Turk. J. Vet. Anim. Sci.
2008; 32(3): 163-171

Research Article

Principal Component and Clustering Analysis of Functional Traits in Swiss Dairy Cattle

Burak KARACAÖREN[1,*], Haja N. KADARMIDEEN[1,2]

[1]Statistical Animal Genetics Group, Institute of Animal Science, Swiss Federal Institute of Technology, ETH Centrum, Zurich CH 8092 - SWITZERLAND

[2]CSIRO Livestock Industries, J. M. Rendel Laboratory, Ibis Ave., Rockhampton QLD 4701 - AUSTRALIA

Received: 09.10.2006

Abstract: The objective of the research was to investigate the relationship among functional traits (body condition score (BCS), milk yield (MY), milking speed (MS), dry matter intake (DMI) and body weight (BW)). Data were from multiparous dairy cows (n = 55) of Chamau research farm of the Swiss Federal Institute of Technology, Switzerland. Principal component analysis with correlation matrix was used to find the relationship among BCS, MY, MS, DMI, BW, and other fixed effects including breed, year at calving, season, parity and year-season interaction. It was found that for all functional traits first 4 principal components explained more than 70% of the total variation. It was found that trading loss of accuracy using principal components scores instead of explanatory variables benefited reduction of dimension of explanatory variables and broke collinearity. Clustering analysis was performed based on different linkage methods and results showed physiological relationships among functional traits; since the data were from an experimental farm where each cow was fed by her MY performance and hence MY was associated with MS and DMI in the same cluster. BCS is correlated with BW and all these functional traits are related with mean lactation curve.

Key Words: Principal component analysis, clustering analysis, functional traits

İsviçre Süt Sığırlarına Ait İşlevsel Karakterlerin Temel Bileşenler ve Kümeleme Yöntemleri ile Analizi

Özet: Bu araştırmanın amacı işlevsel karakterler (vücut kondüsyon puanları, süt verimi, süt hızı, kuru madde tüketimi ve vücut ağırlığı) arasındaki ilişkiyi incelemektir. Veri seti İsviçre Teknik Üniversitesi (ETH) Chamau araştırma çiftliğinden Mayıs 2004 ve Mart 2005 dönemlerinde elde edilmiştir. Temel bileşenler analizi korelasyon matrisi ile kullanılarak sabit etkiler de dahil olmak üzere vücut kondüsyon puanları, süt verimi, süt hızı, kuru madde tüketimi ve vücut ağırlığı arasındaki ilişkiler incelenmiştir. Her karakter için ilk dört temel bileşenin toplam çeşitliliğin % 70'inden fazlasını açıkladığı bulunmuştur. Temel bileşenler regresyonuna ait belirleme katsayısı ile doğrusal regresyona ait belirleme katsayıları karşılaştırılmıştır. Açıklayıcı değişkenler yerine, temel bileşenler puanlarının kullanılması belirleme katsayısının düşmesine yol açsa da, açıklayıcı değişkenlerin sayısının azalması ve bağımlılığın giderilmesini sağlamıştır. İşlevsel karakterler ve doğrusal görünüm karakterleri arasındaki ilişki incelenmiştir; süt hızı ve kuru madde tüketimi hariç diğer karakterler için yine ilk dört temel bileşen çeşitliliği açıklamada yeterli bulunmuştur. Amaçlanan bir doğruluk derecesine ulaşmak için daha fazla sayıda temel bileşen kullanılabilinir. Değişik bağlantı yöntemleri kullanılarak kümeleme analizi yapılmıştır, sonuçlar işlevsel karakterler arasındaki ilişkiyi açık bir şekilde ortaya koymuştur; veri seti deneme çiftliğinden elde edildiğine göre her inek süt verim düzeyine göre yemlenmektedir, dolayısı ile süt verimi, süt hızı ve kuru madde tüketimi aynı küme içerisinde bulunmuştur. Vücut kondüsyon puanları ve vücut ağırlığı ilişkilidir ve bütün bu karakterler ortalama laktasyon eğrisi üzerinden ilişkilidir.

Anahtar Sözcükler: Temel bileşenler analizi, kümeleme analizi, işlevsel karakterler

Introduction

Measuring functional traits such as feed intake and body weight (BW) (especially on a daily basis) is not common in commercial dairy farms because of the need for expensive labour and equipment. The major part of total costs of milk production is due to feed costs. Feed intake is also correlated with other biological functions such as maintenance, growth, reproduction, fetal growth,

* E-mail: burak.karacaoeren@inw.agrl.ethz.ch
* Current adress: Roslin Institute, University of Edinburgh, Roslin BioCentre Midlothian, EH25 9PS, SCOTLAND, UK (E-mail: burak.karacaoeren@bbsrc.ac.uk)

and energy balance (1). Hence, biological and economic efficiency of dairy production might be improved if genetic variation in feed intake and feed efficiency is considered (2). Literature indicates that there is a genetic variation for feed intake (1-3). BW is an important functional trait regulating feed efficiency and energy balance traits in dairy cattle (4), which, in turn, is directly related to regulating energy requirements for milk production versus maintenance. Furthermore, BW, feed (dry matter) intake, and milk production level together form an important cluster of functional traits that determine the amount of fat reserves stored in the body, which is now recorded in many countries as 'body condition score' or BCS (5,6).

A common body condition scoring system has been developed to estimate the average BCS of cows in a herd. This system provides producers a relative score based on an evaluation of fat deposits in relation to skeletal features. The scoring method involves a manual assessment of the thickness of fat cover and prominence of bone at the tail head and loin area. The most widely used body condition scoring system for dairy cattle assigns scores from 1 to 9 in North America and from 1 to 5 in most European countries, with the lowest score meaning emaciated and carrying virtually no fat and the highest score meaning excessive fat.

Milking speed (MS) is another functional trait that relates to the incidence of clinical mastitis, labour time, and electrical power (7,8). In practice, MS is often measured by subjective scoring, whereas in the present study it was measured electronically.

Principal component analysis could be used to find the loadings or factors that explain the highest variation in the data set over dependent variable. The resulting principal components or loadings may decrease the dimension of the explanatory variables and break the possible dependency among explanatory variables, hence the collinearity. When collinearity among explanatory variables exists, linear regression assumptions (linearity, independence, homoscedasticity, and normality) could not be validated. Reduction of the dimension of the explanatory data set could be useful especially when analyzing big datasets. Although it is difficult and subjective, biological interpretation of the principal components could also be useful for the industry. Clustering analysis could be used to classify the variables based on different linkage methods when classes initially not known. One of the main difficulties with clustering analysis is deciding which is the correct linkage method. Different linkage methods; such as, ward, centroid, McQuitty, single, average, complete, median could produce different clusterings hence resulting dendrograms could have different interpretations of physical or biological situation. Mathematical properties, formulas, and comparison of methods were given in Everitt et al. (9). Everitt et al. (9) concluded that no one method can be recommended above all other. Variables within the same cluster are regarded as more related to each other; hence, in addition to principal component analyses, clustering analysis give an idea for reduction of dimension of the data set.

The main objective of this study was to explore the relationships among BCS, MY, MS, DMI, and BW to address the following questions: 1) How much variation in a chosen trait is explained by other functional traits that were included as independent variables in a model? 2) Are there any principal components for functional traits that explain more variation than others? 3) Which traits could be assumed to be more related (or within the same cluster) compared with the others (i.e. both the direction and amount of relationship).

Materials and Methods

Data

Data for MY, MS, DMI, and BW were obtained from multiparous dairy cows (made of Swiss Holsteins and Brown Swiss breeds) stationed at the Chamau research farm of the Swiss Federal Institute of Technology, Switzerland over the period of May 2004-March 2005. The experimental procedures of the farm followed the Swiss Law on Animal Protection and were approved by the Committee for the Permission of Animal Experiments of the Canton of Zug, Zug, Switzerland. The traits, MY, MS, roughage and concentrate intake and BW were recorded daily using automated units by METATRON (American Calan Inc., Northwood, NH, USA). The animals were housed in a free-stall barn. Milk production and other traits were measured twice a day (morning and evening). The concentrate, roughage, minerals, and vitamins were fed according to calculated needs (10).

BCS was measured by the Swiss Holstein Breeding Association as described by Trimberger (11) using the scale 1 to 5, during May 2004-March 2005 for 7 times

approximately at monthly intervals. Type traits were evaluated for Holstein cows only (n = 28). The DMI was calculated by summing the concentrate and roughage intake. Sum of morning and evening measurements of milk production data per day was used for the analysis of MY and DMI while for MS and BW, the measurements were averaged. Results of 7 BCS sessions and correspondence MY, MS, DMI, and BW measurements were averaged for creating the final data set. A summary of the dataset used for analysis is given in Table 1.

Statistical Methods

Forward selection (12) strategy was used to find the explanatory variables for the highest determination of coefficients that worked as new explanatory variables added to the model to achieve maximum determination of coefficients. The model equation is given as

$$y_{ij} = b_i + f + c_j + e_{ij}$$

where y_{ij} is MY (or BCS, MS, DMI, BW) produced by the j^{th} cow of i^{th} breed, b_i is fixed breed effect (n = 55; b_1 = Brown Swiss, n_1 = 27, b_2 = Holstein-Friesian, n_2 = 28), f consists of other fixed effects, including season, age at calving (in months), number of lactation, other functional traits than the dependent one and year-season, c_j is the animal effect, e_{ij} is the random residual variance.

Principal component analysis. Principal component analysis is a method for transforming the variables in a multivariate data set, x_1, x_2, ..., x_p, into new variables, y_1, y_2, ..., y_p which are uncorrelated with each other and account for decreasing proportions of the total variance of the original variables defined as

$$y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$$

$$y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p$$

$$y_p = a_{p1}x_1 + a_{p2}x_2 + ... + a_{pp}x_p$$

with the coefficients being chosen so that $y_1,y_2,...,y_p$ account for decreasing proportions of the total variance of the original variables, $x_1,x_2,...,x_p$ (9). Principal component analysis with correlation matrix (12) was used to find the relationship among BCS, MY, MS, DMI, BW and other fixed effects including breed, year at calving, season, and year-season interaction. Since scales of measurements of the functional traits were different; correlation matrix was used instead of covariance matrix. According to cumulative explanatory proportions number of principal components was chosen and corresponding scores were estimated. Then based on these scores, regression analyses were done again; coefficients of determination based on explanatory variables regression and based on principal component regression scores were compared.

Clustering analysis. Clustering analysis could be used to classify the variables based on different linkage methods when classes initially not known. Clustering analysis (12) using different linkage methods were used for determining the relationships among functional traits (ward, centroid, McQuitty, single, average, complete, median). Distance among clusters defined with corresponding linkage methods as follows; by ward method, increase in sum of squares within clusters; by centroid method, squared Euclidean distance among mean vectors; by McQuitty method, average of the distances with the next cluster (12); by single method minimum distance between pair of objects; by average method; average distance between pair of objects; by complete distance, maximum distance between pair of objects; by median method, squared Euclidean distance between weighted centroids (9). Interpretations were made based on the dendrograms produced by Minitab (12).

Results

The main objective of this study was to conduct a phenotypic analysis exploring relationships and

Table 1. Number of records, N, means, standard deviations, and minimum, maximum for Body Condition Score, Milk Yield, Milking Speed, Body Weight and Dry Matter Intake for some selected days of the first lactation.

Variable	Mean	Std Dev	N	Minimum	Maximum
Body Condition Score	3.21	0.52	55	2.00	4.18
Milk Yield	24.49	9.79	55	5.50	46.35
Milking Speed	2.19	0.82	55	0	4.30
Body Weight	692.66	74.58	55	477.66	860.66
Dry Matter Intake	51.21	7.66	55	29.09	69.38

dependencies among a group of traits that significantly affects the function of dairy cows (functional traits). This investigation was based on data collected from dairy cattle kept in an experimental research farm for mostly nutritional experiments, using exploratory analysis by principal components and clustering analysis. These statistical methods and concepts such as clustering analysis and principal components regression are equally applicable to large volumes of data collected by national animal breeding organizations. Although limited in data size, some of the exploratory analyses were statistically significant and will be useful. The following sections provide results of our findings.

Principal Components Regression on Functional Traits

Principal component analysis was performed using the explanatory variables based on model. Investigation of scree plots (Figure 1) shows that most of the variations are explained by the first 4 principal components.

For the first 4 principal components of all functional traits, eigenvalues, proportions, and cumulatives are given in Table 2.

More than 75% of the total variations are explained using these first 4 principal components for all traits. Table 3 shows results of determination coefficients based on regression analysis with explanatory variables, R-Square, and with the first 4 principal components scores, R-Square^{PC}.

Except for MS, all values of R-Square^{PC} were smaller compared to R-Square. However using principal components instead of explanatory variables reduced the dimension (from 8 explanatory variables to 4 principal components) and broke the collinearity (hence Variance

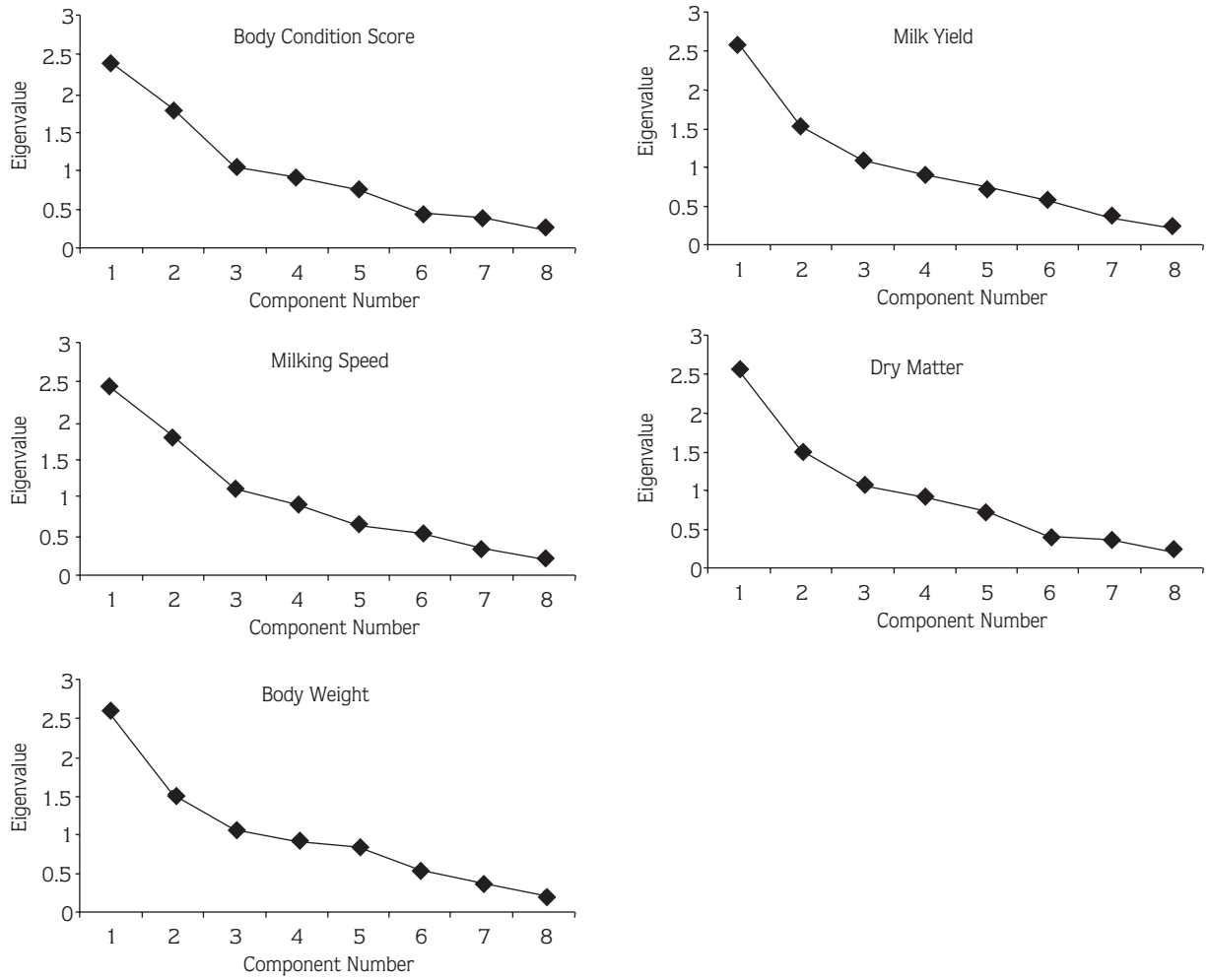

Figure 1. Scree plots of principal component analysis for body condition score, milk yield, milking speed, body weight, and dry matter intake.

Table 2. Eigenvalues, proportions, and cumulatives for the first 4 principal components for Body Condition Score, Milk Yield, Milking Speed, Dry Matter Intake, and Body Weight measured in dairy cattle.

		PC1	PC2	PC3	PC4
Body Condition Score	Eigenvalue	2.41	1.80	1.05	0.91
	Proportion	0.30	0.22	0.13	0.11
	Cumulative	0.30	0.53	0.66	0.77
Milk Yield	Eigenvalue	2.61	1.54	1.09	0.89
	Proportion	0.33	0.19	0.14	0.11
	Cumulative	0.33	0.52	0.66	0.77
Milking Speed	Eigenvalue	2.44	1.81	1.11	0.91
	Proportion	0.31	0.23	0.14	0.11
	Cumulative	0.31	0.53	0.67	0.78
Dry Matter Intake	Eigenvalue	2.34	1.90	1.08	0.96
	Proportion	0.29	0.24	0.13	0.12
	Cumulative	0.29	0.53	0.66	0.79
Body Weight	Eigenvalue	2.56	1.53	1.07	0.91
	Proportion	0.32	0.19	0.11	0.10
	Cumulative	0.32	0.51	0.64	0.76

Table 3. Comparison of determination coefficients using explanatory variables (R-Square) and using the first 4 principal components (R-Square^{PC}).

	R-Square	R-Square^{PC}
Body Condition Score	0.45	0.23
Milk Yield	0.39	0.22
Milking Speed	0.34	0.38
Dry Matter Intake	0.32	0.21
Body Weight	0.44	0.31

R-Square Coefficient of determination obtained from regression based on explanatory analyses.
R-Square^{PC} Coefficient of determination obtained from regression based on the first 4 principal components.

Inflation Factors found 1 for all the functional traits on principal component regression). Hence using principal components instead of explanatory variables gained both reduction of the explanatory data set and broke the collinearity. Comparison of predictions of observations using principal components and actual measurements is shown in Figure 2 for BCS, MY, MS, DMI, and BW. Visual inspection of Figure 2 showed that predictions were reasonably accurate for all the traits since most of the points were lying around the straight line with a slope 1.

Principal Components Regression on Type Traits

Type traits that were evaluated for Holstein cows were analyzed. For each functional trait included in the model, linear type traits that gave the highest determination of coefficients were sought. It was found that minimum 18 variables needed to explain the model fully.

Again investigation of scree plots and cumulative explanation of principal components (results not shown) showed that the first 4 principal components were informative enough. However, R-Square^{PC} MS and DMI were found smaller (0.30 and 0.07, respectively) compared with other functional traits. Hence, in order to achieve certain level of accuracy, more principal components could be used for MS and DMI. R-Square^{PC} was smaller than R-Square (Table 4), but instead of about 18 explanatory variables, only 4 principal components were used (here Variance Inflation Factor became 1).

Figure 3 shows observations and predictions obtained by principal component regression. Again points mostly distributed around the line, and showed that predictions were reasonable except for DMI.

Clustering Analysis of Functional Traits

Dendrograms based on different linkage methods (ward, centroid, McQuitty, single, average, complete, and median) are shown in Figure 4. Since different linkage methods give different clusters, it is hard to interpret which one is correct. However, most of the dendrograms shows the same variables in the same cluster (ward, McQuitty, average, and complete). This is as expected

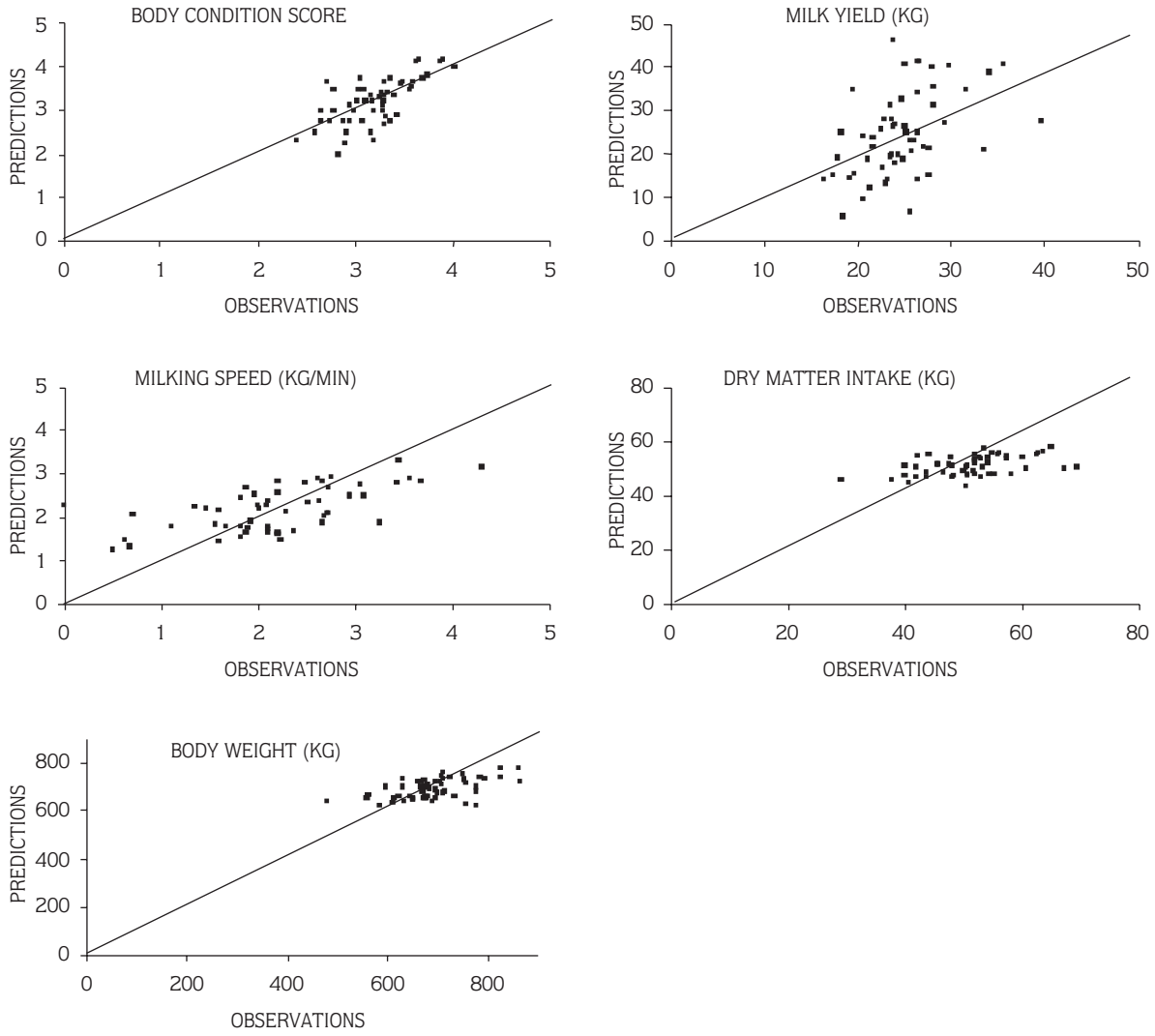

Figure 2. Observations and predictions based on the first 4 loadings of the principal components analysis for body condition score, milk yield, milking speed, dry matter intake, and body weight. Straight line has slope 1 to check the prediction abilities.

Table 4. Comparison of determination coefficients using explanatory variables (R-Square) and using the first 4 principal components ($R\text{-Square}^{PC}$) with additional explanatory variables as type traits.

	R-Square	$R\text{-Square}^{PC}$
Body Condition Score	1	0.64
Milk Yield	1	0.59
Milking Speed	1	0.30
Dry Matter Intake	1	0.07
Body Weight	1	0.59

because the data was from experimental farm where each cow was fed by her MY performance and hence MY was associated with MS and DMI in the same cluster. BCS is correlated with BW and all these functional traits are related with mean lactation curve.

Discussion

This study investigated the relationships among BCS, MY, MS, DMI, and BW using principal components, regression and clustering analyses using dairy cattle kept under experimental conditions. Although, based on experimental data, these results showed important

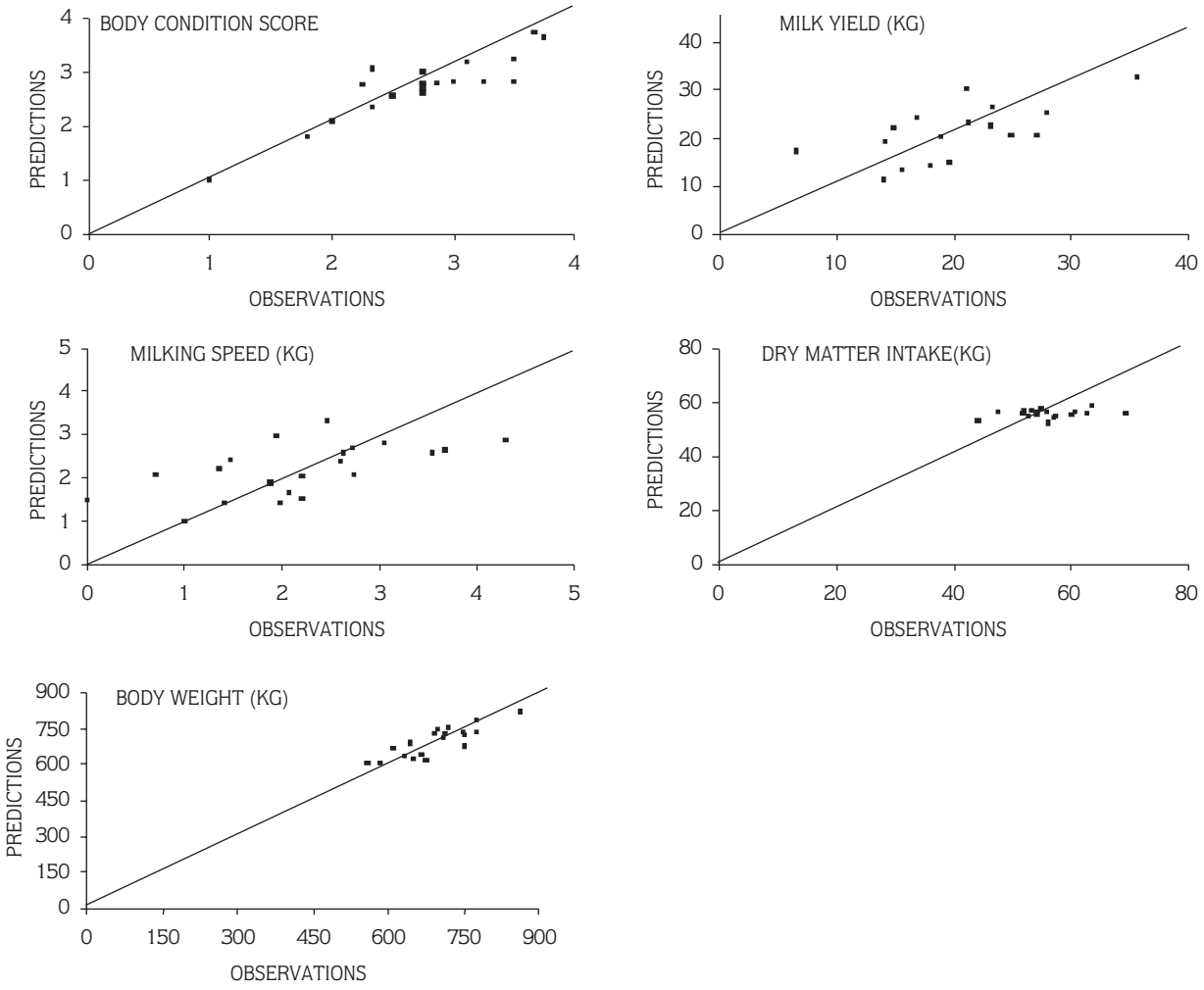

Figure 3. Observations and predictions based on the first 4 loadings of the principal components analysis with type traits for body condition score, milk yield, milking speed, dry matter intake, and body weight. Straight line has slope 1 to check the prediction abilities.

biological associations underlying the phenotypic relationships for many traits that are important for national dairy cattle breeding programs. These results are consistent with the findings of Karacaören et al. (13) who estimated longitudinal genetic correlations among functional traits using 12 years of experimental data.

Principal components could be used for all functional traits and would be useful in both dimension reduction and avoiding collinearity problems, common in the analysis of closely related functional traits such as body measurements or fertility. Type traits as a predictor of MS and DMI were less accurate with only 4 principal components but with increase in the number of principal components, accuracy is expected to increase. Clustering analysis also showed the clearly understandable patterns of physiological relationships among functional traits. However, different linkage methods produced different clusters of traits, with most of the functional traits differentially discriminated in the same clusters. Since the sampling size was not large (as is typical for an experimental farm), results should still be interpreted carefully and confirmed in different cattle populations.

Acknowledgements

Authors thank all personnel in the research station of Swiss Federal Institute of Technology in Chamau, Zug,

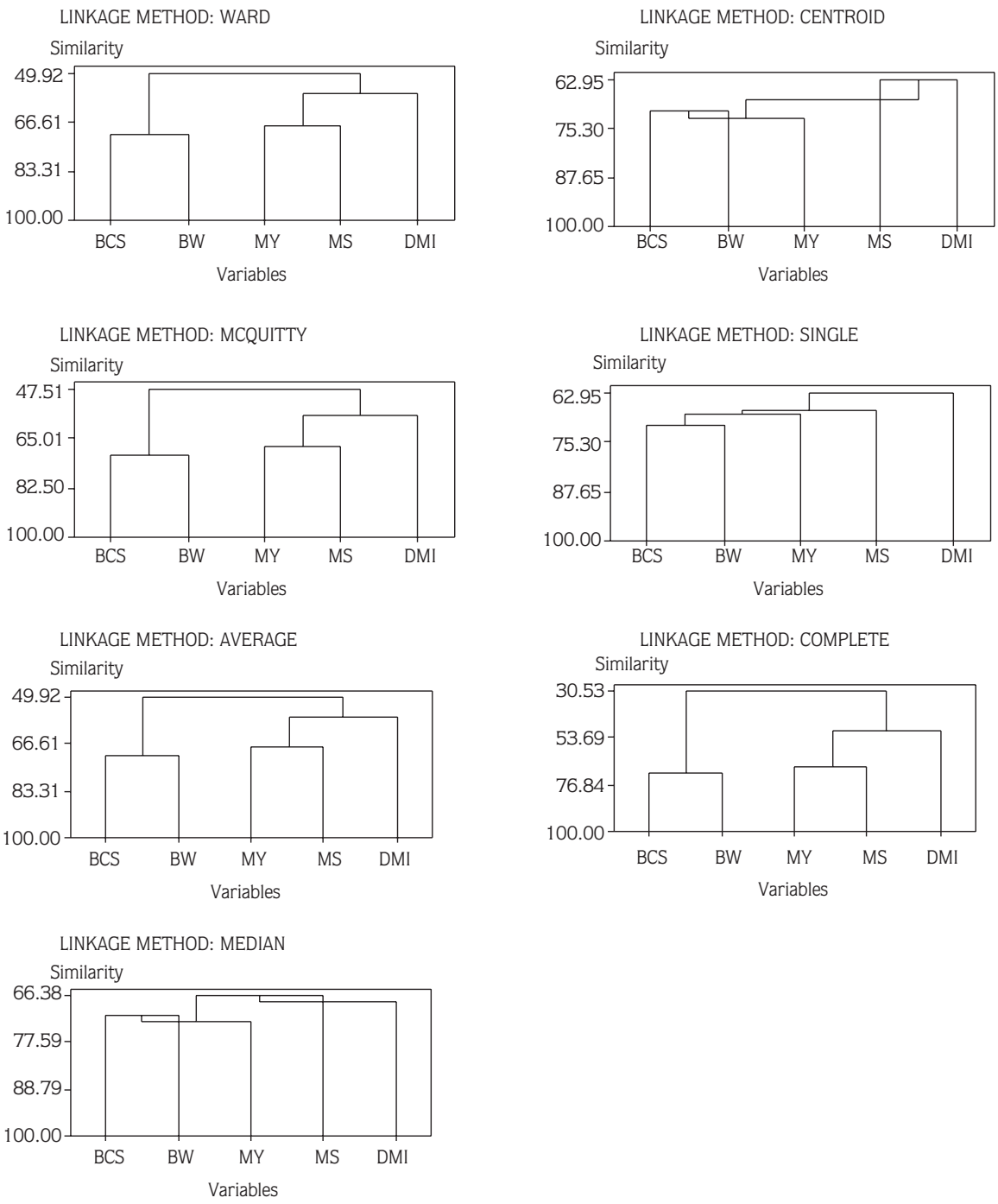

Figure 4. Dendrograms based on different linkage methods for Milk Yield (MY), Milking Speed (MS), Dry Matter Intake (DMI), Body Weight(BW), and Body Condition Score (BCS).

Switzerland for their help in data collection over a number of years. Authors thank Swiss Holstein Association, Patrick Rüttiman, and Dr Silvia Wegmann for providing BCS data, and Dr Trygve Almøy at Norwegian University of Life Sciences, Norway, Dr Luc Janss, and Dr. Kaspar Tschuemperlin for the useful comments.

References

1. Veerkamp, R.F.: Selection for economic efficiency of dairy cattle using information on live weight and feed intake: a review. J. Dairy Sci., 1998; 81:1109-1119.
2. Koenen, E.P.C.: Selection for body weight in dairy cattle. PhD dissertation. Wageningen University, Wageningen, The Netherlands. 2001.
3. Korver, S.: Genetic aspects of feed intake and feed efficiency in dairy cattle: a review. Livest. Prod. Sci., 1988; 20: 1-13.
4. Sondergaard, E., Sorensen, M.K., Mao, I.L., Jensen J.: Genetic parameters of production, feed intake, body weight, body composition, and udder health in lactating dairy cows. Livest. Prod. Sci., 2002; 77: 23-34.
5. Kadarmideen, H.N., Wegmann, S.: Genetic parameters for body condition score and its relationship with type and production traits in Swiss Holsteins. J. Dairy Sci., 2003; 86: 3685-3693.
6. Kadarmideen, H.N.: Genetic correlations among body condition score, somatic cell score, milk production, fertility and conformation traits in dairy cows. Anim. Sci., 2004; 79: 191-201.
7. Boettcher, P.J., Dekkers, J.C.M., Kolstad, B.W.: Development of an udder health index for sire selection based on somatic cell score, udder conformation, and milking speed. J. Dairy Sci., 1998; 81: 1157-1168.
8. Ilahi, H., Kadarmideen, H.N.: Bayesian segregation analysis of milk flow in Swiss dairy cattle using Gibbs sampling. Genet. Sel. Evol., 2004; 36: 563-576.
9. Everitt, B.S., Landau, S. Leese, M.: Cluster Analysis. 4th edn., Arnold Publisher, London. 2001.
10. NRC.: Nutrient Requirements of Dairy Cattle. 7th rev. edn. National Academy Press, Washington, DC. 2001.
11. Trimberger, G.W.: Dairy Cattle Judging Techniques. 2nd edn., Prentice-Hall. 1977.
12. Minitab: Minitab Release 14, Reference manual II. Minitab Inc., State College, PA. 2006.
13. Karacaören, B., Jaffrézic, F., Kadarmideen, H.N.: Genetic parameters for functional traits in dairy cattle from daily random regression models. J. Dairy Sci., 2006; 89: 791-798.