Session G6.7

jorgen.odegard@iha.nlh.no

Mixture models – use and applications within the field of animal breeding

EAAP 2004, Bled, Slovenia

J. Ødegård,* P. Madsen,† D. Gianola,‡* G. Klemetsdal,* J. Jensen,† B. Heringstad,* and I. R. Korsgaard†

*Department of Animal and Aquacultural Sciences, Agricultural University of Norway, P.O. Box 5003, N-1432 Ås, Norway

†Department of Animal Breeding and Genetics, Danish Institute of Agricultural Sciences, Research Centre Foulum, P.O. Box 50, DK-8830 Tjele, Denmark

‡Department of Animal Sciences, University of Wisconsin-Madison, Madison, Wi 53706, USA

Abstract

It is well known that SCS increases as a result of udder infection, further the health status of each cow for each test-day on which SCS is recorded is usually unknown. Hence, the observed SCS can assumed being (at least) a two-component mixture depending on mastitis status. In this study a hierarchical two-component mixture model was developed, assuming that the health class membership associated with each test-day record of SCS was fully determined by an underlying liability variable. The *a priori* probability of mastitis may vary between different sub-groups, and the liability may thus be associated by some fixed and random effects. Based on analysis of simulated data, the model seemingly gives unbiased estimates of all parameters, and also provides a better tool for selection than crudely selecting for lower SCS. The proposed model could easily be extended to handle a wider range of problems related to genetic analyses of mixture traits.

Introduction

For certain traits an unknown underlying group structure may affect distribution of observations. An observation may therefore be drawn from K mutually exclusive and exhaustive distributions (or "groups"). In animal breeding, mixture models have so far primarily been used in QTL-analyses. Other examples of structures that may cause mixture distributions in quantitative traits are factors such as preferential treatment and disease. The latter may cause mixture distributions by mechanisms controlling the relationship between unobserved disease traits (categorical) and continuous traits. For example, SCS may be regarded as a trait sampled from either an "uninfected" or "mastitic" cow, where SCS in the "uninfected" and "mastitic" groups may be normally distributed with different means (e.g., Detilleux and Leroy, 2000), and possibly different variances. Mixture models can be used to categorize the observations into putative disease categories.

Acquiring information about unknown group structures affecting the data could be of great value for several reasons. In some cases, the main purpose is to correct for effects that may cause bias in genetic evaluations (e.g., preferential treatment). In other cases identification of structures, such as disease categories could be used for herd management decisions, medical treatment, and may also improve genetic evaluation of disease traits (e.g., mastitis) by making better use of information from related continuous (mixture) traits (e.g., SCS and electrical conductivity in milk).

So far mixture models for detection of mastitis based on SCS has been developed and analyzed on simulated data (e.g., Ødegård et al., 2003). In these models probability of mastitis

is estimated based on the observed SCS, and observations categorized into "healthy" and "mastitic" classes according to this probability. When calculating probability for mastitis, the model seeks to adjust for "base level" SCS of the cow. However, the *a priori* probability for mastitis is often assumed equal for all observations, which is not realistic for real data. Further, such models do not provide any good criteria for selection for lower incidence of mastitis. A hierarchical mixture model may be needed to implement a more flexible, practical and realistic model. In this model, probability for mastitis may depend on effects such as herd-test-day, stage of lactation, and additive genetic effects, and implies a direct approach for predicting breeding values for liability to mastitis using data coming from mastitis-related mixture traits. In the following we will shortly describe a simple version of this model.

Method

The setting and notation are as in Ødegård et al. (2003). Briefly, the data consists of *n* measurements for a quantitative trait, such as SCS of a cow. A 2-component Gaussian mixture model poses that the *i*th measurement (*i* = animal or record within animal), given location and dispersion parameters (α), and probabilities $\mathbf{P} = \left[P_1, P_2, ..., P_n\right]'$, has the distribution:

$$SCS_i \sim P_i \cdot N[f_i(\boldsymbol{\alpha}), g_i(\boldsymbol{\alpha})] + (1-P_i) \cdot N^*[f^*{}_i(\boldsymbol{\alpha}), g^*{}_i(\boldsymbol{\alpha})] \quad [1]$$

where P_i is the *a priori* probability that SCS_i is drawn from the distribution $\mathrm{N}(\cdot)$, and (1-P_i) is the *a priori* probability that it is drawn from $\mathrm{N}^*(\cdot)$. Ødegård et al. (2003) assumed that $P_i = P$, for all *i*, while in this study, P_i is allowed to differ between observations. Further, $f_i(\boldsymbol{\alpha})$, $g_i(\boldsymbol{\alpha})$, $f^*{}_i(\boldsymbol{\alpha})$, and $g^*{}_i(\boldsymbol{\alpha})$ are functions of the parameter vector $\boldsymbol{\alpha}$. Typically, $f_i(\boldsymbol{\alpha})$ and $f^*{}_i(\boldsymbol{\alpha})$ are linear combinations of fixed and random effects, $g_i(\boldsymbol{\alpha}) = \sigma^2_{e0_{SCS}}$, and $g_i{}^*(\boldsymbol{\alpha}) = \sigma^2_{e1_{SCS}}$ for all *i* = 1,2, ..., *n*. Conditionally on the parameters, α and **P**, the joint density of the data vector **SCS** is:

$$p(\mathbf{SCS}|\mathbf{P}, \boldsymbol{\alpha}) = \prod_{i=1}^{n}\left[P_i \cdot N\left(f_i(\boldsymbol{\alpha}), g_i(\boldsymbol{\alpha})\right) + (1-P_i) \cdot N^*\left(f^*{}_i(\boldsymbol{\alpha}), g^*{}_i(\boldsymbol{\alpha})\right)\right]$$

Estimation by maximum likelihood or by Bayesian approaches are facilitated by augmenting the density above with auxiliary indicator (0, 1) variables Z_i (*i* = 1, 2, …., n). It is assumed that

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathbf{P}, \boldsymbol{\alpha}) = \Pr(\mathbf{Z} = \mathbf{z} \mid \mathbf{P}) = \prod_{i=1}^{n} \Pr(Z_i = z_i \mid P_i) = \prod_{i=1}^{n} P_i^{z_i}(1-P_i)^{z_i}$$

and the joint density of **SCS** and **Z** is given by

$$\mathrm{p}\left(\mathbf{SCS}, \mathbf{z} \middle| \mathbf{P}, \boldsymbol{\alpha}\right) = \mathrm{p}\left(\mathbf{SCS} \middle| \mathbf{z}, \mathbf{P}, \boldsymbol{\alpha}\right)\Pr\left(\mathbf{Z} = \mathbf{z} \middle| \mathbf{P}\right) = \Pr\left(\mathbf{Z} = \mathbf{z} \middle| \mathbf{SCS}, \mathbf{P}, \boldsymbol{\alpha}\right)\mathrm{p}\left(\mathbf{SCS} \middle| \mathbf{P}, \boldsymbol{\alpha}\right). \quad [2]$$

Hence,

$$\Pr\left(\mathbf{Z} = \mathbf{z} \middle| \mathbf{SCS}, \mathbf{P}, \boldsymbol{\alpha}\right) = \frac{\mathrm{p}\left(\mathbf{SCS} \middle| \mathbf{z}, \mathbf{P}, \boldsymbol{\alpha}\right)\Pr\left(\mathbf{Z} = \mathbf{z} \middle| \mathbf{P}\right)}{\mathrm{p}\left(\mathbf{SCS} \middle| \mathbf{P}, \boldsymbol{\alpha}\right)}$$

is the conditional probability distribution of **Z**, given **SCS,** α and **P**. If it can be assumed that $\left(SCS_i, Z_i\right)$, *i=1,...,n*, are mutually independent given α and **P**, then

$$\Pr\left(z_i = 1 \middle| SCS_i, P_i, \boldsymbol{\alpha}\right) = \frac{\mathrm{p}\left(SCS_i \middle| z_i = 1, P_i, \boldsymbol{\alpha}\right)P_i}{\mathrm{p}\left(SCS_i \middle| z_i = 0, P_i, \boldsymbol{\alpha}\right)(1-P_i) + \mathrm{p}\left(SCS_i \middle| z_i = 1, P_i, \boldsymbol{\alpha}\right)P_i} \quad [3]$$

is the posterior probability (given SCS_i, α and **P**) that the draw is made from $\mathrm{N}^*(\cdot)$ (mastitis), whereas the complement is the posterior probability that the draw is from $\mathrm{N}(\cdot)$. Here, SCS_i is the somatic cell score for record *i*.

In this model we postulate the existence of an underlying continuous random variable, called liability (λ), which determines the actual mastitis status for each observation. This is a threshold-liability model (Wright, 1934; Dempster and Lerner, 1950; Gianola, 1982; Gianola and Folley, 1983), which has been used for genetic analysis of clinical mastitis as a binary response (e.g., Heringstad, 2003). Here, the liability is incorporated into what we term a liability-normal mixture (**LNM**) model. In both models true mastitis status goes from 0 to 1 if liability exceeds a given threshold *T*. In the standard threshold liability model, data consists of observed binary responses (say, "presence" or "absence" of clinical mastitis), whereas in the LNM model, data consist of observed SCS. However, distribution of SCS changes from $N(\cdot)$ to $N^*(\cdot)$ according to mastitis status, and putative mastitis status may therefore be inferred based on the observed SCS. In this model the *a priori* probability of IMI+, for a specific observation *i*, is:

$$P_i = \Pr(\lambda_i > T|\boldsymbol{\alpha}) = \Pr(\lambda_i > T|\boldsymbol{\beta}_\lambda, \mathbf{s}_\lambda, \mathbf{p}_\lambda) = \Phi\left(\mathbf{x}_{i\lambda}'\boldsymbol{\beta}_\lambda + \mathbf{z}_{is_\lambda}'\mathbf{s}_\lambda + \mathbf{z}_{ip_\lambda}'\mathbf{p}_\lambda\right) = \Phi(\tilde{\lambda}_i), \quad [4]$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Thus, P_i is not a parameter in this model, but a function of the expected liability. The threshold is assumed to be equal to zero.

Modeling SCS and the liabilities

Conditionally on **Z=z** and parameter vector α, we assume that the SCS and λ variables can be modeled as:

$$\mathbf{y} = \begin{pmatrix} \mathbf{SCS} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{\mathbf{0}_{\mathbf{SCS}}}\boldsymbol{\beta}_{\mathbf{0}_{\mathbf{SCS}}} + \mathbf{M}_\mathbf{z}\mathbf{X}_{\mathbf{1}_{\mathbf{SCS}}}\boldsymbol{\beta}_{\mathbf{1}_{\mathbf{SCS}}} + \mathbf{Z}_{\mathbf{a}_{\mathbf{SCS}}}\mathbf{a}_{\mathbf{SCS}} + \mathbf{Z}_{\mathbf{p}_{\mathbf{SCS}}}\mathbf{p}_{\mathbf{SCS}} + \mathbf{e}_{\mathbf{zSCS}} \\ \mathbf{X}_\lambda\boldsymbol{\beta}_\lambda + \mathbf{Z}_{\mathbf{a}_\lambda}\mathbf{a}_\lambda + \mathbf{Z}_{\mathbf{p}_\lambda}\mathbf{p}_\lambda + \mathbf{e}_\lambda \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{X}_{\mathbf{zSCS}}\boldsymbol{\beta}_{\mathbf{SCS}} + \mathbf{Z}_{\mathbf{a}_{\mathbf{SCS}}}\mathbf{a}_{\mathbf{SCS}} + \mathbf{Z}_{\mathbf{p}_{\mathbf{SCS}}}\mathbf{p}_{\mathbf{SCS}} + \mathbf{e}_{\mathbf{z}\,\mathbf{SCS}} \\ \mathbf{X}_\lambda\boldsymbol{\beta}_\lambda + \mathbf{Z}_{\mathbf{a}_\lambda}\mathbf{a}_\lambda + \mathbf{Z}_{\mathbf{p}_\lambda}\mathbf{p}_\lambda + \mathbf{e}_\lambda \end{pmatrix} \quad [5]$$

$$= \mathbf{X}_\mathbf{z}\boldsymbol{\beta} + \mathbf{Z}_\mathbf{a}\mathbf{a} + \mathbf{Z}_\mathbf{p}\mathbf{p} + \mathbf{e}_\mathbf{z}$$

where; **y** = column vector of SCS and liability variates, $\mathbf{M_z}$ = (*n*×*n*) diagonal matrix of indicator variables, with typical element z_i (*i* = 1, 2, ..., n), $\boldsymbol{\beta}_{\mathbf{0}_{\mathbf{SCS}}}$ = vector of "fixed" effects affecting SCS common to all cows, $\boldsymbol{\beta}_{\mathbf{1}_{\mathbf{SCS}}}$ = vector of "fixed" effects affecting SCS peculiar to cows with mastitis, $\boldsymbol{\beta}_{\mathbf{SCS}} = \left[\boldsymbol{\beta}_{\mathbf{0}_{\mathbf{SCS}}}', \boldsymbol{\beta}_{\mathbf{1}_{\mathbf{SCS}}}'\right]'$, $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_{\mathbf{SCS}}', \boldsymbol{\beta}_\lambda'\right]'$, $\mathbf{a} = \left[\mathbf{a}_{\mathbf{SCS}}', \mathbf{a}_\lambda'\right]'$ = vector of random additive genetic effects, $\mathbf{p} = \left[\mathbf{p}_{\mathbf{SCS}}', \mathbf{p}_\lambda'\right]'$ = vector of random permanent environmental effects, $\mathbf{e}_\mathbf{z} = \left[\mathbf{e}_{\mathbf{zSCS}}', \mathbf{e}_\lambda'\right]'$ = vector of random residuals, and $\mathbf{X}_\lambda, \mathbf{Xz}_{\mathbf{SCS}}, \mathbf{X}_{\mathbf{0}_{\mathbf{SCS}}}, \mathbf{X}_{\mathbf{1}_{\mathbf{SCS}}}, \mathbf{Z}_{\mathbf{a}_{\mathbf{SCS}}}, \mathbf{Z}_{\mathbf{a}_\lambda}, \mathbf{Z}_{\mathbf{p}_{\mathbf{SCS}}}$ and $\mathbf{Z}_{\mathbf{p}_\lambda}$ are incidence matrices of appropriate order, where $\mathbf{X}_{\mathbf{zSCS}} = \left[\mathbf{X}_{\mathbf{0}_{\mathbf{SCS}}}' \quad \left(\mathbf{M}_\mathbf{z}\mathbf{X}_{\mathbf{1}_{\mathbf{SCS}}}\right)'\right]'$, and $\mathbf{e}_{\mathbf{zSCS}} = \left(z_i e_{i0_{SCS}} + (1-z_i)e_{i0_{SCS}}\right)_{i=1,\ldots,n}$, where $e_{i0_{SCS}} \sim N(0, \sigma^2_{0_{SCS}})$, $e_{i1_{SCS}} \sim N(0, \sigma^2_{1_{SCS}})$; $e_{i0_{SCS}}$ and $e_{i1_{SCS}}$, *i* = 1, ..., *n*, are assumed to be mutually independent.

Standard prior distributions were assumed for all location and dispersion parameters (Inverse Wishart for variance-covariance matrices, random effects were assumed normally distributed,

and proper flat priors were assumed for "fixed" effects). Residual correlation between SCS and λ was not estimable with this model, and was therefore set to zero. Estimation was carried out with Gibbs sampling.

Simulation study

The model was tested using simulating data. Four different scenarios were chosen, and each scenario was replicated 20 times. For all scenarios, four generations, each consisting of 800 cows from 10 sires, were simulated. Mastitis frequency was set to 25%. Residual variance for SCS was assumed homogeneous, independent of disease category. For comparison purposes three models were fitted; a standard repeatability model for SCS ignoring the mixture (**IM**), a mixture model for SCS ignoring the structure of the liability (**NM**) (equivalent to Ødegård et al., 2003), and finally a LNM model. The input parameters and means of posterior means for the same parameters estimated with the LNM model are presented in Table 1.

Table 1. *Input (IP) and estimated parameters (EP) for four different scenarios, with standard error (SE). Estimated parameters are reported as means of posterior means for 20 replicates from each scenario.*

Scenario		$\sigma^2_{a_{SCS}}$	$\sigma^2_{a_\lambda}$	$r_{a_{SCS,\lambda}}$	$\sigma^2_{p_{SCS}}$	$\sigma^2_{p_\lambda}$	$r_{p_{SCS,\lambda}}$	$\sigma^2_{e_{SCS}}$
1	IP	0.100	0.120	0.000	0.100	0.120	0.000	0.800
	EP	0.101	0.114	0.039	0.098	0.126	0.031	0.796
	SE	0.012	0.023	0.118	0.012	0.026	0.111	0.015
2	IP	0.100	0.120	0.500	0.100	0.120	0.000	0.800
	EP	0.110	0.129	0.468	0.091	0.109	0.068	0.798
	SE	0.015	0.025	0.083	0.011	0.021	0.122	0.012
3	IP	0.100	0.120	-0.500	0.100	0.120	0.000	0.800
	EP	0.099	0.112	-0.447	0.102	0.125	0.009	0.804
	SE	0.015	0.025	0.098	0.015	0.032	0.118	0.015
4	IP	0.100	0.059	0.000	0.100	0.125	0.000	0.800
	EP	0.094	0.078	0.085	0.105	0.133	0.001	0.804
	SE	0.013	0.017	0.132	0.013	0.030	0.127	0.015

The means of the posterior means for all parameters were not significantly different from their true values. However, if the structure of the underlying liability is ignored, parameters for SCS were confounded with those of the simulated liability. Further, specificity and (even more) sensitivity were slightly reduced (Table 2).

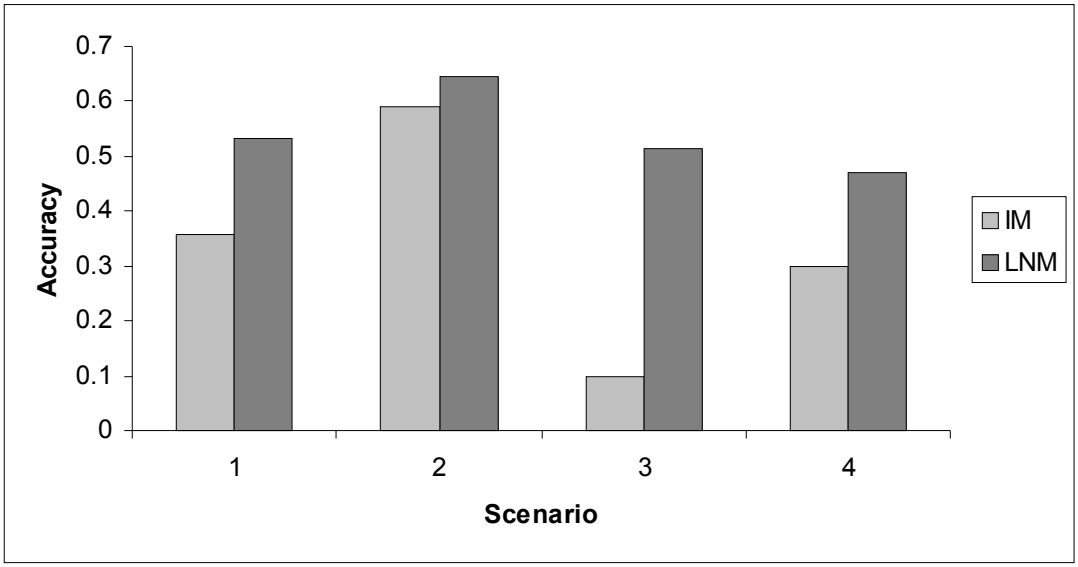
Table 2. *Sensitivity and specificity estimated with a non-hierarchical normal mixture model (NM) and a hierarchical liability - normal mixture model (LNM). Parameters are reported as means of posterior means for 20 replicates from each scenario.*

Scenario		NM	LNM
1	Sensitivity	0.630	0.660
	Specificity	0.879	0.885
2	Sensitivity	0.622	0.664
	Specificity	0.887	0.893
3	Sensitivity	0.618	0.646
	Specificity	0.877	0.882
4	Sensitivity	0.616	0.641
	Specificity	0.890	0.893

An advantage of the LNM model, compared with other mixture models is that it provides EBVs for liability to putative mastitis, which may be directly used in selection. In Figure 1 correlations between true breeding values for liability to mastitis and EBVs from IM and LNM are presented for the different scenarios. Compared to the standard test-day model for SCS, ignoring the mixture, EBVs for liability to mastitis from the LNM model had consistently higher correlations (9-530%) with the true breeding values, particularly when

assuming a negative genetic correlation between "baseline SCS" and liability to putative mastitis (indicating that high SCS in healthy cows reduces risk of infection).

Figure 1. *Correlations between true breeding values for liability to mastitis and predicted breeding values (accuracy) estimated with a standard model for SCS ignoring the mixture (IM) and a liability – normal mixture model (LNM). Average correlations for 20 replicates from each scenario are presented.*



Future aspects

In addition to selecting for cows able to avoid infection, we may also be interested in the cows' ability to recover when infection occurs. This may be achieved by developing a mixture model where size of SCS response to infection has a genetic component, which in turn may be related to probability of recovery from disease. The model could also easily be extended to multivariate mixtures, consisting of multiple mixture traits depending on same mixture variable (e.g., SCS and electrical conductivity in milk), different mixture variables (e.g., SCS and a trait affected by QTL), or both mixtures and non-mixture traits. More advanced mixtures, consisting of more than two components (e.g., healthy, subclinical mastitis, clinical mastitis) may also be developed.

Conclusion

Inferring an underlying liability to mastitis in mixture models for mastitis-related mixture traits probably gives a more realistic and accurate model, both in terms of genetic evaluations and identification of diseased animals. Based on simulation studies the model seemingly gives unbiased estimates of the parameters. The proposed model could easily be extended to handle a wider range of problems related to genetic analyses of mixture traits.

References

Dempster, E. R., and I. M. Lerner. 1950. Heritability of threshold characters. Genetics 35:212-235.

Detilleux, J., P. L. Leroy. 2000. Application of a mixed normal mixture model for the estimation of mastitis-related parameters. J. Dairy Sci. 83:2341-2349.

Gianola, D. 1982. Theory and analysis of threshold characters. J. Anim. Sci. 54:1079-1096.

Gianola, D., and J. L. Foulley. 1983. Sire evaluation for ordered categorical data with a threshold model. Genet. Sel. Evol. 15:201-223.

Heringstad, B., G. Klemetsdal, and J. Ruane. 1999. Clinical mastitis in Norwegian Cattle: Frequency, variance components, and genetic correlation with protein yield. J. Dairy Sci. 82:1325-1330.

Heringstad, B., R. Rekaya, D. Gianola, G. Klemetsdal, and K. A. Weigel. 2003. Genetic change for clinical mastitis in Norwegian Cattle: a threshold model analysis. J. Dairy Sci. 86:369-375.

Ødegård, J., J. Jensen, P. Madsen, D. Gianola, G. Klemetsdal, and B. Heringstad. 2003. Detection of mastitis in dairy cattle by use of mixture models for repeated somatic cell scores: A Bayesian approach via Gibbs sampling. J. Dairy Sci. 86:3694-3703.

Sorensen, D., and D. Gianola. 2002. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer-Verlag, New York, NY.

Wright, S. 1934. An analysis of variability in number of digits in an inbred strain of Guinea pigs. Genetics 19:506-536.