Screening for outliers in multiple trait genetic evaluation

Per Madsen[1], Jukka Pösö[2], Jørn Pedersen[3],

Martin Lidauer[4] and Just Jensen[1]

[1]Centre for Quantitative Genetics and Genomics, Aarhus University, Denmark

[2]Faba co-op, Finland

[3]The Knowledge Centre for Agriculture, Cattle, Denmark

[4]MTT, Agrifood Research Finland, Biotechnology and Food, Genetic Research Group, Finland

Objectives

- Develop, implement and test a simple multivariate method for detection of extreme outliers before data is used in genetic evaluations
- Test the effect of deleting extreme outliers from genetic evaluation

Multivariate outlier

Assumptions:

$$x \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1.0 & 0.6 & 0.8 \\ 0.6 & 1.0 & 0.9 \\ 0.8 & 0.9 & 1.0 \end{bmatrix}, \quad x = \begin{bmatrix} -2 \\ -2 \\ +2 \end{bmatrix}$$

Computing the conditional distribution of $x_3 | x_1 x_2$ gives expectation -2.125 and variance 0.0844

This means that this conditional variable deviates 14.2 SD units from its expectation

Theoretical development

Model:

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

$$\text{where } \mathrm{var}(\mathbf{a}) = \mathbf{A} \otimes \mathbf{G}_0 \text{ and } \mathrm{var}(\mathbf{e}) = \mathbf{I} \otimes \mathbf{R}_0$$

Model for record *i*:

$$\mathbf{y}_i = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$$

$$\mathbf{d}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}$$

$$\mathrm{var}(\mathbf{d}_i) = \mathrm{var}(y_i) + X_i C^{xx} X_i^{'} = Z_i G_0 Z_i^{'} + R_0 + X_i C^{xx} X_i^{'} = D_i$$

where C^{xx} is the part of the inverse coefficent matrix related to fixed effects

Mahanalobis distance

$$M_i = \sqrt{d_i' D_i^{-1} d_i}$$

Under the assumption that $\mathbf{d}_i$ is multivariate normal with zero means and covariance matrix $\mathbf{D}_i$

$$M_i^2 \sim \chi_t^2$$

Approximation to Mahanalobis distance

In large scale genetic evaluation, computation of C^{xx} is not possible

However, D_i is dominated by $Z_i G_0 Z_i^{'} + R_0$

$$\text{Partition } b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

where b_1 contains fixed effects estimated with great accuracy

$$E(y_i) \cong X_i \begin{bmatrix} b_1^p \\ b_2 \end{bmatrix}$$

where b_1^p could be solutions from previous evaluatuion

Approximation to Mahanalobis distance (cont.)

Define *s(i)* as a vector valued function to compute the phenotypic SD of all observed traits in record *i*

The Mahanalobis distance can be approximated as follows:

$$\hat{\mu}_i = \mathrm{X}\begin{bmatrix} \mathrm{b}_1^p \\ 0 \end{bmatrix}$$

$$\mathrm{D}_i^* = diag(s(i))diag(\mathrm{D}_i)^{-1/2}\mathrm{D}_i diag(\mathrm{D}_i)^{-1/2} diag(s(i))$$

$$M_i^2 = (y_i - \hat{\mu}_i)'\mathrm{D}_i^{*-1}(y_i - \hat{\mu}_i)$$

Setting cut-off-points

A very simple tool for setting cut-off-points is the chi-Square plot (Garrett, 1989), where M^2 are ordered and plotted against their corresponding χ^2-values

That is the I^{th} ranked M^2 out of N records, with cumulative probability $p=(I-0.5)/N$ is plotted against $\chi_t^2 = Ci(p,t)$, where $Ci(p,df)$ is the inverse of the cumulative Chi square probability function and *df* is degrees of freedom.

This curve is expected to follow a straight line if:

$$\mathrm{d}_i \sim N(0, \mathrm{D}_i^*)$$

Multivariate outlier editing: NAV Jersey evaluation

Data:

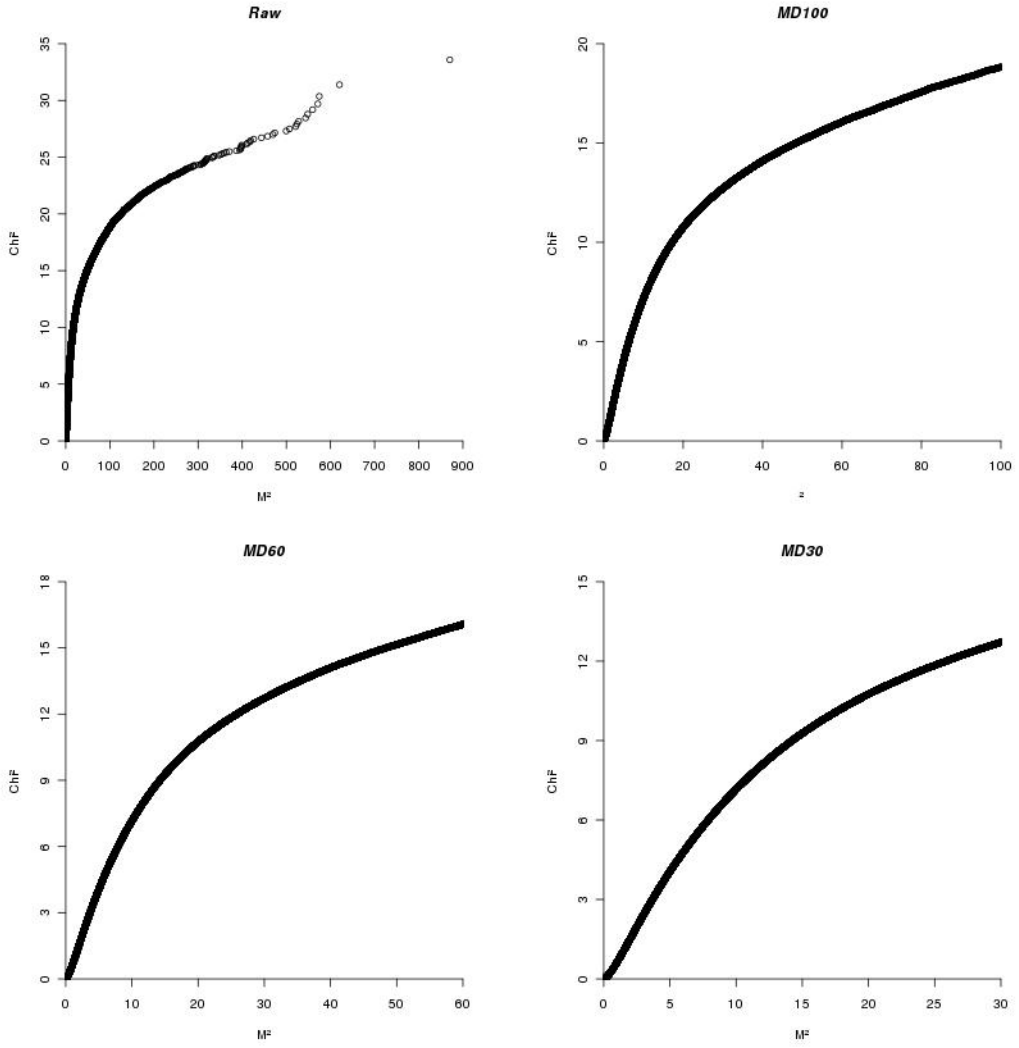
No. of test-day records:	9 884 497
No. of cows:	568 392
No. of traits:	3 (Milk, Protein and Fat)

Model:

The current NAV Jersey Test-day model.

305d EBV's are expressed as indexes standardized to a mean of 100 for a four-years cohort of cows and a standard deviation of 10 for a two-year cohort of bulls

χ^2 plot for different editing rules



If the data is multi variate normal distibuted, it should be a straight line

Scenarios for deleting extreme/outlier records

Situation	Description	No of records deleted (no of cows)	% records deleted
Raw	All data used	0	0
MD100	Records with M^2>100 deleted	801 (788)	0.0081
MD60	Records with M^2>60 deleted	3172 (2991)	0.0321
MD30	Records with M^2>30 deleted	17029 (14156)	0.1723

Predictive ability

Correlation between trait EBV's from "Full" and "Reduced" data for cows having all their records in the last 4 years (no records in "Reduced" dataset)

Predictive ability for different categories of cows

Edit rule	Data used for prediction: Raw			
	Cows having records classified by M^2			
	No limit	M^2 >100	M^2 >60	M^2 >30
No. of cows	96698	226	854	3593
Trait				
Milk	0.58	0.53	0.51	0.51
Protein	0.59	0.55	0.55	0.56
Fat	0.55	0.52	0.52	0.53

Predictive ability for cows having record(s) with $M^2 > 30$ deleted

	Data used in prediction			
	Raw	MD100	MD60	MD30
# of cows	3593	3593	3593	3593
Trait				
Milk	0.51	0.51	0.52	0.53
Protein	0.56	0.57	0.57	0.59
Fat	0.53	0.54	0.54	0.57

BULL: Change in indices between evaluation on "Raw" and "MD30"

Trait	Number of bulls by magnitude of change in index									
	-4	-3	-2	-1	0	1	2	3	4	5
Milk	0	1	6	84	13733	124	8	0	0	1
Protein	0	3	6	114	13441	373	16	2	2	0
Fat	5	2	23	421	12406	1040	49	9	1	1

Cows: Change in indices between evaluation on "Raw" and "MD30"

Trait	Number of cows by magnitude of change in index								
	-17 - -4	-3	-2	-1	0	1	2	3	4 – 32
Milk	115	243	992	8073	728446	9960	1135	322	265
Protein	146	315	1428	10199	718814	16302	1547	447	894
Fat	571	969	3140	23037	682622	31329	4332	1823	1788

INTERBULL validation test 3

	Raw	MD100	MD60	MD30
Milk	-5.15 ns	-5.42 ns	-5.62 ns	-5.06 ns
Protein	-0.18 ns	-0.17 ns	-0.17 ns	-0.15 ns
Fat	-0.23 ns	-0.22 ns	-0.22 ns	-0.20 ns

Conclusions

- An outlier detection rule based on an approximate Mahanalobis distance is easy to implement
- Application of such a rule requires determination of an optimum cut-off-point
- A series of analysis using the same structure as the INTERBULL 3 validation test can be applied to determine this optimum
- Use of such a rule will increase the accuracy of predicted breeding values for the animals involved and will also remove potential bias in contemporary animals

Predictive ability for different categories of cows

Edit rule	Data used for prediction: MD100		
	Cows having records classified by M^2		
	$M^2 > 100$	$M^2 > 60$	$M^2 > 30$
# of cows	226	854	3593
Trait			
Milk	0.60	0.53	0.51
Protein	0.62	0.57	0.57
Fat	0.57	0.53	0.54

Predictive ability for different categories of cows

Edit rule	Data used for prediction: MD60	
	Cows having records classified by M^2	
	$M^2 > 60$	$M^2 > 30$
# of cows	854	3593
Trait		
Milk	0.54	0.52
Protein	0.59	0.57
Fat	0.55	0.54

Predictive ability for different categories of cows

Edit rule	Data used for prediction: MD100		
	Cows having records classified by M^2		
	$M^2 > 100$	$M^2 > 60$	$M^2 > 30$
# of cows	226	854	3593
Trait			
Milk	0.60	0.53	0.51
Protein	0.62	0.57	0.57
Fat	0.57	0.53	0.54

Predictive ability for different categories of cows

Edit rule	Data used for prediction: MD60	
	Cows having records classified by M^2	
	$M^2 > 60$	$M^2 > 30$
# of cows	854	3593
Trait		
Milk	0.54	0.52
Protein	0.59	0.57
Fat	0.55	0.54