FURTHER INSIGHTS INTO TESTS OF VARIANCE COMPONENTS AND MODEL SELECTION

C. Delmas[1], JL. Foulley, [1]C. Robert-Granié[2]

[1]INRA-SGQA, 78350 Jouy-en-Josas, [2]INRA-SAGA, BP 27, 31326 Castanet-Tolosan France

INTRODUCTION

Since the pioneer works of Henderson (1973, 1984), linear mixed effects models are the cornerstone of statistical analysis in animal breeding. Later Laird and Ware (1982) and Diggle (1988) developped random effects models to analyse longitudinal data. In that framework it is often interesting to test the significance of some fixed or random coefficients in a model (see for example Foulley *et al.,* 2000). For that purpose, likelihood ratio tests method appears to be a natural and relevant technique. However, though tests of fixed effects are classical and well known, tests of variance components are nonstandard and often misunderstood. The purpose of this article is to discuss the problem of likelihood ratio tests for variance components and to give some new interpretations in terms of model selection. Our discussion is based on the results for nonstandard testing situations by Self and Liang (1987) and Stram and Lee (1994).

LIKELIHOOD RATIO TESTS

In many situations it is desired to test that a parameter lies in an r-dimensional subset of the p-dimensional parameter space with r<p, versus the alternative that it lies in the complement of this subset in the parameter space. If, under the null hypothesis, the true value of the parameter is an interior point of the parameter space then, under suitable regularity conditions, the asymptotic distribution of the likelihood ratio test (LRT) statistic is a chi-square with p-r degrees of freedom. This result is due to Wilks (1938).

In the case of variance components testing such classical conditions are not satisfied because the true value of the parameter under the null hypothesis is a boundary point of the parameter space. Therefore the nullity of the first partial derivatives of the log-likelihood function is no longer a necessary condition of optimality since the maximum can be attained on the boundary of the parameter space. In that case of maximization under constraints the necessary conditions of optimality are the Karush, Khün, Tucker conditions.

Under some mild regularity conditions, Self and Liang (1987) proved that the asymptotic distribution of the LRT statistic is that of :

$$\sup_{\theta\in(C_\Omega-\theta_0)}[-(Z-\theta)^T I(\theta_0)(Z-\theta)]-\sup_{\theta\in(C_{\Omega_0}-\theta_0)}[-(Z-\theta)^T I(\theta_0)(Z-\theta)]$$

where Ω (resp. Ω_0) stands for the parameter space (resp. the parameter space under the null hypothesis), θ_0 denotes the true value of the parameter under the null hypothesis, $I(\theta_0)$ denotes the Fisher information at θ_0 supposed positive definite, C_Ω (resp. C_{Ω_0}) stands for the tangent plane to Ω (resp. Ω_0) at θ_0 and Z is multivariate Gaussian with mean 0 and variance $I^{-1}(\theta_0)$. The above expression can be rewritten as :

$$\inf_{\tilde{\theta} \in K_0} \| \tilde{Z} - \tilde{\theta} \|^2 - \inf_{\tilde{\theta} \in K} \| \tilde{Z} - \tilde{\theta} \|^2$$

where $\tilde{Z}$ is centred multivariate Gaussian with identity variance and K and K_0 are the corresponding adequate subsets. In most cases the distribution of such an expression can be shown to be a mixture of chi-square distributions.

Stram and Lee (1994) applied this result to the problem of variance components testing. In the general framework of a linear model with (q+1) random coefficients (RC) they studied the test of q RC versus (q+1). Under the null hypothesis, the q RC were assumed to be linearly independent. Then the asymptotic distribution of the LRT statistic was proved to be a 50 :50 mixture of two chi-square distributions with q and (q+1) degrees of freedom.

FURTHER INTERPRETATIONS

The complete parameter space is a kind of paraboloïd in the (q+1)-dimensional euclidian space (where the origin would be the null hypothesis; where the q covariances of the last random coefficient with the others would be on the first q axes and its variance on the last axis) . Thus, in the complete model, the maximum likelihood estimator of the parameter can be an interior point of the parameter space or a boundary point. Asymptotically, since the tangent plane of the parameter space at the origin is an half-space, these two alternatives have the same probability 0.5. Let us denote n the number of parameters under the null hypothesis. When the maximum likelihood estimator is an interior point of the parameter space, then the likelihood of the complete model with n+(q+1) parameters is higher than the likelihood of any sub-model. On the contrary, when it is a boundary point, this means that the likelihood of the sub-model with n+q parameters is higher than the likelihood of the complete model. In this sub-model, the last RC is a linear combinaison of the first q. Thus the q regression coefficients are the q extra parameters. Note that there are also sub-models with n+i parameters for all i=1 to (q-1). In those sub-models the last RC is also a linear combinaison of the first q but some regression coefficients are zero. Those sub-models cannot be found by a maximum likelihood procedure since they correspond to parameters that lie in the interior of the boundary but on some sub-spaces of inferior dimension. Thus they are of probability zero. All these arguments explain more clearly the 50: 50 mixture of chi-square distributions with q and (q+1) degrees of freedom. They finally lead us to be aware of the existence of many sub-models with fewer parameters that were hidden at the first stage of the analysis but that must not be forgotten. Note that the number of all the sub-models plus the complete model is 2^q .

When the LRT leads us to accept the null hypothesis, we exactly know the model accepted. On the opposite when the LRT is significant we are led to reject the null hypothesis but we do not know which model to accept among all the possible 2^q (sub)-models. If the maximum likelihood estimator is on the boundary of the parameter space it is clear that a sub-model with at most n+q parameters is preferable to the complete model. When it is an interior point then we do not know if the complete model with n+(q+1) parameters is significantly better than the other sub-models with n+i parameters with i in 1 to q. Thus, at this stage of the analysis, we are confronted with a problem of model selection. Many approaches can be used to solve this problem. A first one could be based on the ideas of Akaïke (1974) or Schwarz (1978). We will not go further in this direction in this paper. We prefer to emphasize the second method that could be based on recursive testing procedures. We begin with k=0. The first step is to test a linear model with k RC versus (k+1) RC. The asymptotic distribution of the LRT statistic can be proved to be a 50:50 mixture of chi-square distributions with p and (p+1) degrees of freedom where p denotes the number of linearly independent RC under the null hypothesis. If the test is not significant we accept the null hypothesis and the algorithm is stopped. If it is significant, the second step is to test the set of all the sub-models (when they exist) versus the complete model. The asymptotic distribution of the LRT statistic can be proved to be a 50 :50 mixture of chi-square distributions with 0 and 1 degree of freedom. If the test is significant we go back to the first step with k=k+1. If it is not we test the nullity of all the regression coefficients by classical chi-square procedures since, in those cases, 0 is not a boundary point of the parameter space. When a sub-model is accepted we go back to the first step with k=k+1. At the end of this algorithm an « ideal » model has been selected.

A numerical application to facial growth data (Pothoff and Roy (1964)) was studied. Four linear models have been considered for the analysis of these data.

Model 1 : $Y_{ijk} = \mu + \alpha_i + \beta_i t_j + \varepsilon_{ijk}$

Model 2 : $Y_{ijk} = \mu + \alpha_i + \beta_i t_j + a_{ik} + \varepsilon_{ijk}$

Model 3 : $Y_{ijk} = \mu + \alpha_i + \beta_i t_j + a_{ik} + b_{ik} t_j + \varepsilon_{ijk}$

Model 4 : $Y_{ijk} = \mu + \alpha_i + \beta_i t_j + a_{ik} + b_{ik} t_j + c_{ik} t_j^2 + \varepsilon_{ijk}$

The subscript i stands for the sex, j for the measurement period with t_j the corresponding time. The subscript k stands for the intra-sex individual. $u_{ik} = (a_{ik}, b_{ik}, c_{ik})^T$ for all i and k are assumed $N(O,D)$ independent with $D = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} \\ \sigma_{ab} & \sigma_b^2 & \sigma_{bc} \\ \sigma_{ac} & \sigma_{bc} & \sigma_c^2 \end{pmatrix}$. ε_{ijk} for all i, j and k are assumed independent $N(O,\sigma_e^2)$ and independent of u_{ik} for all i and k. Results are given in Table 1. Note that the testing procedures lead us to choose model 2. In the case of model 4, neither the SAS estimation procedure nor the AS-REML algorithm converge in the parameter space whereas the PX-EM algorithm does. Thus it seems that some available algorithms are not always well adapted to maximization under constraints.

Table 1.: Summary of the results obtained in the analysis of growth data

	Model 1	Model 2	Model 3	Model 4	
				PX-EM	*SAS*
σ_a^2		337.27	835.50	789.57	1977.71
σ_{ab}			-46.53	-39.38	-266.75
σ_b^2			4.42	3.47	46.60
σ_{ac}				0.93	5.07
σ_{bc}				-0.055	-0.963
σ_c^2				0.0011	0
Eigenvalues of D				791 1.5 10e-8	2013.7 10.43 -0.02
-2L	884.04	843.65	842.36	840.71	841.78
P-value		1vs2 : 0	2vs3 : 0.39	3vs4 :0.54	

CONCLUSION

Through a detailed analysis we have emphasized the existence of sub-models with few parameters that have a real importance in variance components testing and model selection. We have proposed a solution to take them into account in a close study. Nevertheless all the theoretical results that must be used to test under nonstandard conditions appear to be hardly implemented since the available algorithms are not always adapted to maximization under constraints. The main issue adressed in this paper is the question of dimensionality of the variance matrix of the RC. To that respect the eigen functions and values proposed by Kirkpatrick and Heckman (1989) might be helpful tools for further research directions.

REFERENCES

Akaïke, H. (1974) *IEEE Trans. Automat. Control* **19 :** 716-723.

Diggle, P.J. (1988) *Biometrics* **44 :** 959-971.

Foulley, J.L., Jaffrézic, F., Robert-Granié, C. (2000) *Genet. Sel. Evol.* **32 :** 129-141.

Henderson, C.R. (1973) *Proceedings of the animal breeding and genetics symposium in honor of Dr J. Lush, Am. soc. Animal Science-Am. Dairy Science Assoc., Champaign,* 10-41.

Henderson, C.R. (1984) Applications of linear models in animal breeding*, Univ. of Guelph*.

Kirkpatrick, M. and Heckman, N. (1989) *J. Math. Biol.* **27 :** 429-450.

Laird, N.M. and Ware, J.H. (1982) *Biometrics* **38 :** 963-974.

Pothoff, R.F. and Roy, S.N. (1964) *Biometrika* **51 :** 313-326.

Schwarz, G. (1978) *Ann. Stat.* **6 :** 461-464.

Self, S.G. and Liang, K.L. (1987) *J. Am. Stat. Ass.* **82 :** 605-610.

Stram, D.O. and Lee, J.W. (1994) *Biometrics* **50 :** 1171-1177.

Wilks, S.S. (1938) *Ann. Math. Stat.* **9 :** 60.