A NEW METHODOLOGY TO ANALYZE MONTHLY SOMATIC CELL COUNTS

J.C. Detilleux[1], P.L. Leroy[1]

[1]Faculty of Veterinary Medicine, University of Liege, Liege, Belgium

SUMMARY

A generalized linear finite mixture model and an expectation-maximization algorithm to fit the model to monthly somatic cell count data are discussed. The proposed algorithm can be readily implemented in available animal breeding packages. This model accounts for non-genetic effects affecting somatic cell counts differently in healthy and in infected cows.

Keywords : Finite mixture model, somatic cell counts, EM algorithm.

INTRODUCTION

Several studies have shown variability among cows for natural resistance or susceptibility to udder infections. Selection for resistance to intramammary infection (IMI) may therefore be possible but direct measures of IMI are not readily available. Usually, data on indirect indicators of IMI are on somatic cell counts (SCC). One important difficulty in using SCC to find animals most resistant to mastitis is that factors known to influence SCC are different in healthy from infected cows. If, in both infected and non infected cows, milking frequency, milk fractions, and management factors influence monthly SCC, in infected cows, SCC may also vary with type of mastitis pathogens, stage of lactation, and age of the cow (Detilleux et al. 1997). Two different statistical models should therefore be used to analyze SCC in healthy and infected cows. But, because intramammary infection status is generally unknown, one model is usually applied indifferently to SCC obtained from infected or non infected cows.

The objective of this study was to develop a new methodology for analyzing monthly SCC by considering SCC distribution among infected and non infected cows separately.

MATERIALS AND METHODS

The proposed approach is to apply to SCC data a generalized linear finite mixture model for which the following density function is assumed:

$$g(y) = \sum_{d=0}^{1} \pi_d g_d(y),$$

where

π_d = Pr (D = d) = the probability to be infected (d = 1) or not (d = 0),

and

$$g_d(y) = \frac{1}{2\pi^{.5N_d}} \frac{1}{|V|^{.5}} \exp\{-.5\ (y - X_d\beta_d)' V^{-1} (y - X_d\beta_d)\},$$

with y = SCC vector data; β_0 = vector of fixed effects for animals free of IMI; β_1 = vector of fixed effects for animals with IMI; X_0 and X_1 are design matrices relating β_0 and β_1 to y ; V = Var(y) which includes complete known genetic relationship between animals (V = Z A Z' σ^2_a + I σ^2_e).

Let $\theta = \left[\pi_1, \beta_0, \beta_1, \sigma_a^2, \sigma_e^2\right]$, then the likelihood equations are (Everitt and Hand 1981) :

$$0 = \sum_{d=0}^{1} \tau_d \frac{\partial \log(\pi_d)}{\partial \theta} + \sum_{d=0}^{1} \tau_d \frac{\partial \log(g_d(y))}{\partial \theta},$$

where

$$\tau_d = \frac{\pi_d\, g_d(y)}{\pi_0 g_0(y) + \pi_1 g_1(y)}$$

is the posterior probability to be infected (d = 1) or not (d = 0), knowing SCC.

Because the first term of the likelihood equation is a function of the prevalence of the disease (π_1) only, and the second term is a function of the parameters of the component distributions only, the likelihood equations may be split into 2 equations (Jansen, 1993) :

$$0 = \sum_{d=0}^{1} \tau_d \frac{\partial \log(\pi_d)}{\partial \theta},$$

and

$$0 = \sum_{d=0}^{1} \tau_d \frac{\partial \log(g_d(y)}{\partial \theta}.$$

The maximum likelihood estimates of the parameters can be obtained using the EM algorithm (Dempster, 1979). In the E step, the posterior probabilities τ_d are evaluated given the current parameter estimates for θ. In the M step, likelihood equations are solved by fixing τ_d. From the first part of the likelihood, mastitis prevalence is estimated as :

$$p_1 = \frac{\tau_1}{\tau_0 + \tau_1}$$

From the second part, it can be shown that REML estimates for β_0 and β_1, and for σ^2_a and σ^2_e are:

$$b_0 = (X_0'V^{-1}X_0)^- \; X_0'V^{-1}y,$$

$$b_1 = (X_1'V^{-1}X_1)^- \; X_1'V^{-1}y,$$

$$s_a^2 = p_0 s_{a0}^2 + p_1 s_{a1}^2,$$

$$s_e^2 = p_0 s_{e0}^2 + p_1 s_{e1}^2,$$

where s_{a0}^2, s_{a1}^2, s_{e0}^2, s_{e1}^2 are REML estimates of additive and error variances from 2 distributions with likelihoods :

$$g_0(y) = \frac{1}{2\pi^{.5N_0}} \frac{1}{|V|^{.5}} \exp\{-.5\,(y - X_0\beta_0)'V^{-1}(y - X_0\beta_0)\}$$

$$g_1(y) = \frac{1}{2\pi^{.5N_1}} \frac{1}{|V|^{.5}} \exp\{-.5\,(y - X_1\beta_1)'V^{-1}(y - X_1\beta_1)\}$$

Then, we go back to update the posterior probabilities and the EM cycle is repeated until convergence is reached. It can be seen from those equations that the algorithm can be implemented easily into programs traditionally used in animal breeding when estimating variance components and genetic breeding values. Also, prevalence of IMI can be estimated and different SCC breeding values can be computed for healthy or infected animals, separately .

CONCLUSIONS

Analyses of SCC as candidate for selection against mastitis resistance may be improved with generalized linear finite mixture models. With these models, REML estimates of additive genetic values are obtained more precisely, by correcting for known non-genetic effects affecting differently infected and healthy animals, without knowing the prevalence of the disease in the population. Because the likelihood equations of mixture model can be treated as likelihoods of 2 non-mixture models, the proposed algorithm can be readily implemented in available animal breeding packages.

REFERENCES

Dempster, A.P., Laird, N.M ., and Rubin, D.R. (1977) *J. Royal Stat Soc B* **39** : 1-38

Detilleux, J. C., Jacquinet E., Harvengt A., and leroy P. L. (1997) *Ann. Méd. Vét.* **141** : 199-205.

Everitt, B.S., and Hand, D.J. (1981) 'Finite mixture distributions' 1st ed. Cambridge University Press, London.

Jansen, R.C. (1993) *Biometrics* **49** : 227-231