



J. Dairy Sci. 97:7476–7486

<http://dx.doi.org/10.3168/jds.2014-7982>

© American Dairy Science Association®, 2014. Open access under [CC BY-NC-ND license](#).

Lameness detection challenges in automated milking systems addressed with partial least squares discriminant analysis

E. Garcia,*†¹ I. Klaas,* J. M. Amigo,† R. Bro,† and C. Enevoldsen*

*Centre for Herd-oriented Education, Research and Development (HERD), Department of Large Animal Sciences, Grønnegaardsvej 2, DK-1870

†Department of Food Science, Spectroscopy and Chemometrics, Rolighedsvej 30, DK-1958, University of Copenhagen, Frederiksberg C, Denmark

ABSTRACT

Lameness causes decreased animal welfare and leads to higher production costs. This study explored data from an automatic milking system (AMS) to model on-farm gait scoring from a commercial farm. A total of 88 cows were gait scored once per week, for 2 5-wk periods. Eighty variables retrieved from AMS were summarized week-wise and used to predict 2 defined classes: nonlame and clinically lame cows. Variables were represented with 2 transformations of the week summarized variables, using 2-wk data blocks before gait scoring, totaling 320 variables ($2 \times 2 \times 80$). The reference gait scoring error was estimated in the first week of the study and was, on average, 15%. Two partial least squares discriminant analysis models were fitted to parity 1 and parity 2 groups, respectively, to assign the lameness class according to the predicted probability of being lame (score 3 or 4/4) or not lame (score 1/4). Both models achieved sensitivity and specificity values around 80%, both in calibration and cross-validation. At the optimum values in the receiver operating characteristic curve, the false-positive rate was 28% in the parity 1 model, whereas in the parity 2 model it was about half (16%), which makes it more suitable for practical application; the model error rates were, 23 and 19%, respectively. Based on data registered automatically from one AMS farm, we were able to discriminate nonlame and lame cows, where partial least squares discriminant analysis achieved similar performance to the reference method.

Key words: lameness detection in automatic milking system, animal welfare, pattern recognition, partial least squares discriminant analysis

INTRODUCTION

Automatic milking systems (AMS), also called robotic milking, were implemented in the 1990s to

reduce labor costs in dairy herds. By 2010, almost 10,000 farms had adopted AMS worldwide (de Koning, 2011); more than 2,000 are located in the Netherlands, whereas Denmark, Norway, and Sweden have around 1,000 farms each (Bisaglia et al., 2012; Landin and Gyllenswärd, 2012). However, increasing numbers are foreseen in northwest Europe (Steenefeld et al., 2012). The frequency of the cows' voluntary visits to the AMS is a major determinant of production efficiency (Ketelaarde Lauwere et al., 1996; Borderas et al., 2008; Lyons et al., 2013). Thus, an alarm from the AMS is generated when the cows' milking parameters deviate markedly from the expected pattern.

Recurrent evidence exists that painful conditions in the claws will reduce AMS visits (Klaas et al., 2003; Bach et al., 2007; Jacobs and Siegford, 2012). In the case of subtle pain though, the cow may merely reduce the number of AMS visits sporadically and she may eat less, leading to decreased milk production and compromised health and fertility.

Detecting even subtle painful conditions could be important for the herd manager interested in early and accurate intervention. Pain in the claws is difficult to assess under field conditions. Usually, the cows will avoid pain by changing their walking behavior (i.e., become lame). Signs of lameness have been associated with substantial financial losses (Sprecher et al., 1997; Blowey, 1998; Green, 2009) and constitute important indicators of reduced cow welfare (von Keyserlingk et al., 2009). In the traditional milking parlor, the personnel can detect behavioral changes visually when collecting cows for milking or when cows are leaving the parlor at least twice per day. In AMS, individual daily inspection is needed to detect subtle signs of lameness and this will be time consuming and, thus, costly. Therefore, it is highly relevant to develop automated systems to identify cows experiencing lameness.

Automatic milking systems generate large amounts of data on milking, feeding, and physical activity parameters. Disease treatments may be recorded and more constant cow characteristics, such as breed, age, and stage of lactation, are updated automatically (Jacobs and Siegford, 2012). These data are often the basis for

Received January 24, 2014.

Accepted August 20, 2014.

¹Corresponding author: efg@sund.ku.dk

the alarm lists, which should be addressed daily by the herd manager. However, lameness detection systems currently available seem far from being implemented worldwide on commercial farms, as studies on these systems often rely on relatively small sample sizes and a limited number of farms (Rutten et al., 2013). Logistic regression and linear discriminant analysis have been applied in animal and veterinary science for classification of animals, as diseased versus nondiseased, based on potential predictors (Greiner and Gardner, 2000; Heald et al., 2000; Nielsen et al., 2012). The AMS provides a series of potential predictors often larger in number than the cows available in most farms. Traditional approaches, such as the abovementioned logistic regression and discriminant analysis, may be inefficient or biased due to multicollinearity and overfitting (Ye and Zhao, 2010; Serrano-Cinca and Gutiérrez-Nieto, 2013). Multivariate methods (e.g., principal component analysis) have also been used for analysis in the animal science field (Bro et al., 2002; Dumas et al., 2005; Miekley et al., 2013), often focusing on data reduction (Sloth et al., 2003; Gorzecka et al., 2011). Based on automated data collection in dairy herds, pattern recognition has been the objective in several studies using principal component analysis, neural networks, or classification trees (Nielen et al., 1995; Klaas et al., 2004; Cavero et al., 2008; Ghotoorlar et al., 2012; Piwczyński et al., 2013).

Partial least squares discriminant analysis (PLS-DA) is a common tool used in classification in cases where multicollinearity is an issue (Vong et al., 1988; Wold et al., 2001; Chong and Jun, 2005). It allows investigation of hundreds or thousands of variables by using visualization tools to screen and understand complex data. Traditional applications of univariate analyses aim at detecting single or few predictors (e.g., logistic regression). Instead, PLS-DA comes as an attractive approach to finding latent patterns in a truly multivariate phenomenon, where many variables are correlated with each other but none is a good lameness indicator alone.

The aim of this investigation was to explore robotic milking-related variables potentially associated with clinical lameness. The objectives of this feasibility study were to (1) explore the usefulness of PLS-DA for lameness detection based on automated recordings of cow activity and milking process from an AMS herd and (2) suggest relations between these patterns and signs of lameness.

MATERIALS AND METHODS

Farm

We selected a commercial Danish dairy farm with 150 milking cows, free cow traffic, and 2 robotic milk-

ing units [voluntary milking system (VMS); DeLaval International AB, Tumba, Sweden] corresponding to 2 groups. There was a separate section with deep bedding (straw) for fresh cows and another one for dry cows. Only cows in the 2 freestall groups (robots) were lameness scored to ensure scoring under the same conditions. The farm had freestalls with mattresses and shavings, and a slatted floor maintained by a cleaning robot 8 times per day. The milking cows were automatically fed a TMR 7 times per day at 0200, 0600, 1000, 1200, 1400, 1800, and 2200h. On average, 67 cows were assigned to each robot. The cow breeds were Danish Holstein (13%), Danish Red (21%), and crossbred (66%). The cows were trimmed by a hoof trimmer every 4 mo and also by the staff at drying-off. At the beginning of the study, lactation number ranged from 1 to 7, 40% were first parity, and cows were milked on average 2.3 times per day, with a 9.2-h median milking interval (interquartile range: 5.4 h) and producing a median of 11.0 kg of milk per milking (interquartile range: 5.6 kg). Cows were, on average, at 153 DIM (range: 2 to 632 DIM).

Data Collection

Gait scoring of all milking cows was done by the first author weekly for 5 wk in autumn 2012 and for 5 wk in spring 2013 inside the freestalls by gently encouraging each cow to walk along the alleys. Asymmetric gait was assessed using a 4-point scale adapted from DairyCo (Kenilworth, UK; Reader et al., 2011): score 1 = even, long, and fluid strides (nonlame); score 2 = uneven steps, but the limbs favored were not obvious (nonlame); score 3 = 1 or more limbs favored obviously (lame); and score 4 = very reluctant to put weight on 1 or more limbs (severely lame). The first and the second author did an agreement study in the first week of the trial, whereas all gait scores used in the models were from the first author. The first author had limited experience in lameness scoring and the second author had several years of experience, although with a 5-point scoring system. The overall error rate of both intra- and interobserver agreement (between the first and second author) was around 15 to 20% when using a dichotomized classification of lame versus nonlame (detailed information presented in the Results section). We calculated the kappa statistic as an index of observer agreement with linear weighting and unweighted, respectively, for the 4- and binary-category results (Sim and Wright, 2005). Daily data obtained from the farm database (VMS Client 2009, v. 8.40; DeLaval International AB) was summarized week-wise, where the week was defined as d 1 to 7, with the gait scoring done on d 7. For the milking data, in every week (day 1 to 7),


		Days in milk - week median	
		Lactation no. (parity)	
		Consumed concentrate variance (kg ²)	
		Consumed concentrate sum (kg)	
		No. of milkings variance	
		No. of milkings sum	
		No. of refusals variance	
		No. of refusals sum	
		No. of total visits variance (milkings + refusals)	
		No. of total visits sum	
		No. of milkings with kicks variance	
		No. of milkings with kicks sum	
		Milking duration variance (min ²)	
		Milking duration sum (min)	
		Maximum milking duration variance (min ²)	
		Minimum milking duration variance (min ²)	
		Milk yield variance (L ²)	
		Milk yield sum (L)	
		No. of incomplete milkings variance	
		No. of incomplete milkings sum	
		No. of milkings with not milked teat variance	
		No. of milkings with not milked teat sum	
		Performance index (DeLaval) variance	
		Performance index (DeLaval) median	
		Average milk flow variance (L ² /min ²)	
		Average milk flow median (L/min)	
		Peak flow variance (L ² /min ²)	
		Peak flow median (L/min)	
		Milking interval variance (h ²)	
		Milking interval median (h)	
		Mastitis index (DeLaval) variance	
		Mastitis index (DeLaval) median	
		No. of kicks variance	
		No. of kicks sum	
		Activity index variance at 0000 h	
		...	
		Activity index variance at 2300 h	
		Activity index median at 0000 h	
		...	
		Activity index median at 2300 h	
cow 1	week 1	1	
...	...		
i	j		
		80	

Figure 1. Data matrix $\mathbf{X}_{\text{week } j}$ list of variables retrieved and structure of the data, according to cow and week. Days in milk and lactation number were not included in the models. The rows are structured from the first to the i th cow and from the first to the j th week. The DeLaval indices are from DeLaval International AB (Tumba, Sweden).

variance and median was calculated for each original variable, or variance and sum for some original variables (see Figure 1 for detailed variable description). The same was done to the activity data (variance and median), but each hour (0000 to 2300 h) was considered a different variable. Other summary measures have been considered (e.g., mean, interquartile range, and week difference), but the above approach achieved models with lower error rates. A DeLaval activity tag (DeLaval International AB) attached to the collar neck of the cow registered activity data every hour (24 h per day) using a radio link. The activity index is a sum of the binary registrations (0/1) done every 14.11 s, which means it could range between 0 and 255 each hour (Larsson, 2007). We assumed that high activity during estrus could mask a lameness condition. Then, all cows with heat alarms and insemination were flagged and the activity data from these days were excluded from the data summaries described above (e.g., for a cow in heat for 3 d, we still keep activity data for 4 d in a given week). After data management, 1,373 weekly observations were available for all milking cows. After inspection of raw data, observations with missing data on the gait scoring were deleted, as well as some observations that had consecutive zero values in the activity data (indicating activity sensor malfunctioning), resulting in 1,112 observations. Only the gait scores 1, 3, and 4 were included to identify which variables could be the strongest predictors. Cows with gait score 2 were not included in the models, because they comprised both

cows coping with the slippery floor and cows developing lameness or being slightly affected. Given that only 9 observations had gait score 1 among parity 3 or higher (all other 191 observations scored as lame), the models were built exclusively based on parity 1 (50 cows) and parity 2 (38 cows), in total using 332 weekly observations from 88 cows.

Preprocessing

Milk and activity variables for the current week are defined as the $\mathbf{X}_{\text{week } j}$ matrix, with j being the j th week from 1 to 5 in each year. Additionally, a new matrix of data was appended column-wise with the same variables (milk and activity) but with data from the previous week ($\mathbf{X}_{\text{week } j-1}$). Concatenating $\mathbf{X}_{\text{week } j}$ and $\mathbf{X}_{\text{week } j-1}$ we get \mathbf{X}_{new} .

Each original variable was represented in 2 different ways in the final data matrix. To avoid data with zero values, an offset of 1 was added to each element of the data (\mathbf{X}_{new}), and the natural logarithm was calculated to reduce positive skewness and to improve linearity between variables (Dallal, 2012), which defined the \mathbf{X}_{log} matrix as follows:

$$\mathbf{X}_{\text{log}} = \log(1 + \mathbf{X}_{\text{new}}). \quad [1]$$

Second, the \mathbf{X}_{new} matrix was appended column-wise to the previous matrix:

$$\mathbf{X}_{\text{c}} = \text{concatenate}[\mathbf{X}_{\text{log}}, \mathbf{X}_{\text{new}}]. \quad [2]$$

Finally, the combined matrix \mathbf{X}_c was then autoscaled separately in each variable to define the final matrix $\mathbf{X}_{\text{final}}$ as follows:

$$\mathbf{X}_{\text{final}} = \text{autoscale}(\mathbf{X}_c), \quad [3]$$

where the autoscale function subtracts the mean of each column (variable) to the observed value and divides it by the standard deviation, a standard procedure when variables have different units and offsets (Ballabio and Consonni, 2013).

The final matrix $\mathbf{X}_{\text{final}}$ was analyzed containing 2 submatrices with transformed data \mathbf{X}_{log} and original data \mathbf{X}_{new} (in total, 320 variables), as DIM and lactation number were not used as inputs to the models (32 milk variables, 24 activity variance variables, 24 activity median variables) \times 2 wk (current and previous) \times 2 preprocessing methods = 320 variables).

Analytical Strategy

Classification models were developed with PLS-DA (PLS-DA), which is a method that models the variation of several variables using fewer, so-called latent variables. These latent variables are weighted averages of the original variables and have the property that they are well suited for both describing the variation in the data and for classification (Ballabio and Consonni, 2013). The classification is obtained by prediction of a dependent variable, which in this case is simply a dummy matrix defining 2 classes: lame (gait scores 3 and 4) and nonlame (gait score 1). The PLS-DA model is composed by a score (of a given sample) and a loading vector for each latent variable. For nomenclature purposes, we use the term “score” (alone) to refer to the PLS-DA models, whereas we use “gait score” when referring to lameness assessment. The loadings represent the relationship between the original variables and the latent variables, and the scores represent the coordinates of the samples; that is, its position in the multidimensional space with respect to the latent variable space. This modeling approach does not assume any causal relationships but finds a combination of events or pattern that happens to be present when the cow is lame (e.g., low activity, milk yield, and number of AMS visits).

Model Building and Validation

Parity was an important confounder (results not shown), and the prevalence of lameness increased with increasing parity and stage of lactation. To evaluate the predictive power of all other variables within parity group, we built separate models for parity 1 and 2,

and stage of lactation was left out of the initial models. The effect of stage of lactation on the model error and performance was tested in the final models. The data excluded for representing a transition lameness degree (score 2), was used to generate predictions to evaluate how many alarms could be expected on this group of cows using the respective models (parity 1 or 2). As the number of observations was insufficient for dividing the data into calibration and validation sets, cross-validation (CVAL) was done instead, predicting the available weekly gait scores of each cow at a time (leave-one-cow-out, all the weeks), so that the model predicting the cow left out did not include any observation from this cow. In other words, several models equal to the number of cows available were built, each time leaving one of the cows out. Finally, the CVAL classification error and model output parameters were calculated.

Model control was based on the diagnostics Q residuals and Hotelling’s T^2 (Ballabio and Consonni, 2013). The Q residuals measure the unexplained variance (error), whereas Hotelling’s T^2 measures the variation within the model and is the distance from the center (origin) of the model. Outlier detection was done by visual inspection of Q residuals versus Hotelling’s T^2 plot. We modeled the Y response from 0 (nonlame) to 1 (lame) using the default threshold of 0.5. That is, cows with a predicted value ≥ 0.5 are considered a lameness alarm. Then, varying the threshold, model performance can be evaluated by visual inspection of receiver operating characteristic (ROC) curves, comparing the calibration with the CVAL ROC curves (Ballabio and Consonni, 2013). The calibration ROC curve shows the fitted model (regression), which is expected to be too optimistic due to overfitting. Thus, the CVAL ROC curve is considered more reliable, because its results come from n cross-validation models predicting unknown samples from a given cow (where n is equal to the number of cows). This means each cow’s predictions, which can be true or false, are combined for all the cows from these n models to generate the CVAL confusion matrix. The classification error rate, sensitivity, and false-positive rate were also assessed and compared between parity groups, and also within each model by comparing calibration and CVAL results. The closer the 2 ROC curves are (calibration and CVAL), the more robust the model is expected to be on future unknown samples.

Partial least squares discriminant analysis aims at identifying patterns distinguishing the 2 classes, lame and nonlame. After the first latent variable is modeled, the residuals will be modeled by the second latent variable and so on. The number of latent variables was chosen based on minimal CVAL classification error. As the registered AMS data are noisy and might contain

irrelevant variables, it is useful to perform variable selection to exclude variables that might weaken the gait score prediction. Variable influence on projection (VIP) is a summary of the importance of one variable both in **X**- and **Y**-variation (Wold et al., 2001). Starting from a full PLS-DA model with all the variables, visual inspection of loadings and VIP was done to perform backward variable selection, removing iteratively the variables that had the lowest loadings and VIP (Chong and Jun, 2005). At the final stage of variable selection, variables responsible for high residuals were removed. The analysis was done with PLS_Toolbox software (v.7.0.2; Eigenvector Research Inc., Wenatchee, WA) using the MATLAB programming language (v.7.14; The MathWorks Inc., Natick, MA).

RESULTS

The weekly lameness prevalence varied between 24 and 41% for all gait scored cows in 2012. At wk 1 of 2012, parity 1 and 2 had a lameness prevalence of 10 and 16%, respectively, whereas the weekly prevalence in 2013 varied between 15 and 27% and 19 and 37%, respectively. With 4 gait categories, the first and second author agreed on the same gait score in 53% of the 135 cows scoring independently. The weighted kappa was then 0.43 (95% CI: 0.32 to 0.55), indicating moderate agreement (Sim and Wright, 2005). Correspondingly, the intraobserver agreement was 52% (first author) and 54% (second author) when scoring 85 cows twice in the same day, and the respective weighted kappa values were 0.39 (95% CI: 0.23 to 0.54) and 0.30 (95% CI: 0.12 to 0.48), indicating fair agreement (Sim and Wright, 2005). If the weighted kappa statistic was calculated excluding gait score 2 category, the agreement was then moderate to substantial (Sim and Wright, 2005), achieving 0.76 (95% CI: 0.62 to 0.90), 0.51 (95% CI: 0.23 to 0.78), and 0.50 (95% CI: 0.23 to 0.76), respectively, for interobserver agreement and repeatability of the first and second author. Merging the 4 categories into a binary assessment (lame vs. non-lame) and excluding gait score 2, the interobserver agreement would return a kappa of 0.81 (95% CI: 0.67 to 0.96), corresponding to almost perfect agreement (Sim and Wright, 2005). Hence, with an agreement on 91, 85, and 78% of the cases (lame vs. not lame, excluding gait score 2), respectively, for interobserver agreement and repeatability of the first and second author, we can alternatively estimate that the average error of the reference method was around 15%.

Parity 1 PLS-DA Model

A model with 4 latent variables explained 47% of the predictor variation and 45% of the lameness variation,

using 17 out of the original 320 variables. The cross-validated sensitivity and specificity were, respectively, 79 and 77% (calibration: 84 and 82%), as shown in the ROC curves in Figure 2a. In CVAL, the model generated 73 true positives, 87 true negatives, 28 false positives, and 20 false negatives, which corresponded to a classification error of 23%, compared with 17% in calibration where overfitting is usually expected. No measurements were considered outliers. Inclusion of DIM in the final model did not change the results (cross-validated model error: 23%). The samples and variables are presented respectively in Figure 3a1 and 3b1, with the variables being shaded or colored according to VIP.

The reference gait scoring has an error around 15%, and the given model error is 23%. These 23% represent the actual error and the reference error. Even if the model were perfect, it would show an error of 15% due to the reference error. Hence, the real error of the model is approximately $\sqrt{23^2 - 15^2} = 17$, which is as good as the reference method (DiFoggio, 1995). With respect to being lame, the following variables had the highest positive values for regression coefficients, between 0.2 and 0.3: activity index median at 2200 h in the last week, performance index variance, and activity index variance at 0600 h in the last week and at 1100 h in the current week. If we observe the score plot of the 2 most important latent variables in Figure 3a1, a separation between lame and nonlame cows could be noted from the upper left to the lower right quadrant. In Figure 3b1, the corresponding loadings identify the variables that are causing this grouping in the direction of these 2 cow clusters. The interpretation could be that, compared with their nonlame herd mates, lame cows generally had more unstable milking performance and activity (higher variance) while being more active in the evening. Further detailed interpretation of other variables could be done (also at higher dimensions or latent variables), with the corresponding scores and loading plots, but the 2 first latent variables show the best separation between the 2 classes while explaining 39% of the lameness variation. Still, we see from the loadings plot (Figure 3b1) that performance index variance is negatively correlated with the feeding distributed in the robot with respect to the variation shown in these 2 latent variables, as they almost lie on an imaginary straight line crossing the origin. We also see that activity index median at 1100 and 1200 h are correlated because they lie very close in the projection space, at least in latent variable 1 versus 2, and so on.

Parity 2 PLS-DA Model

The parity 2 model was more complex, using 28 out of the original 320 variables (Figure 3b2). A model

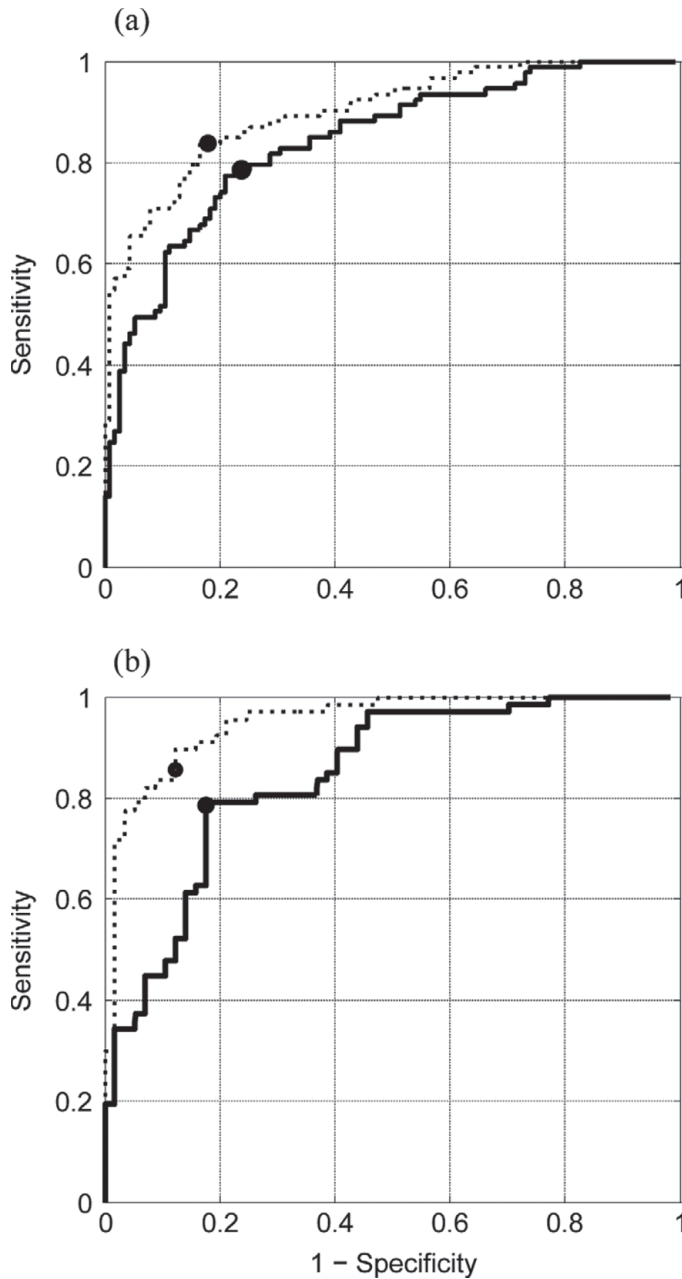


Figure 2. Receiver operating characteristic (ROC) curves for parity 1 (a) and parity 2 (b) models, respectively, with cross-validated (CVAL) sensitivity 79 and 79%, CVAL specificity 77 and 83%, and CVAL classification error 23 and 19%, respectively. Dashed line = calibration; solid line = cross-validation; circles = and optimal values.

with 6 latent variables explained 52% of the predictor variation and 58% of the lameness variation, having cross-validated sensitivity and specificity, respectively, of 79 and 83% (calibration: 85 and 88%), as shown in the ROC curves in Figure 2b. In CVAL, the model generated 53 true positives, 47 true negatives, 10 false positives, and 14 false negatives, which corresponded to a classification error of 19 versus 14% in calibration.

No outliers were removed. Inclusion of DIM in the final model did not change the results considerably (cross-validated model error: 21%). The samples and variables can be assessed in Figure 3a2 and 3b2, respectively, with the variables colored according to VIP.

Similarly, if we correct for the reference error, the actual model error could be considered negligible compared with the reference. As for being lame, the following variables had the highest positive values for regression coefficients, between 0.2 and 0.3: activity index median at 0500 h (both weeks), activity index variance at 1300 h in the last week and at 1600 h in the current week, and average milk flow variance in the last week. In Figure 3a2, a separation between lame and nonlame cows could also be distinguished in the score plot from the upper right to the lower left quadrant, whereas in Figure 3b2, the loadings show how the variables contribute to the samples positioning. The interpretation could be that, compared with their nonlame herd mates, lame cows generally had higher activity in early morning, a more unstable activity pattern in the daytime (higher variance), and a poorer milking performance (low performance index median and high variance of performance index and average milk flow) for the most informative latent variables 1 and 2, which explain 51% of the lameness variation. Again, only the most important variables are highlighted, but it can be derived that in latent variable 1 versus 2 (Figure 3b2) a negative correlation is found between average milk flow median and milking interval median, activity index median at 1200 h in the last week and activity index median at 0400 h, activity index median at 0600 h and activity index median at 1500 h, and so on.

Excluded Data: Score 2 Sample Prediction

With the parity 1 model, the prediction of parity 1 score 2 cows resulted in 159 alarms (55%), whereas 129 events were predicted as nonlame. Predictions of parity 2 score 2 cows with the parity 2 model returned 65 alarms (42%) and 88 nonlame events. The PLS-DA scores for these samples were scattered on all quadrants of the projection space (figure not shown); that is, we did not observe clustering with a specific group, either lame or nonlame samples.

DISCUSSION

The objective of this study was to extract relevant information from AMS data for lameness detection. Separate models were developed for parity 1 and parity 2, achieving a non-error rate around 80% when discriminating between nonlame (score 1 out of 4) and lame cows (score 3 or higher). The models relied ex-

clusively on milking- or physical activity-related data. Thus, such results support our hypothesis of hidden behavioral and performance patterns useful to detect lame cows on AMS dairy farms.

Prior to model building, we were interested in estimating the reference method error. Our raw intra- and interobserver agreement, as defined by kappa statistics, is within the range of previous studies. Intraobserver and interobserver kappa values range from 0.30 to 0.68 and 0.15 to 0.59, respectively (Thomsen et al., 2008), or from 0.47 to 0.70 and 0.11 to 0.75, respectively (Schlageter-Tello et al., 2013). Surprisingly, we found a higher agreement in interobserver kappa values than in intraobserver values. The interobserver assessment was made at the same point in time, whereas the intraobserver assessment was made at 2 different time points; thus, cows might have shown variation in their walking behavior. Therefore, the error rate of the reference method includes observer error and cow variation, which is important to consider in model performance discussion.

The parity 1 group had the lowest weekly lameness prevalence. Yet, the parity 1 model identified 73 of 93 lame events, based only on 4 latent variables. Similarly, high specificity led to classification of 87 of 115 nonlame events. The parity 2 group, on the other hand, which had higher weekly lameness prevalence, generated a model with higher sensitivity and specificity than parity 1, detecting 53 of 67 lame events and 47 of 57 nonlame events, while using 6 latent variables. In both models, the first 2 and most informative latent variables differed. In the parity 1 model, latent variable 1 was mostly related to high activity, having the highest loadings in activity index median in daytime hours, whereas latent variable 2 was mostly related to milking information, consumed concentrate, activity variance, and activity in the evening. In contrast, latent variable 1 from the parity 2 model had the highest loadings in milking-related variables, whereas latent variable 2 had the highest loadings in activity data. In summary, activity data was more informative than milk data in parity 1 to classify lame cows, whereas in the parity 2 model it was the opposite. It could simply be the case that younger and, therefore, more playful cows (parity 1) generate richer information in activity data than older cows (Løvendahl and Chagunda, 2010). These differences between the 2 models support the hypothesis that parity 1 and 2 cows differ in production, physiology, and behavior, exhibiting different patterns. As a result, future studies might benefit from modeling parity groups independently to identify the strongest variables for lameness detection. Nonetheless, the 2 models showed common features: (1) higher activity median at specific time points, (2) lower performance

index variance was associated with nonlame cows, and (3) higher performance index variance was associated with lame cows.

As we had several variables with similar values of regression coefficients, we can conclude that all are important to achieve the discrimination in this particular farm. In addition, median-based variables might be good descriptors of the nonlame class, whereas variance-based variables might best describe the lame class pattern, as activity median-based variables pulled the nonlame cows to one corner of the projection space and variance-based variables pulled the lame cows to the opposite corner. This finding held true both in parity 1 and parity 2 models (Figure 3).

Interestingly, nonlame cows had higher values on activity median variables at feeding times 1800 h in parity 1 and at 0600, 1200, and 2200 h in parity 2, where lame cows correspondingly had lower values. This suggests that lame cows might be choosing less crowded time periods to eat, as they might be less eager to compete for food due to painful conditions, which is in agreement with previous studies (Blackie et al., 2011; Yunta et al., 2012).

Inclusion of DIM in the final models did not cause any improvements. Thus, the information on DIM, even if useful to lameness detection, probably is correlated with information from other variables and, therefore, already captured in the model.

A common way to assess within-sample variation is simply to visually inspect the score plots, colored or numbered by the sample name, and check for subject clusters. Hence, 2 observations of the same subject should then lie close together in the projection space or, in extremis, on top of each other (e.g., pure replicate without instrumental error). In our case, even though some cow clusters were identified, the within-cow variation was not similar across different cows. Some cows clustered with the same gait score, indicating those observations were similar. However, a few cows clustered with different gait scores, and others did not cluster at all, being far away from each other, although having the same gait score. In some cases, a different stage of lactation could be the most obvious explanation for a longer distance within cow, but certainly cows not clustering and with the same gait score are most likely the ones contributing to the model error. Lame cows undetected by the model could have had a lameness episode of short duration, which might not influence the behavior and production data significantly, but still be easily detected with gait scoring. The models discriminated correctly individual gait score changes between lame and nonlame, meaning that a change was detected apart from the general lame/nonlame trend, even though we did not modeled cows individually.

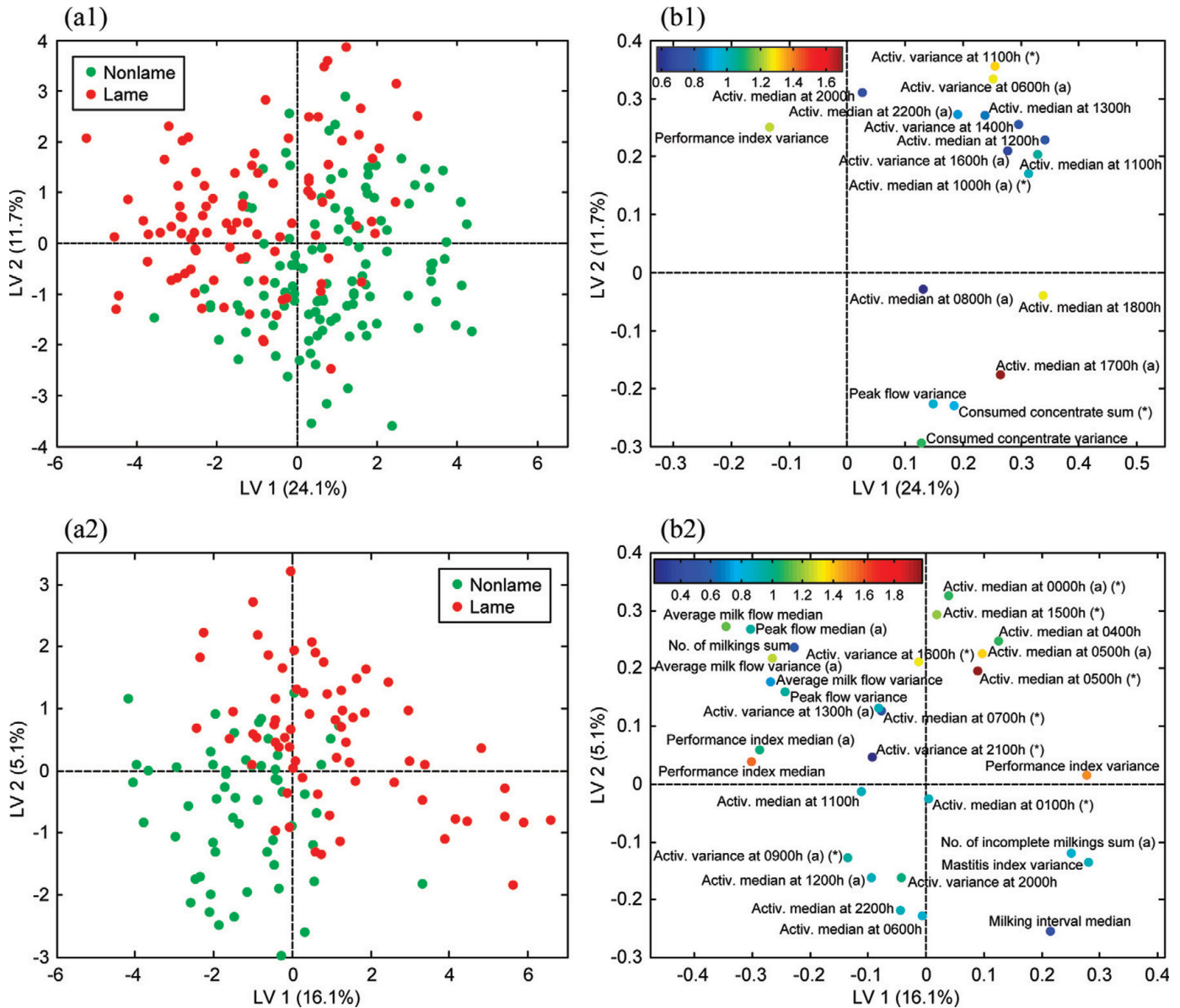


Figure 3. Scores (a1 and a2) and loadings (b1 and b2) of the first 2 latent variables (LV) retained in the models parity 1 (a1 and b1) and parity 2 (a2 and b2), obtained with partial least squares discriminant analysis. Parity 1 had 4 latent variables with 47% explained variance and parity 2 had 6 latent variables with 52% explained variance. Variables from the second matrix (\mathbf{X}_{new}) are followed by (*), whereas all the others are from the first matrix (\mathbf{X}_{log}). Variables referring to the last week's data are followed by (a); otherwise, they refer to the current week's data. The scores are shaded or colored according to the clinical scoring (reference method). The loadings are shaded or colored according to the variable influence on projection (VIP) shown in the color bar, which is a summary measure of importance for explaining the variation in the variables and in the lameness response. In the loadings plot, activity index is abbreviated as Activ. Color version available in the online PDF.

Our estimates of predictive values can be considered promising if compared with sensors developed specifically for lameness detection, which reported a range between 22 to 80% for sensitivity and specificity, with only 4 sensor systems reporting higher than 80% (Rutten et al., 2013). Moreover, although a high specificity (>80%) could be achieved in some studies (Ito et al., 2010; Miekley et al., 2012), this generally led to lower sensitivity.

A combination of high sensitivity and specificity has been achieved in few studies (Poursaberi et al., 2010; Maertens et al., 2011; de Mol et al., 2013; Van Hertem et al., 2013); these were based on data from 1 farm and a relatively small sample size, also true in our case, which means the application of models on unknown cows and farms will require further validation studies. In a Danish study, Jónsson (2011) achieved 76% sensitivity and 74% specificity when combining feeding,

activity, and visits to the robot data. In contrast to our study, Jónsson (2011) concluded that activity data did not add much predictive power using neck activity sensor data, but here it was a useful predictor for lameness detection, especially in parity 1 cows. Two studies (de Mol et al., 2013; Van Hertem et al., 2013) indicated that milk and feeding data combined with sensor data improved lameness detection accuracy. Interestingly, de Mol et al. (2013) conducted a model validation, which achieved a similar model performance in the test set when compared with their calibration set. With a different setup, based on herd health records and assuming strict veterinary monitoring and identification of lameness by the cow pusher, Van Hertem et al. (2013) reached 89% sensitivity and 85% specificity using a logistic regression model with milk, rumination, and neck sensor activity data as input. Even with a study design different from the 2 previous studies, we also benefited from transformations and variance scaling instead of using raw data. Direct comparison with these studies is not appropriate due to differences in study design (e.g., monthly lameness scoring) as well as the absence of reference method error estimates.

Altogether, the models herein are robust and could fairly well classify all observations from 1 unknown cow, if the model were built based on all the others, because CVAL ROC curves were close to calibration.

Limitations

Our data may not be representative of the farm at another point in time or of another farm with data that behave differently due to local factors (e.g., management, housing, and so on). Also, different robot brands, software, and activity sensors might be present whenever a study includes more than 1 farm. We assumed estrus behavior could mask lameness due to high activity, but this is difficult to test in practice. Therefore, we acknowledge that removing activity data respective to estrus period might introduce bias on the activity estimate of that particular week.

Farmers might not tolerate the false-alarm rate: on average, 2.8 cows would not be lame for every 10 cows on the alarm list in the parity 1 model. Conversely, with 1.6 false in every 10 alarms, the parity 2 model may be more suitable to practical application than the parity 1 model. The model threshold can be adjusted to achieve higher specificity as depicted in Figure 2, at the expense of sensitivity.

Exclusion of cows with gait score 2 might be the greatest limitation in our study. We noted that many cows gait scored as 2 were not motivated to walk or to walk fast enough to enable proper assessment. By

excluding almost 50% of the observations, we cannot expect to have modeled all the variation in this farm, and then the direct application of such models in this farm could lead to higher error rates. Possibly, some cows are lame in all 4 feet, but they can hide it walking slowly, whereas some cows might just walk more slowly or more stiffly because they try to cope with slippery flooring. The prediction results of score 2 cows confirm this is not a homogeneous category, as they did not cluster in a specific region of the projection space. If score 2 samples lay between nonlame and lame samples, we could hypothesize that this was a well-defined class representing a gray area of lameness degree. This was not the case, and because intermediate gait score categories show the lowest kappa values (Schlageter-Tello et al. 2013), we need to assume that some score 2 cows could be lame or severely lame, and others were truly not lame. Therefore, including cows with an unnatural gait in the group of nonlame cows would disturb what could be considered as an ideal nonlame status or control group. Our novel findings should be interpreted cautiously due to the small data set of only 1 farm, but certainly deserve extended investigation.

Final Remarks

Both models had a cross-validated classification error (~20%) that might not be considered accurate enough by some herd managers. However, compared with the performance of mastitis alarms based on electrical conductivity, with reported sensitivities of 55 to 89% and specificities of 56 to 99%, which do not meet the International Organization for Standardization standard (Rutten et al., 2013), these results encourage further study of the relationships between lameness and AMS data. It is noteworthy that the results presented were based on data not measuring lameness directly, in contrast to data from force sensors, pain measurements, accelerometer data, kinematics, or video image analysis. Yet, with multivariate data analysis and simple data preprocessing, we achieved sensitivity and specificity values that could be valuable for screening obvious and severe lameness cases. A further advantage of the multivariate methods applied was the direct interpretation by visual inspection of the samples and variable plots. To the best of our knowledge, this is the first study using PLS-DA to build a lameness detection model. In a previous work using another latent variable method (principal component analysis), Miekley (2013) reported high error rates. Moreover, we did not find other studies in the literature reporting a correlation between lameness and variance of the DeLaval performance index. Finally, PLS-DA seemed to be an

interesting tool to finding intricate data patterns of behavior and milking information that might not have been studied before.

CONCLUSIONS

The aim of this study was to explore robotic milking-related variables associated with clinical lameness and build a model to discriminate nonlame and lame cows. Promising sensitivity and specificity values were achieved when performing calibration and cross-validation, using 2 separate models built for parity 1 and parity 2 of a commercial dairy farm. Activity data and milking-related information from AMS were useful for lameness classification. Multivariate data analysis was valuable to investigate AMS data by unveiling data patterns of nonlame and lame cows that support previous research. Taking the error of the reference method into account, we achieved a low real model error. Reducing the reference method error rate and performing further model validation is recommended. In conclusion, the present findings support the hypothesis that AMS data could be useful for lameness detection, where PLS-DA seems an effective method to handle and interpret data from different sources.

ACKNOWLEDGMENTS

We thank the University of Copenhagen (Frederiksberg C, Denmark), namely the Chemometrics Analysis Center (CHANCE) and Centre for Herd-oriented Education, Research and Development (HERD), and the Karla and Svend O. Kochs Foundation (Frederiksberg, Denmark) for funding this study. We thank the farmer, the veterinary practitioner, and the staff for their close collaboration, and also the reviewers for valuable feedback on this manuscript.

REFERENCES

- Bach, A., M. Dinarés, M. Devant, and X. Carré. 2007. Associations between lameness and production, feeding and milking attendance of Holstein cows milked with an automatic milking system. *J. Dairy Res.* 74:40–46.
- Ballabio, D., and V. Consonni. 2013. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Anal. Methods* 5:3790–3798.
- Bisaglia, C., Z. Belle, G. van den Berg, and J. C. A. M. Pompe. 2012. Automatic vs. conventional feeding systems in robotic milking dairy farms: A survey in the Netherlands. Pages 1–6 in International Conference of Agricultural Engineering, CIGR-AgEng2012, Valencia. Federación de Gremios de Editores de España, Madrid, Spain.
- Blackie, N., J. Amory, E. Bleach, and J. Scaife. 2011. The effect of lameness on lying behaviour of zero grazed Holstein dairy cattle. *Appl. Anim. Behav. Sci.* 134:85–91.
- Blowey, R. W. 1998. Welfare aspects of foot lameness in cattle. *Ir. Vet. J.* 51:203–207.
- Borderas, T. F., A. Fournier, J. Rushen, and A. M. B. de Passillé. 2008. Effect of lameness on dairy cows' visits to automatic milking systems. *Can. J. Anim. Sci.* 88:1–8.
- Bro, R., F. van den Berg, A. Thybo, C. M. Andersen, B. M. Jørgensen, and H. Andersen. 2002. Multivariate data analysis as a tool in advanced quality monitoring in the food production chain. *Trends Food Sci. Technol.* 13:235–244.
- Cavero, D., K.-H. Tölle, C. Henze, C. Buxadé, and J. Krieter. 2008. Mastitis detection in dairy cows by application of neural networks. *Livest. Sci.* 114:280–286.
- Chong, L.-G., and C.-H. Jun. 2005. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* 78:103–112.
- Dallal, G. E. 2012. Transformations: Logarithms. The Little Handbook of Statistical Practice. Accessed Sep. 24, 2014. <http://www.jerrydallal.com/LHSP/logs.htm>.
- de Koning, C. J. A. M. 2011. Milking machines | Robotic milking. Pages 952–958 in *Encyclopedia of Dairy Sciences*. 2nd ed. John W. Fuquay ed. Academic Press, San Diego, CA.
- de Mol, R. M., G. André, E. J. B. Bleumer, J. T. N. van der Werf, Y. de Haas, and C. G. van Reenen. 2013. Applicability of day-to-day variation in behavior for the automated detection of lameness in dairy cows. *J. Dairy Sci.* 96:3703–3712.
- DiFoggio, R. 1995. Examination of some misconceptions about near-infrared analysis. *Appl. Spectrosc.* 49:67–75.
- Dumas, M.-E., C. Canlet, L. Debrauwer, P. Martin, and A. Paris. 2005. Selection of biomarkers by a multivariate statistical processing of composite metabonomic data sets using multiple factor analysis. *J. Proteome Res.* 4:1485–1492.
- Ghotoorlar, S. M., S. M. Ghamsari, I. Nowrouzian, S. M. Ghotoorlar, and S. S. Ghidary. 2012. Lameness scoring system for dairy cows using force plates and artificial intelligence. *Vet. Rec.* 170:126.
- Gorzecka, J., N. C. Friggens, C. Ridder, and H. Callesen. 2011. A universal index of uterine discharge symptoms from calving to 6 weeks postpartum. *Reprod. Domest. Anim.* 46:100–107.
- Green, L. 2009. Lameness in dairy cows: Piecing together the evidence base and looking forward. Page 1–7 in *Proc. Cattle Lameness Conf. 2009*. University of Bristol, the Dairy Group and University of Nottingham, Sutton Bonington, UK.
- Greiner, M., and I. A. Gardner. 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev. Vet. Med.* 45:3–22.
- Heald, C. W., T. Kim, W. M. Sischo, J. B. Cooper, and D. R. Wolfgang. 2000. A computerized mastitis decision aid using farm-based records: An artificial neural network approach. *J. Dairy Sci.* 83:711–720.
- Ito, K., M. A. G. von Keyserlingk, S. J. LeBlanc, and D. M. Weary. 2010. Lying behavior as an indicator of lameness in dairy cows. *J. Dairy Sci.* 93:3553–3560.
- Jacobs, J. A., and J. M. Siegford. 2012. Invited review: The impact of automatic milking systems on dairy cow management, behavior, health, and welfare. *J. Dairy Sci.* 95:2227–2247.
- Jónsson, R. I. 2011. Detection of oestrus and lameness in dairy cows. Doctoral Thesis. Technical University of Denmark (DTU), Kongens Lyngby, Denmark.
- Ketelaar-de Lauwere, C. C., S. Devir, and J. H. M. Metz. 1996. The influence of social hierarchy on the time budget of cows and their visits to an automatic milking system. *Appl. Anim. Behav. Sci.* 49:199–211.
- Klaas, I. C., C. Enevoldsen, M. Vaarst, and H. Houe. 2004. Systematic clinical examinations for identification of latent udder health types in Danish dairy herds. *J. Dairy Sci.* 87:1217–1228.
- Klaas, I. C., T. Rousing, C. Fossing, J. Hindhede, and J. T. Sørensen. 2003. Is lameness a welfare problem in dairy farms with automatic milking systems? *Anim. Welf.* 12:599–603.
- Landin, H., and M. Gyllenswärd. 2012. Ratta rätt I robot—Mjölknings, juverhälsa och hygien. Pages 41–47 in *Djurhälso- & Utfodringskonferensen 2012*, Uppsala, Sweden. Svensk Mjölk, Hållsta, Sweden.
- Larsson, P. 2007. Economic analysis of the De Laval activity meter system for heat detection—A case study on farm level. MS Thesis in Business Administration. Dept. of Economics, Faculty of

- Natural Resources and Agricultural Sciences, Swedish University of Agriculture Science, Uppsala, Sweden.
- Løvendahl, P., and M. G. G. Chagunda. 2010. On the use of physical activity monitoring for estrus detection in dairy cows. *J. Dairy Sci.* 93:249–259.
- Lyons, N. A., K. L. Kerrisk, N. K. Dhand, and S. C. Garcia. 2013. Factors associated with extended milking intervals in a pasture-based automatic milking system. *Livest. Sci.* 158:179–188.
- Maertens, W., J. Vangeyte, J. Baert, A. Jantuan, K. C. Mertens, S. De Campeneere, A. Pluk, G. Opsomer, S. Van Weyenberg, and A. Van Nuffel. 2011. Development of a real time cow gait tracking and analysing tool to assess lameness using a pressure sensitive walkway: The GAITWISE system. *Biosystems Eng.* 110:29–39.
- Miekley, B. 2013. Electronic monitoring of mastitis and lameness: An application and evaluation of control methods. Doctoral Thesis. Christian-Albrechts-Universität zu Kiel, Kiel, Germany.
- Miekley, B., I. Traulsen, and J. Krieter. 2012. Detection of mastitis and lameness in dairy cows using wavelet analysis. *Livest. Sci.* 148:227–236.
- Miekley, B., I. Traulsen, and J. Krieter. 2013. Principal component analysis for the early detection of mastitis and lameness in dairy cows. *J. Dairy Res.* 80:335–343.
- Nielen, M., Y. H. Schukken, A. Brand, S. Haring, and R. T. Ferwerda-Van Zonneveld. 1995. Comparison of analysis techniques for on-line detection of clinical mastitis. *J. Dairy Sci.* 78:1050–1061.
- Nielsen, T. D., I. L. Vesterbæk, A. B. Kudahl, K. J. Borup, and L. R. Nielsen. 2012. Effect of management on prevention of *Salmonella* Dublin exposure of calves during a one-year control programme in 84 Danish dairy herds. *Prev. Vet. Med.* 105:101–109.
- Piwczyński, D., Z. Nogalski, and B. Sitkowska. 2013. Statistical modeling of calving ease and stillbirths in dairy cattle using the classification tree technique. *Livest. Sci.* 154:19–27.
- Poursaberi, A., C. Bahr, A. Pluk, A. Van Nuffel, and D. Berckmans. 2010. Real-time automatic lameness detection based on back posture extraction in dairy cattle: Shape analysis of cow with image processing techniques. *Comput. Electron. Agric.* 74:110–119.
- Reader, J. D., M. J. Green, J. Kaler, S. A. Mason, and L. E. Green. 2011. Effect of mobility score on milk yield and activity in dairy cattle. *J. Dairy Sci.* 94:5045–5052.
- Rutten, C. J., A. G. J. Velthuis, W. Steeneveld, and H. Hogeveen. 2013. Invited review: Sensors to support health management on dairy farms. *J. Dairy Sci.* 96:1928–1952.
- Schlageter-Tello, A., E. Bokkers, P. Groot Koerkamp, T. Van Hertem, S. Viazzi, C. Romanini, I. Halachmi, C. Bahr, D. Berckmans, and C. Lokhorst. 2013. Within and between observer agreement for specific levels in a five levels locomotion score for dairy cows. Pages 88–89 in 17th Intl. Symp. 9th Intl. Conf. Lameness in Rumin., Bristol, UK. University of Bristol, Bristol, UK.
- Serrano-Cinca, C., and B. Gutiérrez-Nieto. 2013. Partial least square discriminant analysis for bankruptcy prediction. *Decis. Support Syst.* 54:1245–1255.
- Sim, J., and C. C. Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys. Ther.* 85:257–268.
- Sloth, K. H. M. N., N. C. Friggens, P. Løvendahl, P. H. Andersen, J. Jensen, and K. L. Ingvarsten. 2003. Potential for improving description of bovine udder health status by combined analysis of milk parameters. *J. Dairy Sci.* 86:1221–1232.
- Sprecher, D. J., D. E. Hostetler, and J. B. Kaneene. 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. *Theriogenology* 47:1179–1187.
- Steeneveld, W., L. W. Tauer, H. Hogeveen, and A. G. J. M. Oude Lansink. 2012. Comparing technical efficiency of farms with an automatic milking system and a conventional milking system. *J. Dairy Sci.* 95:7391–7398.
- Thomsen, P. T., L. Munksgaard, and F. A. Tøgersen. 2008. Evaluation of a lameness scoring system for dairy cows. *J. Dairy Sci.* 91:119–126.
- Van Hertem, T., E. Maltz, A. Antler, C. E. B. Romanini, S. Viazzi, C. Bahr, A. Schlageter-Tello, C. Lokhorst, D. Berckmans, and I. Halachmi. 2013. Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. *J. Dairy Sci.* 96:4286–4298.
- von Keyserlingk, M. A. G., J. Rushen, A. M. de Passillé, and D. M. Weary. 2009. Invited review: The welfare of dairy cattle—Key concepts and the role of science. *J. Dairy Sci.* 92:4101–4111.
- Vong, R., P. Geladi, S. Wold, and K. Esbensen. 1988. Source contributions to ambient aerosol calculated by discriminant partial least squares regression (PLS). *J. Chemometr.* 2:281–296.
- Wold, S., M. Sjöström, and L. Eriksson. 2001. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 58:109–130.
- Ye, F., and A. Zhao. 2010. What you see may not be what you get—A brief introduction to overfitting. Cancer Biostatistics Workshop, Vanderbilt University, Nashville, TN. Accessed Sep. 24, 2014. http://biostat.mc.vanderbilt.edu/wiki/pub/Main/AlexZhao/Overfitting_Cancer_workshop_04162010.pdf.
- Yunta, C., I. Guasch, and A. Bach. 2012. Short communication: Lying behavior of lactating dairy cows is influenced by lameness especially around feeding time. *J. Dairy Sci.* 95:6546–6549.