J. Dairy Sci. 101:11159–11164
https://doi.org/10.3168/jds.2018-15123

Short communication: Multivariate outlier detection for routine Nordic dairy cattle genetic evaluation in the Nordic Holstein and Red population

H. Gao,*[1] P. Madsen,* J. Pösö,† G. P. Aamand,‡ M. Lidauer,§ and J. Jensen*
*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark
†Faba Co-Op, FIN-01301 Vantaa, Finland
‡Nordic Cattle Genetic Evaluation, DK-8200 Aarhus, Denmark
§Natural Resources Institute Finland (Luke), FIN-31600 Jokioinen, Finland

ABSTRACT

It is of practical importance to ensure the data quality from a milk-recording system before use for genetic evaluation. A procedure was developed for detection of multivariate outliers based on an approximation for Mahalanobis distance and was implemented in the Nordic Holstein and Red population. The general target of this procedure is based on the Nordic Cattle Genetic Evaluation yield model, which is a 9-trait model for milk, protein, and fat in the first 3 lactations. The procedure is based on the phenotypic correlation structure as a function of days in milk (DIM) and on computation of trait means and standard deviations within a production year, lactation, and DIM. For each record in the data, a Mahalanobis distance value was computed based on the trait mean and the covariance matrix for the actual production year, lactation, and DIM. A set of cutoff values, ranging from 10 to 100 with steps of 10, for discarding multivariate outliers was investigated. Prediction accuracy was calculated as the Pearson correlations between estimated breeding values predicted by full data set and estimated breeding values predicted by reduced data set for cows without records in the reduced data set and with 1 or more records deleted due to the editing rules on Mahalanobis distance. The results showed that, averaged over all scenarios, gains of 0.005 to 0.048 on prediction accuracy have been obtained by deleting the multivariate outliers. The improvements were more profound for progeny of young bulls compared with progeny of proven bulls. It is easy to implement this multivariate outlier-detection procedure in the routine genetic evaluation for different dairy cattle breeds; however, an optimal cutoff value for Mahalanobis distance needs to be defined to achieve an acceptable compromise between genetic evaluation accuracy and data deletion.

Key words: prediction accuracy, Mahalanobis distance, data deletion

Short Communication

The prediction accuracy of breeding values is of great importance for animal improvement programs. Animals are ranked based on their predicted breeding values and selection decisions are made to boost genetic improvement. Many studies have focused on statistical modeling to improve the accuracy of breeding values (Henderson, 1990; Gianola, 2000; Gianola and Rosa, 2015); however, besides those advanced statistical techniques and computing procedures, another profound and often neglected factor is the quality of data from milk recording system, which is highly complex. Data collections are generally from multiple streams, such as thousands of farms, laboratories, or regions in animal breeding. A joint cattle genetic evaluation system was implemented in Finland, Sweden, and Denmark in 2002 (http://www.nordicebv.info/about-nav/). In this context, outlier detection needs to be strengthened in routine genetic evaluation system.

In routine genetic evaluation systems, records require quality filtering before entering the evaluation procedures, and these editing guidelines differ by countries or associations according to their own data structure. Traditionally, quality filtering in animal breeding has been conducted on a per-trait basis by excluding observations with low univariate density. For normally distributed traits, this is based on estimation of location and dispersion parameters, which is equivalent to excluding a record if it deviates more than a preset number of standard deviations from the mean. However, genetic evaluations are mostly conducted using a multiple-trait model to take advantage of the genetic relationships among the traits and estimate EBV for several traits simultaneously (Henderson and Quaas, 1976; Schaeffer, 1984; Meyer, 1991; Jensen et al., 1997).

Received May 28, 2018.
Accepted August 3, 2018.
[1] Corresponding author: hongding.gao@mbg.au.dk

We observed cows with anomalous EBV because of 1 multivariate outlier, and this multivariate outlier also affected the EBV of cows from the same contemporary group. In the multivariate case, records that are multivariate outliers may not necessarily be univariate outliers (Madsen et al., 2012); therefore, the methods for detection of univariate outliers are not sufficient.

In classical statistical literature, the metric used for testing multivariate deviation is Mahalanobis distance (**MD**), where each observation obtains a numerical measurement to the centroid of the data set and the correlation of the data sets is taken into account via the covariance matrix (Mahalanobis, 1925, 1936). For each record, the Mahalanobis distance is defined as

$$\mathrm{MD}_i = \sqrt{\mathbf{d}_i^{'}\mathbf{S}^{-1}\mathbf{d}_i},$$

where $\mathbf{d}_i$ is the vector of deviations from expectation for the ith record, and $\mathbf{S}$ is the sample covariance matrix. For a p-dimension multivariate normally distributed data, the values of MD_i^2 are approximately following the chi-squared distribution with p degrees of freedom. However, MD is not directly applicable for large-scale data sets used in genetic evaluation, as it requires computation of expectations for all observation as well as the covariance matrix for all observations. Applying this step directly to complex models can lead to extremely tedious computations.

The objective of our study was to develop a procedure for the detection of the presence of multivariate outliers and implement in the data pipeline for the genetic evaluation in the Nordic Holstein and Red Cattle population. The data we used consisted of 7,195,490 Holstein cows with 144,808,774 test-day records, and 3,900,835 Nordic Red cows with 87,021,681 test-day records of the first 3 lactations. The data were used in the official February 2015 genetic evaluation of production traits for Sweden, Denmark, and Finland run by Nordic Cattle Genetic Evaluation (NAV). Each record consisted of 1 observation of 3 traits on milk, fat, and protein yield. Days in milk ranged between 8 and 365.

To make computation of MD feasible for implementation in the data pipeline for genetic evaluation, the following approximation were introduced. (1) Deviation from expectations was replaced by deviation from functions fitted through the country × TP × lactation × DIM subclasses means, where time period (TP) was defined as 12-mo periods back to the most recent recording month in the data. To ensure a smooth change in expectation over DIM, a function on DIM was fitted for each country × TP × lactation subclass. (2) The sample covariance matrix was replaced by the phenotypic correlation matrix derived from the variance components used in the evaluation model and scaled by fitted standard deviation for country × TP × lactation × DIM:

$$\mathrm{MD}_{ctlsi} = \sqrt{\mathbf{d}_{ctlsi}^{'}\mathbf{S}_{ctls}^{-1}\mathbf{d}_{ctlsi}},$$

where $\mathbf{d}_{ctlsi}$ is the vector of deviation from fitted country × TP × lactation × DIM means for the ith record, and $\mathbf{S}_{\mathrm{ctls}}$ is the corresponding fitted covariance matrix for country × TP × lactation × DIM. A detailed description of the procedure is given in Madsen et al. (2012).

A set of MD^2 cutoff values representing different scenarios (ranging from 10 to 100 by steps of 10) were applied to the full data set to discard the outliers (i.e., records with a MD^2 value larger than the cutoff value were considered as multivariate outliers and removed from the data set). The lower (higher) the cutoff value for MD^2, the stricter (looser) the editing rule of the detection. The number of deleted records for each scenario are presented in Table 1 for Holstein and Nordic Red cattle data. Across all the scenarios, the general profile for the deleted records could be classified as (1) anomalous yield records compared with the stage of lactation and also compared with previous and following records of the cow; or (2) extreme protein-to-fat ratio, which can be clearly observed along with an elevated SCC.

Based on the idea of Interbull method 3 for model validation (Boichard et al., 1995), a reduced data set was generated by removing the records from the last 4 yr under each scenario, and the routine NAV random regression test-day model (Lidauer et al., 2015) with 9 traits for milk, fat, and protein in first 3 lactations was applied both for the full and reduced data sets in each scenario. The EBV were combined across the first 3 lactations using weights of 0.5, 0.3, and 0.2, respectively. All the genetic evaluations were performed using MiX99 package (Vuori et al., 2006).

Accuracies of genetic evaluations were measured as the correlations between EBV predicted by full data set and EBV predicted by the reduced data set for cows without records in the reduced data set and with 1 or more records deleted due to the editing rules on MD^2. The differences between before and after removing outliers can be assessed through the comparison with the correlations on the raw data set (no editing rule applied) for the same validation cows. Unbiasedness of genetic evaluation was computed as the regression of EBV from full data set on EBV from reduced data set for the same cows. The cows in the validation data set were divided into 2 groups: (1) progeny of the proven bulls (**Group$_{\mathbf{pr}}$**) and (2) progeny of the bulls having all their progeny with records in the last 4 yr (**$\mathbf{Group_{yo}}$**). The whole validation procedure was conducted separately for each group.

In the Holstein population, for validation cows in Group_{pr} the improvement of prediction accuracies after deleting the multivariate outliers ranged from 0.012 to 0.023 (average = 0.018) for milk, from 0.020 to 0.060 (average = 0.037) for fat, and from 0.009 to 0.032 (average = 0.019) for protein across the 10 scenarios. For validation cows in Group_{yo} prediction accuracies slightly declined for milk and protein after deleting the multivariate outliers; however, improvement in prediction accuracy for fat ranged from 0.016 to 0.085 (average = 0.048). The regression coefficients of EBV from full data set regressed on EBV from reduced data set were slightly lower than the expected value of 1.0 for milk and protein, and regression coefficients slightly increased in most scenarios, particularly for milk, indicating less prediction bias was achieved by deleting the outliers for validation cows in Group_{pr}. However, small declines of regression coefficients were found for validation cows in Group_{yo}.

In the Red population, for the validation cows in Group_{pr} gains in prediction accuracy from 0.004 to 0.022 (average = 0.012) were achieved by deleting the multivariate outliers for milk, whereas we found gains from 0.005 to 0.021 (average = 0.013) for fat and from 0.003 to 0.016 (average = 0.009) for protein. For the validation cows in Group_{yo}, we found gains in prediction accuracy from 0.001 to 0.018 (average = 0.005) for milk, from 0.021 to 0.042 (average = 0.029) for fat, and from 0.008 to 0.041 (average = 0.019) for protein. The regression coefficients of EBV from the full data set on EBV from the reduced data set were slightly higher than the expected value of 1.0 for milk for cows in Group_{pr}, indicating a small underestimation of EBV for the validation cows in Group_{pr} compared with other traits. Regression coefficients increased from 0.005 to 0.024 (average = 0.013) and from 0.001 to 0.039 (average = 0.013) for milk and protein for cows in Group_{yo}, respectively.

Figures 1 and 2 illustrate the trend of the prediction accuracies and unbiasedness versus scenarios for Holstein and Red, respectively. It is clearly shown in the figures, for most scenarios, that the genetic evaluation performed better after deleting the multivariate outliers from the raw data set in terms of prediction accuracy and unbiasedness of EBV. Compared with the Group_{pr}, the gains in prediction accuracy were higher among the 10 scenarios for Group_{yo} across the traits. Large differences in prediction accuracy and bias were observed between Group_{pr} and Group_{yo} for Holstein population (Figure 1). This indicates less accurate and more biased predictions in Group_{yo}, which could be mainly due to the biased EBV introduced by preferential treatment of bull dams. In a practical situation, an optimal cutoff value needs to be identified according to the practical data structure. An approach to identity the cutoff value was proposed by Garrett (1989), where a chi-squared plot by plotting the MD^2 against the quantiles of χ^2 values was used to reflect the deviation of the current data distribution from the multivariate normality, and different cutoff values need to be tested until a straight line is visualized. However, the cutoff must reach a reasonable compromise between improvement of prediction accuracies and stringency of detecting outliers. One compromise we made to allow us to keep more data in the routine evaluation using the Nordic test-day model was to reduce the residual covariances between traits to have a smaller multivariate effect in case 1 record was deviated significantly from its expectation; this procedure was found to be useful to increase the predictability of EBV for cows with extreme records (Lidauer et al., 2015). In our study, as illustrated in

Table 1. Number of cows, number of records, and proportion of records deleted based on square of Mahalanobis distance in each scenario by test-day record data for Nordic Holstein and Nordic Red cattle

	Holstein			Red		
Scenario[1]	No. of cows deleted	No. of records deleted	Proportion of deleted records (%)	No. of cows deleted	No. of records deleted	Proportion of deleted records (%)
100	9,969	10,131	0.007	5,585	5,786	0.007
90	12,841	13,080	0.009	7,499	7,820	0.009
80	16,902	17,288	0.012	10,548	11,107	0.013
70	22,964	23,637	0.016	15,507	16,510	0.019
60	32,604	33,767	0.023	23,688	25,689	0.030
50	49,544	51,988	0.036	39,226	43,646	0.050
40	81,800	87,792	0.061	70,865	82,428	0.095
30	156,387	176,287	0.122	143,535	180,699	0.208
20	388,325	499,083	0.345	349,267	514,607	0.591
10	1,758,989	3,388,398	2.340	1,247,596	2,734,095	3.142

[1] Records with square of Mahalanobis distance (MD^2) larger than the cutoff value (column one) were deleted.

Figures 1 and 2, especially for Holstein, the prediction accuracy and unbiasedness improved dramatically across scenarios for Group$_{yo}$, indicating more stringent scenarios were preferred. Similar validation results were observed between scenario 20 and scenario 10; however, by using the most stringent cutoff value (scenario 10), the achievement was at the expense of deleting 2 to 3% of the raw data (Table 1). Thus, scenario 20 was preferred in this context to keep as much data as possible and also to have EBV for more animals. Similar results were reported by (Gao et al., 2014) based on the Nordic Jersey population, where gains of 0.022, 0.024, and 0.019 on prediction accuracies were achieved on average for milk, fat, and protein in Group$_{pr}$, and gains of 0.039, 0.020, and 0.030 were achieved on average for milk, fat, and protein in Group$_{yo}$, respectively.

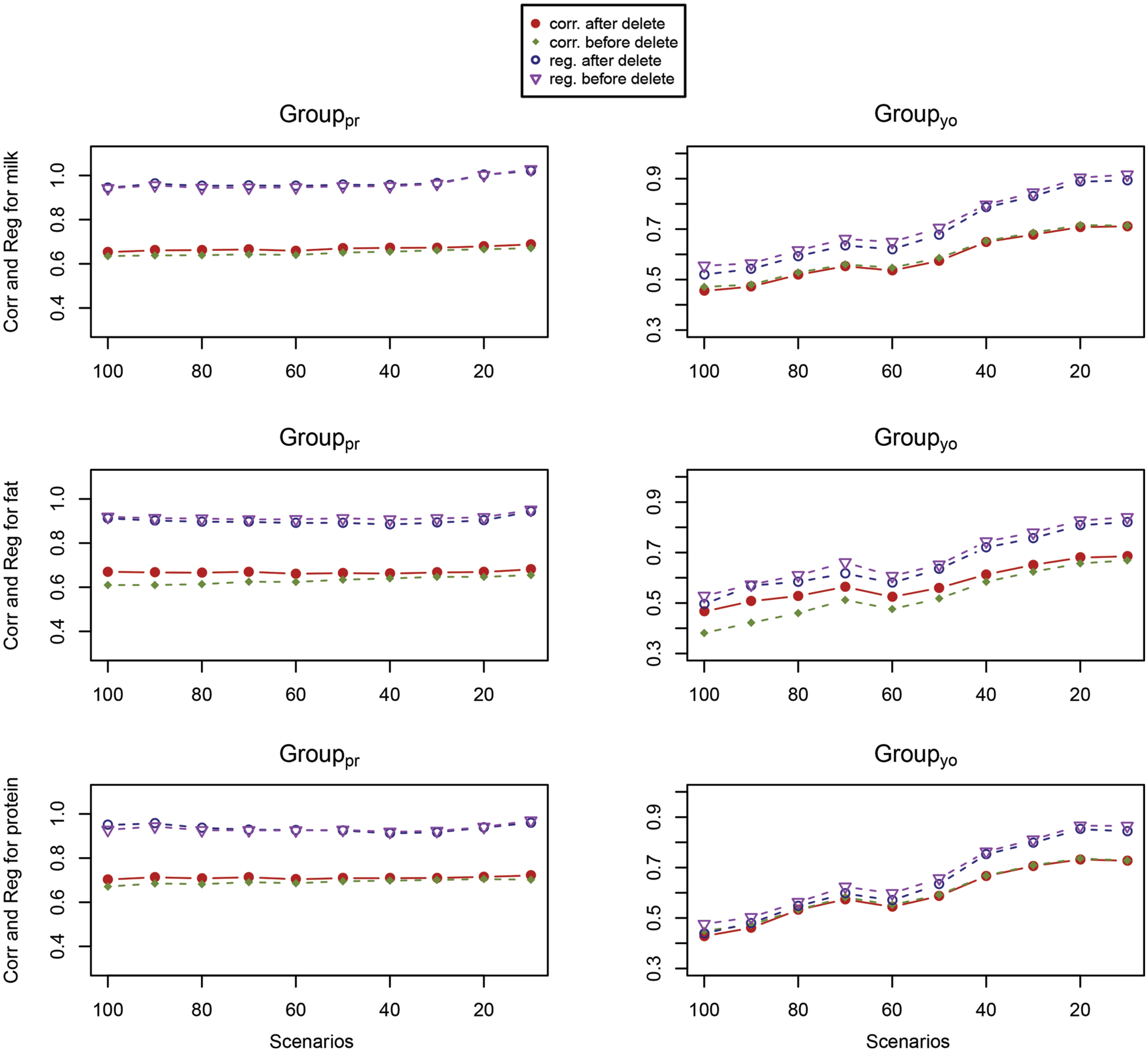

Figure 1. Correlation coefficients (Corr) between EBV from full and reduced data sets and regression coefficients (Reg) of EBV from the full data set regressed on EBV from reduced data set (y-axis) for validation cows in Group$_{pr}$ and Group$_{yo}$ before and after deleting the multivariate outliers in relation to cutoff values (x-axis) for milk, fat, and protein for Nordic Holstein population. Group$_{pr}$ = validation cows as progeny of the proven bulls; Group$_{yo}$ = validation cows as progeny of the bulls having all their progeny with records in the last 4 yr. Color verson available online.

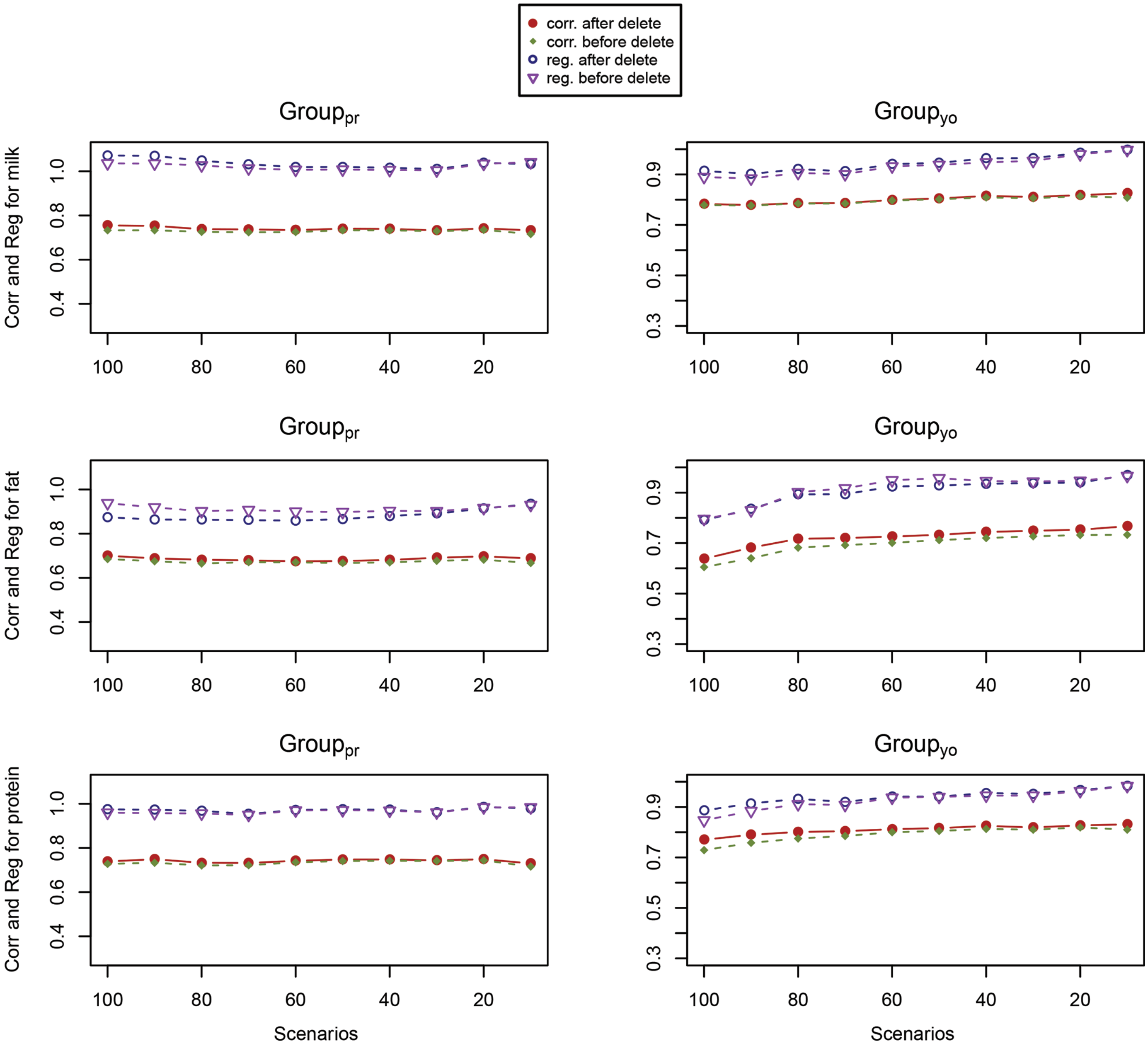

Figure 2. Correlation coefficients (Corr) between EBV from full and reduced data sets and regression coefficients (Reg) of EBV from the full data set regressed on EBV from the reduced data set (y-axis) for validation cows in Group$_{pr}$ and Group$_{yo}$ before and after deleting the multivariate outliers in relation to cutoff values (x-axis) for milk, fat, and protein for Nordic Red population. Group$_{pr}$ = validation cows as progeny of the proven bulls; Group$_{yo}$ = validation cows as progeny of the bulls having all their progeny with records in the last 4 yr. Color version available online.

In the present study, to avoid the potential negative influences from the extreme outliers in the upstream of genetic evaluation, we proposed a procedure for detection of multivariate outliers in the raw data sets from the Nordic Holstein and Nordic Red cattle population, and demonstrated that, under most of the study scenarios, the prediction accuracies can be improved and the prediction bias can be reduced by simply applying this procedure for screening the multivariate outliers. The improvement we found was more profound for progeny of young bulls compare with progeny of proven bulls; however, the distinction between errors and records of limited value is not clear cut as implied in this procedure. Therefore, it would be useful to investigate effects of using heavy-tailed residual distributions, such as the t-distribution (Strandén and Gianola, 1999).

Moreover, the procedure we proposed has a shortcoming, because when we compute the sample means and covariance matrix the outliers were involved and would affect the distance measurement; thus, a robust method is more preferred (Rousseeuw and Vanzomeren, 1990).

In conclusion, the procedure of multivariate outlier detection is a useful tool to improve the quality of genetic evaluation, and this procedure could be easily used in the routine genetic evaluations and extended to other dairy breeds based on an optimal cu-off value for Mahalanobis distance to achieve an acceptable compromise between the genetic evaluation accuracy and data deletion.

ACKNOWLEDGMENTS

The authors thank SEGES, Danish Agriculture and Food Council F.m.b.A., Aarhus, Denmark; Växa Sverige, Uppsala, Sweden; Faba Co-Op, Vantaa, Finland; and the Nordic Cattle Genetic Evaluation for providing data.

REFERENCES

Boichard, D., B. Bonaiti, A. Barbat, and S. Mattalia. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. J. Dairy Sci. 78:431–437.

Gao, H., P. Madsen, J. Pösö, J. Pedersen, M. Lidauer, and J. Jensen. 2014. Multivariate outlier detection in genetic evaluation in Nordic Jersey cattle. Page 835 in Proc. 10th World Congr. Genet. Appl. Livest. Prod., Vancouver, Canada. http://www.wcgalp.org/system/files/proceedings/2014/multivariate-outlier-detection-genetic-evaluation-nordic-jersey-cattle.pdf.

Garrett, R. G. 1989. The chi-square plot—A tool for multivariate outlier recognition. J. Geochem. Explor. 32:319–341.

Gianola, D. 2000. Statistics in animal breeding. J. Am. Stat. Assoc. 95:296–299.

Gianola, D., and G. J. M. Rosa. 2015. One hundred years of statistical developments in animal breeding. Annu. Rev. Anim. Biosci. 3:19–56.

Henderson, C. R. 1990. Statistical methods in animal improvement: Historical overview. Pages 2–14 in Advances in Statistical Methods for Genetic Improvement of Livestock. Springer-Verlag, Heidelberg, Germany.

Henderson, C. R., and R. L. Quaas. 1976. Multiple trait evaluation using relatives records. J. Anim. Sci. 43:1188–1197.

Jensen, J., E. A. Mäntysaari, P. Madsen, and R. Thompson. 1997. Residual maximum likelihood estimation of (co) variance components in multivariate mixed linear models using average information. J. Indian Soc. Agric. Stat. 49:215–236.

Lidauer, M. H., J. Poso, J. Pedersen, J. Lassen, P. Madsen, E. A. Mantysaari, U. S. Nielsen, J. A. Eriksson, K. Johansson, T. Pitkanen, I. Stranden, and G. P. Aamand. 2015. Across-country test-day model evaluations for Holstein, Nordic Red Cattle, and Jersey. J. Dairy Sci. 98:1296–1309.

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 Bull., Cork, Ireland. https://journal.interbull.org/index.php/ib/article/view/1269/1308.

Mahalanobis, P. C. 1925. Analysis of race-mixture in Bengal. J. Asiat. Soc. Bengal 23:301–333.

Mahalanobis, P. C. 1936. On the generalised distance in statistics. Proc. Indian Nat. Inst. Sci. Proc. Nation. Inst. Sci. India. 12:49–55.

Meyer, K. 1991. Estimating variances and covariances for multivariate animal-models by restricted maximum-likelihood. Genet. Sel. Evol. 23:67–83.

Rousseeuw, P. J., and B. C. Vanzomeren. 1990. Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc. 85:633–639.

Schaeffer, L. R. 1984. Sire and cow evaluation under multiple trait models. J. Dairy Sci. 67:1567–1580.

Strandén, I., and D. Gianola. 1999. Mixed effects linear models with t-distributions for quantitative genetic analysis: A Bayesian approach. Genet. Sel. Evol. 31:25–42.

Vuori, K., I. Strandén, M. Lidauer, and E. A. Mäntysaari. 2006. MiX99—Effective solver for large and complex linear mixed models. Page 254 in Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil. http://www.wcgalp.org/proceedings/2006/mix99-effective-solver-large-and-complex-linear-mixed-models.