time \( t \) given state \( j \) at time \( t-1 \) and the observation sequence \( y_1 y_2 \ldots y_T \).
4. Simulation

Survey of SCC
Pathogen
Severity of response
Genetics
Data sets
Accuracy
4a. Survey  (de Haas et al., 2002, 2004)

All lactations

No clinical case

No case (clin. or subclin.)
4b. Pathogen

Coli

Aureus
4c. Severity of response

Severe (>50%)

Moderate (<50%)

Jalil Mehrzad\textsuperscript{a,b,c}, Luc Duchateau\textsuperscript{a}, Christian Burvenich\textsuperscript{a*}

4d. Genetics

- 3 discrete generations with 400 cows per generation
- sires selected from 30 different bulls
- each cow was replaced by a daughter
- mating at random

- breeding values for base animals $\sim N(0, I \sigma_a^2)$
- additive variance of 0.15 or 0.25.
- breeding values for non-base animals $\sim N(\text{mid-parent}, \sigma_a^2/2)$
- No selection, no inbreeding
4e. Simulated data sets

\[ \mu_0^t \text{ for } t = 1 \text{ to } T \]

\[ \mu_1^t \text{ for } t = 1 \text{ to } T \]
Simulated data sets:

- % infected cows = 20, 50%
- % *E. coli* among infected cows = 0, 50, 100%
- high and moderate responders: $\mu_1^n$
- $\sigma^2_0 = 1.0$ or $1.4 \rightarrow$ residuals IMI- $\sim N(0, \sigma^2_0)$
- $\sigma^2_a = 0.15$ or $0.25$

**Gibbs sampler**

- 1000 iterations
- 200 burn-in
- 10 replications
4f. Accuracy

- $\text{corr}(a, \hat{a})$

- differences: $\theta - \hat{\theta}$ for $\mu_0, \mu_1, \sigma_a^2, \sigma_0^2, \sigma_1^2$

- sensitivity: $\text{SE} = \sum_{k=1}^{N} \sum_{t=1}^{T} \text{pr}(\hat{z}_k^t = 1 | z_k^t = 1)$

- specificity: $\text{SP} = \sum_{k=1}^{N} \sum_{t=1}^{T} \text{pr}(\hat{z}_k^t = 0 | z_k^t = 0)$

- proba of correct classification:

$$\text{PCC} = \sum_{k=1}^{N} \sum_{t=1}^{T} \text{pr}(z_k^t = 1 \cap \hat{z}_k^t = 1) \cup (z_k^t = 0 \cap \hat{z}_k^t = 0)$$
Proportion of *E. coli* among infected cows

- Sensitivity
- Specificity
- Correct classification

Proportion of *E. coli* among infected cows
High SE = SNOT

→ No further test
Theoretical correlation
SP and PCC decreased when % *E. coli* among infected increased.

Proportion of *E. coli* among infected cows
**Difference between true and estimated means**

- **Proportion of infected cows**: 20%, 50%
- **Difference**: IMI-, IMI+

**Biases increased when disease prevalence is low**

**Difference between true and estimated variances**

- **% of infected cows**: 20%, 30%, 40%, 50%, 70%
- **Difference**: -1.2, -1, -0.8, -0.6, -0.4

**Biases increased when disease prevalence is high**
Conclusions
- **Same amount of data**
  - Increased accuracy of MLE
  - Resistance and tolerance
  - Transition probabilities

- **Simplification of reality**
  - Age, season, herd, ..
  - ‘Isolated’ proba of IMI-
  - Genetic relationship between cows for IMI
Genetic relationship among IMI and SCC
Thank you!