Modelling individual patterns of somatic cell scores to derive mammary infection status

J. Detilleux
University of Liège
Faculty of veterinary medicine

jdetilleux@ulg.ac.be

Outline

- Introduction
- Finite mixture model
- Hidden Markov model
- Simulation
- Conclusions

1. Introduction

Biological marker

Accuracy of available marker

Resistance/tolerance

1a. Biological marker

- Reduction of bovine mastitis prevalence
  Opposition to disease transmission

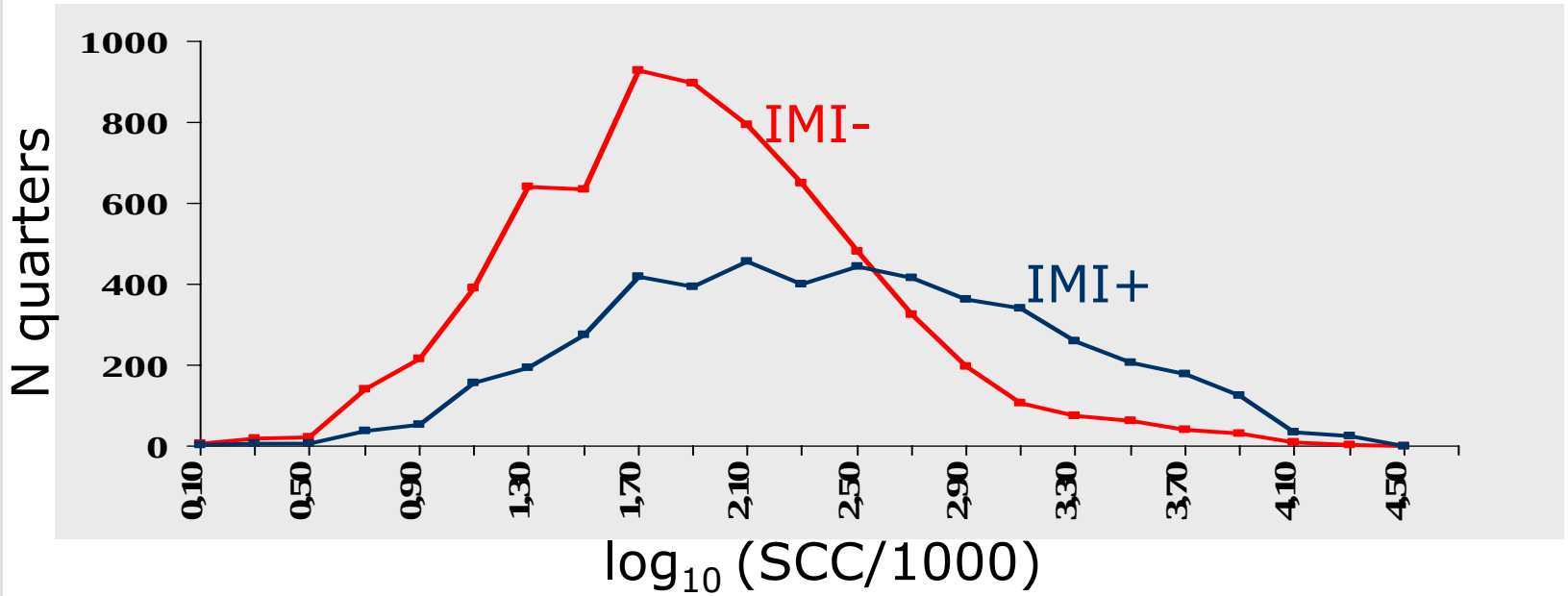
- Early detection of mammary infection (IMI)

  - 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 pr(IMI)

1b. Accuracy

Frequency distribution of SCC for IMI- or IMI+ quarters



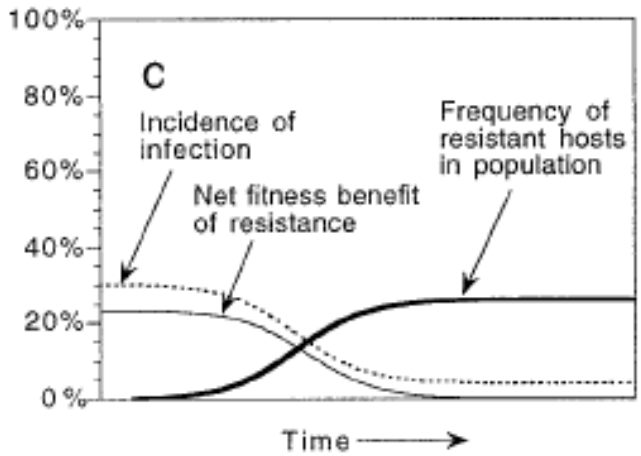
Imperfect detectability

→ misleading info on transmission dynamics
(prevalence, incidence, association with disease, …)

1c. Resistance/ tolerance

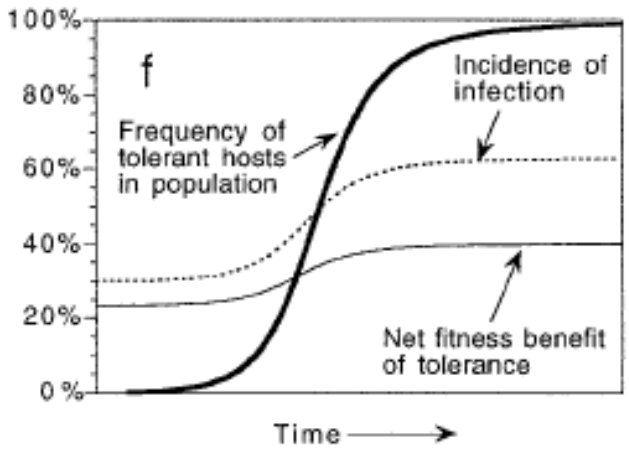
B. A. Roy[1] and J. W. Kirchner[2] *Evolution*, 54(1), 2000, pp. 51–63

- Resistance = low proba of infection



+ herd immunity
- natural selection

- Tolerance = little fitness loss after infection



- disease spread
+ natural selection

2. Finite mixture model

General formulation

Likelihood

EM algorithm

2a. Formulation of FMM

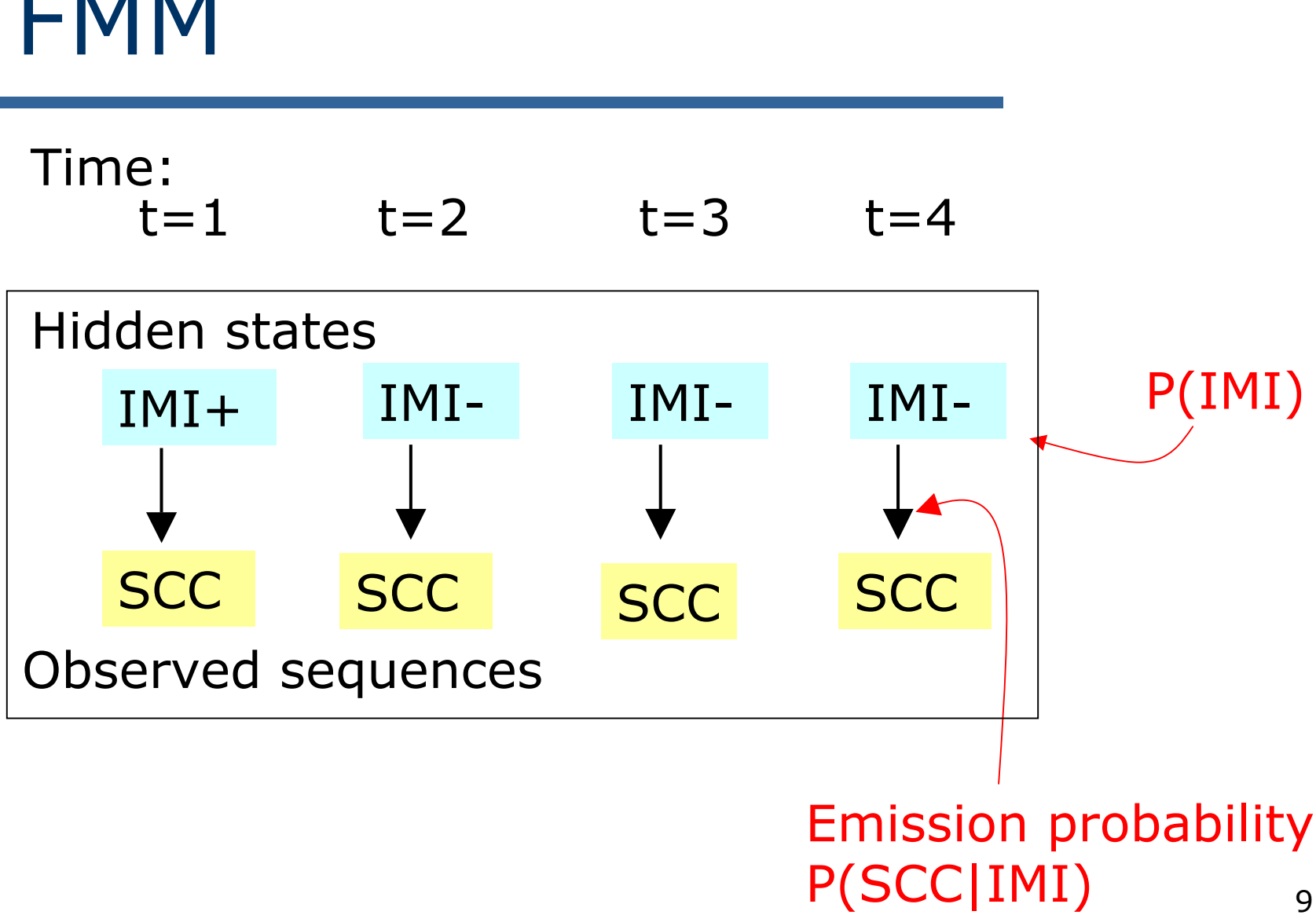
$$Pr(SCC) = pr(SCC \,\Omega\, IMI^{-}) + pr(SCC \,\Omega\, IMI^{+})$$

$$Pr(SCC) = pr(IMI^{-}) * pr(SCC|IMI^{-}) + [1 - pr(IMI^{-})] * pr(SCC|IMI^{+})$$

- Resistance
- Surrogate

- Tolerance
- Accuracy

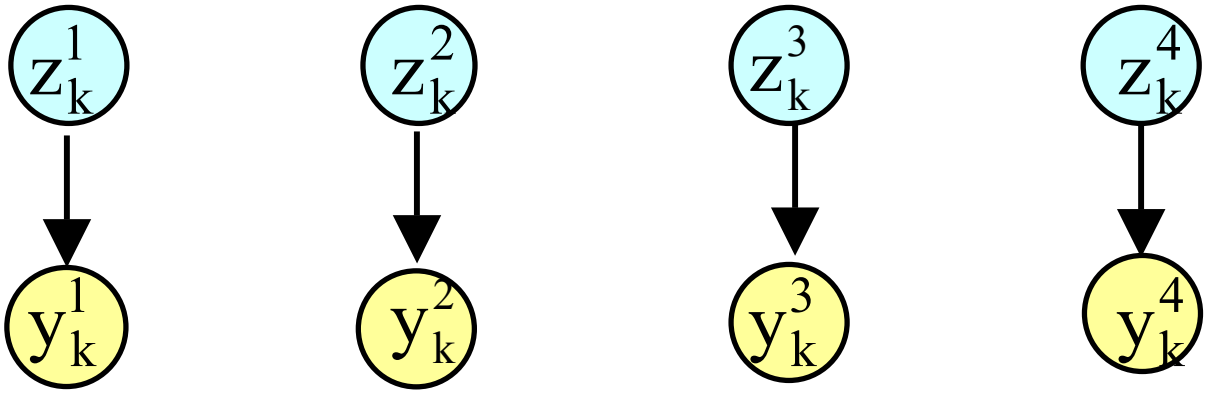
FMM


Time:
t=1
t=2
t=3
t=4
Hidden states
IMI+
IMI-
IMI-
IMI-
SCC
SCC
SCC
SCC
Observed sequences
P(IMI)
Emission probability:
P(SCC|IMI)

FMM

Time:

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

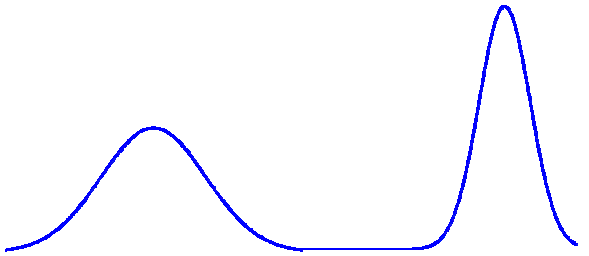


$$p(\underline{y}_k) = p(y_k^1, y_k^2, \ldots, y_k^T)$$
$$= \prod_{t=1}^{T} p(y_k^t \mid z_k^t = 0)p(z_k^t = 0) + p(y_k^t \mid z_k^t = 1)p(z_k^t = 1)$$

FMM

$F = (\pi_\kappa, \mu_0, \mu_1, \sigma^2_0, \sigma^2_1)$

$$p(y_k^t \mid z_k^t = 0) \sim N(\mu_0^t, \sigma_0^2)$$
$$p(y_k^t \mid z_k^t = 1) \sim N(\mu_1^t, \sigma_1^2)$$



$$p(z_k^t = 0) \sim Bi(\pi_k)$$

cows are independent: $p(\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N) = \prod_{k=1}^{N} p(\underline{y}_k)$

2b.Maximum likelihood estimation

Likelihood for one sequence (one cow):

$$p(\underline{y}_k \mid F) = p(y_k^1, y_k^2, \ldots, y_k^T \mid F) = \sum_{\underline{z}} p(y_k^1, y_k^2, \ldots, y_k^T \mid \underline{z}, F) \times p(\underline{z} \mid F)$$

$$p(y_k^1, y_k^2, \ldots, y_k^T \mid \underline{z}, F)$$

$$= p(y_k^1, y_k^2, \ldots, y_k^T \mid z_k^1, z_k^2, \ldots, z_k^T, F)$$

$$= p(y_k^1 \mid z_k^1, z_k^2, \ldots, z_k^T, F)\, p(y_k^2 \mid y_k^1, z_k^1, z_k^2, \ldots, z_k^T, F) \ldots$$

$$= p(y_k^1 \mid z_k^1, F)\, p(y_k^2 \mid z_k^2, F) \ldots p(y_k^T \mid z_k^T, F)$$

$$= \prod_{t=1}^{T} p(y_k^t \mid z_k^t, F)$$

FMM

$$p(\underline{z} \mid F) = p(z_k^1, z_k^2, \ldots, z_k^T \mid F)$$

$$= p(z_k^1 \mid F)\, p(z_k^2 \mid z_k^1, F)\, p(z_k^3 \mid z_k^2, z_k^1, F) \ldots p(z_k^T \mid z_k^{T-1}, z_k^{T-2}, \ldots, z_k^1, F)$$

$$= p(z_k^1 \mid F)\, p(z_k^2 \mid F)\, p(z_k^3 \mid F) \ldots p(z_k^T \mid F) = \prod_{t=1}^{T} p(z_k^t \mid F)$$

$$p(y_k^1, y_k^2, \ldots, y_k^T \mid F) = \sum_{\underline{z}} \prod_{t=1}^{T} p(y_k^t \mid z_k^t, F)\, p(z_k^t \mid F)$$

$$= \prod_{t=1}^{T} p(y_k^t \mid z_k^t = 0, F)\, p(z_k^t = 0 \mid F) + \prod_{t=1}^{T} p(y_k^t \mid z_k^t = 1, F)\, p(z_k^t = 1 \mid F)$$

Likelihood for N cows:

$$p(\underline{y} \mid F) = \prod_{k=1}^{N} p(\underline{y}_1, .., \underline{y}_N \mid F)$$

2c. EM algorithm

E step:

$$Q^{(p-1)} = E[\log[p(\underline{y}_1,\ldots,\underline{y}_N;\underline{z}_1,\ldots,\underline{z}_{Nk}^{T} \mid \Theta) \mid \underline{y}_1,\ldots,\underline{y}_N;\Theta^{(p-1)}]$$

$$= \sum_{k=1}^{N}\sum_{t=1}^{T}\left[\begin{array}{l} E[z_k^t = 0 \mid \underline{y}_k]\{\log(\pi\{+\log p(y_k^t \mid z_k^t = 0)\} \\ +E[z_k^t = 1 \mid \underline{y}_k]\{\log(1-\pi)+\log p(y_k^t \mid z_k^t = 1)\} \end{array}\right]$$

$$E[z_k^t = i \mid \underline{y}_k] = \frac{\pi_k\ N(\mu_i^t, \sigma_i^2)}{\pi_k\ N(\mu_0^t, \sigma_0^2) + (1-\pi_k)N(\mu_1^t, \sigma_1^2)} = \zeta_{i,k}^t$$

It is the probability of being in state i at time t, given the SCS sequence $y_1\ y_2 \ldots y_T$.

FMM

M step: $\Theta^{(p)} = \text{argmax } Q^{(p-1)}$

$$\hat{\pi}_k = \frac{\sum_{t=1,T} \zeta^t_{0,k}}{T}$$

= expected frequency in state IMI- for the cow

$$\mu^t_i = \frac{\sum_{k=1,N} \zeta^t_{i,k}\, y^t_k}{\sum_{k=1,N} \zeta^t_{i,k}}$$

$$\hat{\sigma}^2_i = \frac{\sum_{k=1,N} \sum_{t=1,T} \zeta^t_{i,k} (y^t_k - \hat{\mu}^t_i)^2}{\sum_{k=1,N} \sum_{t=1,T} \zeta^t_{i,k}}$$

= weighted mean and variance

3. Hidden Markov model

General formulation

Likelihood

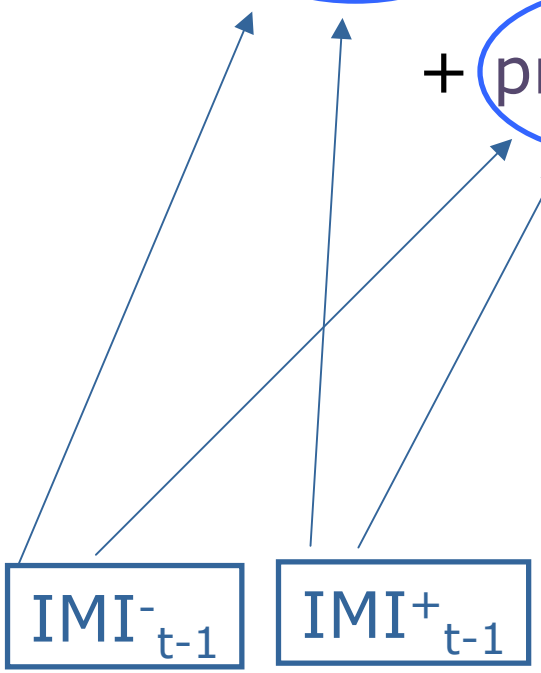
EM algorithm

2a. Formulation

$$Pr(SCC)_t = pr(IMI^-)_t * pr(SCC|IMI^-)_t + pr(IMI^+)_t * pr(SCC|IMI^+)_t$$

Tolerance

Resistance

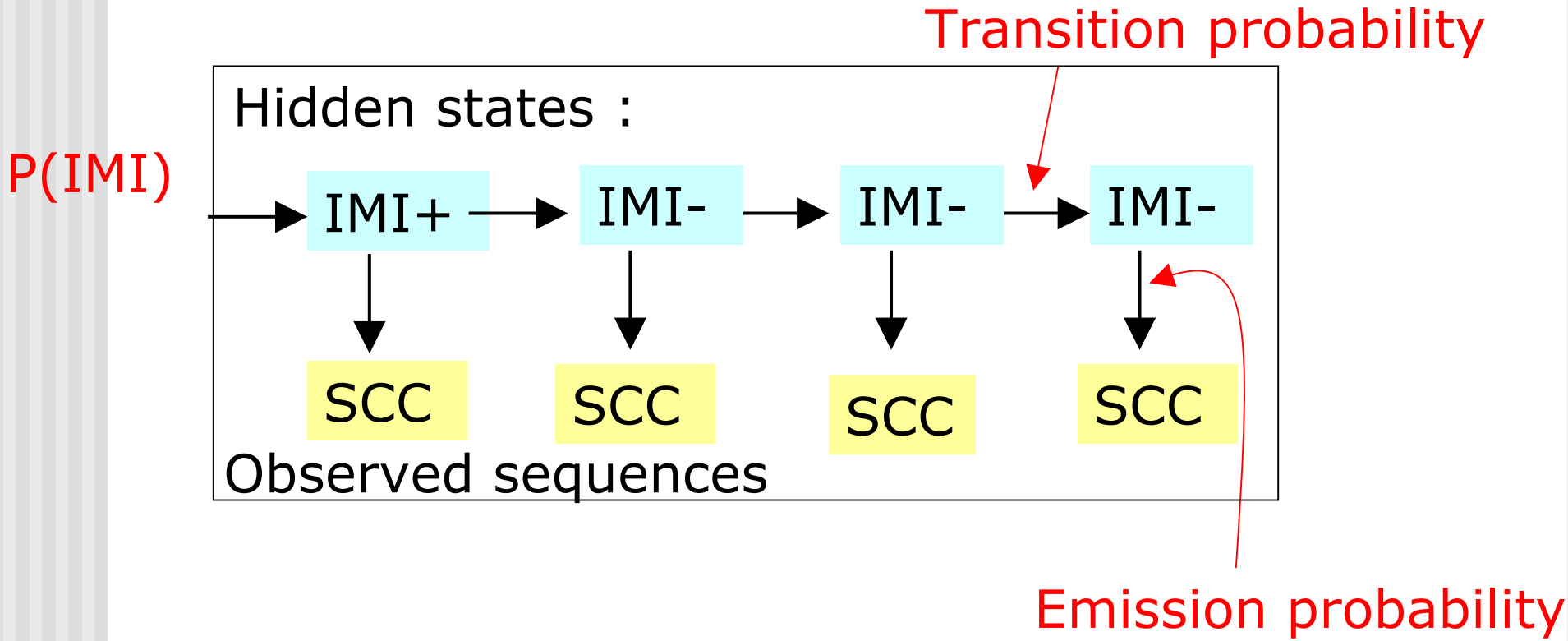


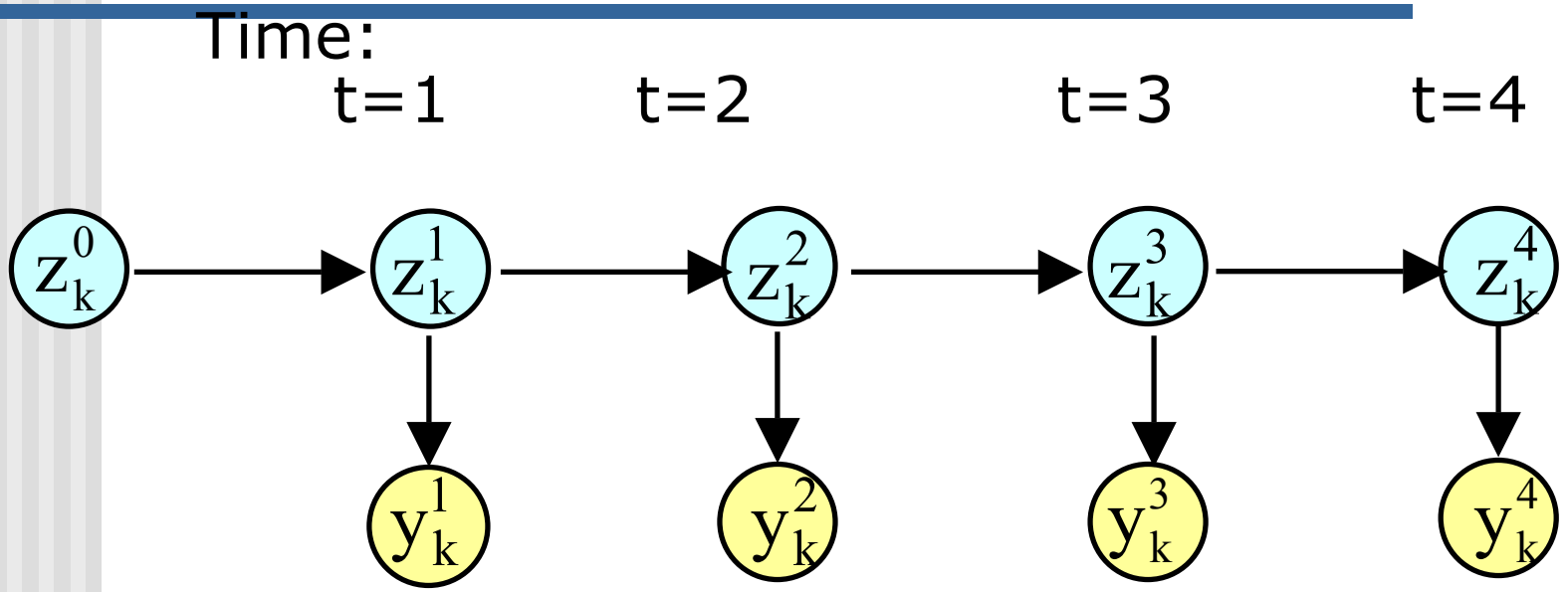
Recovery: $IMI^+_{t-1} \rightarrow IMI^-_t$

Recurrence: $IMI^-_{t-1} \rightarrow IMI^+_t$

Persistence$^+$: $IMI^+_{t-1} \rightarrow IMI^+_t$

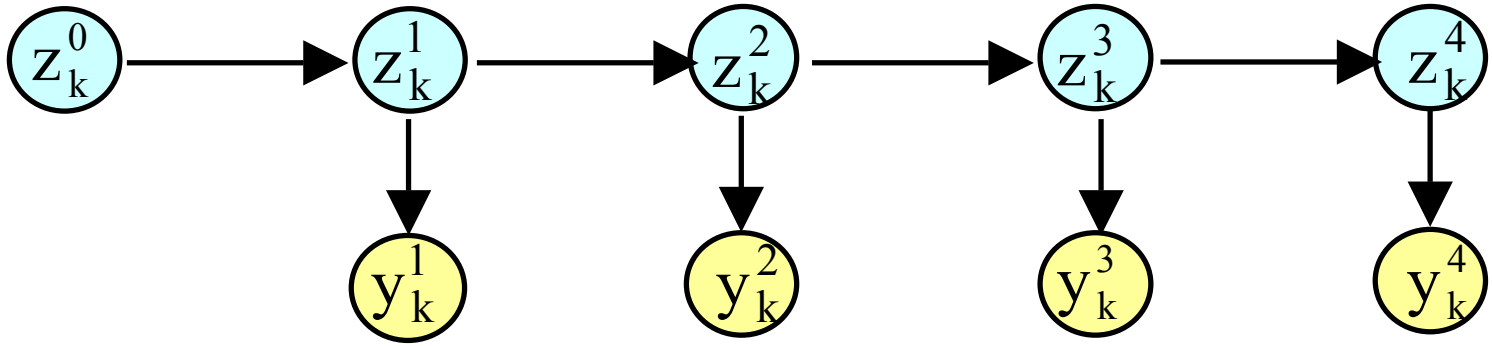
Persistence$^-$: $IMI^-_{t-1} \rightarrow IMI^-_t$


Time:
t=1
t=2
t=3
t=4
Transition probability
Hidden states :
P(IMI)
IMI+
IMI-
IMI-
IMI-
SCC
SCC
SCC
SCC
Observed sequences
Emission probability



y_k^t = value of biomarker at t^{th} time on k^{th} cow

z_k^t = 0 if IMI- z_k^t = 1 if IMI+



$$p(\underline{y}_k) = p(y_k^1, y_k^2, \ldots, y_k^T)$$

$$= p(z_k^0 = 0)$$

$$+ \prod_{t=1}^{T} p(y_k^t \mid z_k^t = 0)[p(z_k^t = 0 \mid z_k^{t-1} = 0) + p(z_k^t = 0 \mid z_k^{t-1} = 1)]$$

$$+ p(y_k^t \mid z_k^t = 1)[p(z_k^t = 1 \mid z_k^{t-1} = 0) + p(z_k^t = 1 \mid z_k^{t-1} = 1)]$$

- Output independence: observations are independent given the unknown IMI state

$$p(y_k^t \mid z_k^t, y_k^{t-1}, y_k^{t-2}, ..) = p(y_k^t \mid z_k^t)$$

- Time is discrete
- Markov property: the next state depends only on the current state

$$p(z_k^{t+1} \mid z_k^t, z_k^{t-1}, ..., z_k^1) = p(z_k^{t+1} \mid z_k^t)$$

"Conditioned on the present,the past & future are independent"

- Stationarity: transition probabilities are time invariant

$$p(z_k^{t+1} = i \mid z_k^t = j) = a_k^{ij}$$

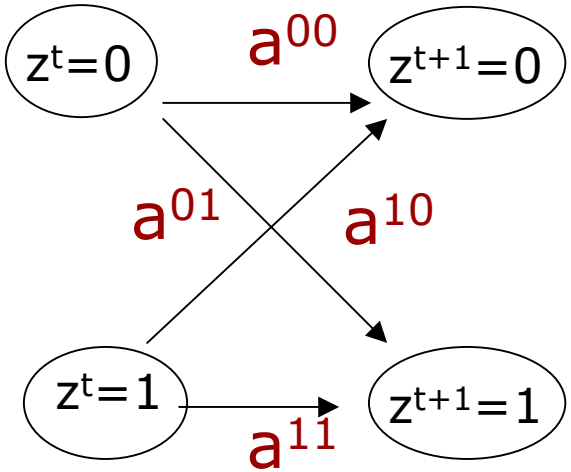
$H = (A, \mu_0, \mu_1, \sigma^2_0, \sigma^2_1, \lambda_k)$

$$p(y_k^t \mid z_k^t = 0) \sim N(\mu_0^t, \sigma_0^2)$$

$$p(y_k^t \mid z_k^t = 1) \sim N(\mu_1^t, \sigma_1^2)$$

$$p(z_k^t = 1) \sim Bi(\lambda_k)$$

cows are independent: $p(\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N) = \prod_{k=1}^{N} p(\underline{y}_k)$



$$p(z_k^t = i \mid z_k^{t-1} = j) = a_k^{ij}$$

$$\rightarrow A_k = \begin{pmatrix} a_k^{00} & a_k^{01} \\ a_k^{10} & a_k^{11} \end{pmatrix}$$

3b. Maximum likelihood estimation

Likelihood for one sequence (one cow):

$$p(\underline{y}_k \mid H) = p(y_k^1, y_k^2, \ldots, y_k^T \mid H) = \sum_{\underline{z}} p(y_k^1, y_k^2, \ldots, y_k^T \mid \underline{z}, H) \times p(\underline{z} \mid H)$$

$$p(y_k^1, y_k^2, \ldots, y_k^T \mid \underline{z}, H)$$

$$= p(y_k^1, y_k^2, \ldots, y_k^T \mid z_k^1, z_k^2, \ldots, z_k^T, H)$$

$$= p(y_k^1 \mid z_k^1, z_k^2, \ldots, z_k^T, H)\, p(y_k^2 \mid y_k^1, z_k^1, z_k^2, \ldots, z_k^T, H) \ldots$$

$$= p(y_k^1 \mid z_k^1, H)\; p(y_k^2 \mid z_k^2, H) \ldots p(y_k^T \mid z_k^T, H)$$

$$= \prod_{t=1}^{T} p(y_k^t \mid z_k^t, H)$$

as in FMM

$$p(\underline{z} \mid H) = p(z_k^0, z_k^1, \ldots, z_k^T \mid H)$$

$$= p(z_k^0 \mid H)\, p(z_k^1 \mid z_k^0, H)\, p(z_k^2 \mid z_k^1, z_k^0, H) \ldots$$

$$= p(z_k^0 \mid H)\, p(z_k^1 \mid z_k^0, H)\, p(z_k^2 \mid z_k^1, H) \ldots$$

$$= p(z_k^0 \mid H) \prod_{t=2}^{T} p(z_k^t \mid z_k^{t-1}, \ldots$$

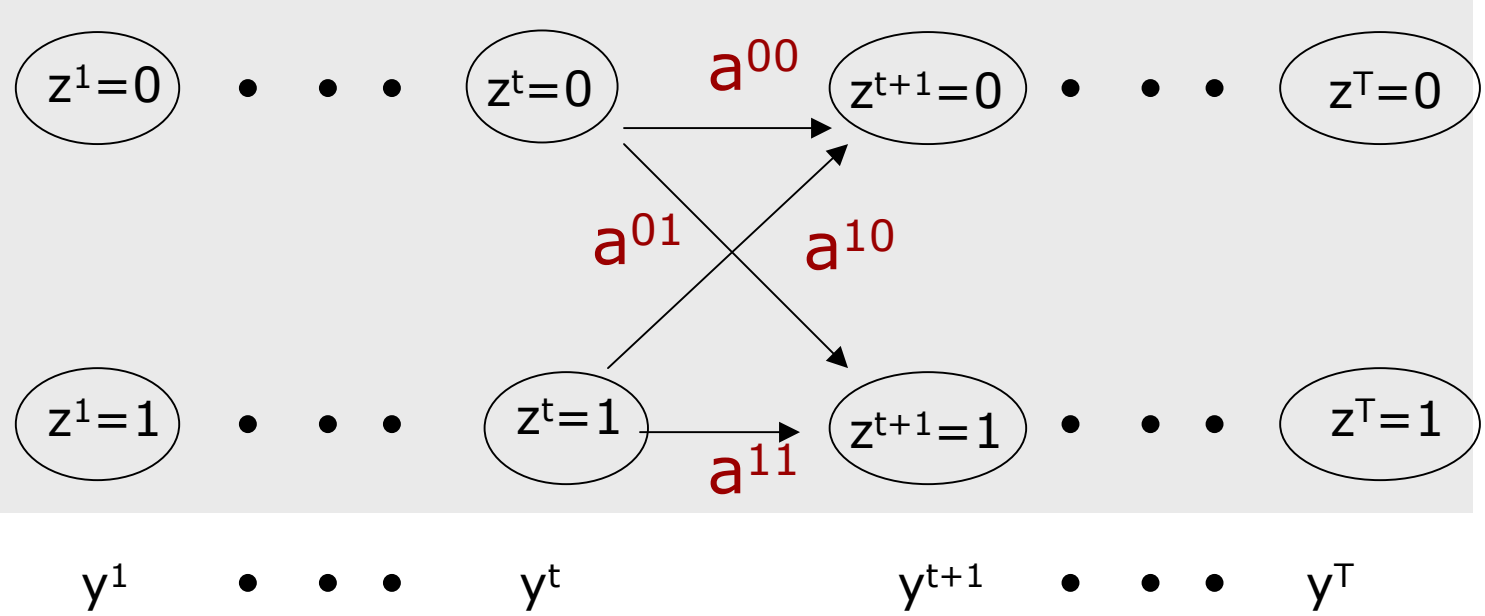
$$p(y_k^1, \ldots, y_k^T \mid H)$$

$$= \sum_{\underline{z}} p(z_k^0 \mid H) \prod_{t=1}^{T} p(z_k^t \mid z_k^{t-1}, H)\, p(y_k^t \mid z_k^t, H)$$

nb hidden sequences per cow= 2^T → total number of operations ~ $4N^T$

Likelihood for N cows: $p(\underline{y} \mid F) = \prod_{k=1}^{N} p(\underline{y}_1, .., \underline{y}_N \mid H)$

Forward-backward algorithm



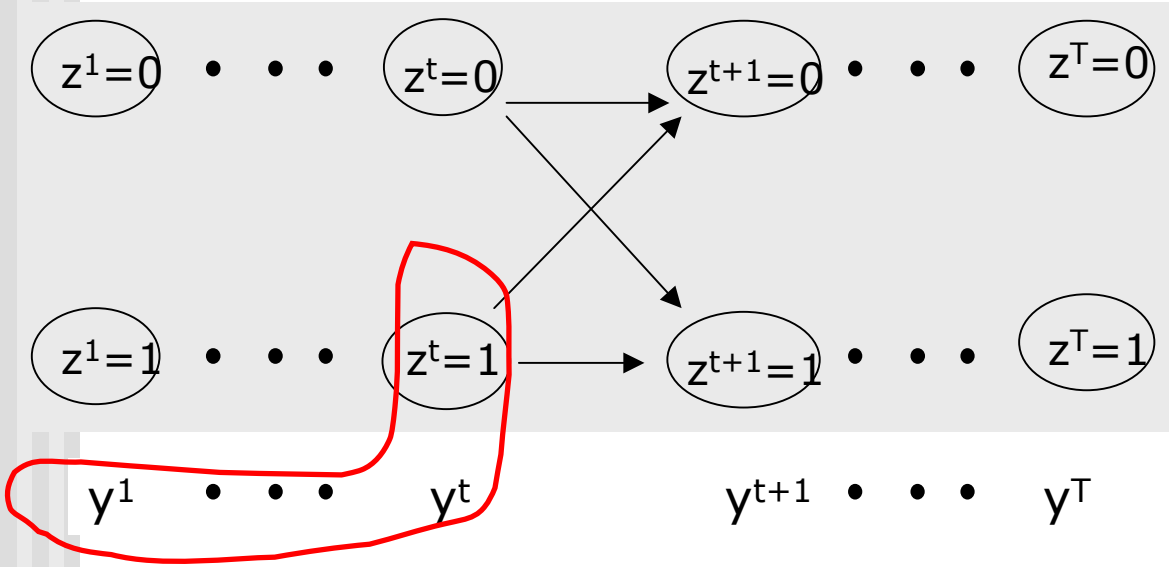
The algorithm takes on the order of 4T computations

- **Forward probabilities** : proba that, given H, at time t, the state is i and the sequence of partial observation (y_1 … y_t) has been generated

$$\alpha_k^t(i) = p(y_k^1 \ldots y_k^t, z_k^t = i)$$

$$\alpha_k^0(0) = p(y_k^0 \mid z_k^0 = 0)\, p(z_k^0 = 0)$$

$$\alpha_k^{t+1}(0) = [\alpha_k^t(0)\, a_k^{00} + \alpha_k^t(1)\, a_k^{01}]\, p(y_k^t | z_k^t = 0)$$

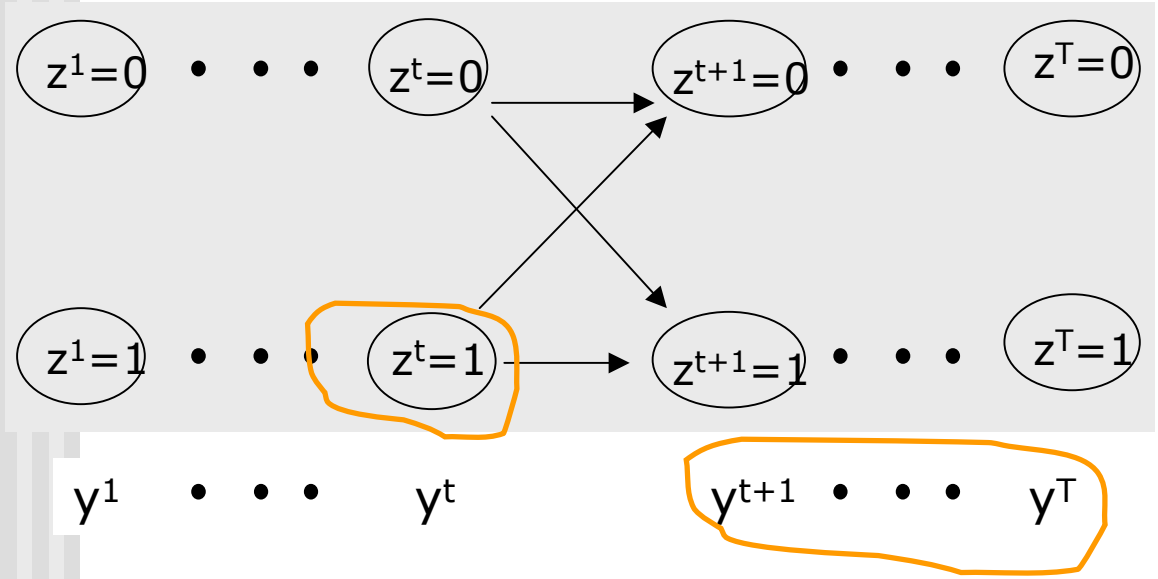


3 steps
induction (t=0)
recursion (increasing t)
termination (t=T)

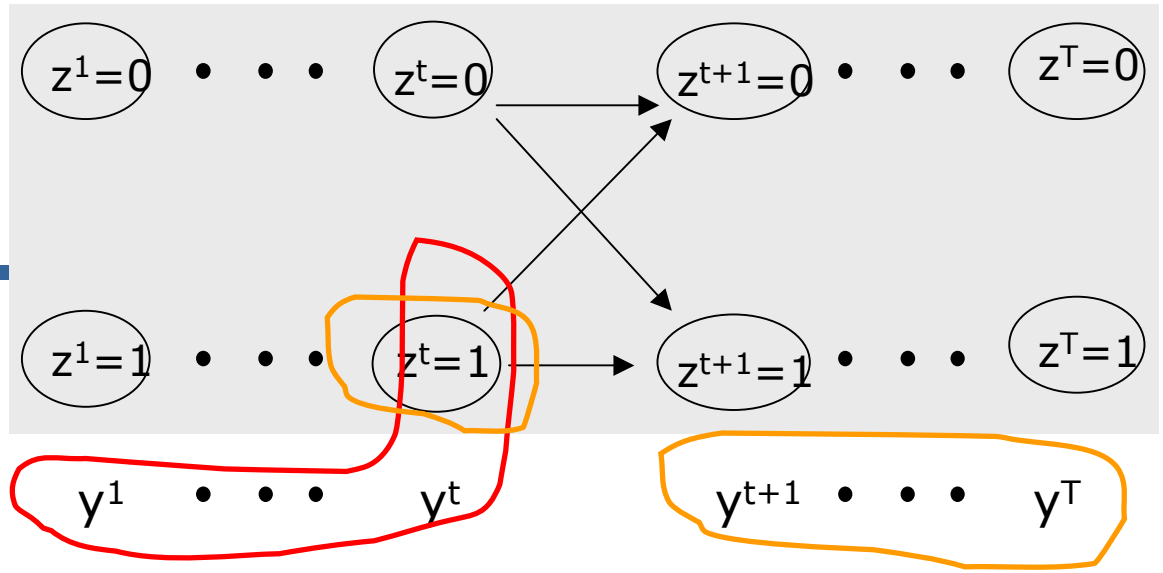
- **Backward probabilities :** proba that, given H and given the state i at time t, a sequence of partial observation (y_{t+1} … y_T) has been generated

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

$$\beta_k^t(0) = [a_k^{00}\, p(y_k^{t+1} \mid z_k^{t+1} = 0)\beta_k^{t+1}(0)] + [a_k^{10}\, p(y_k^{t+1} \mid z_k^{t+1} = 1)\beta_k^{t+1}(1)]$$



3 steps
induction (t=T)
recursion (decreasing t)
termination (t=0)



$$\alpha_k^t(1)\,\beta_k^t(1) = p(y_k^1 \ldots y_k^t, z_k^t = 1 \mid H)\; p(y_k^T \ldots y_k^{t+1} \mid z_k^t = 1)$$

$$= p(y_k^1 \ldots y_k^t, y_k^{t+1} \ldots y_k^T z_k^t = 1 \mid H)$$

$$= p(\underline{y}_k, z_k^t = 1 \mid H)$$

$$\alpha_k^t(0)\,\beta_k^t(0) = p(\underline{y}_k, z_k^t = 0 \mid H)$$

$$\Rightarrow \quad \alpha_k^t(0)\,\beta_k^t(0) + \alpha_k^t(1)\,\beta_k^t(1) = p(\underline{y}_k \mid H)$$

3c. EM algorithm

E step:

$$Q^{(p-1)} = E[\log[p(\underline{y}_1,\ldots,\underline{y}_N;\underline{z}_1,\ldots,\underline{z}_N \mid \Theta) \mid \underline{y}_1,\ldots,\underline{y}_N;\Theta^{(p-1)}]$$

$$= \sum_{k=1}^{N} \left[\begin{array}{l} E[z_k^0 = 1 \mid \underline{y}_k]\log(\lambda_k) + E[z_k^0 = 0 \mid \underline{y}_k]\log(1-\lambda_k) \\ + \sum_{t=1}^{T}\sum_{i,j}^{0,1} E[z_k^t = i, z_k^{t+1} = j \mid \underline{y}_k]\log(a_k^{ij}) \\ + \sum_{t=1}^{T}\sum_{i,j}^{0,1} E[z_k^t = i \mid \underline{y}_k]\log N(\mu_i^t, \sigma_i^2) \end{array} \right]$$

$$E[z_k^t = i \mid \underline{y}_k] = p[z_k^t = i \mid \underline{y}_k] = \gamma_{i,k}^t$$

$$= \frac{p[\underline{y}_k, z_k^t = i]}{p[\underline{y}_k, z_k^t = 0] + pr[\underline{y}_k, z_k^t = 1]}$$

$$= \frac{\alpha_k^t(i)\,\beta_k^t(i)}{\alpha_k^t(0)\,\beta_k^t(0) + \alpha_k^t(1)\,\beta_k^t(1)}$$

It is the probability of being in state i at time t, given the observation sequence $y_1\ y_2 \ldots y_T$.

$$E[z_k^t = i, z_k^{t+1} = j \mid \underline{y}_k] = p[z_k^t = i, z_k^{t+1} = j \mid \underline{y}_k] = \xi_{ij,k}^t$$

$$= \frac{p[\underline{y}_k, z_k^t = i, z_k^{t+1} = j]}{p[\underline{y}_k, z_k^t = 0] + p[\underline{y}_k, z_k^t = 1]}$$

$$= \frac{\alpha_k^t(i)\ a_k^{ij}\ \beta_k^{t+1}(j)\ N(\mu_j^{t+1}, \sigma_j^2)}{\alpha_k^t(0)\beta_k^t(0) + \alpha_k^t(1)\beta_k^t(1)}$$

It is the probability of being in state i at time t and in state j at time t+1, given the observation sequence $y_1\ y_2 \ldots y_T$.

M step: $\Theta^{(p)} = \text{argmax } Q^{(p-1)}$

$$\hat{\lambda}_k = \gamma^0_{1,k}$$

= expected frequency of state IMI- at start

$$\hat{a}^{ij}_k = \frac{\sum_{t=1,T} \xi^t_{ij,k}}{\sum_{t=1,T} \gamma^t_{j,k}}$$

= expected number of transitions from state j to state i divided by the expected number of transitions out of state j

= weighted means and variances

$$\hat{\mu}^t_i = \frac{\sum_{k=1,N} \gamma^t_{i,k}\, y^t_k}{\sum_{k=1,N} \gamma^t_{i,k}}$$

$$\hat{\sigma}^2_i = \frac{\sum_{k=1,N} \sum_{t=1,T} \gamma^t_{i,k}\, (y^t_k - \hat{\mu}^t_i)^2}{\sum_{k=1,N} \sum_{t=1,T} \gamma^t_{i,k}}$$

4. Simulation

Survey of SCC

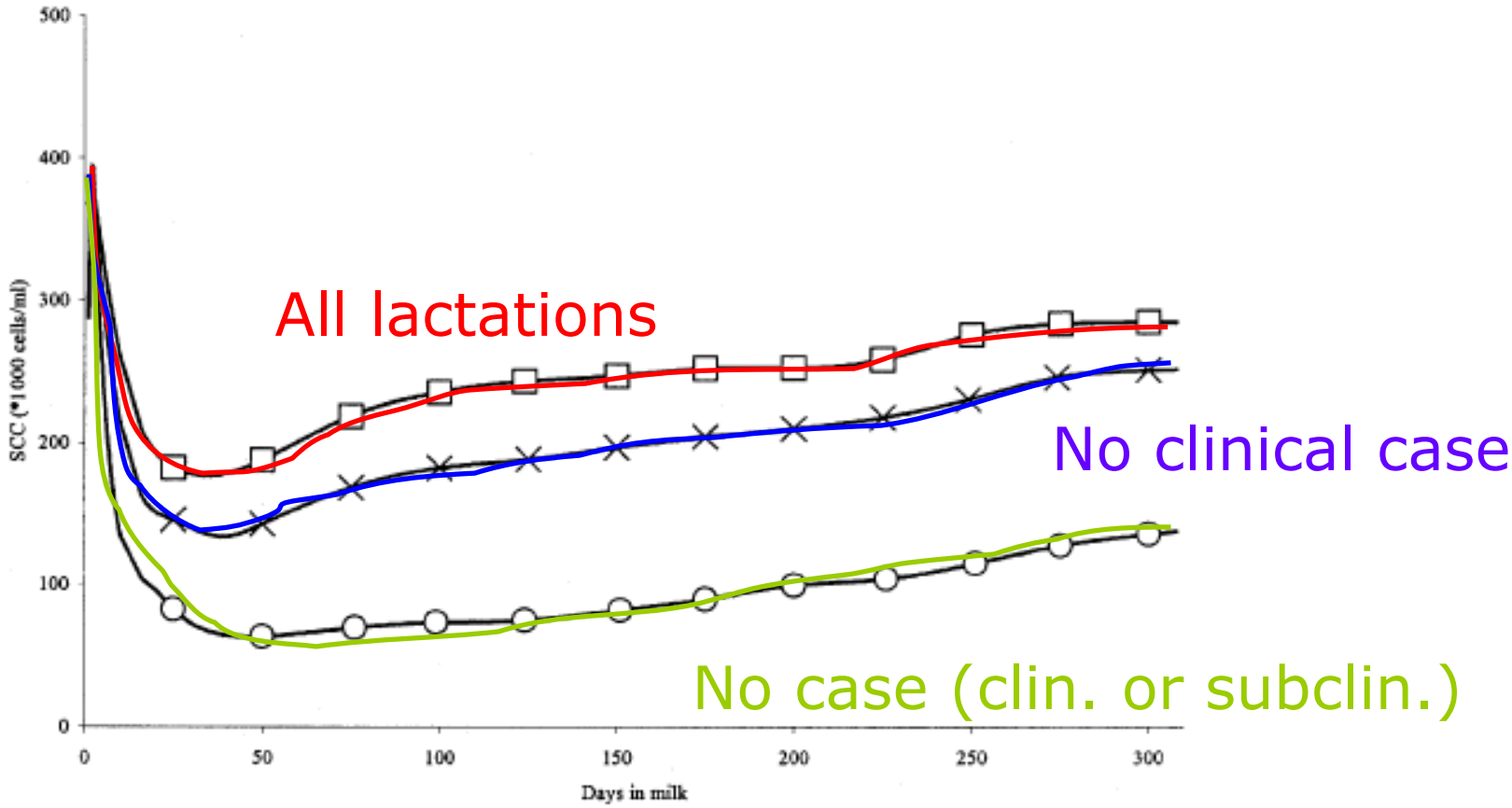
Pathogen

Severity of response

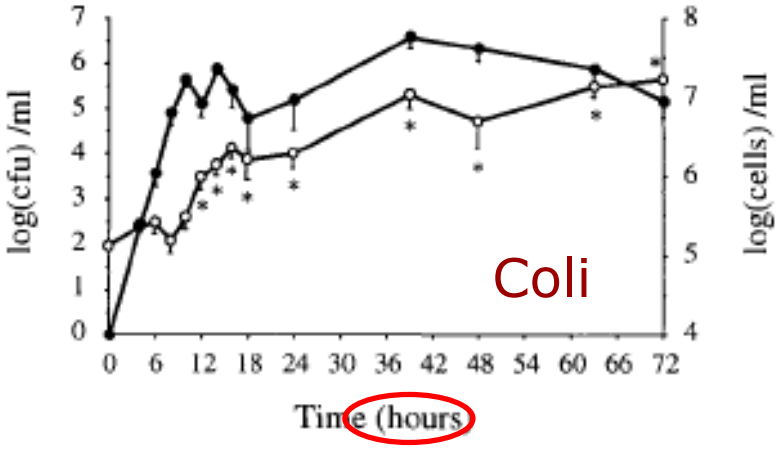
Data sets

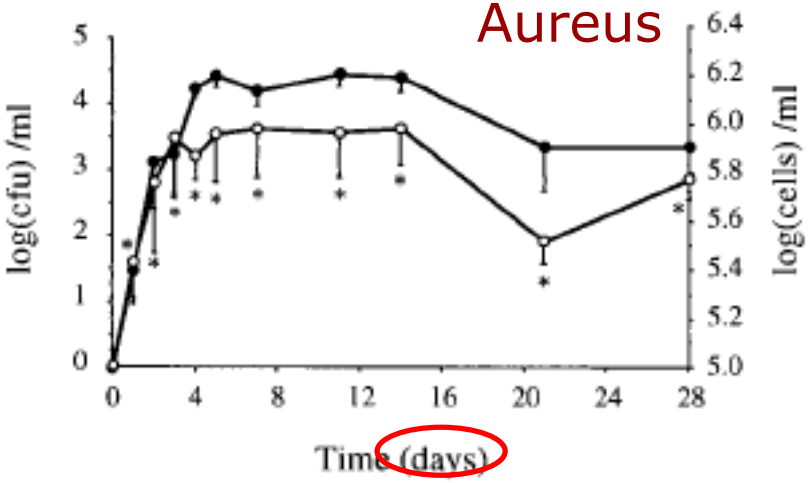
Accuracy of MLE

4a. Survey (de Haas et al., 2002, 2004)



4b. Pathogen

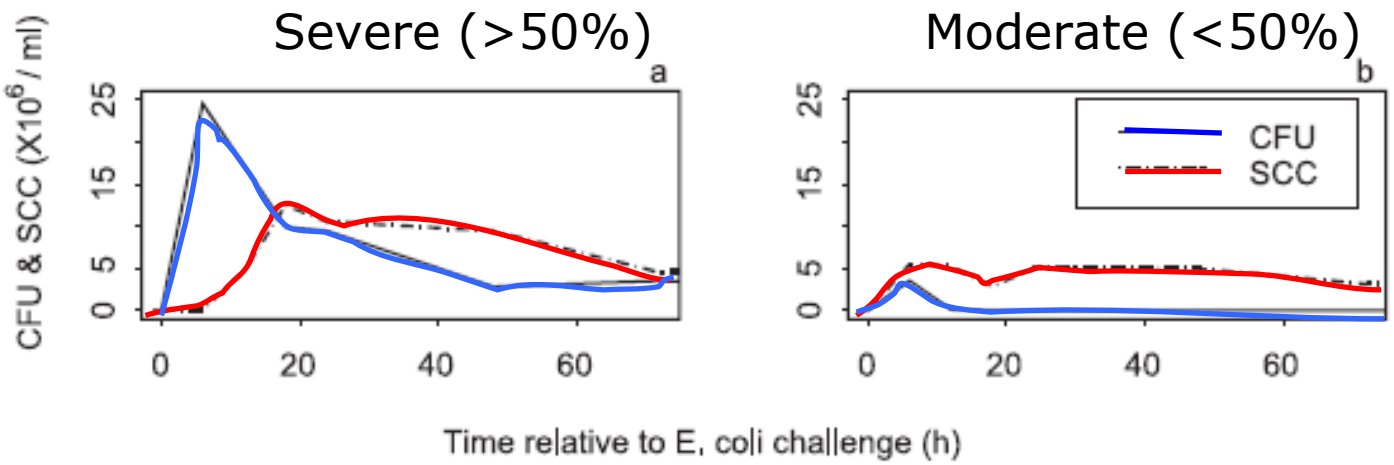




CÉLINE RIOLLET, PASCAL RAINARD,* AND BERNARD POUTREL

CLINICAL AND DIAGNOSTIC LABORATORY IMMUNOLOGY, Mar. 2000, p. 161–167

4c. Severity of response



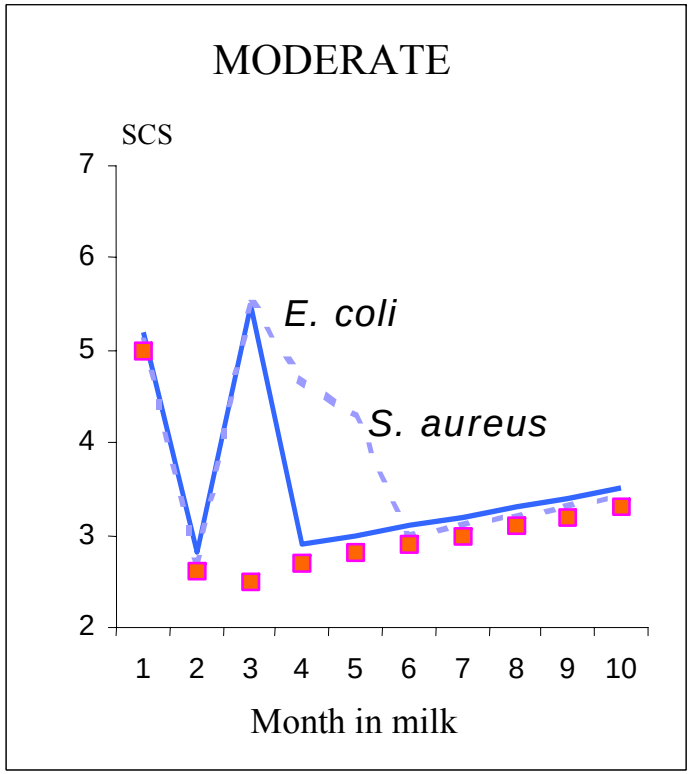
Jalil MEHRZAD[a,b,c], Luc DUCHATEAU[a], Christian BURVENICH[a*]

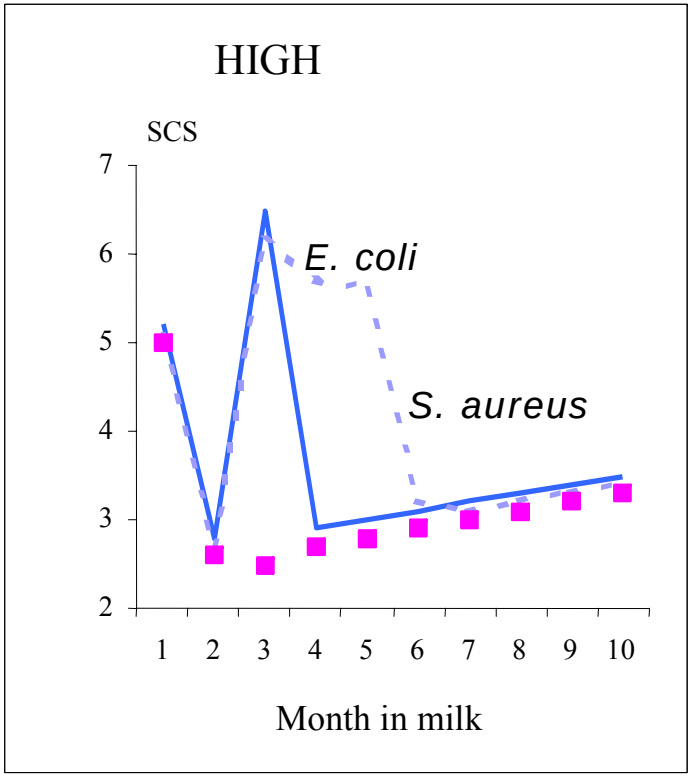
Vet. Res. 36 (2005) 101–116

4d. Simulated data sets

μ_0^t for t = 1 to 10

μ_1^t for t = 1 to 10





3 simulated data sets:

% infected cows = 20, 50%

% *E. coli* among infected cows = 0,50,100%

high and moderate responders: μ_1^t

$\sigma_0^2 = 1.0 \text{ or } 1.4$

EM algorithm for both HMM and FMM:

same priors

500 iterations

10 replications

Accuracy of MLE: $\theta - \hat{\theta}$

4e. Accuracy of MLE

Bias in μ_0^t (3 to 5)

FMM : 0.15 (0.03)

HMM : 0.19 (0.02)

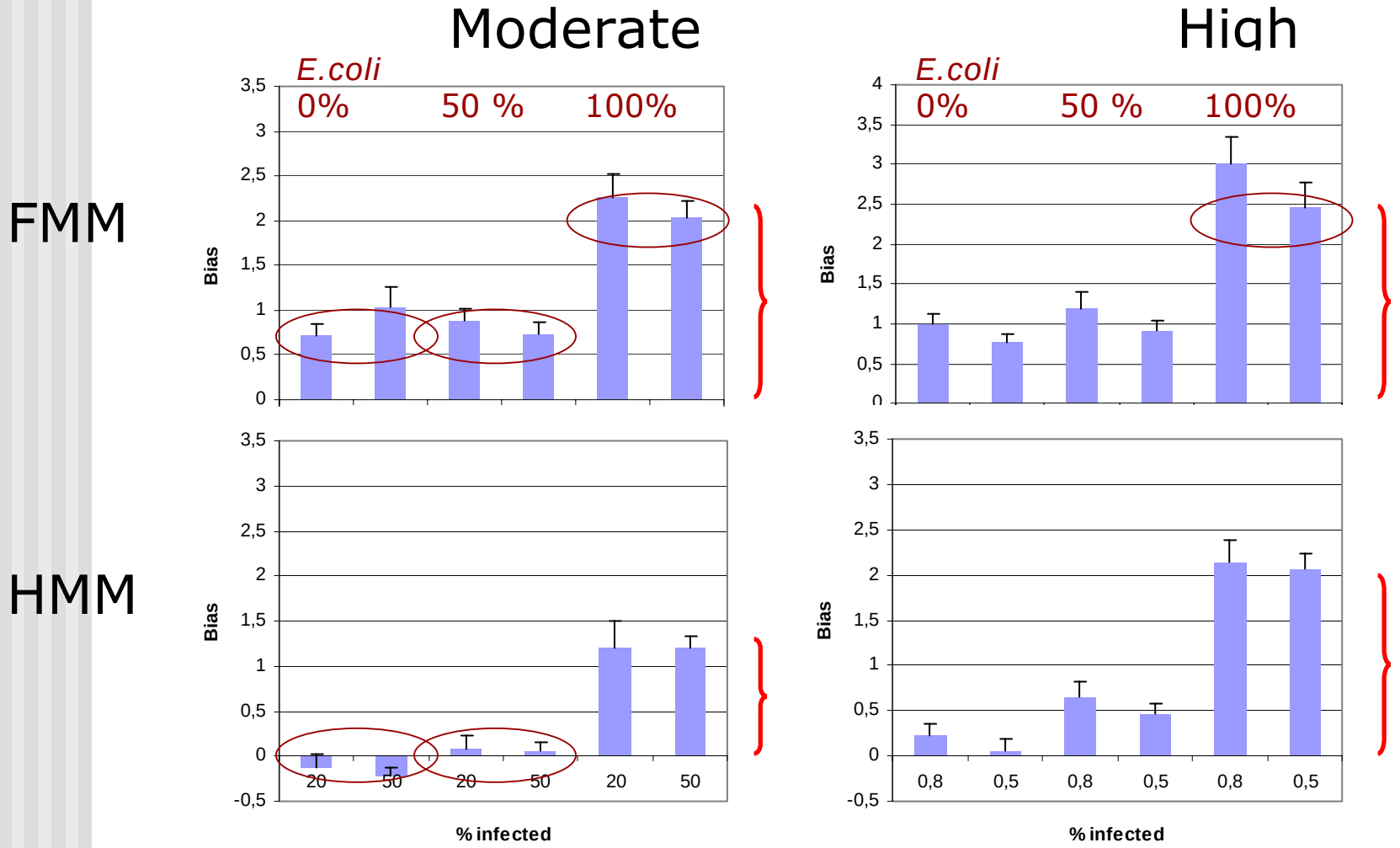
Bias in σ_0^2 (1.0 or 1.4)

FMM : 0.22 (0.04)

HMM : 0.24 (0.05)

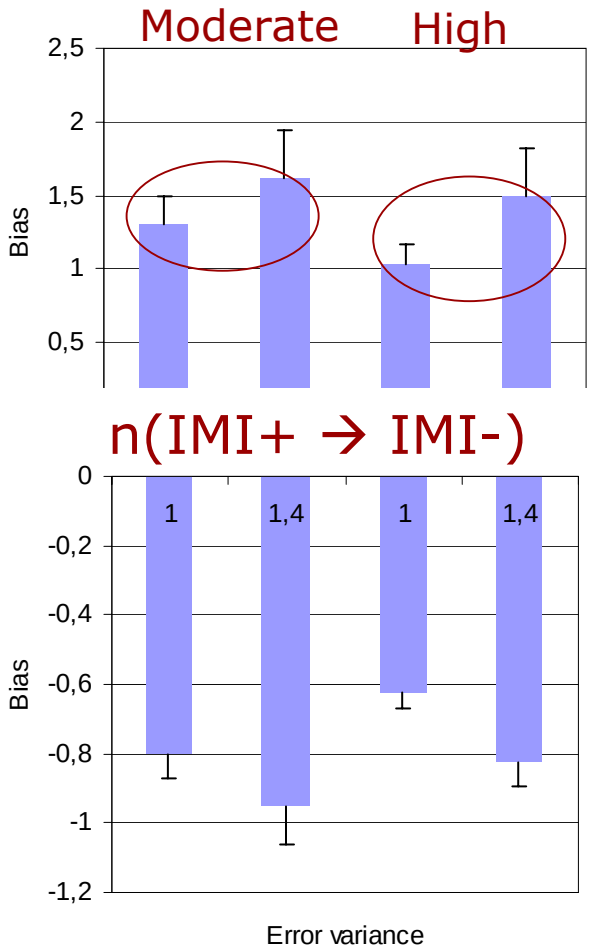
Similar differences between true values and MLE for parameters of the IMI- distribution

Bias in μ_1^t (5 to 7)

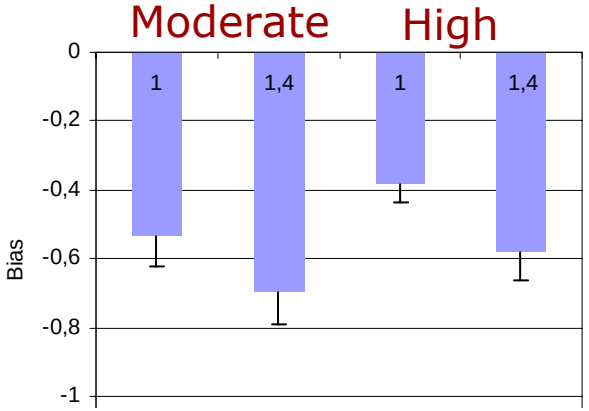


Bias in I

n(IMI- → IMI-)

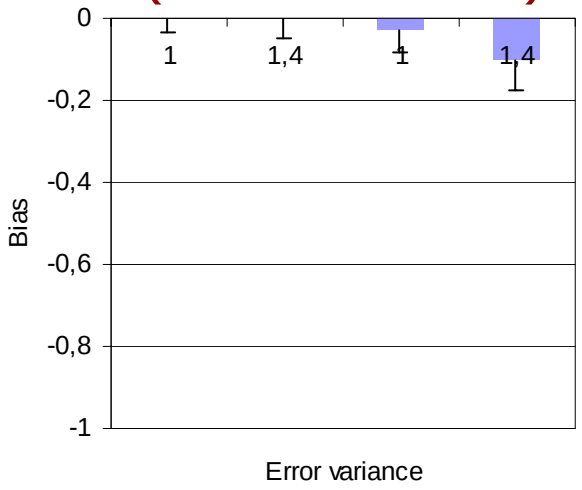


n(IMI- → IMI+)



n(IMI+ → IMI-)

n(IMI+ → IMI+)



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

Thank you for your attention