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

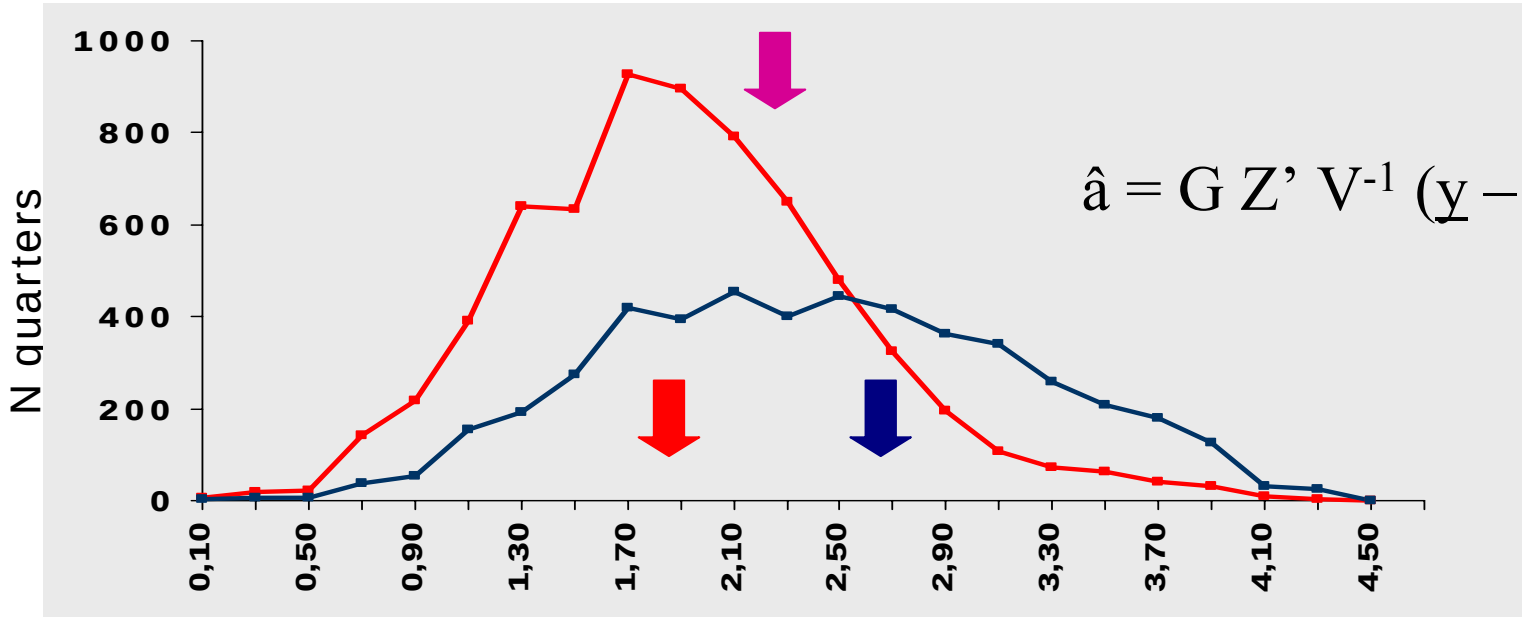
1a. Biological marker

- 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 pr(IMI)

1b. Accuracy

Frequency distribution of SCC for IMI- or IMI+ quarters



$$\hat{a} = G\,Z'\,V^{-1}\,(\underline{y} - \underline{\mu})$$

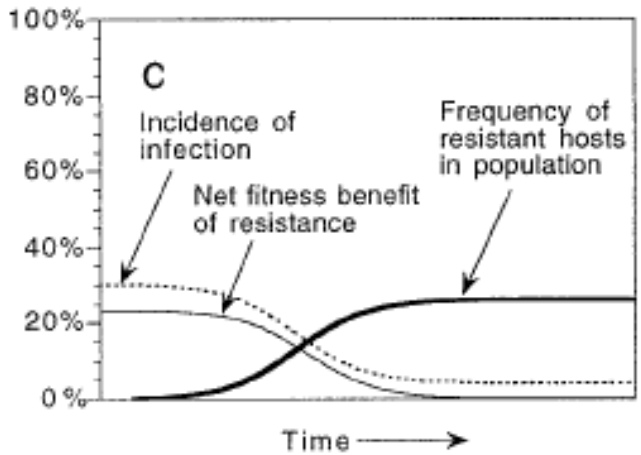
Imperfect detectability

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

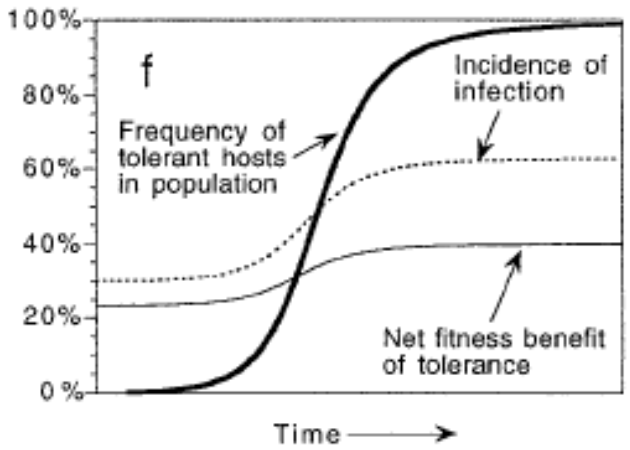
1c. Resistance/ tolerance

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

- Resistance = low proba of infection



- Tolerance = little fitness loss after infection



Breeding goals ?

+ herd immunity
- natural selection

- disease spread
+ natural selection

2. Mixed hidden Markov model

General formulation

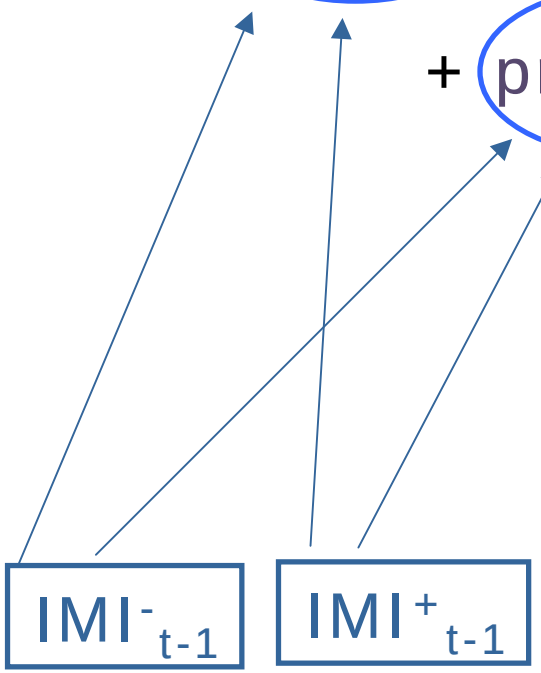
Bayes estimation

2a. Formulation

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

Tolerance

Resistance

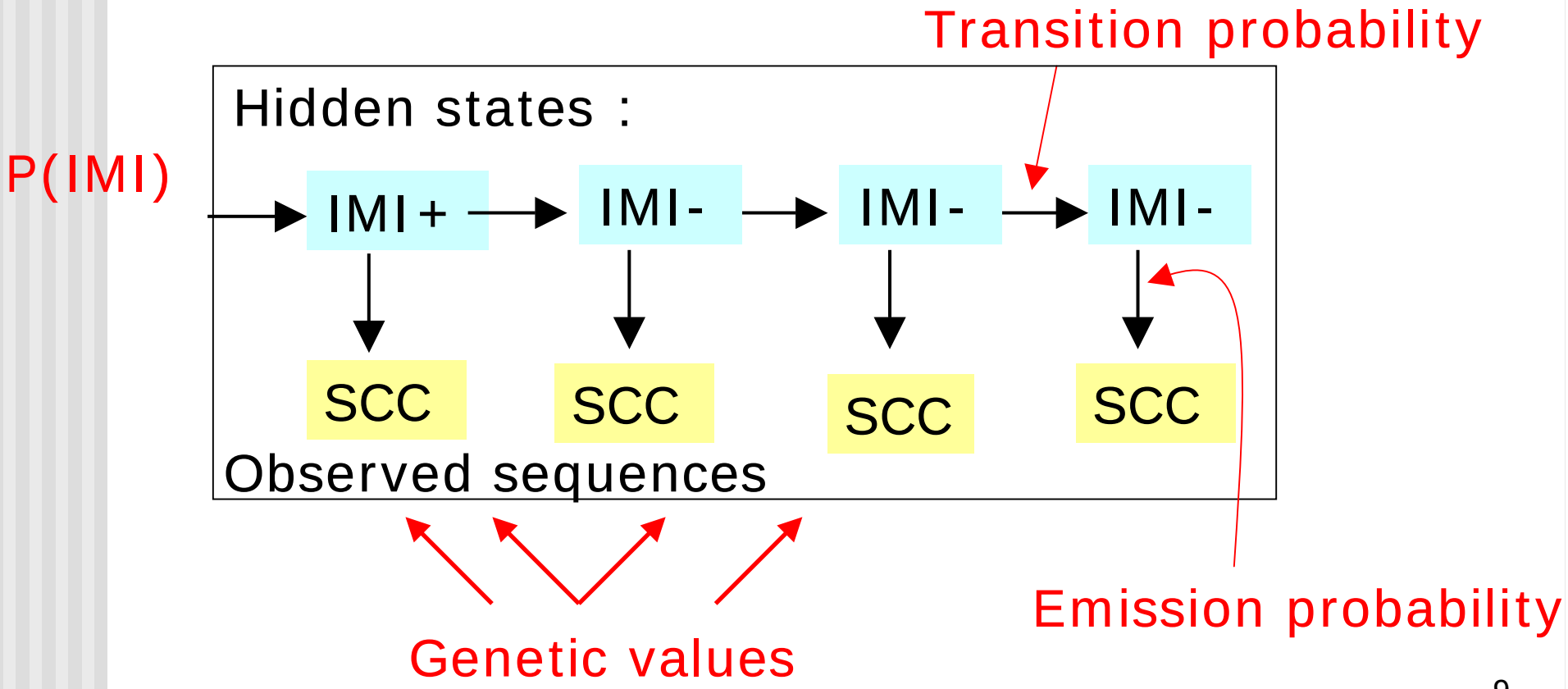


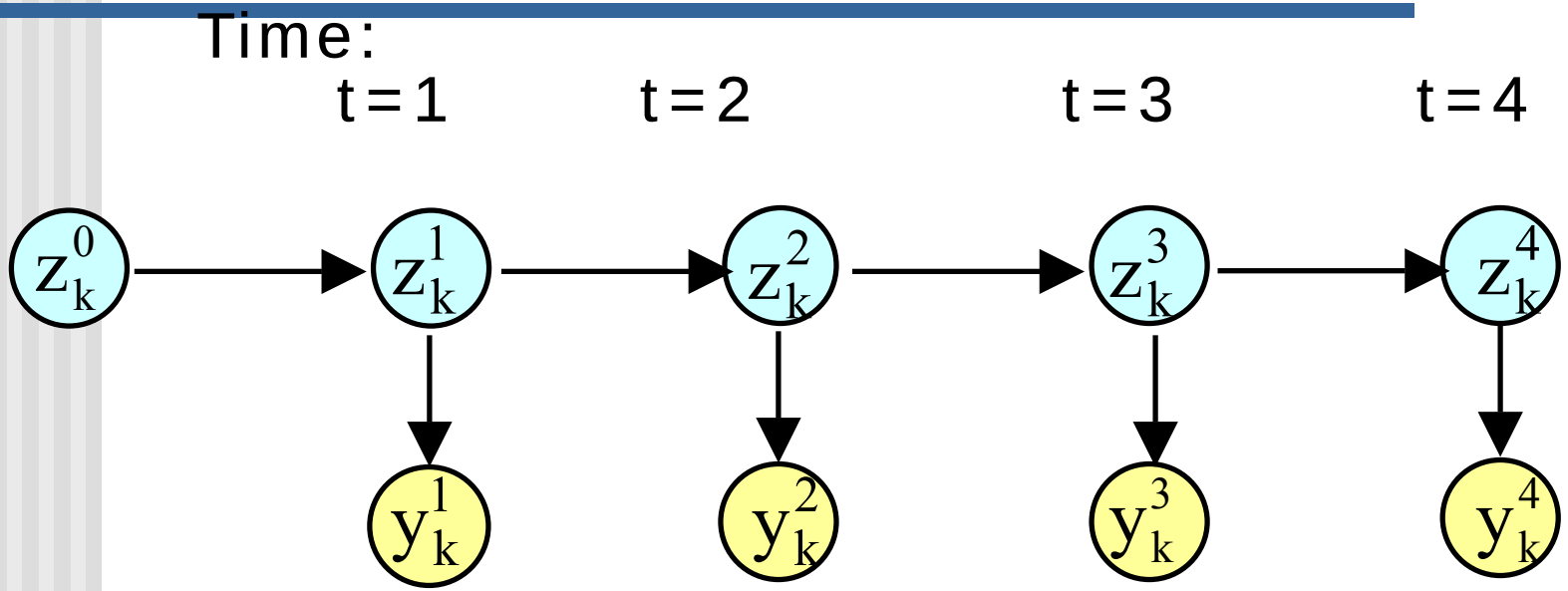
Recovery: $IMI^+_{t-1} \rightarrow IMI^-_t$

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
Genetic values



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+

- 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

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

$$p(\underline{y} \mid \underline{\mu}_0, \underline{\mu}_1, \sigma_0^2, \sigma_1^2, \underline{a}, \underline{z}) \sim N(M_0\underline{\mu}_0 + M_1\underline{\mu}_1 + Z\underline{a}, R)$$

$\underline{y}$ = (NT X 1) vector of data
$\underline{z}$ = (NT X 1) vector of hidden states
$\underline{\mu}_0$ = (T X 1) vector of fixed effects for data on a IMI- cow
$\underline{\mu}_1$ = (T X 1) vector of fixed effects for data on a IMI+ cow
$\underline{a}$ = (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 $\underline{a}$ to $\underline{y}$.

$R \sim J_0\sigma_0^2 + J_1\sigma_1^2$ J_i = (NT X NT) matrix with elements = 1 if $z_k^t = i$

2b. Estimation

Prior distributions

$$\underline{\mu}_0 \sim N(\underline{1}\, m_0, I\, s_0^2)$$
$$\underline{\mu}_1 \sim N(\underline{1}\, m_1, I\, s_1^2$$
$$\underline{a} \sim N(0,\ A\ \sigma_a^2)$$

$$\sigma_a^2 \sim \chi^{-2}(s_a^2)$$
$$\sigma_0^2 \sim \chi^{-2}(s_0^2)$$
$$\sigma_1^2 \sim \chi^{-2}(s_1^2)$$

$$z_k^0 \sim \mathrm{Br}(\lambda_k)$$
$$(z_k^t = 0 \mid z_k^{t-1} = 0) \sim \mathrm{Br}(\pi_k^{00})$$
$$(z_k^t = 1 \mid z_k^{t-1} = 0) \sim \mathrm{Br}(\pi_k^{01})$$

$$\lambda_k, \pi_k^{00}, \pi_k^{01} \sim \mathrm{Un}[0,1]$$

Joint posterior distribution

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

$$p(\theta \mid \underline{y}) = \sum_{\underline{z}} p(\underline{y} \mid \theta)\, p(\theta)$$

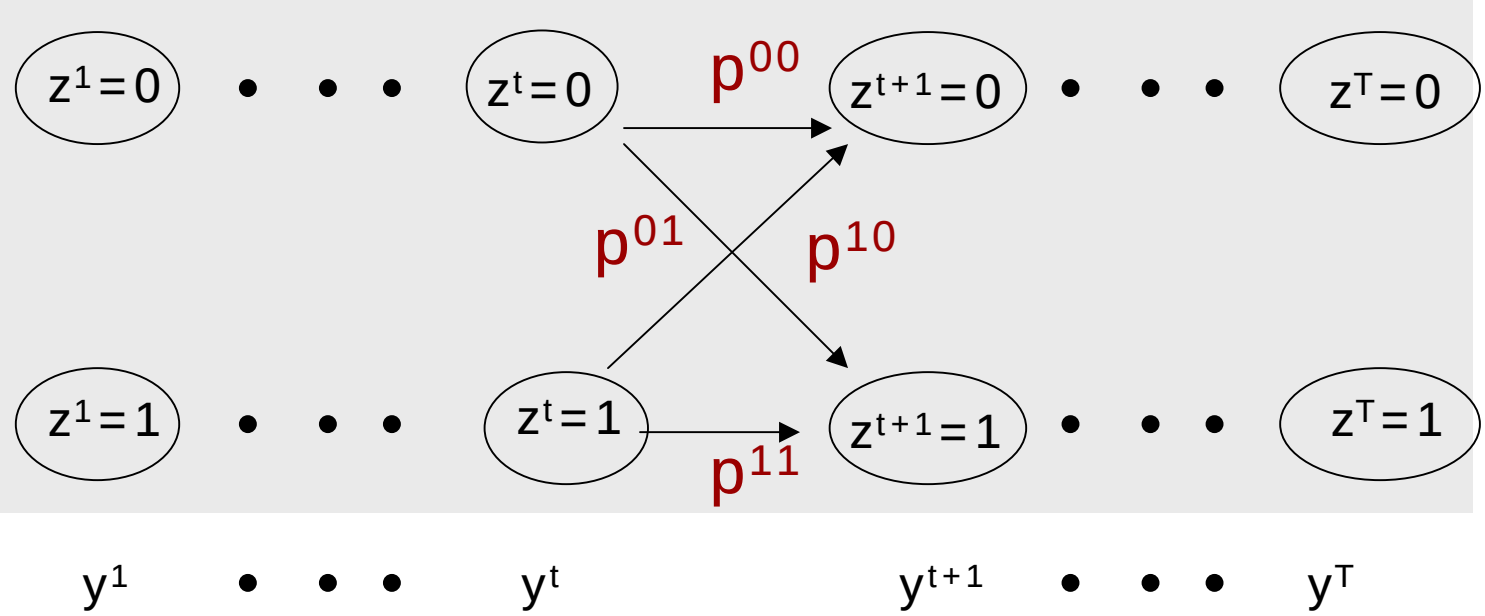
nb hidden sequences per cow= 2^T

→ total number of operations ~ $4N^T$

Use of the trellis stucture of HMM

→ Forward-backward algorithm

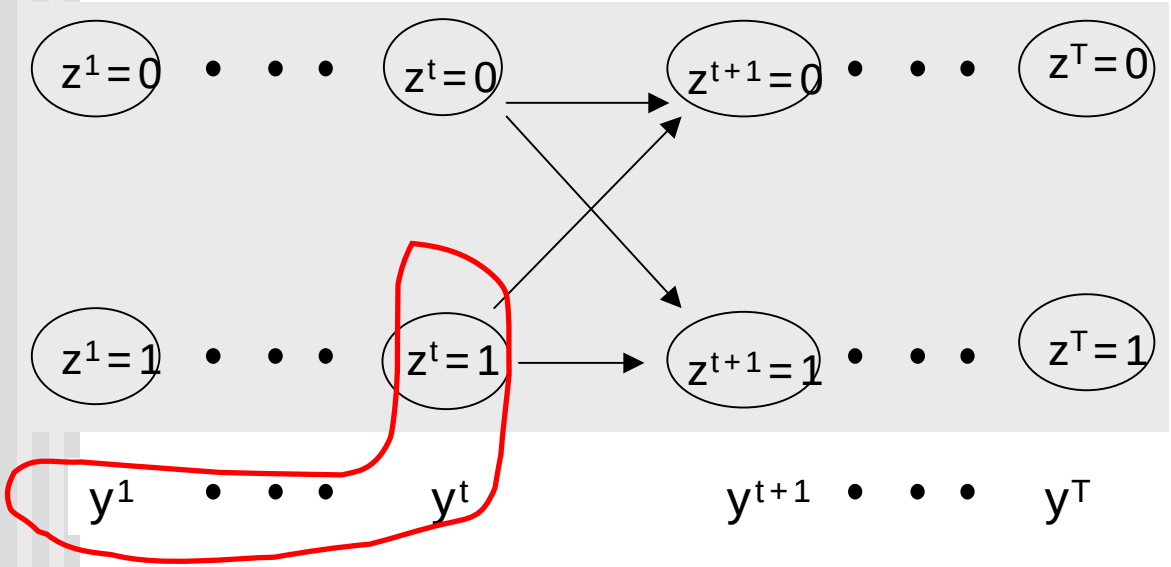
Forward-backward algorithm



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

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



3 steps

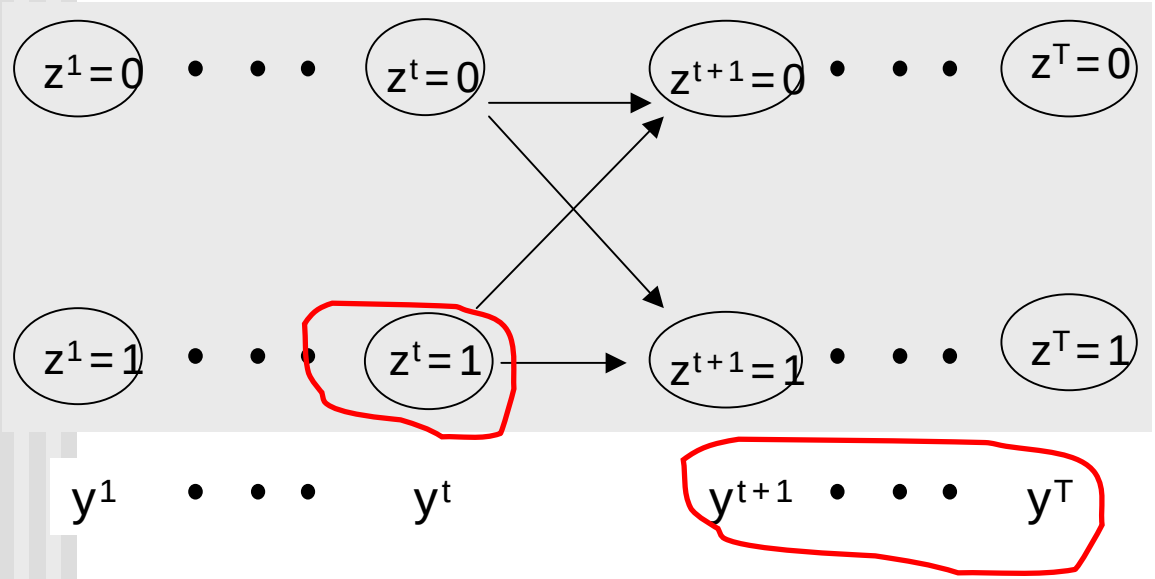
induction (t=0)

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

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



3 steps

induction (t=T)

recursion (decreasing t)

termination (t=1)

Fully conditional distributions

$$\mu_i^t \succ N\left(\frac{s_i^2 \sum_k^N (y_k^t - a_k)\, I_k^{i,t} + m_i\, \sigma_i^2}{(s_i^2 \sum_k^N n_{i,k}^t) + \sigma_i^2}, \frac{s_i^2\, \sigma_i^2}{(s_i^2 \sum_k^N n_{i,k}^t) + \sigma_i^2} \right)$$

$$\underline{a} \sim N(\tilde{\theta}_1, C_{11}^{-1}) \qquad \tilde{\theta}_1 = C_{11}^{-1}\left[r_1 - C_{12}\theta_2\right]$$

$$(\sigma_a^2) \sim \chi_{N+\nu}^{-2} \quad (\underline{a}' A^{-1} \underline{a} + \nu\, s_a^2)$$

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

$$\lambda_k \sim \text{beta}(I_k^{0,1} + 1; I_k^{1,1} + 1)$$

$$\pi_k^{00} \sim \text{beta}(n_k^{00} + 1, n_k^{10} + 1) \qquad \pi_k^{01} \sim \text{beta}(n_k^{01} + 1, n_k^{11} + 1)$$

$$\underline{z}_k^1 \sim \mathrm{Br}(\zeta_{0,k}^1) \qquad \zeta_{0,k}^0 = p[z_k^0 = 0 \cap \underline{y}_k] = \alpha_k^1(0)\ \beta_k^1(0)$$

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

$$\underline{z}_k^t \sim \mathrm{Br}(\xi_{ij,k}^t) \qquad \xi_{ij,k}^t = p\,[z_k^t = i \mid z_k^{t-1} = j,\ \underline{y}_k]$$

$$= \frac{\alpha_k^{t-1}(j)\ \ \pi_k^{ij}\ \ \beta_k^t(i)\ \mathrm{pr}(y_k^t \mid z_k^t = i)}{\alpha_k^{t-1}(j)\ \beta_k^{t-1}(j)}$$

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

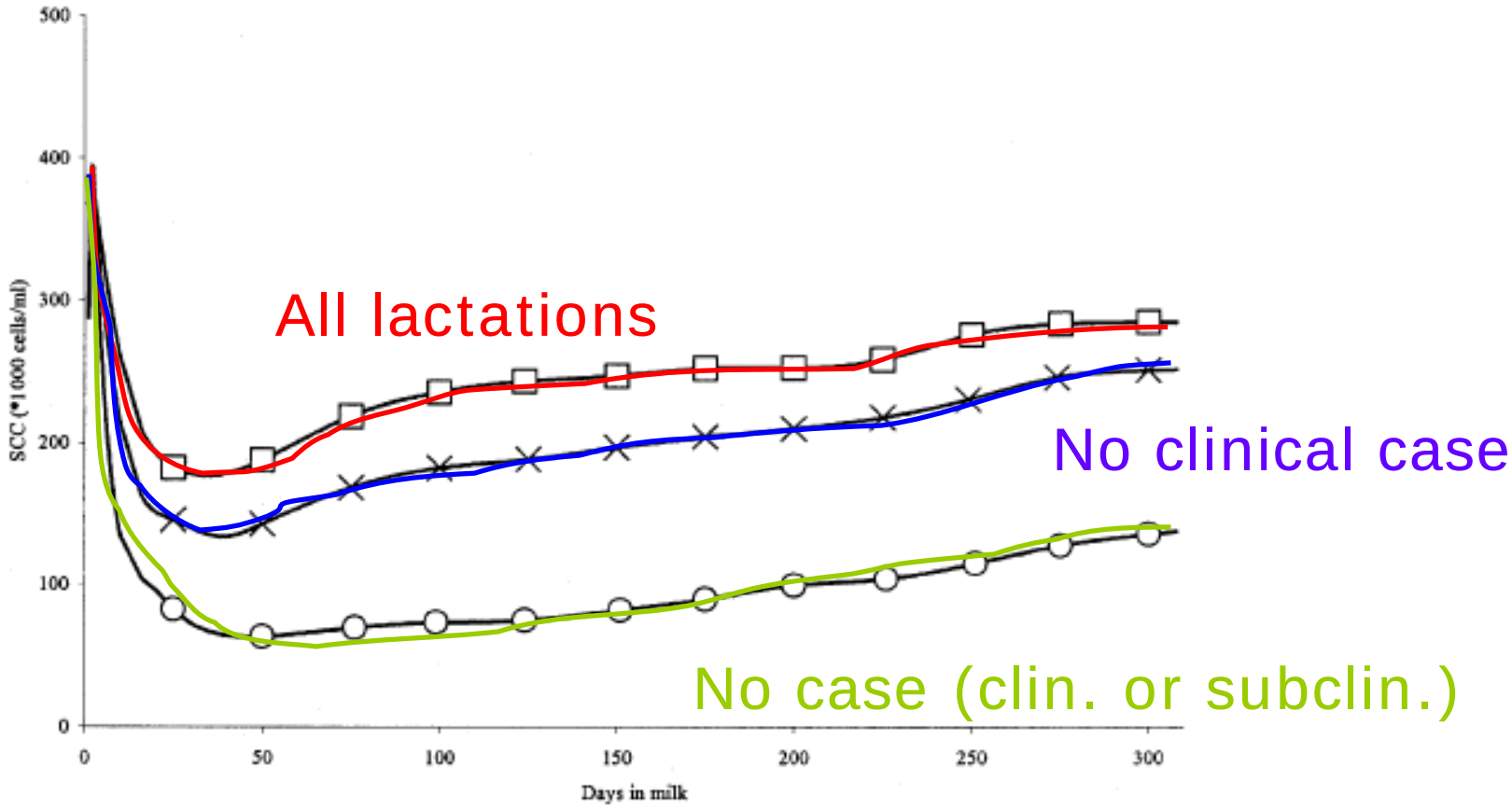
Severity of response

Genetics

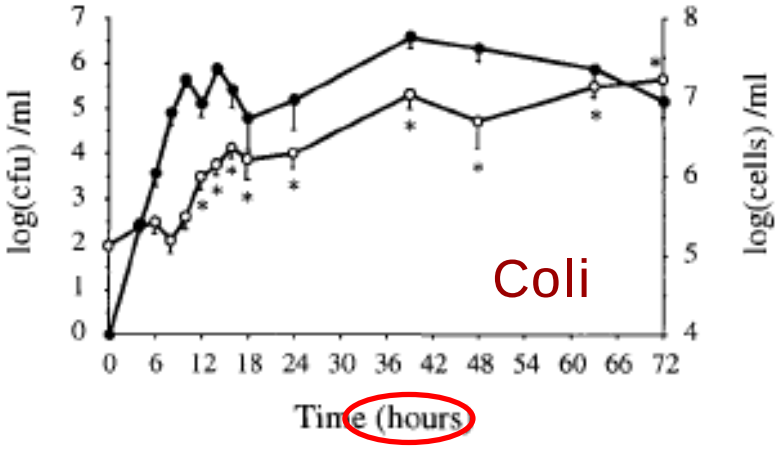
Data sets

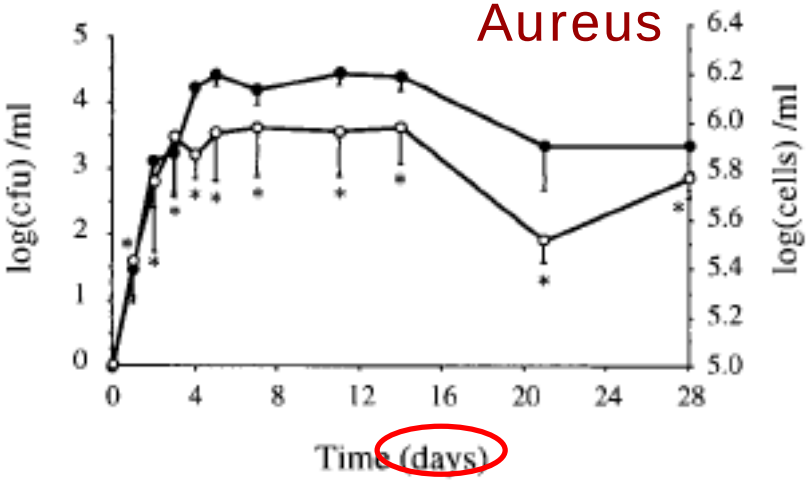
Accuracy

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



4b. Pathogen

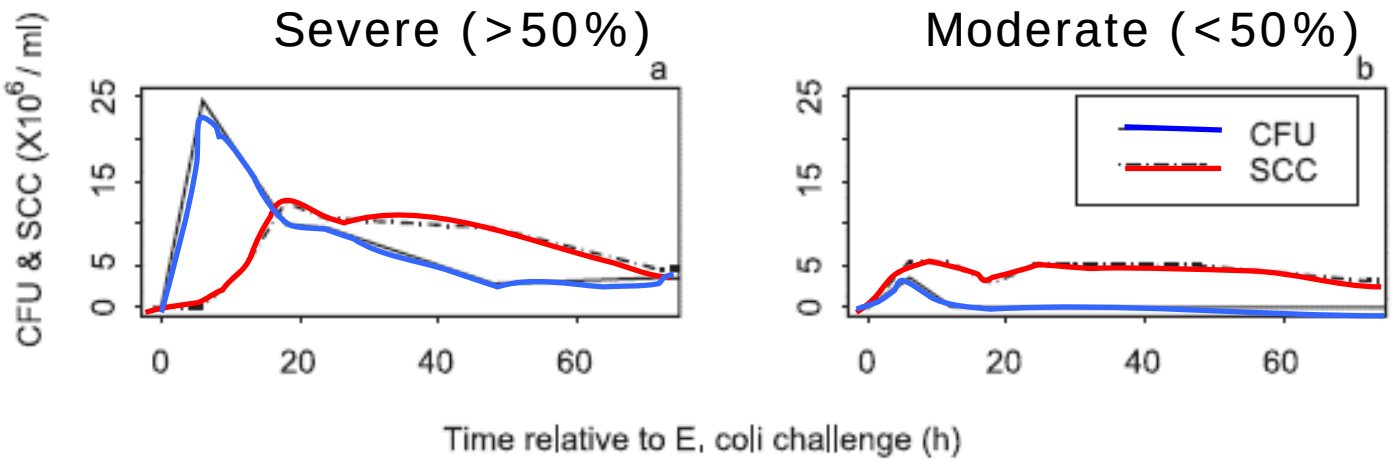




CÉLINE RIOLLET, PASCAL RAINARD,* AND BERNARD POUTREL

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. Genetics

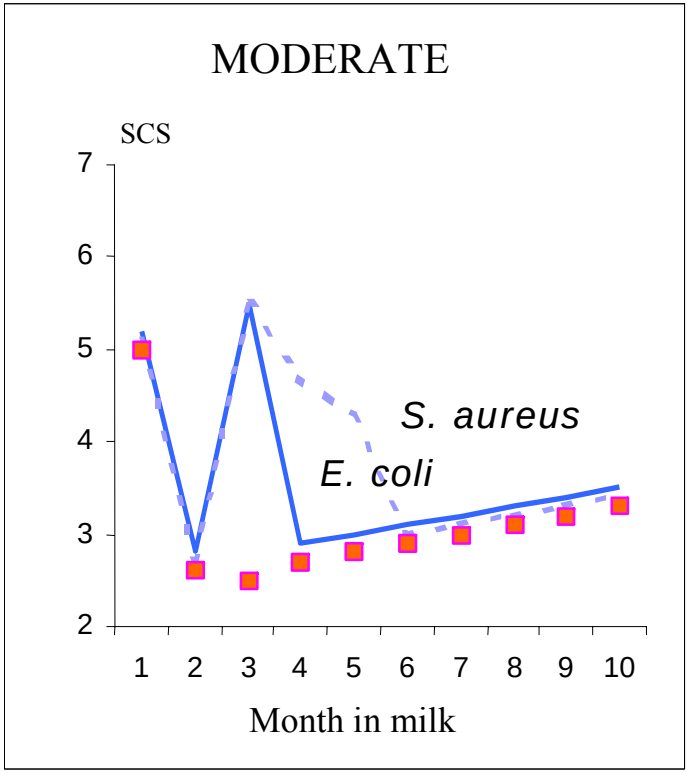
- 3 discrete generations with 400 cows per generation
- sires selected from 30 different bulls
- each cow was replaced by a daughter
- mating at random

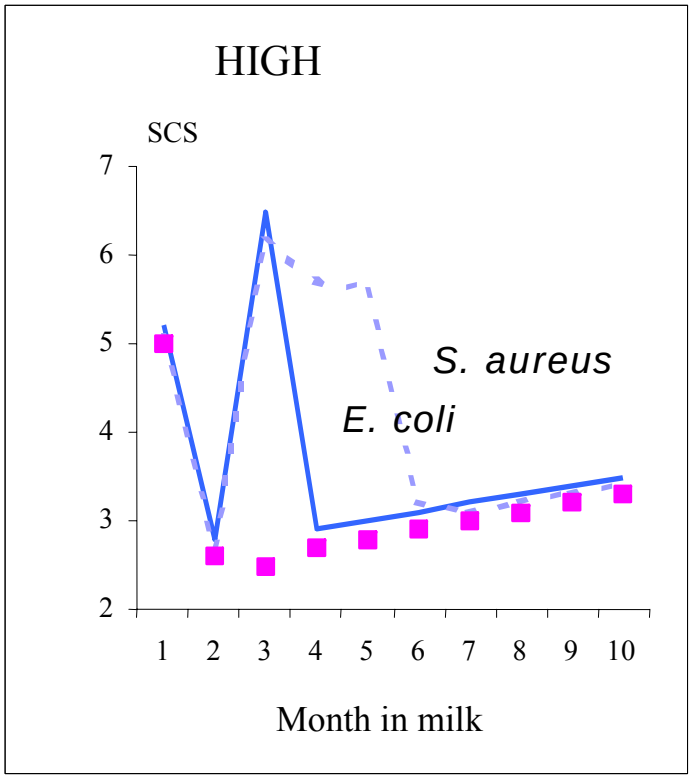
- breeding values for base animals ~N(0, I σ_a^2)
- additive variance of 0.15 or 0.25.
- breeding values for non-base animals ~N(mid-parent,$\sigma_a^2/2$)
- No selection , no inbreeding

4e. Simulated data sets

μ_0^t for t = 1 to T

μ_1^t for t = 1 to T





Simulated data sets:

% infected cows = 20, 50%

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

high and moderate responders: μ_1^t

$\sigma_0^2 = 1.0 \text{ or } 1.4$ → residuals IMI- ~N(0,σ_0^2)

$\sigma_a^2 = 0.15 \text{ or } 0.25$

Gibbs sampler

1000 iterations

200 burn-in

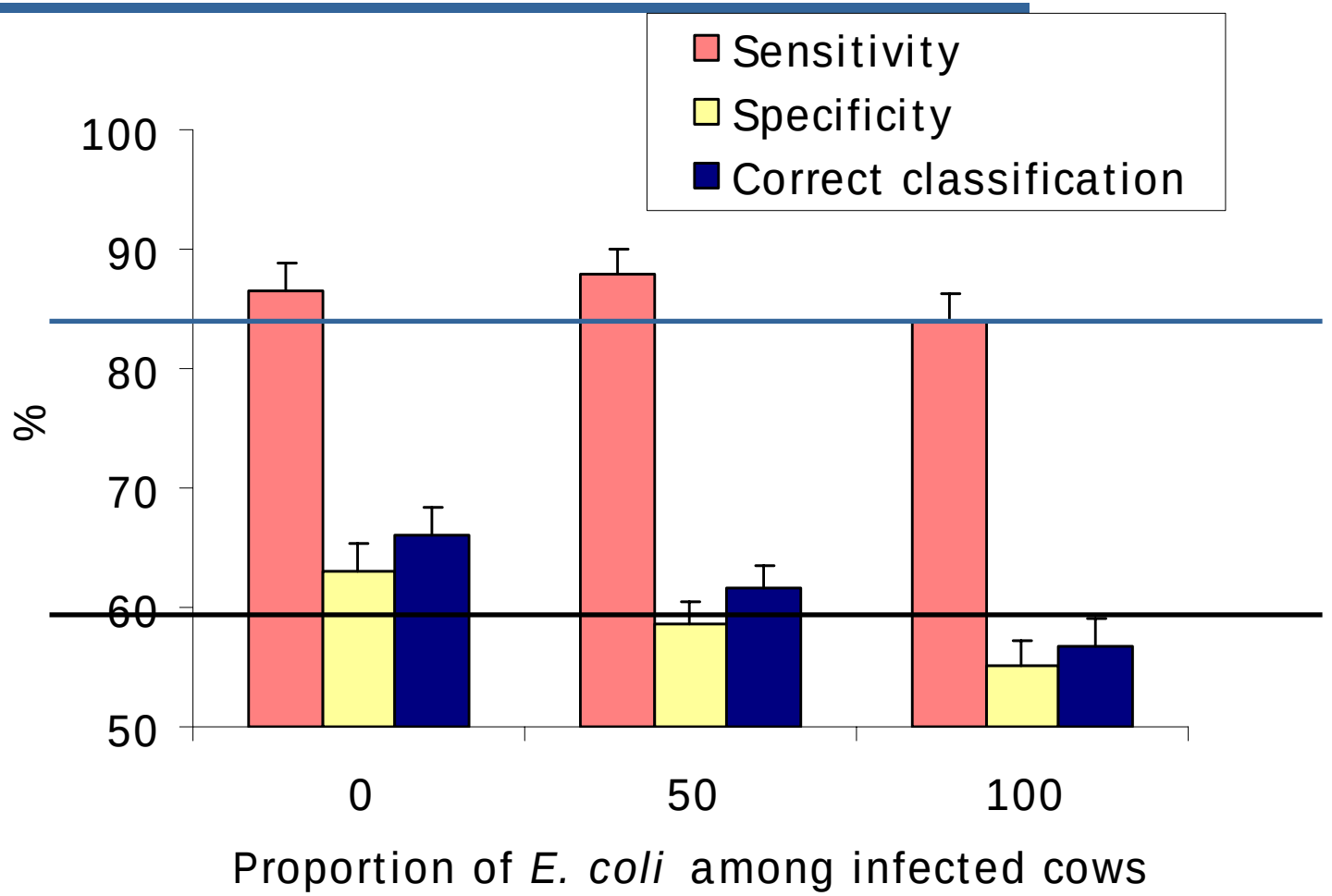
10 replications

4f. Accuracy

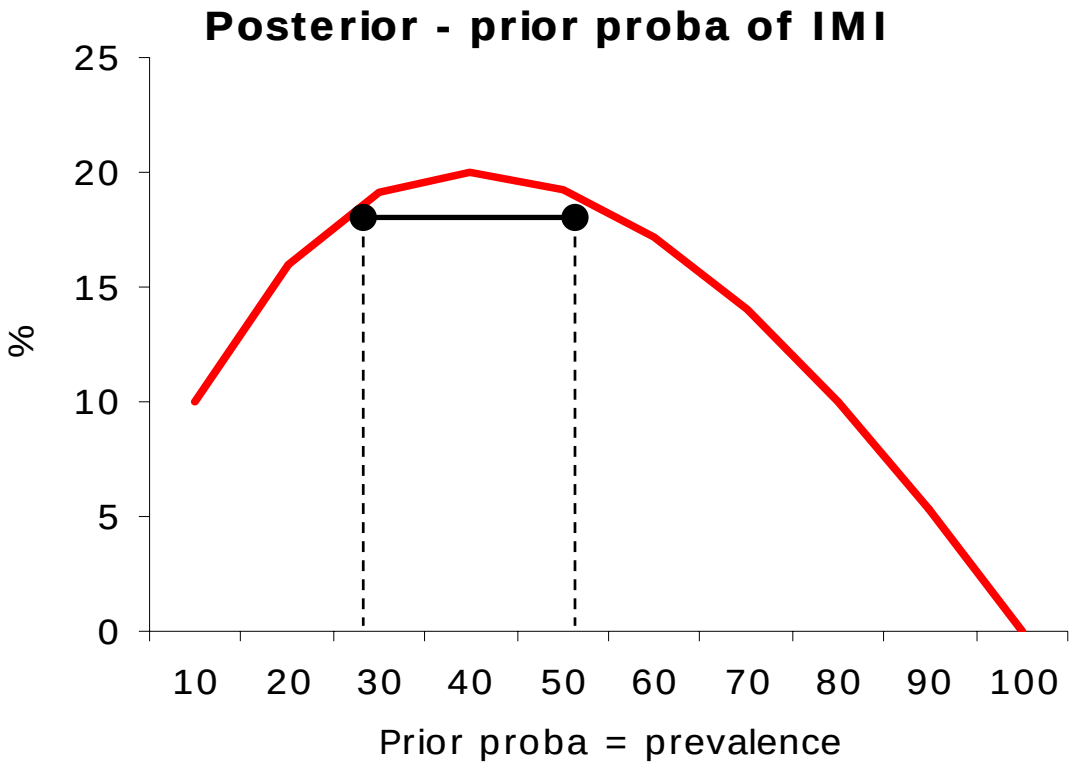
- corr(a,â)
- differences : $\theta - \hat{\theta}$ for μ_0, μ_1, $\sigma_a^2, \sigma_0^2, \sigma_1^2$

- sensitivity: $SE = \sum_{k=1,N} \sum_{t=1,T} p\left(\hat{z}_k^t = 1 \middle| z_k^t = 1\right)$
- specificity: $SP = \sum_{k=1,N} \sum_{t=1,T} pr\left(\hat{z}_k^t = 0 \middle| z_k^t = 0\right)$
- proba of correct classification:

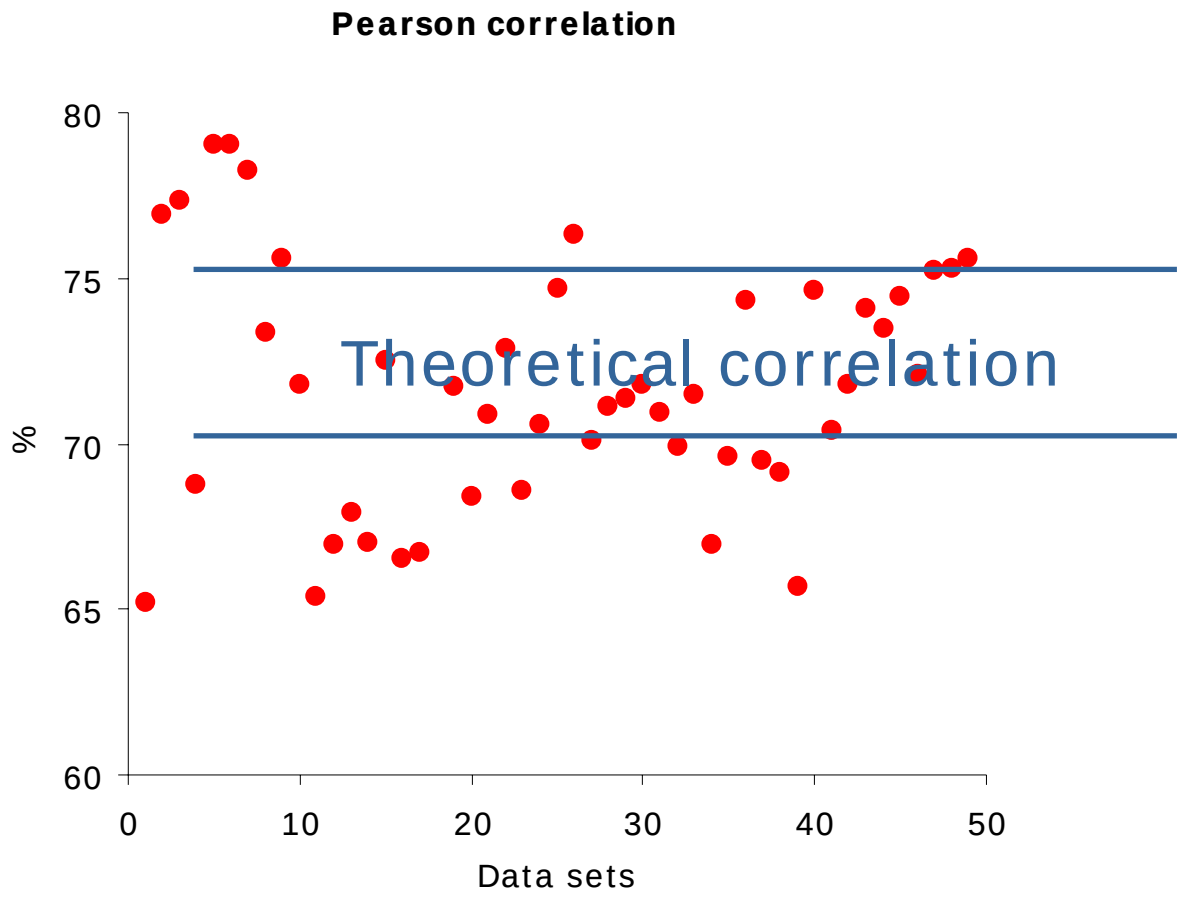
$$PCC = \sum_{k=1,N} \sum_{t=1,T} pr\left[\left(z_k^t = 1 \cap \hat{z}_k^t = 1\right) \cup \left(z_k^t = 0 \cap \hat{z}_k^t = 0\right)\right]$$


Sensitivity
Specificity
Correct classification
100
90
80
70
60
50
%
0
50
100
Proportion of E. coli among infected cows

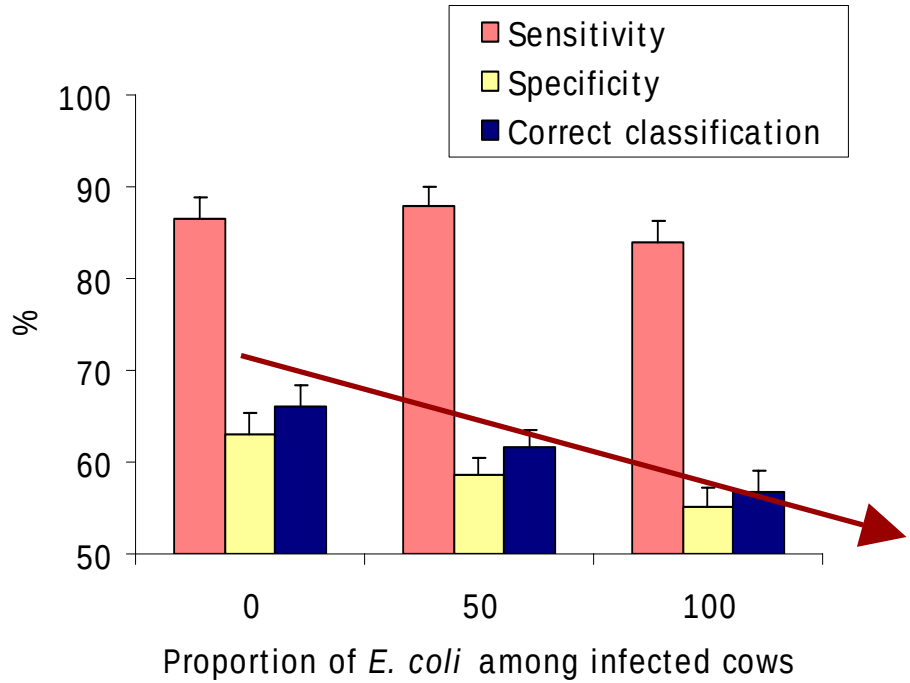
High SE = SNOUT

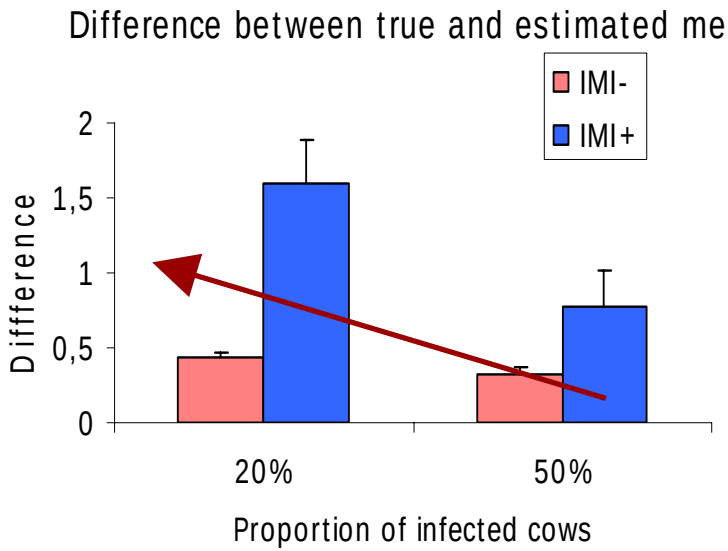


→ No further test

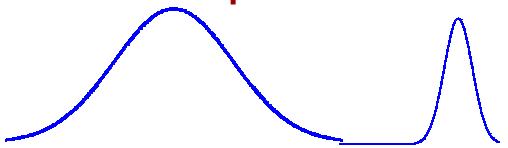

Pearson correlation
80
75
70
65
60
%
0
10
20
30
40
50
Data sets
Theoretical correlation

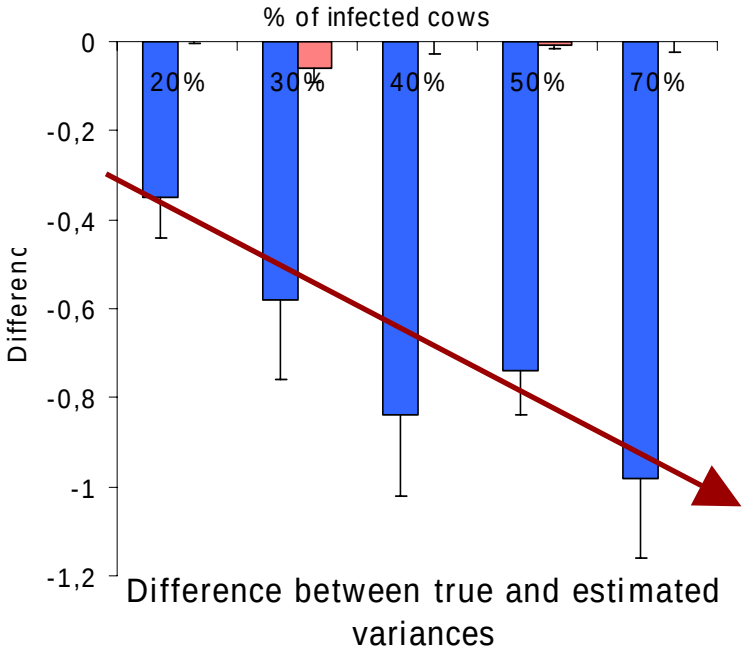
SP and PCC decreased when % *E. coli* among infected increased





Biases increased when disease prevalence is low



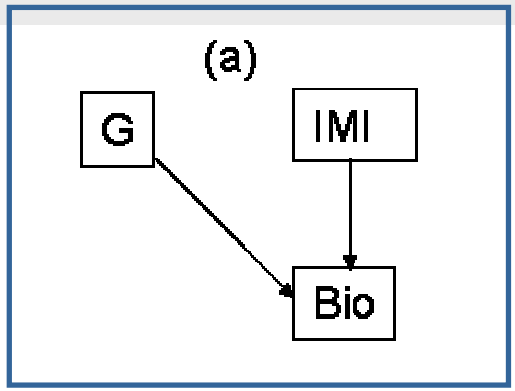


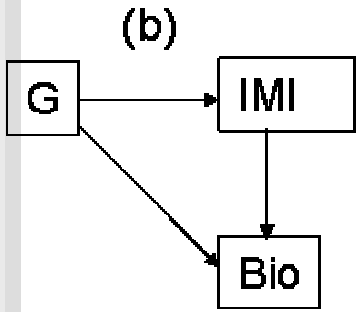
Biases increased when disease prevalence is high

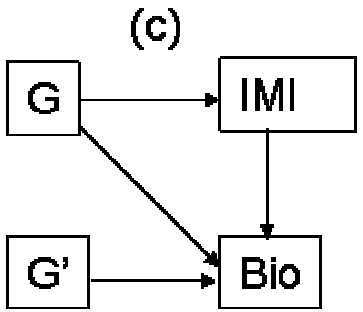
Conclusions

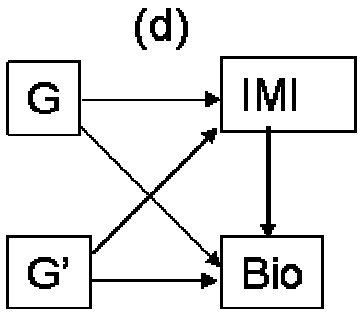
- Same amount of data
  - Increased accuracy of MLE
  - Resistance and tolerance
  - Transition probabilities
- Simplification of reality
  - Age, season, herd, ..
  - ‘Isolated’ proba of IMI-
  - Genetic relationship between cows for IMI

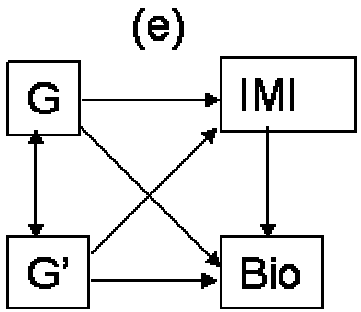
Genetic relationship among IMI and SCC












Thank
you!
ARTIE.COM