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 Research, Genetic Research Group*

Abstract

Use of multivariate models in genetic evaluation requires a multivariate method for detecting erroneous outliers that cannot be detected using univariate methods. A simple rule for detecting outliers based on an approximated Mahanalobis distance was applied to Jersey data from the routine Nordic genetic evaluation in dairy cattle. Application of such is simple to implement and increased the accuracy of predicted breeding values for animals that has one or more records edited. Potential biases in evaluations for contemporary animals were also reduced. Optimum editing rules can be determined using the same data structures as used in the standard INTERBULL test for model verification.

Introduction

In all routine genetic evaluation procedures new data goes through a cleaning step before being included in the dataset used for genetic evaluation. No standard guidelines exist for development of such procedures. The procedures used needs to be computationally efficient and be relatively simple to implement. Simple rules are also advantageous in communication to end users of genetic evaluations.

Traditionally, procedures used for data cleaning and outlier detection have been implemented on a per trait basis such that observations with low univariate density have been excluded. For normal data this is equivalent to exclude an observation if it deviates more than a preset number of standard deviation units from the expectation.

However, in the multivariate case such simple univariate procedures may not be sufficient. Consider a simple example of data

$$\text{from } N(\mathbf{0},\Sigma) \text{ where } \Sigma = \begin{bmatrix} 1.0 & 0.6 & 0.8 \\ 0.6 & 1.0 & 0.9 \\ 0.8 & 0.9 & 1.0 \end{bmatrix}$$

Assume that we observe a vector of data

$$\mathbf{x} = \begin{bmatrix} -2 \\ -2 \\ +2 \end{bmatrix}.$$

All the observations deviate either -2 or +2 standard deviation units from the expectation and are well within the range of acceptable normally distributed data when ignoring the underlying multivariate distribution.

However, if we for illustration compute the conditional distribution of $x_3 \mid x_1\ x_2$ we find that this conditional variable has expectation -2.125 and variance 0.0844. This means that this conditional variable deviates 14.2 SD units from its expectation. A deviation which is well outside of what would be expected from normally distributed data. This is a clear indication that the observation is a multivariate outlier.

The purpose of this study was to develop, implement and test a simple multivariate method for detection of extreme outliers before data is used in genetic evaluations. Different strategies for eliminating outliers are compared using the standard INTERBULL methods for national model verification and by assessment of the predictive ability of the model used.

Materials and Methods

Theoretical development

Consider the following multivariate mixed linear model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad (1)$$

where $\text{var}(\mathbf{a}) = \mathbf{A} \otimes \mathbf{G}_0$ and $\text{var}(\mathbf{e}) = \mathbf{I} \otimes \mathbf{R}_0$, and $\mathbf{y}$, $\mathbf{a}$ and $\mathbf{e}$ are assumed to be multivariate normally distributed. The covariance matrices $\mathbf{G}_0$ and $\mathbf{R}_0$ are assumed known as usual in genetic evaluation practice. For simplicity we do not consider repeated observations or similar in the model. However, extensions to such examples are straightforward.

The model for record i:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{a} + \mathbf{e}_i \qquad (2)$$

where $\mathbf{y}_i$ is a vector of length t, the number of traits included in the analysis, and $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{e}_i$ are corresponding sub matrices and sub vectors of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{e}$, respectively.

If we compute residual deviations for a single multivariate record as:

$$\mathbf{d}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}} \qquad (3)$$

$$\begin{aligned} \text{var}(\mathbf{d}_i) &= \text{var}(\mathbf{y}_i) + \mathbf{X}_i\mathbf{C}^{xx}\mathbf{X}_i^{'} \\ &= \mathbf{Z}_i\mathbf{G}_0\mathbf{Z}_i^{'} + \mathbf{R}_0 + \mathbf{X}_i\,\mathbf{C}^{xx}\mathbf{X}_i^{'} \\ &= \mathbf{D}_i \end{aligned}$$

where $\mathbf{C}^{xx}$ is the sub matrix corresponding to the fixed effects in the inverse of the coefficient matrix of the MME assuming that there is no inbreeding.

We can now test for extreme observations by computing $x_{ij} = {}^{d_{ij}}\!/\!_{\sqrt{\mathbf{D}_{jj}}}$ where subscript j indicates trait j. This quantity will have a standard normal distribution and two sided critical values from the standard normal distribution can be used to determine cut off points for observations considered to be univariate outliers.

In the classical statistical literature testing for multivariate deviation refers to a measure called Mahanalobis distance:

$$\mathbf{M}_i = \sqrt{\mathbf{d}_i^{'}\mathbf{D}_i^{-1}\mathbf{d}_i} \qquad (4)$$

as a measure of multivariate distance (Penny, 1996) and (Pena and Prieto, 2001). As shown by (Krzanowski, 2000):

$$\mathbf{M}_i^2 \sim \chi_t^2 \qquad (5)$$

under the assumption of $\mathbf{d}_i$ being multivariate normal with zero mean and covariance matrix $\mathbf{D}_i$. Thus, actual values of $\mathbf{M}_i^2$ for record i can be compared with critical values for a chi-square distribution with *t* degrees of freedom.

Approximation

In practice the computation of $\mathbf{M}_i^2$ is not feasible as it requires solutions for the MME before outliers can be detected and this would double the computing time used for genetic evaluation. Furthermore, in most practical application it is not possible to compute $\mathbf{C}^{xx}$ since very large matrices usually are involved. In many cases $\mathbf{D}_i$ is dominated by $\mathbf{Z}_i\mathbf{G}_0\mathbf{Z}_i^{'} + \mathbf{R}_0$ so a first approximation could be to ignore the effects of estimation errors in the fixed effects when $\mathbf{D}_i$ is computed.
Here we propose a further simplification so that solutions for the MME from earlier routine runs can be utilized.

Partition $\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}$ where $\mathbf{b}_1$ contains effects estimated with great accuracy and are included in every run of the model and $\mathbf{b}_2$ contains effects with many levels and where new levels usually are introduced with each new run of the model. Examples could be effects of age, lactation or days in milk for effects in $\mathbf{b}_1$ and effects such as herd-year-season or herd-test-day for effects in $\mathbf{b}_2$.

For each record we compute: $E(\mathbf{y}_i) \cong \mathbf{X}_i \begin{bmatrix} \hat{\mathbf{b}}_1^p \\ \mathbf{0} \end{bmatrix}$ where $\hat{\mathbf{b}}_1^p$ is the solution for $\mathbf{b}_1$ from a previous run of the model.

Similarly define *s(i)* as a vector valued function to compute the phenotypic SD of all observed traits in record *i*. For a simple animal model this may be the population SD after correcting the data for effects in $\mathbf{b}_1$.

Now, the Mahanalobis distance can be approximated using the following procedure:

(1) $$\hat{\mu}_i = \mathbf{X} \begin{bmatrix} \hat{\mathbf{b}}_1^p \\ 0 \end{bmatrix}$$

(2)

$$\mathbf{D}_i^* = diag(s(i))diag(\mathbf{D}_i)^{-½}\mathbf{D}_i diag(\mathbf{D}_i)^{-½} diag(s(i))$$

(3) $$M_i^2 = (\mathbf{y}_i - \hat{\mu}_i)^{'} \mathbf{D}_i^{*-1}(\mathbf{y}_i - \hat{\mu}_i)$$

Intuitively the above procedure adjusts the variance of the observation to take account of the effects in $\mathbf{b}_2$ that are not corrected for. The distribution of M^2 will approximately be χ_t^2.

Outliers and extreme observations

The density of the multivariate normal distribution in principle extends to the whole real line in *t* dimensions and it is therefore difficult to distinguish between real outliers and extreme observations belonging to the distribution of the data under analysis.

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 l^{th} ranked M^2 out of N records, with cumulative probability *p*=(*l*-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 which in this case is equal to the number of traits (t). This curve is expected to follow a straight line if the data in $\mathbf{d}_i$ are $N(0,\mathbf{D}_i^*)$. Extreme observations will deviate from this line and can be used to determine a cut-off-point above which observations are deemed to be outliers.

Example

The procedure developed was tested on the data from the Jersey breed used in the Inter-Nordic genetic evaluation for dairy cattle run by Nordic Cattle Genetic Evaluation (NAV). All recorded Jersey cows in Denmark and Sweden were included in the analysis.

The data included 9 884 497 test day records of 568 392 cows, where each record consisted of one observation on milk, fat and protein. The model used for analysis was the routine model used by NAV for genetic evaluation. Evaluations from the model are expressed as 305D 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.

Expectations in (1) were computed based on effects due to country and days in milk (dim). The standard deviation of corrected data *s(i)* were also computed within country and dim.

Mahanalobis distances (*M*^{2}) were computed for all test-days records and plotted against expected χ^2 -values as shown in Figure 1. It is clearly seen from the figure that the relation is nonlinear and thus the distribution of data corrected for $\hat{\mathbf{b}}_1^p$ is extra-normal or contains outliers.

Genetic evaluation omitting outliers

Four different editing rules for excluding records based on their *M*^{2} were applying:

1. Raw (No limits),
2. MD100 (records with *M*^{2} > 100 were deleted)
3. MD60 (records with *M*^{2} > 60 were deleted)
4. MD30 (records with *M*^{2} > 30 were deleted)

This corresponds to applying more and more stringent editing rules. If a record was marked as an outlier all three traits were deleted. Number of records deleted in the different situations are shown in Table 1.

Across the three situations with limits on M^2, the general characteristic for the edited records were either:

1. High yield compared to stage of lactation and also compared to previous and following records for that particular cow
2. Extreme *protein* to *fat* ratio, often observed for records with an elevated Somatic Cell Score

The four editing situations were analyzed using the NAV Jersey routine model either on the full dataset or on a reduced dataset by deleting records from the last four years. This corresponds to the setup for the standard INTERBULL test method 3 (IB3) for model validation.

Predictability

The predictive ability of the models using different editing rules for outliers were assessed by computing the correlation between evaluations from the full and the reduced dataset used in the IB3 test. Correlations were computed for cows having all their records in the last four years of data. Computations were conducted in four groups: all cows and for cows that had at least one observation deemed as outlier for each of the three editing criteria. Computations were done on the full dataset (Raw) and for datasets using the three different editing rules.

Results are shown in Table 2. The correlation between predicted breeding values for *milk, protein* and *fat* were 0.58, 0.59 and 0.55 respectively using the Raw data were no editing were done. When cows were classified as having at least one record deemed as outlier according to the three criteria the correlations dropped. For *milk* the correlation dropped to 0.53. 0.51 and 0.52 as the criterion for deeming an observation as outlier became more stringent. Note that when using the Raw data no observations were deleted. The cows were only classified as having at least one observation that was deemed as outlier according to the three different editing rules.

When editing was applied the correlation between early prediction based on pedigree index and future evaluations increased. For *milk* the correlation increased from 0.53 to 0.60 when applying the rule M^2>100. Similar trends were observed for *protein* and *fat.*

Applying a more stringent editing rule (M^2>60) of course included the less stringent rule. Comparing the correlation between early prediction and late prediction the correlation for *milk* increase from 0.51 to 0.53 and similar trends were observed for *protein* and *fat.*

When we applied the most stringent rule (M^2>30) this can be compared with the classification to where analyses were performed using less stringent rules but cows were classified according to M^2>30.

Again for *milk* increased from 0.51 in Raw, to 0.51 in MD100, 0.52 in MD60 and finally 0.53 in MD30. That is applying more stringent editing rules increased the correlation between early and late prediction for the cows involved. A similar trend was observed for *protein* and *fat.* In all cases the highest correlation was obtained using the most stringent rule (M^2>30). This indicates that even more stringent rules should be applied. More analysis is needed to determine the optimum cut-off point. A closer look at Figure 1 indicates that a cut-off-point in the range 20-25 may be optimal.

Applying rules for editing multivariate outliers of course influences the animals with records edited but will also influence their contemporaries. The changes were computed for bulls and for cows separately and the results are shown in Table 3 for Bulls and in Table 4 for cows. Most bulls only have minor changes but a single bull exists that changes more than 5 index units which are more than ½ a genetic standard deviation of index units. For cows a large proportion of animals have changes and for a few cows there is very considerable change.

Trend validation

The IB3 test is not optimal for this problem since it focuses on verifying model estimates of genetic trend and not on the predictive ability of the model. Therefore, the IB3 test was only made as a check-up to see if removing outliers does affect IB3 test results
The IB3 test results are shown in Table 5. All results were non-significant and differences between different cut-off-point for M^2 were small. This is because only a very small fraction of all data were deleted. However, there was a trend towards test criterion closer to its expectation (0) when applying more and more stringent editing rules towards outliers.

Conclusions

Application of multivariate models for genetic evaluations changes the problem of screening the input data for erroneous outliers from a univariate to a multivariate problem. An outlier detection rule based on 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.

Reference

Garrett, R.G. 1989. The chi-square plot: a tool for multivariate outlier recognition. *Journal of Geochemical Exploration 32:1,* 319-341.

Krzanowski, W.J. 2000. *Principles of multivariate analysis.* Oxford University Press Oxford.

Pena, D. & Prieto, F.J. 2001. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics 43:3,* 286-310.

Penny, K. I. 1996. Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 45:1,* 73-81.

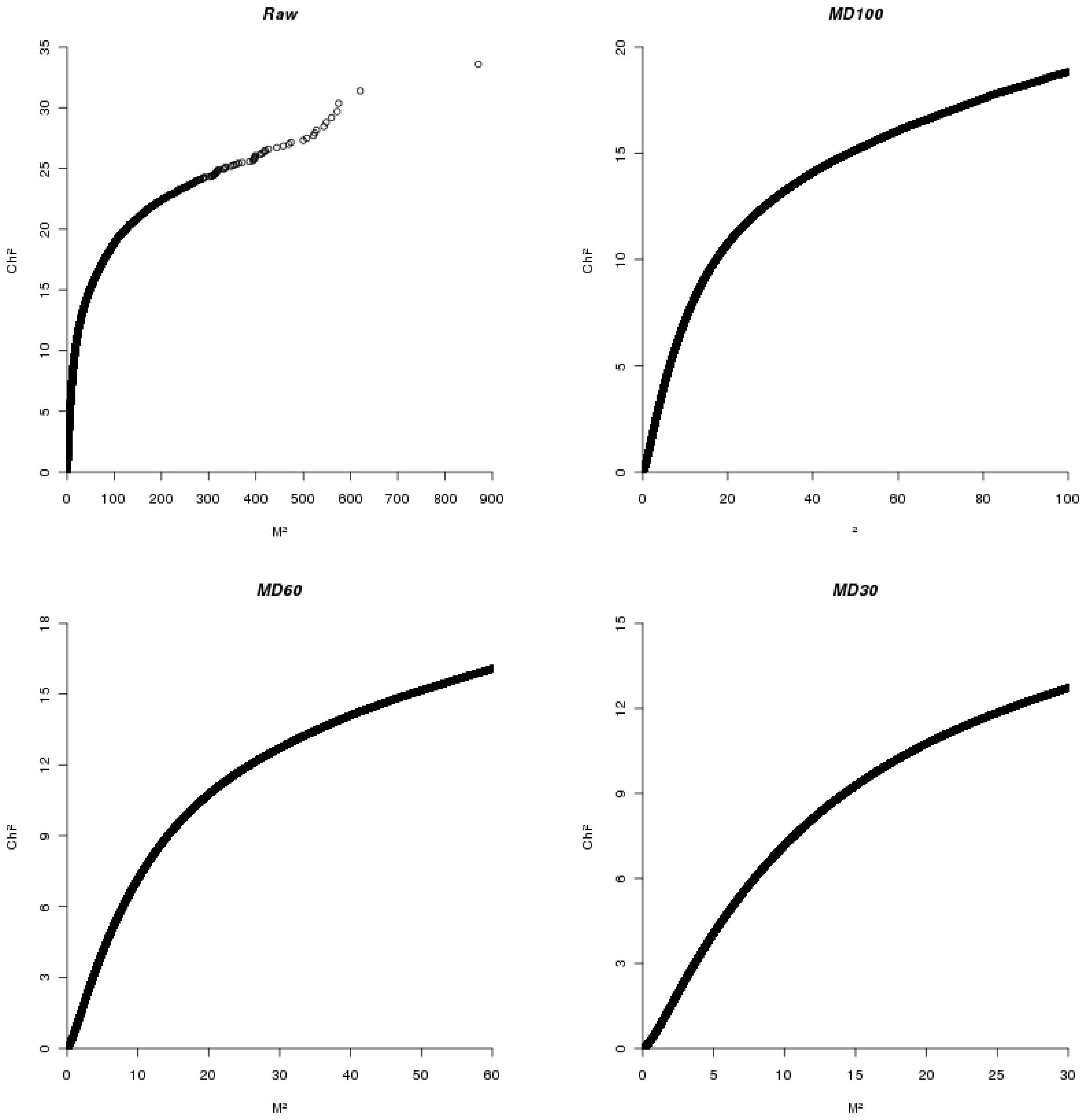

Figure 1. χ^2 plot for the 4 editing situations.

Table 1. Strategies applied for outliers detection and number of records deleted[1)].

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

[1)] All three traits are deleted

Table 2. Correlations between EBV's based on full and reduced datasets for cows having all their records in the last 4 years of data. Correlation are computed for all cows (no limit) and for cows classified by having at least one record deleted as outlier based on M^2 value.

	Data used in prediction[1)]									
	Raw				MD100			MD60		MD30
Limit on	No	>100	> 60	>30	>100	>60	>30	>60	>30	>30
No. of	96698	226	854	3593	226	854	3593	854	3593	3593
Trait										
Milk	0.58	0.53	0.51	0.51	0.60	0.53	0.52	0.54	0.52	0.53
Protein	0.59	0.55	0.55	0.56	0.62	0.57	0.57	0.59	0.57	0.59
Fat	0.55	0.52	0.52	0.53	0.57	0.53	0.54	0.55	0.54	0.57

1) Raw all data, MD100, MD60 and MD30 records with MD^2 > 100, 60 or 30 deleted

Table 3. Distribution for change in indices for bulls between "Raw" and MD30.

Trait	Number of bulls by magnitue 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

Table 4. Distribution for change in indices for cows between "Raw" and MD30.

Trait	Number of cows by magnitude of change in index											
	-17 - -6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6 - 32
Milk	22	26	67	243	992	8073	728446	9960	1135	322	122	51	92
Protein	18	26	102	315	1428	10199	718814	16302	1547	447	176	612	106
Fat	104	130	337	969	3140	23037	682622	31329	4332	1823	882	434	472

Table 5. Results of INTERBULL test 3 when applying different rules for editing outliers.

	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