Proceedings, 10th World Congress of Genetics Applied to Livestock Production

# Multivariate outlier detection in genetic evaluation in Nordic Jersey cattle

**H. Gao*, P. Madsen*, J. Pösö†, J. Pedersen‡, M. Lidauer§ and J. Jensen***
*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark, †FABA Co-op, Helsinki, Finland, ‡The Knowledge Centre for Agriculture, Cattle, Aarhus, Denmark, §MTT Agrifood Research, Jokioinen, Finland.

**ABSTRACT:** A procedure was developed for detection of multivariate outliers based on an approximation for Mahanalobis Distance (MD) and was implemented in the Nordic Jersey population. Evaluations are carried out by Nordic Cattle Genetic Evaluation (NAV), who uses a 9 trait model for milk, protein and fat in the first 3 lactations. It is based on the phenotypic correlation structure as a function of days in milk (DIM) and on computation of trait means and standard deviations (SDs) within classes of production year (PY), lactation and days in milk (DIM). For each record in the data, MD is computed based on trait means and co-variance matrix for the actual PY, lactation and DIM. Accuracy of EBV's is improved for animals having extreme outlier record(s) deleted compared to EBV's based on data not filtered for MD.
**Keywords:** Mahalanobis distance; Genetic evaluation; Multivariate outlier

## Introduction

The accuracy of national genetic evaluation relies on a chain of events, and one of the profound factors is quality of data from the milk recording system. In all routine genetic evaluation procedures, the data needs some quality filtering prior to entering evaluation. However, there are no standard guidelines for development of such procedures and different countries have different editing guidelines according to their own data situations. Traditionally, quality filtering has been implemented on a per trait basis by excluding observations with low univariate density. For normally distributed traits this is equivalent to excluding an observation if it deviate more than a preset number of SD units from the mean. However, in the multivariate case such simple univariate procedures may not be sufficient since records which are multivariate outliers may not necessarily be univariate outliers (Madsen et al., 2012).

In classical statistical literatures, the metric used for testing multivariate deviation is Mahanalobis distance (MD) (Mahalanobis, 1925; 1936). However, MD is not directly applicable for large datasets used in genetic evaluation as it requires computation of expected means and co-variance matrix for all observations. For complex models this is tedious. The objective of this study was to develop a procedure for detection the presence of multivariate outliers that can be used in the data pipeline for the genetic evaluation in the Nordic Jersey population.

## Materials and Methods

**Data.** The data used in this study consisted of 606,283 Jersey cows with 10,598,803 test-day records in the first three lactations, which was used in the official November 2013 genetic evaluation of production traits for Denmark and Sweden run by Nordic Cattle Genetic Evaluation (NAV). Each record comprised one observation on milk, fat and protein, and the days in milk (DIM) ranged between 8 and 365 days. All data were classified into 12 mo. time periods so that the most recent data were in a full time period. The Swedish data was allocated into 18 TP and the Danish data into 23 TP.

**Mahanalobis distance calculation.** Mahalanobis distance is a metric used for testing multivariate deviation which was defined as following:

$$\mathrm{MD}_i = \sqrt{(x_i - \mu)^T S^{-1} (x_i - \mu)}$$

where $i$ stands for the i$^{th}$ observation in the group of values from 1 to $n$, $\mu$ is the sample mean value and $S$ is the sample covariance matrix. For a $p$-dimension dataset with multivariate normally distributed, the values of $\mathrm{M}^2 = \mathrm{MD}^2$ are following the chi-square distribution with $p$ degrees of freedom: $\mathrm{M}_i^2 \sim \chi_p^2$. In this study, the means and SDs were calculated by country, TP, lactation and DIM. A function was used to fit the raw means and SDs within country, TP, lactation and DIM. The fitted SDs were used to scale the phenotypic covariance matrix for the particular DIM and the fitted means were used as the reference when computing the MD value for each record. A detailed description of the procedure is given in (Madsen et al., 2012).

**Scenarios.** A set of cut-off values representing different scenarios (ranging from 10 to 100 by step 10) on square of MD ($M^2$) were performed on the full data set to discard the outliers, which means the records with $M^2$ larger than the cut-off value were removed as multivariate outliers. The lower the cutoff value for $M^2$, the stricter is the editing rule of the detection, and the numbers of deleted records for each scenario are presented in Table 1.

**Table 1. Number of cows and records deleted based on $M^2$ for each scenario**

| Scenario | Deletion | No. of cows with records deleted | No. of records deleted | Prop. deleted (%) |
|---|---|---|---|---|
| 100 | $M^2 > 100$ | 905 | 933 | 0.0088 |
| 90 | $M^2 > 90$ | 1198 | 1245 | 0.0117 |
| 80 | $M^2 > 80$ | 1614 | 1694 | 0.0160 |
| 70 | $M^2 > 70$ | 2320 | 2451 | 0.0231 |
| 60 | $M^2 > 60$ | 3413 | 3651 | 0.0344 |
| 50 | $M^2 > 50$ | 5317 | 5815 | 0.0549 |
| 40 | $M^2 > 40$ | 8894 | 10077 | 0.0951 |
| 30 | $M^2 > 30$ | 16749 | 19994 | 0.1886 |
| 20 | $M^2 > 20$ | 40410 | 54010 | 0.5096 |
| 10 | $M^2 > 10$ | 173604 | 334520 | 3.1562 |

**Genetic evaluation.** Based on the idea of Interbull method 3 for model validation, a reduced dataset was generated by removing the records from the last four years under each scenario, the routine NAV random regression test-day model (http://www.nordicebv.info/Routine+evaluation/Routineevaluation.htm?wbc_purpose=%2f) with 9 traits for milk, fat and protein in first 3 lactations was applied both for the full and reduced datasets in each scenario. EBVs were combined across the first three lactations using weights 0.5, 0.3, 0.2, respectively. All the genetic evaluations were performed using Mix99 package (Lindauer and Strandén, 1999).

**Validation.** The accuracies of genetic evaluation were measured as the correlations between EBVs based on full and reduced datasets for cows without records in the reduced dataset and with one or more record(s) deleted due to the editing rules on $M^2$. The differences between before and after removing outliers can be assessed through the comparison with the correlations on the raw dataset (no editing rule applied) for the same validation individuals. Un-biasedness of genetic evaluation was computed as the regression of EBVs from full dataset on EBVs from reduced dataset for the same cows. The cows in the validation dataset were divided into two groups: 1) progeny of the proven bulls ($Group_{pr}$); 2) progeny of the bulls having all their progenies with records in the last four years ($Group_{yo}$). The whole validation procedure was carried out separately on each group.

## Results and Discussion

The correlations and regression coefficients for the cows in $Group_{pr}$ and $Group_{yo}$ for milk, fat and protein are shown in Table 2 to Table 4, respectively. In general, the genetic evaluation performed better after deleting the multivariate outliers from the raw dataset in terms of accuracy and un-biasedness of EBV. The differences between the correlations based on "clean" datasets and raw datasets among the 10 scenarios ranged from 0.017 to 0.027 (average 0.022) for milk, from 0.018 to 0.031 (average 0.024) for fat, from 0.015 to 0.029 (average 0.019) for protein. The regression coefficients of EBVs from full dataset on EBVs from reduced dataset were slightly lower than the expected value of 1.0 for most of the scenarios for milk and protein. For the validation cows in $Group_{yo}$, the proportion of cows in this group reduced to 11% on average, difference between the two correlations varied from 0.012 to 0.089 (average 0.039) for milk, from 0.006 to 0.049 except scenario 60 with -0.016 (average 0.02) for fat, from 0.009 to 0.057 (average 0.030) for protein. Compared with the $Group_{pr}$, the gains in prediction accuracy were higher in general among the 10 scenarios for $Group_{yo}$ across the three analyzed traits. The regression coefficients significantly deviated from 1.0 indicating more bias than the predictions on $Group_{pr}$, and this could be mainly due to the bias EBVs introduced by preferentially treated bull dams.

The proposed procedure is very simple and can avoid the negative influence of extreme outliers. However,

**Table 2. Correlations and regression coefficients of genetic evaluation before and after removing the multivariate outliers on the cows in $Group_{pr}$ and $Group_{yo}$ for milk**

| | $Group_{pr}$ | | | | | $Group_{yo}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | No. | Corr.[1] | Corr.raw[2] | Reg.[3] | Reg.raw[4] | No. | Corr. | Corr.raw | Reg. | Reg.raw |
| 100 | 235 | 0.774 | 0.751 | 0.993 | 0.943 | 26 | 0.597 | 0.508 | 0.622 | 0.647 |
| 90 | 327 | 0.766 | 0.749 | 1.000 | 0.962 | 34 | 0.550 | 0.496 | 0.579 | 0.610 |
| 80 | 400 | 0.761 | 0.739 | 0.986 | 0.943 | 46 | 0.503 | 0.444 | 0.519 | 0.519 |
| 70 | 573 | 0.733 | 0.706 | 0.922 | 0.887 | 66 | 0.513 | 0.470 | 0.559 | 0.563 |
| 60 | 820 | 0.736 | 0.713 | 0.952 | 0.923 | 101 | 0.369 | 0.338 | 0.450 | 0.438 |
| 50 | 1281 | 0.714 | 0.693 | 0.964 | 0.940 | 161 | 0.537 | 0.504 | 0.619 | 0.610 |
| 40 | 2044 | 0.692 | 0.670 | 0.963 | 0.950 | 266 | 0.555 | 0.532 | 0.641 | 0.637 |
| 30 | 3606 | 0.703 | 0.682 | 0.988 | 0.978 | 483 | 0.509 | 0.483 | 0.631 | 0.614 |
| 20 | 8115 | 0.718 | 0.697 | 1.024 | 1.018 | 1068 | 0.549 | 0.526 | 0.682 | 0.681 |
| 10 | 31434 | 0.752 | 0.728 | 1.045 | 1.045 | 4157 | 0.591 | 0.579 | 0.741 | 0.762 |

[1]Correlations after removing the outliers.
[2]Correlations before removing the outliers.
[3]Regression coefficients after removing the outliers.
[4]Regression coefficients before removing the outliers.

**Table 3. Correlations and regression coefficients of genetic evaluation before and after removing the multivariate outliers on the cows in $Group_{pr}$ and $Group_{yo}$ for fat**

| | $Group_{pr}$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | No. | Corr. | Corr.raw | Reg. | Reg.raw | No. | Corr. | Corr.raw | Reg. | Reg.raw |
| 100 | 235 | 0.752 | 0.722 | 1.070 | 1.088 | 26 | 0.703 | 0.678 | 0.558 | 0.552 |
| 90 | 327 | 0.724 | 0.702 | 1.004 | 1.029 | 34 | 0.639 | 0.633 | 0.524 | 0.528 |
| 80 | 400 | 0.724 | 0.706 | 1.012 | 1.050 | 46 | 0.474 | 0.448 | 0.410 | 0.420 |
| 70 | 573 | 0.743 | 0.724 | 1.027 | 1.067 | 66 | 0.413 | 0.364 | 0.379 | 0.366 |
| 60 | 820 | 0.724 | 0.704 | 1.007 | 1.037 | 101 | 0.322 | 0.338 | 0.320 | 0.344 |
| 50 | 1281 | 0.720 | 0.700 | 0.989 | 1.013 | 161 | 0.356 | 0.317 | 0.367 | 0.340 |
| 40 | 2044 | 0.715 | 0.687 | 0.972 | 0.988 | 266 | 0.432 | 0.400 | 0.453 | 0.438 |
| 30 | 3606 | 0.715 | 0.690 | 0.988 | 1.002 | 483 | 0.400 | 0.381 | 0.424 | 0.423 |
| 20 | 8115 | 0.702 | 0.677 | 0.983 | 0.987 | 1068 | 0.412 | 0.399 | 0.444 | 0.451 |
| 10 | 31434 | 0.725 | 0.694 | 1.030 | 1.023 | 4157 | 0.504 | 0.494 | 0.585 | 0.596 |

**Table 4. Correlations and regression coefficients of genetic evaluation before and after removing the multivariate outliers on the cows in $Group_{pr}$ and $Group_{yo}$ for protein**

| | $Group_{pr}$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | No. | Corr. | Corr.raw | Reg. | Reg.raw | No. | Corr. | Corr.raw | Reg. | Reg.raw |
| 100 | 235 | 0.808 | 0.787 | 1.001 | 0.958 | 26 | 0.525 | 0.468 | 0.426 | 0.470 |
| 90 | 327 | 0.790 | 0.775 | 0.974 | 0.947 | 34 | 0.506 | 0.479 | 0.442 | 0.483 |
| 80 | 400 | 0.781 | 0.761 | 0.968 | 0.938 | 46 | 0.478 | 0.433 | 0.413 | 0.423 |
| 70 | 573 | 0.773 | 0.752 | 0.938 | 0.921 | 66 | 0.475 | 0.435 | 0.439 | 0.452 |
| 60 | 820 | 0.755 | 0.736 | 0.936 | 0.924 | 101 | 0.352 | 0.343 | 0.352 | 0.368 |
| 50 | 1281 | 0.747 | 0.730 | 0.954 | 0.951 | 161 | 0.447 | 0.412 | 0.436 | 0.428 |
| 40 | 2044 | 0.733 | 0.715 | 0.940 | 0.942 | 266 | 0.462 | 0.435 | 0.456 | 0.449 |
| 30 | 3606 | 0.739 | 0.723 | 0.954 | 0.952 | 483 | 0.400 | 0.374 | 0.418 | 0.401 |
| 20 | 8115 | 0.740 | 0.722 | 0.973 | 0.968 | 1068 | 0.420 | 0.401 | 0.450 | 0.445 |
| 10 | 31434 | 0.764 | 0.735 | 1.002 | 0.999 | 4157 | 0.501 | 0.491 | 0.575 | 0.588 |

the distinction between errors and records of limited value is not clear cut as is implied in this procedure. Therefore, it would be useful to investigate effects of using fat tailed residual distributions such as the t-distribution (Stranden and Gianola, 1999).

## Conclusion

The objective of this study was to investigate the effect of filtering of multivariate outliers in genetic evaluation for Jersey cattle population. Gains on prediction accuracy were achieved by screening the multivariate outliers from the raw dataset and prediction biases were reduced. The improvement of prediction accuracy is more profound for progeny of young bulls. Finally, it is also important to extend this procedure to the other dairy breeds and implemented using an optimal cut-off value for $M^2$ to achieve an acceptable compromise between genetic evaluation accuracy and data deletion.

## Literature Cited


Lindauer, M. and I. Strandén. 1999. Fast and flexible program for genetic evaluation in dairy cattle. Pages 20-25 in Proc. Interbull Bulletin, Tuusula, Finland.

Madsen, P., J. Pösö, J. Pedersen, M. Lidauer, and J. Jensen. 2012. Screening for outliers in multiple trait genetic evaluation. Pages 85-91 in Proc. Interbull Bulletin, Cork, Ireland.

Mahalanobis, P. C. 1925. Analysis of race-mixture in Bengal. Journal and Proceedings of the Asiatic Society of Bengal (23):301-333.

Mahalanobis, P. C. 1936. On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta) 2:49-55

Stranden, I. and D. Gianola. 1999. Mixed effects linear models with t-distributions for quantitative genetic analysis: a Bayesian approach. Genetics Selection Evolution 31(1):25-42.