Well-Connected, Informative Sub-Sets of Data

Hossein Jorjani
Interbull Centre
Department of Animal Breeding and Genetics
Swedish University of Agricultural Sciences
S - 750 07, Uppsala, Sweden
Hossein.Jorjani@hgen.slu.se

Abstract
Two measures of connectedness for individual bulls and one measure of connectedness for countries are presented in this paper. For individual bulls the spread of a bull's daughters in different countries is compared with a hypothetical, yet realistic and common situation in which all daughters have their records in only one country. Standard deviation of number of daughters or sum of the absolute differences from the mean number of daughter have been used to measures the spread. Both of these two methods could provide an "effective number of proofs". For countries the measure is based on the proportion of cows sired by common bulls between the two countries. These measures would facilitate the choice of country combinations and / or bulls in estimation of genetic correlations between countries.

Introduction

In estimating genetic correlations between countries, as implemented at the Interbull Centre, we need to work with sub-sets of data in form of specific country combinations, using only those bulls that have multiple proofs in different countries plus full-sibs or ¾ sibs of such bulls. Hitherto the choice of country combinations has been based on our prior knowledge about genetic ties between specific countries and the fact that bulls from some countries are good link-providers. However, the current practice eludes automatization, and further, maybe considered as subjective by some. Therefore, it is desirable to find a quantitative method that can help us to accomplish the task of choosing a well-connected sub-set of data to be used in the estimation process.

Desirable properties of such a method were outlined in a previous study (Jorjani, 1999) as:

1) It should be quantitative, preferably bound between 0 and 1;
2) It should be able to yield a measure of connectedness for individual bulls (in contrast with connectedness between management groups, i.e. countries in our case);
3) It should be able to take into account some kind of weighting factor, examples of which are number of daughters, national reliabilities, nationally obtained (genetic or phenotypic) parameters or a combination of such factors;
4) It should be able to reflect genetic relationships between bulls by incorporating some information from the relationship matrix; and finally,
5) It should be able to avoid all other steps that are included in the international bull evaluations so that it could be used in the screening process (included in the usual data checks).

The aims of this paper are to report the results obtained from two new measures of connectedness for bulls and to examine the problems in using a previously introduced measure of connectedness, i.e. genetic similarity (Rekaya et al., 1999).

Connectedness for Individual Bulls

In order to distinguish between connectedness caused by pedigree links and by the presence of proofs in more than one country, I will refer to the former as genetic connectedness (GC) and to the latter as statistical connectedness

(SC). Mark (1999) proposed to measure SC in a two-way layout, e.g. bull-country by looking at the spread of daughters of a bull across countries and the spread of bulls across countries in the following manner:

$$SC = 1 - \sum_{i=1}^{b} \frac{\sum_{j=1}^{c} | n_{ij} - \bar{n}_{i.} |}{4(c-1)b\ \bar{n}_{i.}} - \sum_{j=1}^{c} \frac{\sum_{i=1}^{b} | n_{ij} - \bar{n}_{.j} |}{4(b-1)c\ \bar{n}_{.j}}$$

in which *b* is the number of bulls, *c* is the number of countries and n_{ij} is the number of daughters of bull *i* in country *j*.

The first part of the right hand side of the above equation was modified in two different ways to obtain connectedness values for individual bulls. In the first modification sum of the absolute values of differences (AD) and in the second modification standard deviation of number of daughters of a bull in different countries (SD) were used as measures of spread of a bull's number of daughters across countries. In both cases the maximum possible spread was used to standardize the connectedness value, resulting to two measures defined as SC_AD and SC_SD, as follows:

$$(SC\_AD)_i = 1 - \frac{(\sum_{j=1}^{c} | n_{ij} - \frac{\bar{n}_{i.}}{c} |)/c}{\frac{| n_{i.} - \frac{n_{i.}}{c} | + (c-1) | \frac{n_{i.}}{c} |}{c-1}}$$

and

$$(SC\_SD)_i = 1 - \frac{\sqrt{(\sum_{j=1}^{c} n_{ij}^2 - \frac{n_{i.}^2}{c})/c}}{\sqrt{\frac{(n_{i.} - \frac{n_{i.}}{c})^2 + (c-1)(\frac{n_{i.}}{c})^2}{c-1}}}$$

It is easy to see that for a bull with only one proof SC_AD yields a value equal to *1/c*. Further, it is possible to use both methods to obtain an effective number of proofs for each bull by multiplying them by the number of countries.

Each of these two measures of connectedness can be weighted by some values related to the total number of daughters. In this study I chose to use $\sqrt{n_{ij}}$ as the weighting factor.

Connectedness for Countries

To measure connectedness for countries the concept of genetic similarity, introduced by Rekaya et al. (1999) was used. Genetic similarity between two countries is the fraction of cows sired by bulls with proofs in both counties in proportion to the total number of cows in the two countries:

$$GS_{ij} = \frac{\sum_{k=1}^{2} \sum_{l=1}^{cb_{ij}} n_{lk}}{\sum_{k=1}^{2} \sum_{l=1}^{tb_{ij}} n_{lk}}$$

in which N_{lk} is the number of daughters for each bull, CB_{ij} is the number of bulls with proofs in both countries *i* and *j*, and TB_{ij} is the total number of bulls in countries *i* and *j*.

Results

Table 1 shows number of bulls and proofs for the trait milk yield in the six breeds of evaluation.

Table 1. Summary of number of bulls and their number of proofs

# of proofs	# of bulls					
	AYR	BSW	GUE	HOL	JER	SIM
1	9377	7872	1309	85516	6546	24657
2	236	346	66	4004	343	733
3	56	83	15	1164	72	123
4	17	59	6	522	38	44
5	6	25		347	17	23
6	8	13		191	8	7
7	2	14		123	6	
8		7		81	5	
9		2		72		
10		1		54		
11				47		
12				35		
13				37		
14				27		
15				19		
16				20		
17				9		
18				16		
19				4		
20				4		

Source: INTERBULL, March 2000 test-run

To illustrate the effects of these two measures of connectedness an excerpt from the result for GUE is shown in Table 2.

Table 2. Some examples of the statistical connectedness values

Np	N_{ij}	SC		WSC	
		AD	SD	AD	SD
1	15	0.250	0.134	0.97	0.52
1	468	0.250	0.134	5.41	2.90
1	1175	0.250	0.134	8.57	4.59
1	6168	0.250	0.134	19.63	10.52
2	51	0.500	0.500	3.57	3.57
2	633	0.348	0.243	8.75	6.11
2	1396	0.282	0.170	10.52	6.35
2	10726	0.253	0.137	26.15	14.18
3	243	0.750	0.696	11.69	10.85
3	686	0.551	0.470	14.43	12.31
3	1414	0.360	0.260	13.55	9.76
3	1431	0.582	0.576	22.01	21.79
4	504	0.772	0.722	17.33	16.22
4	9907	0.359	0.259	35.70	25.83
4	2947	0.385	0.290	20.92	15.72
4	2847	0.469	0.387	25.03	20.64

Np=number of proofs; N_{ij}=number of daughters, SC=statistical connectedness, WSC=weighted statistical connectedness, AD & SD= absolute difference and standard deviations (for details see text)

As it can be seen in Table 2, both measures of statistical connectedness are quite sensitive to the imbalances in distribution of daughters between countries. A noteworthy observation is that some bulls with very low number of daughters get high connectedness values, while other bulls with the same number of proofs and a large number of daughters receive low connectedness values, as if the latter bulls are punished, some times severely, for strong imbalance. This seems, however, to be a fortunate consequence, because strong imbalances are usually possible only when semen from a proven bull, with possibility of having a second batch of daughters, are imported. In other words, simultaneous progeny testing is rewarded in these measures of connectedness.

Summary of the results obtained from the two measures of connectedness for the trait milk yield in the six breeds of evaluation are presented in Table 3. Although the effective number of proofs for individual bulls are not shown in Table 2, however, the result shown in Table 3 indicates that the effective number of proofs obtained from SC_AD is potentially a more useful than just simple number of proofs.

Table 3. Number of bulls and proofs, and averages of the various measures of connectedness

	AYS	BSW	GUE	HOL	JER	SIM
1	9702	8422	1396	92292	7035	25587
2	10177	9434	1510	107680	7815	26825
3	1.049	1.120	1.082	1.167	1.111	1.048
4	406.4	319.1	259.8	541.5	428.4	336.7
5	8	10	4	26	8	8
6	0.129	0.109	0.263	0.042	0.135	0.128
7	1.033	1.088	1.053	1.104	1.077	1.028
8	2.132	1.615	3.233	0.793	1.869	1.737
9	0.071	0.066	0.150	0.034	0.081	0.071
10	0.571	0.656	0.601	0.888	0.650	0.568
11	1.201	1.142	1.903	0.854	1.267	1.000

1- Number of bulls, 2- number of proofs, 3- mean number of proofs / bull, 4- average number of daughters / bull, 5- number of countries participating in the INTERBULL evaluations, 6- average SC_AD, 7- average effective number of proofs using SC_AD, 8- average weighted number of proofs, 9- average SC_SD, 10- average effective number of proofs using SC_SD, 11- average weighted number of proofs.

Genetic Similarity

Results of implementing genetic similarity concept have been presented before and I will elaborate more on it in the Discussion. Here only one example from AYR is presented (Table 4).

Table 4. Degree of genetic similarity between pair-wise country combinations in Ayrshire bull populations

	FIN	NOR	SWE	USA	NZL	AUS	GBR
CAN	.008	.000	.046	.382	.145	.297	.225
FIN		.011	.096	.004	.007	.005	.003
NOR			.052	.000	.000	.026	.000
SWE				.013	.006	.136	.007
USA					.065	.117	.102
NZL						.357	.210
AUS							.251

Comparison of these results with the last year's results indicate clearly that genetic similarity has improved markedly in the AYR populations, especially for SWE and AUS.

Discussion

We have already seen that dividing countries into well-connected sub-groups of countries by using the concept of genetic similarity leads to higher estimates of genetic correlations. However, the problem with the matrix of genetic similarity is that it cannot be examined visually if the number of countries is large. Therefore, the work on this subject should continue until we can find a quantitative method, suitable for automatization, which can divide countries into groups of well-connected sub-set. One suggestion could be to use eigenvalues / eigenvectors or some kind of decomposition, e.g. Cholesky or LU, to better understand the nature of this matrix.

In order to see the effects of choosing a sub-set of data on estimated genetic correlations a test-run, involving AYR, was performed. To choose the subset only bulls with a SD_AD value of larger than 0.2 were used. The preliminary results indicate that within each group of well-connected countries the estimated genetic correlations showed an increase, while in poor-connected group of countries, or between the two groups of countries the estimated genetic correlations showed a decrease. This is understandable, because in case of well-connected countries by using SC_AD only more informative bulls with a higher average number of proofs enter the analysis. In the poor connected group using SC_AD leads to further loss of already scant information and therefore it becomes even more difficult to detect covariances.

On the surface it seems that statistical connectedness (SC) and genetic similarity (GS) are two independent measures. However, this may not be entirely true. Because in both SC_AD and SC_SD a bull's number of daughters plays a roll and it is obvious that the more daughters a bull with multiple proofs has, the more he contributes to GS. This becomes even more so if the weighted values of SC_AD or SC_SD are used.

To reiterate the problem I can say that, for example, in the Holstein population there are a large number of countries and a large number of bulls. To estimate genetic parameters and breeding values in a reasonably short window of time we need to choose a sub-set of data. Up to now we have been restricted by large number of bulls coming from link-provider countries and therefore have been forced to reduce number of countries that are included in the analysis. This in turn leads to problems in getting a positive definite correlation matrix. There seems to be two alternatives to solve this problem.

1) To create one phantom country, with all bulls with a high effective number of proofs, to be present in all country combinations.
2) To identify link-provider countries by looking at the average values of each country's corresponding row in the GS matrix, and from these countries only use bulls with the highest effective number of bulls.

Of course in any case one can use weighted values of SC instead of un-weighted ones.

Acknowledgements

Fruitful discussions with colleagues at the INTERBULL Centre, especially Thomas Mark and financial support from USDA/NAAB is gratefully appreciated.

References

Jorjani, H. 1999. Connectedness in dairy cattle populations. *INTERBULL Bulletin 25*, 21-24

Mark, T. 1999. International genetic evaluation of dairy sires for clinical mastitis and somatic cell count. *MSc thesis, Royal Veterinary & Agricultural University of Copenhagen.* 111pp.

Rekaya, R., Weigel, K.A. & Gianola. D. 1999. Bayesian estimation of a structural model for genetic covariances for milk yield in five regions of the USA. *50th Meeting of the EAAP. Zurich. Switzerland.* 22-26 August 1999.