Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Polynomial order selection in random regression models *via* penalizing adaptively the likelihood

J.D. Corrales[1,2], S. Munilla[2] & R.J.C. Cantet[2,3]

1 Grupo de Genética, Mejoramiento y Modelación Animal, GaMMA, Universidad de Antioquia, Medellín, Colombia
2 Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, Buenos Aires, Argentina
3 Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina

Keywords
Legendre polynomial; model selection; penalizing adaptively the likelihood; random regressions.

Correspondence
R.J.C. Cantet, Av. San Martín 4453 (C1417DSE), Buenos Aires, Argentina.
Tel: 54-11-4524-8000, ext. 8184;
Fax: 54-11-4524-8735;
E-mail: rcantet@agro.uba.ar

Received: 10 June 2014;
accepted: 28 October 2014

Summary

Orthogonal Legendre polynomials (**LP**) are used to model the shape of additive genetic and permanent environmental effects in random regression models (**RRM**). Frequently, the Akaike (**AIC**) and the Bayesian (**BIC**) information criteria are employed to select LP order. However, it has been theoretically shown that neither AIC nor BIC is simultaneously optimal in terms of consistency and efficiency. Thus, the goal was to introduce a method, 'penalizing adaptively the likelihood' (**PAL**), as a criterion to select LP order in RRM. Four simulated data sets and real data (60 513 records, 6675 Colombian Holstein cows) were employed. Nested models were fitted to the data, and AIC, BIC and PAL were calculated for all of them. Results showed that PAL and BIC identified with probability of one the true LP order for the additive genetic and permanent environmental effects, but AIC tended to favour over parameterized models. Conversely, when the true model was unknown, PAL selected the best model with higher probability than AIC. In the latter case, BIC never favoured the best model. To summarize, PAL selected a correct model order regardless of whether the 'true' model was within the set of candidates.

Introduction

The random regression model (RRM) is used in dairy cattle for the genetic evaluation of production traits that change over time. In a RRM, the shape of the lactation curve is accounted for by an average trajectory plus a set of random regression coefficients that define individual deviates related to the additive genetic and permanent environmental effects. Orthogonal Legendre polynomials (LP) are commonly used to model the covariance structure between the random regression coefficients for test-day records. In this context, accurate prediction of the additive genetic and permanent environmental effects in the RRM requires using the proper order of the Legendre polynomial. When the trait evaluated is milk yield, LP of the same order (usually in between 3 and 5) for both type of effects are typically used (Pool & Meuwissen 2000; Strabel *et al.* 2005; Herrera *et al.* 2013). However, the LP order does not have to be equal for both types of effects (Pool *et al.* 2000; Liu *et al.* 2006; Bignardi *et al.* 2009). For example, Liu *et al.* (2006) obtained a better fit with a model of order 5 for the additive genetic effects and order 7 for the permanent environmental effects.

Different criteria have been used to find the polynomial order of the model with the best fit and parsimony. The two criteria most frequently used are the Akaike information criterion (AIC; Akaike 1974) and the Bayesian information criterion (BIC; Schwarz 1978) (Bignardi *et al.* 2009). Both AIC and BIC are based on minimizing the expected estimated Kullback–Leibler distance as a fundamental basis for model selection (Burnham & Anderson 2004). In longitudinal studies, the use of AIC has been criticized due to its tendency to favour the model with the

  doi:10.1111/jbg.12130

highest order when sample size gets very large (McQuarrie *et al.* 1997). On the other hand, BIC behaves poorly when the true model is not among the candidates (Burnham *et al.* 2011). In practice, the true model is rarely known, so that it is not clear which criteria should be used. To overcome this problem, Stoica & Babu (2013) introduced a novel rule for model order selection based on 'penalizing adaptively the likelihood' (referred to by its acronym, PAL). In simpler terms, an adaptive criterion is one that uses information from the previous step to increase its selective performance. The PAL allows selecting a model order when the best model order is unknown. The goal of this study was to introduce the PAL criterion to select LP order for additive genetic and permanent environmental effects in RRM. The performance of the procedure is assessed through a simulation experiment and its implementation is illustrated using milk yield at first lactation data from Colombian Holstein cows.

Methods

In matrix form, the model equation for a RRM can be represented as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}\,\boldsymbol{u} + \boldsymbol{W\gamma} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is the $N \times 1$ vector of observations, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effects and $\boldsymbol{u}((N_A m_1) \times 1)$ and $\boldsymbol{\gamma}$ $((N_D m_2) \times 1)$ are the vectors of random regression coefficients for the additive genetic and the permanent environmental effects, respectively. In this notation, N_A stands for the number of animals in the pedigree file, N_D is the number of cows with records and m_1 and m_2 are the orders of the Legendre polynomials for the corresponding function. Finally, $\boldsymbol{e}$ is the random vector of error terms, whereas $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ are the incidence matrices for fixed effects, breeding values and permanent environmental random coefficients, respectively (see Schaeffer 2004, for more details). The additive genetic, permanent environmental and residual variance–covariance matrices are

$$\mathrm{Var}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\gamma}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & 0 & 0\\ 0 & \boldsymbol{P}_e & 0\\ 0 & 0 & \boldsymbol{R}\end{bmatrix},$$

where $\boldsymbol{G} = \boldsymbol{A} \otimes \boldsymbol{K}_\mathrm{A}$, being $\boldsymbol{A}\,(N_\mathrm{A} \times N_\mathrm{A})$ the additive genetic relationship matrix among animals, and $\boldsymbol{K}_\mathrm{A}$ a square matrix that contains covariances among the random regression coefficients for the additive genetic effects. As usual, the symbol $\otimes$ stands for the Kronecker product operator (Searle 1982). Additionally, $\boldsymbol{P_e} = \boldsymbol{I}_{N_\mathrm{D}} \otimes \boldsymbol{K}_\mathrm{Pe}$, with $\boldsymbol{I}_{N_\mathrm{D}}$ an identity matrix of order N_D and $\boldsymbol{K}_\mathrm{Pe}$ a square matrix with covariances among the random regression coefficients for the permanent environmental effects. Finally, $\boldsymbol{R} = \mathrm{Diag}\{\sigma^2_{e_k}\}$ is a diagonal matrix with the same variance component $\sigma^2_{e_k}$ within the residual class, k.

Assuming that the data vector $\boldsymbol{y}$ follows a multivariate normal distribution, the expression for minus two times the log of the restricted maximum likelihood $(-2 \ln L)$ is

$$\begin{aligned}-2\ln L = {} & const + N_A \ln|\boldsymbol{K}_\mathrm{A}| + m_1 \ln|\boldsymbol{A}| \\ & + N_\mathrm{D} \ln|\boldsymbol{K}_\mathrm{Pe}| + \ln|\boldsymbol{C}| + \ln|\boldsymbol{R}| + y'\boldsymbol{P}y.\end{aligned} \tag{2}$$

In (2), $\boldsymbol{C}$ is the coefficient matrix of the mixed model equations, $\ln|\cdot|$ denote the log determinant of the corresponding matrix $(\cdot)$, and $\boldsymbol{y'Py}$ is the error sum of squares of the model (Meyer & Hill 1997). Let $\boldsymbol{V}$ be the variance–covariance matrix of the data vector $\boldsymbol{y}$, then $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{V}^{-1}\boldsymbol{X}'$. Function (2) is the cornerstone of the two most frequently used methods to select the appropriate order for the Legendre polynomials in a RRM model: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). More precisely, these two methods are based on minimizing the penalized likelihood (Burnham & Anderson 2002)

$$\min_n[-2\ln L + n\omega], \tag{3}$$

where n is the number of parameters in the model and ω is a 'penalty' coefficient. If the restricted likelihood function (2) is used into (3), the penalty coefficient is $\omega_\mathrm{AIC} = 2a_n^*$ for AIC (Müller *et al.* 2013) with $a_n^* = (N-p)/(N-p-n-1)$, being N the number of test-day records and p the number of parameters to be estimated for fixed effects. In turn, $\omega_\mathrm{BIC} = \ln(N)$ for BIC.

Most importantly, these two methods differ in the way they select the 'best' model among those under consideration (Yang 2005): whereas AIC ranks models based on an efficiency criterion (*i.e.* the best model is the one that minimize the error variance asymptotically; Casella & Berger 2002, pp. 470–473), BIC is known to be a consistent criterion (*i.e.* when the sample size tend to infinity, the probability that BIC chooses the best model approaches to one; Casella & Berger 2002, pp. 467–470). If the true model is among those under consideration, BIC would be the best criterion to choose. However, if that is not the case, AIC would be preferable (Burnham & Anderson 2004). Still, AIC is criticized in the literature of longitudinal analysis as it tends to choose models with higher order when the sample size grows unbounded (Shibata 1981; McQuarrie *et al.* 1997). Moreover, Yang (2005)

showed theoretically that neither AIC nor BIC is simultaneously optimal in terms of consistency and efficiency.

To summarize these ideas, consider a set of nested models $M_1 \subset M_2 \subset \ldots \subset M_{\tilde{n}}$, where the subscript indicates the number of parameters, and a true model M_{n_0} with n_0 parameters. By 'true' model (M_{n_0}), we mean the model with the smallest possible number of parameters that is closest (in a Kullback–Leibler sense) to the model that generates the data (Davidson & Mackinnon 2004). In practice, as it is uncertain if the true model is within the set of models under consideration, it is unclear which criterion should be used. To overcome this problem, Stoica & Babu (2013) introduced an alternative approach based on penalizing adaptively the likelihood (PAL criterion). Notice that the negative log-likelihood term in (3) decreases with increasing n, whereas the penalty term increases. The intuition behind their approach to obtain an ideal penalty term is explained by Stoica & Babu (2013) as follows:

(i)When the number of parameters is smaller than in the 'true' model, that is $n < n_0$, a small penalty would make (3) to decrease with increasing n.

(ii)Whereas, if the number of parameters is larger than in the 'true' model, that is $n > n_0$, the penalty term should increase with increasing n, so that (3) increases.

To accomplish these principles, Stoica & Babu (2013) chose the penalty term that multiplies n in PAL to be equal to

$$\omega_{\mathrm{PAL}} = \ln(\tilde{n})\left(\frac{\ln(r_n+1)}{\ln(\rho_n+1)}\right), \qquad (4)$$

where $\tilde{n}$ is the largest number of parameters for the model within the set being considered (for example, if the set of models is $M_1 \subset M_2 \subset \ldots M_{48}$, then $\tilde{n} = 48$), and

$$r_n = 2\ln L_{n-1} - 2\ln L_1 \text{ and } \rho_n = 2\ln L_{\tilde{n}} - 2\ln L_{n-1},$$

are generalized likelihood ratios between model M_{n-1} and the reduced model M_1 or the complete model $M_{\tilde{n}}$, respectively. We assume that M_1 is only a 'reference model' and $r_2 = 0$. As a result, the PAL criterion for model order selection is defined as

$$\mathrm{PAL} = -2\ln L_n + n\ln(\tilde{n})\frac{\ln(r_n+1)}{\ln(\rho_n+1)}. \qquad (5)$$

The best model according to this criterion is the one with the lowest value of PAL. A small example is included in the Appendix 1 to describe how to calculate PAL. When the true model is within the set of candidates, PAL selects the same model order as BIC, otherwise PAL favours a similar model as AIC (see Stoica & Babu (2013), for details). Given these properties, the PAL criterion appears to be an attractive method to assess the LP order of additive genetic and permanent environmental effects for RRM. In the next section, we examine the performance of PAL, when compared to AIC and BIC, by means of a simulation experiment. Additionally, the PAL criterion was applied in a real scenario using daily milk yield data from the Colombian Holstein population.

Simulation experiment

Data for the simulation were created by sampling records with the structure and the pedigree of the Colombian Holstein data set (Table 1). Fixed effects of herd-test-day, age of the cow (as linear and quadratic regressions) and the phenotypic trajectory were included in the model. In the following description, LPm_1 m_2 refers to the order of the Legendre polynomial for additive genetic (m_1) and permanent environmental (m_2) effects, respectively. Four scenarios were considered based on the fraction of records in the data set that were simulated from a single true model: (i) **TM**: data set with all (100%) records simulated from either LP33, LP44, LP55 and LP66 orders; (ii) **NS**: 95% of the records were generated from the true model and 5% were randomly chosen from the other three models simulated. For example, while considering model LP55 as the true one, 5% of records were randomly drawn from either models LP33, LP44 or LP66; (iii) **NR**: 95% of the records were sampled from the true model and 5% were chosen at random from the real Holstein data set; and (iv) **UT:** data set completely simulated but with 'unknown' true model. By 'unknown', we mean that records on any replicate came from four equally represented (25% each) orders of Legendre polynomials: LP35, LP45, LP55 and LP65. For this latter scenario, the 'best' model was chosen using mean square error of prediction, which was calculated as follows:

$$MSEP = \sum_{i=1}^{l}\frac{(\hat{a}_i - a_i)^2}{l}.$$

In the above formula, l is the number of cows with records, a_i corresponds to the true breeding value and $\hat{a}_i$ is the *BLUP* (a_i). At each replicate, true breeding values were sampled in equal proportion from each of the four models tested: LP35, LP45, LP55 and LP65; whereas predicted breeding values were calculated from the single model that was fit from the four

described earlier. The model with the minimum value of MSEP was considered the best model. For each scenario, a total of 100 replicates were simulated and analysed using AIC, BIC and PAL.

Analysis of a milk yield data set

Data were 60 513 first lactation records of test-day milk yields from 6675 Holstein cows, collected from January 1989 to June 2008 in 164 herds from the Colombian Holstein Association. Test-day records were taken within the period from 5 to 305 days of lactation or 19 to 48 months (when looking at age at first calving). Milk yield ranged from 5.1 to 48.4 kg. The minimum number of cows to form any contemporary group was 7. The pedigree file contained 17 062 animals. Table 1 provides a description of the data used in the study.

As before, let LP be the polynomial order of the RRM. Then, two subindexes are used to indicate additive genetic effects (first) and permanent environmental effects (last): LPm_1 m_2. Orders are such that $m_1 = 3,\ldots, 6$ and $m_2 = 3,\ldots, 6$. All model orders within the set plus the simplest model fitting the intercept term only for additive genetic and permanent environmental effects (LP11) were tested with AIC, BIC and PAL. Fixed effects were herd-test-day, age of the cow (with linear and quadratic terms) and the phenotypic trajectory. All models had six intervals for residual variance (6-35, 36-95, 96-125, 126-215, 216-245, 246-305 DIM) and LP of order 5 that account for the phenotypic trajectory. The values of -2 ln L, AIC, BIC, the likelihood ratio test (LRT) and the estimates of the covariance components for the 17 different random regression models were calculated using REML and the 'average information' algorithm (Gilmour *et al.* 1995), by means of the package Wombat (Meyer 2007). The reduced model (M_1) to calculate the PAL was LP11, and the model with the highest order was LP66 ($\tilde{n} = 48$). Additionally, for the Holstein data, the alternative models were also compared by means of their predicted ability using the weighted MSEP (wMSEP) as described by Odegård *et al.* (2003). The procedure is as follows: two data sets are generated by excluding observations from the initial data set. Next, the MSEP is calculated for each subset and, finally, the wMSEP is computed as the average of both estimates of MSEP.

Estimated heritabilities at day t of the lactation curve were calculated using the following formula (Van Der Werf *et al.* 1998; Jakobsen *et al.* 2002):

$$\hat{h}_t^2 = \frac{\hat{\sigma}^2_{a(t)}}{\hat{\sigma}^2_{a(t)} + \hat{\sigma}^2_{pe(t)} + \hat{\sigma}^2_{e(t)}},$$

where $\hat{\sigma}^2_{a(t)}$, $\hat{\sigma}^2_{pe(t)}$ and $\hat{\sigma}^2_{e(t)}$ are the additive genetic, permanent environmental and residual variances at day t, respectively.

Table 1 Descriptive features of the data set

Item	Value
Number of test-day records	60 513
Number of cows with records	6675
Number of animal in pedigree	17 062
Number of contemporary groups	4211
Mean of milk yield (kg)	18.80 ± 5.95
Mean age at first parity	31.67 ± 4.61

Results

The probability of selecting the model with the correct number of parameters using AIC, BIC and PAL under different scenarios simulated is presented in Figure 1. When the true model was among the candidates, PAL and BIC selected with probability one the correct order of the Legendre polynomials for the additive genetic and permanent environmental effects. Instead, not always AIC selected the correct order. When a 5% noise was added to the data set with the true model, AIC tends to overestimate the correct order. Conversely, when the true model was unknown, PAL selected the best model with higher

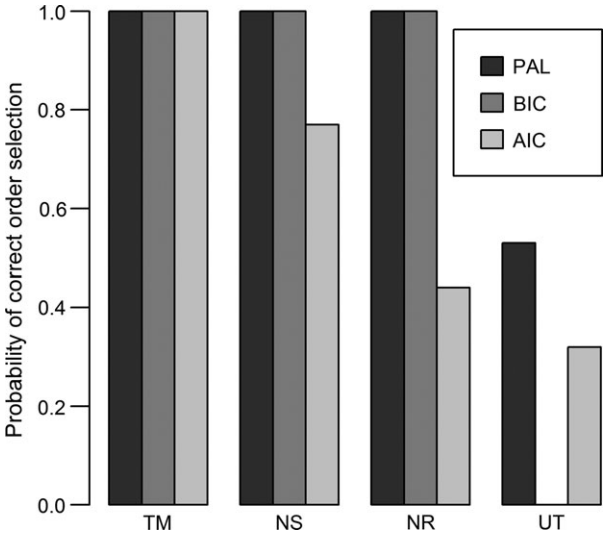

Figure 1 Probability of correct order selection in four scenarios using the penalizing adaptively the likelihood (PAL) criterion, Bayesian information criterion (BIC) and Akaike information criterion (AIC) criterion.

probability than AIC. In this latter case, BIC never chose the best model.

Information about the number of covariance parameters, −2 ln-likelihood, PAL, AIC and BIC for the RRM with different LP order fitted using the Colombia Holstein data set is presented in Table 2. The lowest value of AIC corresponded to LP66 and was followed by LP56. These models had the largest number of parameters over the set of models considered in this study. In contrast, the choice of models using PAL was the same as those selected using BIC: LP36 was best, followed by the LP46. The LP order for additive genetic effects was lower than the one for permanent environmental effects. Based on likelihood ratio test (LRT), the model with more parameters (LP66) was best, but differences in predictive abilities among models of order 6 for permanent environmental effects were small. Based on wMSEP, models LP66 and LP36 were ranked approximately equal.

Model LP36 (order 3 and 6 for the additive genetic and permanent environmental Legendre polynomial, respectively) was selected according to the PAL criterion. Estimated residual variances from LP36 were equal to 5.11, 3.57, 3.35, 3.06, 2.72 and 2.63, for the 6 to 35, 36 to 95, 96 to 125, 126 to 215, 216 to 245 and 246 to 305 DIM, respectively. Figure 2 displays the estimated variances of milk production for additive and permanent environmental effects for the models selected by PAL and BIC defined over time. The lowest estimate of the additive genetic variance was 3.11 kg^2 at day 305, whereas the highest value was 8.10 kg^2 at day 102. Corresponding values for permanent environmental variances were 8.53 kg^2 at day 17 and 11.27 kg^2 at day 305. The estimates of heritability (h^2) on the trajectory of milk production obtained were 0.39 at day 142 for the highest point, and the lowest value was 0.18 at day 305.

Discussion

In this study, we introduced the PAL criterion for selecting the order of Legendre polynomials in random regression models for milk production and compared its performance against standard methods (*i.e.* AIC and BIC) through a simulation experiment. All three methods selected the best model when a 100% of records where generated by the true model. However, when noise was introduced in the simulations, PAL and BIC behaved selecting the correct model, whereas AIC tended to overestimate the model order. These results are consistent with previous studies showing that AIC tends to overfit, whereas PAL and BIC outperform AIC when the true model with some noise is in the candidate set (Schwarz 1978; Shibata 1981; Stoica & Babu 2013). Conversely, when data were simulated under no true model (UT), PAL

Table 2 Model selection criteria for analyses with different orders of Legendre polynomials (LP) for additive genetic (m_1) and permanent environmental (m_2) effects (bold values correspond to the best model for each criterion).

		Selection criteria[1]					
LPm_1 m_2	Number of parameters	−2 ln *L*	AIC	BIC	PAL	LRT*	wMSEP
1) LP66	48	180 814	**180 910**	181 342	181 406	**(1–2) 18**	**1.5898**
2) LP56	42	180 832	180 916	181 294	181 248	(2–3) 20	1.5905
3) LP46	37	180 852	180 926	181 259	181 163	(3–4) 36	1.5987
4) LP36	33	180 888	180 954	**181 250**	**181 083**	(4–8) 380	1.5992
5) LP65	42	180 936	181 020	181 397	181 189	(5–6) 282	1.6608
6) LP55	36	181 218	181 290	181 614	181 433	(6–7) 22	1.7988
7) LP45	31	181 240	181 302	181 580	181 423	(7–8) 28	1.8000
8) LP35	27	181 268	181 322	181 565	181 407	(8–12) 618	1.7964
9) LP64	37	181 122	181 196	181 528	181 332	(9–10) 268	1.7442
10) LP54	31	181 390	181 452	181 730	181 550	(10–11) 465	1.8729
11) LP44	26	181 855	181 907	182 141	181 989	(11–12) 31	2.0265
12) LP34	22	181 886	181 930	182 128	181 986	(12–16) 1297	2.0269
13) LP63	33	181 287	181 353	181 649	181 465	(13–14) 305	1.8483
14) LP53	27	181 592	181 646	181 889	181 728	(14–15) 451	1.9905
15) LP43	22	182 043	182 087	182 285	182 143	(15–16) 1140	2.1263
16) LP33	18	183 183	183 219	183 381	183 183	(16–17) 9310	2.4090
17) LP11	3	192 493					5.6990

[1]REML log-likelihood (−2 ln *L*), AIC = Akaike information criterion, BIC= Bayesian information criterion, PAL = penalizing adaptive the likelihood, LRT = likelihood ratio test between models (*i.e.* 1–2 means the comparison between model 1 and model 2).

*$p < 0.01$ and wMSEP= weighted mean of MSEP from two independent samples.

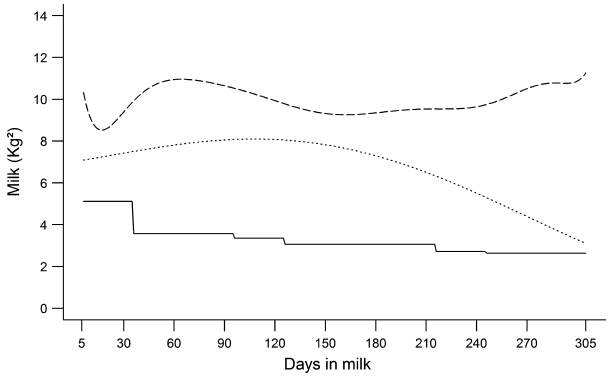

Figure 2 Permanent environmental variance (dashed line), additive genetic variance (dotted line) and residual variance (solid line) of daily milk yield in first lactation.

performs better than AIC, and much better than BIC, when assessing the best model as the one that minimizes MSEP. That AIC could be more effective to choose the best model than BIC when the true model was unknown can be explained by the fact that the number of models does not grow very fast in dimension, and the MSEP of the model selected by AIC approaches asymptotically the minimum value from the set of candidate models (Shibata 1981; Yang 2005).

Although the use of Legendre polynomials in RRM allows a more flexible shape of the lactation curve, high-order polynomials are frequently impossible to implement in large populations. This is due to the requirement of a sizeable computer capacity and also to the possibility of obtaining negative correlations between distant test-day (Pool & Meuwissen 2000; Jamrozik *et al.* 2001). For this reason, the simulated data sets analysed in this study considered only models with LP66 as the maximum order of Legendre polynomials for both additive genetic and permanent environmental effects. On the other hand, orders lower than 3 do not fit well deviations from a typical lactation curve. Therefore, it was considered that the best model was within the range LP33 to LP66, and all models in the interval were fitted to the data.

In practical implementation of RRMs, the first task the analyst must perform is to evaluate which model order (for both additive genetic and permanent environmental effects) is most supported by the data. Selecting the order is often difficult principally because statistical criteria are not clearly defined (Bignardi *et al.* 2009). López-Romero & Carabano (2003), Liu *et al.* (2006) and Bignardi *et al.* (2009) used both AIC and BIC to choose the best polynomial order for RRM, but the choice of LP order was not consistent. Our simulation results suggest that PAL is a good criterion to choose the order of the Legendre polynomials in any implementation of RRM to milk production data. The PAL provides with a rule for choosing among different LP orders, and in particular when the results produced by AIC differ from those produced by BIC. Stoica & Babu (2013) pointed out that there is no theoretical proof of the superiority of PAL over AIC and/or BIC so far. However, it can be employed with the idea that the use of PAL safely subsumes using AIC and BIC together, as the decision will stick with the criterion that is consistent with the framework of inference for the given data and the models compared.

As it can be inferred from our implementation with a real data set of milk yield, PAL also allowed to differentiate between the best order for the additive and the permanent environmental effects. It has been previously observed that LP order for permanent environmental effects tended to be higher than for additive effects (Pool & Meuwissen 2000; López-Romero & Carabano 2003; Carabaño *et al.* 2007). For example, Liu *et al.* (2006) used log-likelihood and information-theoretical measures for order selection and found LP57 as the best model. In turn, Bignardi *et al.* (2009) used both AIC and BIC to end up choosing LP7.12, whereas López-Romero & Carabano (2003) selected a model with LP order 2–3 for additive genetic effects and 5–6 for permanent environmental effects. In our implementation, the estimates of the genetic parameters from the model selected by PAL were of similar magnitude with those found in previous research by Jakobsen *et al.* (2002), López-Romero & Carabano (2003) and López-Romero *et al.* (2003). The residual variances at the beginning of the lactation were larger than those at other intervals. Similar results were found by López-Romero *et al.* (2003) when evaluating the heterogeneity of residual variance of a RRM. We also found that the variances for permanent environmental effects were higher than those for additive genetic effects. A possible explanation lies in the fact that under the typical grazing conditions in Colombia, milk production is highly affected by environmental effects.

To conclude, our findings suggest that PAL is a promising adaptive model selection criterion while assessing the Legendre polynomial order in RRM. Two practical considerations are most important when applying PAL. First, a nested structure for the models to be compared is required. However, in the context of assessing Legendre polynomial order in RRM, a nested structure arises naturally by adding

orders to the polynomials. Second, within the set of models under consideration, reduced and full models are needed to compute the generalized likelihood ratios in the PAL formula.

Acknowledgements

Funding for this research was provided by grants of COLCIENCIAS (Departamento Administrativo de Ciencia, Tecnología e Innovación, Francisco José de Caldas fellowship 497/2009, Colombia). and Universidad de Antioquia (CODI Sostenibilidad 2014 E01808) from Colombia. and CONICET (PIP 833/2013), both grants from Argentina. The authors would like to thank Asociación Holstein de Colombia for providing the data for the study.

References

Akaike H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Bignardi A.B., El Faro L., Cardoso V.L., Machado P.F., Albuquerque L.G. (2009) Random regression models to estimate test-day milk yield genetic parameters Holstein cows in Southeastern Brazil. *Livest. Sci.*, **123**, 1–7.

Burnham K.P., Anderson D.R. (2002) Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn.. Springer Verlag, New York.

Burnham K.P., Anderson D.R. (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.*, **33**, 261–304.

Burnham K.P., Anderson D.R., Huyvaert K.P. (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav. Ecol. Sociobiol.*, **65**, 23–35.

Carabaño M.J., Díaz C., Ugarte C., Serrano M. (2007) Exploring the use of random regression models with legendre polynomials to analyze measures of volume of ejaculate in Holstein bulls. *J. Dairy Sci.*, **90**, 1044–1057.

Casella G., Berger R. (2002) Statistical Inference, 2nd edn. Duxbury, Thomson Learning, CA.

Davidson R., Mackinnon J.G. (2004) Econometric Theory and Methods. Oxford University Press, New York.

Gilmour A.R., Thompson R., Cullis B.R. (1995) Linear mixed models algorithm for average information REML: an efficient in linear mixed models variance parameter estimation. *Biometrics*, **51**, 1440–1450.

Herrera A.C., Munera O.D., Cerón-Muñoz M.F. (2013) Variance components and genetic parameters for milk production of Holstein cattle in Antioquia (Colombia) using random regression models. *Rev. Col. Cienc. Pecu.*, **26**, 90–97.

Jakobsen J.H., Madsen P., Jensen J., Pedersen J., Christensen L.G., Sorensen D. (2002) Genetic parameters for milk production and persistency for Danish Holsteins estimated in random regression models using REML. *J. Dairy Sci.*, **85**, 1607–1616.

Jamrozik J., Gianola D., Schaeffer L.R. (2001) Bayesian estimation of genetic parameters for test day records in dairy cattle using linear hierarchical models. *Livest. Prod. Sci.*, **71**, 223–240.

Liu Y.X., Zhang J., Schaeffer L.R., Yang R.Q., Zhang W.L. (2006) Short communication: Optimal random regression models for milk production in dairy cattle. *J. Dairy Sci.*, **89**, 2233–2235.

López-Romero P., Carabano M. (2003) Comparing alternative random regression models to analyse first lactation daily milk yield data in Holstein – Friesian cattle. *Livest. Prod. Sci.*, **82**, 81–96.

López-Romero P., Rekaya R., Carabaño M. (2003) Assessment of homogeneity vs. heterogeneity of residual variance in random regression test-day models in a Bayesian analysis. *J. Dairy Sci.*, **86**, 3374–3385.

McQuarrie A., Shumway R., Tsai C. (1997) The model selection criterion AICu. *Stat. Probab. Lett.*, **34**, 285–292.

Meyer K. (2007) WOMBAT – A tool for mixed model analyses in quantitative genetics by REML. *J. Zhejiang Univ. Sci. B*, **8**, 815–821.

Meyer K., Hill W.G. (1997) Estimation of genetic and phenotypic covariance functions for longitudinal or "repeated" records by restricted maximum likelihood. *Livest. Prod. Sci.*, **47**, 185–200.

Müller S., Scealy J.L., Welsh A.H. (2013) Model selection in linear mixed models. *Stat. Sci.*, **28**, 135–167.

Odegård J., Jensen J., Klemetsdal G., Madsen P., Heringstad B. (2003) Genetic analysis of somatic cell score in Norwegian cattle using random regression test-day models. *J. Dairy Sci.*, **86**, 4103–4114.

Pool M.H., Meuwissen T.H.E. (2000) Reduction of the number of parameters needed for a polynomial random regression test day model. *Livest. Prod. Sci.*, **64**, 133–145.

Pool M.H., Janss L.L., Meuwissen T.H. (2000) Genetic parameters of legendre polynomials for first parity lactation curves. *J. Dairy Sci.*, **83**, 2640–2649.

Schaeffer L.R. (2004) Application of random regression models in animal breeding. *Livest. Prod. Sci.*, **86**, 35–45.

Schwarz G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Searle S. (1982) Matrix Algebra Useful for Statistics. John Wiley & Sons, New York.

Shibata R. (1981) An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

Stoica P., Babu P. (2013) Model order estimation via penalizing adaptively the likelihood (PAL). *Signal Process.*, **93**, 2865–2871.

Strabel T., Szyda J., Ptak E., Jamrozik J. (2005) Comparison of random regression test-day models for Polish Black and White cattle. *J. Dairy Sci.*, **88**, 3688–3699.

Van Der Werf J.H.J., Goddard M.E., Meyer K. (1998) The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. *J. Dairy Sci.*, **81**, 3300–3308.

Yang Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, **92**, 937–950.

Appendix 1 A small example showing how to calculate PAL

Consider the following simulated data to illustrate the calculation of PAL. The true model was LP33; however, data from three other models are included in the table below. The third column includes the logarithm of the likelihood function, as can be obtained from expression (2).

LPm_1 m_2	Number of parameters	ln L	ω_{PAL}	PAL
1) LP11	8	−97 864		
2) LP22	12	−93 755	0	187 510
3) LP33	18	−92 247	3.6656	184 560
4) LP44	26	−92 226	8.0792	184 662

The calculus of ω_{PAL} in column four was performed from expression (4), which is transcribed here for convenience:

$$\omega_{PAL} = \ln(\tilde{n})\left(\frac{\ln(r_n+1)}{\ln(\rho_n+1)}\right).$$

The value $\tilde{n}$ is the number of parameters in the model (within the set) with the highest number, and here is equal to $\tilde{n} = 26$ for LP44. Then, to obtain $\omega_{\text{PAL-LP33}}$, we first calculate the values of r_{LP33} and ρ_{LP33} from (5) as follows:

$$\begin{aligned} r_{LP33} &= 2\ln L_{\text{LP22}} - 2\ln L_{\text{LP11}} \\ &= 2\times(-93\,755) - 2\times(-97\,864) = 8218; \end{aligned}$$

$$\begin{aligned} \rho_{LP33} &= 2\ln L_{\text{LP44}} - 2\ln L_{\text{LP22}} \\ &= 2\times(-92\,226) - 2\times(-93\,755) = 3016. \end{aligned}$$

So that

$$\omega_{\text{PAL-LP33}} = \ln(26)\frac{\ln(8218+1)}{\ln(3016+1)} = 3.6656$$

As a result, the PAL criterion for model order selection is obtained from (6) as

$$\text{PAL(LP33)} = -2\ln L_{\text{LP33}} + 18\omega_{\text{PAL-LP33}}.$$

Finally,

$$\begin{aligned} \text{PAL(LP33)} &= -2\times(-92\,247) + 18\times 3.6656 \\ &= 184\,560. \end{aligned}$$

The calculation of PAL for models LP44 and LP22 is performed in a similar fashion.

