DEGREE OF CONNECTEDNESS IN MIXED MODELS

J.J. Tosh and J.W. Wilton
Centre for Genetic Improvement of Livestock
University of Guelph, Guelph, Ontario, N1G 2W1, Canada

SUMMARY

Degree of connectedness is defined for each factor in a model to quantitatively describe the design of data. In this simulation study, estimates of residual variance and of solutions and standard errors obtained from analyses of disconnected data and of groups of connected subclasses of the same data were compared. The factor with the greatest number of levels controlled formation of groups, and degree of connectedness of the complete design. Analysis of all data together appeared appropriate when degree of connectedness was high.

INTRODUCTION

Field data often have a very small proportion of the total possible subclasses filled, and a highly unbalanced design. The data contain connected subsets of design points represented by filled subclasses. All linear contrasts of fixed effects within a connected subset are estimable. Fernando et al. (1983) argue that random effects are always estimable so contrasts among them can be made without consideration of connectedness. This is true if levels of the factor are randomly sampled from a population but disconnectedness arises with non-random sampling. Schaeffer (1975) and Petersen (1978) consider connectedness when analyzing random effects. When some comparisons among levels of a factor are not valid, solutions are difficult to interpret. Disconnectedness can also interfere with the row space or rank of equations so that degrees of freedom for tests of hypotheses differ from those expected. To avoid these problems, each group of connected subclasses is analyzed separately, but with thousands of levels of random factors, few efforts are made to determine whether the design is connected. The objectives of this study were to quantitatively define degree of connectedness, and to compare analyses of data disconnected to varying degrees with analyses of connected groups of the same data, using mixed models, in order to determine conditions where analysis of all data together is acceptable.

DEGREE OF CONNECTEDNESS

Degree of connectedness is determined after filled subclasses are grouped into connected subsets. Weeks and Williams (1964) describe a method to group nearly-identical N-tuples in an N-way cross-classification. Pairwise comparisons among levels of each factor are recorded for each connected group. Let the number of unique pairs across all groups be n_{pi} for the i-th factor. Degree of connectedness of the i-th factor is defined as

$$c_i = n_{pi}/[n_i(n_i-1)/2]$$

where n_i is the total number of levels of the i-th factor ($n_i \geq 2$). Total degree of connectedness for the complete design considers all factors and is

defined as

$$c_T = (\Sigma n_{pi})/(\Sigma[n_i(n_i-1)/2]).$$

Degree of connectedness under these definitions is the probability of making valid pairwise comparisons among factor levels when analyzing all data together. Values of degree of connectedness range from 0 (all levels of the factors are disconnected) to 1 (complete connectedness).

Example: Consider three factors (n_A=2, n_B=4, n_C=6) in a design with filled subclasses that form two groups of connected 3-tuples.

Group 1	(1,1,1)	(1,2,2)	(2,1,6)
	(1,1,6)	(1,2,4)	(2,2,2)
	(1,2,1)	(2,1,1)	(2,2,6)
Group 2	(1,3,3)	(1,4,3)	(2,3,5)
	(1,3,5)	(2,3,3)	(2,4,5)

The pairwise comparisons are

	Group 1	Group 2	n_{pi}	$n_i(n_i-1)/2$	c_i
Factor A	1-2	1-2	1	1	1.00
Factor B	1-2	3-4	2	6	0.33
Factor C	1-2, 1-4, 1-6, 2-4, 2-6, 4-6	3-5	7	15	0.47

and the degree of connectedness of the complete design is c_T=0.45.

ANALYSIS OF A MIXED MODEL WITH DIFFERENT DEGREES OF CONNECTEDNESS

Data for a balanced 3-way cross-classified design were simulated. A mixed model with two fixed factors (n_A=10, n_B=20) and a random factor (n_C=200) was used. Each cell had one observation. The model in matrix notation was

$$\mathbf{y}=\mathbf{Xb}+\mathbf{Zu}+\mathbf{e}$$

where **y** is a 40000 X 1 vector of observations,
b is a 30 X 1 vector of assumed fixed effects,
u is a 200 X 1 vector of assumed random effects ~N(0,48),
e is a 40000 X 1 vector of random residual effects ~N(0,72).
The design matrices **X** and **Z** were composed of zeroes and ones. Expected value of **y** was **Xb**. Expected values of random vectors were null vectors. $Var(\mathbf{u})=\mathbf{I}\sigma_u^2$ and $var(\mathbf{e})=\mathbf{I}\sigma_e^2$, and covariances were assumed null.

To mimic unbalanced field data, 1% of the total number of subclasses (n=400) were chosen systematically. Subclasses eligible for selection were restricted to those representing only certain levels of each factor, or $(1/2)^3$=12.5% of the total subclasses. Five sets of data were chosen with attempts to produce different degrees of connectedness. The five sets of observations had equal numbers of levels (n_A=5, n_B=10, n_C=100).

For each set of data, connected data points were grouped by the method of Weeks and Williams (1964), and degrees of connectedness were calculated (Table 1). Mixed model equations (Henderson, 1973) were solved for each connected group and for all groups together in each data set. The model was similar to

Table 1. Characteristics of the five sets of data analyzed.

Data Set (n=400)	Number of Groups of Connected Subclasses	Connectedness c_A	c_B	c_C	c_T
1	10	0.600	0.111	0.091	0.092
2	7	0.300	0.356	0.242	0.244
3	7	0.200	0.822	0.354	0.357
4	2	1.000	0.444	0.495	0.496
5	4	0.400	0.644	0.859	0.856

that used to simulate the completely cross-classified data. The rank of the coefficient matrix was used to check for interference of connectedness. Unexpected dependencies were observed when some levels of both fixed factors only existed in the same connected group. In 3-way models, if one factor is connected and there is more than one group the other two factors always coexist. Sets of data with unexpected dependencies were not analyzed.

Estimates of residual variance, solutions for pairwise comparisons among levels of each factor, and standard errors of solutions were obtained from analyses of each group of connected subclasses and of all groups together for each of the five data sets. Results were expressed as deviations from those obtained from analysis of the simulated completely cross-classified data (n=40000). Residual variances were estimated by

$$\hat{\sigma}_e^2 = (y'y - \hat{b}'X'y - \hat{u}'Z'y)/(n - r(X))$$

where n is the number of observations in the analysis. Sums of squares of the deviations of solutions and of the standard errors of comparisons were divided by the number of valid pairs in each analysis. Connected groups contained only pairs that were estimable but all groups together could have solutions to pairs that were not estimable. Weighted means for groups within each data set were calculated; the weight was n for residual variances, and was number of valid pairs in each group for sums of squares of deviations of solutions and standard errors. Weighted means of analyses of groups of connected subclasses were compared to single results from analysis of all data together.

RESULTS AND DISCUSSION

Degree of connectedness of the random factor was negatively associated with number of groups of connected subclassses (Table 1). The number of levels of this factor was many times greater than the other factors in the model. Formation of the groups of connected subclasses, and the equation for total degree of connectedness, was dominated by the random factor.

Estimates of residual variance depend on n, and on the design of data and r(X) which are influenced by degree of connectedness. Analysis of all data together, whether connected or not, increased n providing an estimate of residual variance that had a smaller deviation from that obtained from the completely cross-classified data (Table 2). Analysis of all data together did not appear to have a greater advantage when there was a larger number of groups of connected subclasses. More work is needed to investigate the effect of connectedness on rank of equations and variance estimates.

Sums of squares of deviations of solutions (Table 2) were related to

Table 2. Results of two methods of analysis (CG=groups of connected subclasses analyzed separately; AG=all groups analyzed together) expressed as deviations from results of completely cross-classified data.

Data Set	Analysis*	Absolute Deviation of $\hat{\sigma}_e^2$	Sum of Squares of Deviations of Solutions Factor A	Factor B	Factor C	Sum of Squares of Deviations of Standard Errors# Factor A	Factor B	Factor C
1	CG	14.09	8.42	10.36	1.07	0.0226	0.0156	0.0068
	AG	8.66	0.71	36.91	1.30	0.0076	0.0392	0.0005
2	CG	7.62	10.09	6.24	0.28	0.0058	0.0084	0.0028
	AG	2.52	30.83	2.65	0.18	0.0000	0.0009	0.0004
3	CG	12.54	2.95	0.64	0.26	0.0032	0.0053	0.0022
	AG	2.54	16.32	0.19	0.09	0.0000	0.0014	0.0001
4	CG	3.70	1.48	2.87	0.05	0.0088	0.0044	0.0003
	AG	1.49	0.22	2.76	0.05	0.0097	0.0072	0.0001
5	CG	10.95	3.90	7.56	0.09	0.0146	0.0173	0.0011
	AG	3.34	23.74	1.21	0.02	0.0265	0.0041	0.0001

* For CG analyses, results are weighted means of groups.
# These numbers are not yet multiplied by $\hat{\sigma}_e^2$.

degree of connectedness. When degree of connectedness was high, the sum of squares was lower for the analysis of all data together. This was a better method of analysis since n was greater and the probability of making incorrect pairwise comparisons was low. With a low degree of connectedness individual groups of subclasses provided better solutions because only estimable pairwise comparisons were made. The method of analysis that is recommended depends on the degree of connectedness of the factor of interest.

Sums of squares of deviations of standard errors (Table 2) did not show an association with degree of connectedness or number of observations in the analysis. For the random factor, the sum of squares was consistently lower when all data was analyzed together. Analysis of all data together gave better standard errors of prediction.

Random factors have greater numbers of levels than fixed factors, and determine the degree of connectedness of the complete design. In animal breeding, the random animal solutions are of interest. If animals are highly connected in actual field data, analysis of all data together is recommended.

REFERENCES

FERNANDO, R.L., GIANOLA, D., and GROSSMAN, M. 1983. J.Dairy Sci. 66:1399-1402.

HENDERSON, C.R. 1973. Proc.Anim.Breed. & Genet.Symp. in honor of Dr. Jay L. Lush. A.S.A.S. and A.D.S.A., Champaign, IL. p. 10-41.

PETERSEN, P.H. 1978. Acta Agric.Scand. 28:360-362.

SCHAEFFER, L.R. 1975. Biometrics 31:969-977.

WEEKS, D.L., and WILLIAMS, D.R. 1964. Technometrics 6:319-324.