Use of linear splines to simplify longitudinal analyses

I. Misztal[1], J. Bohmanova[1], S. Tsuruta[1] and H. Iwaisaki[2], [1]University of Georgia, Athens, 30605, USA, [2]Niigata University, Niigata-shi 950-2181, Japan

This study investigated the use of linear splines as alternatives to polynomials in random regression models. With linear splines, parameters for all effects except permanent environment and residual can be the same as in multiple trait models, simplifying validation and computations. Also, artifacts at boundaries are less likely. One comparison involved simulated data in beef cattle involving weights at days 1, 205±45 and 365±50. Models included were multiple trait, random regression with cubic Legendre polynomials, and random regression with linear splines and 3 knots. Variance components in the three models were equivalent at days 1, 205 and 365. The multiple trait model was the least accurate because it did not account for variability in days for random effects. Both random regression models gave nearly identical results, but the model with splines was simpler and converged much faster. In another comparison involving field data on beef cattle, variance components for a similar model were estimated by a multiple trait and by a random regression model with linear splines. Large percentage of records for birth and weaning weights were missing. The model with splines gave more realistic estimates of heritabilities and correlations. Random regression models with linear splines are simpler and safer alternatives than models with polynomials.

Introduction

Many analyses by random regression analyses use Legendre polynomials. These polynomials are able to model a variety of curves for variances and covariances, and they have better numerical properties than regular polynomials. However, they also have many undesirable properties. Fit at the extremes of the trajectory may be poor. Curves at points of the trajectory with few records may contain artifacts. When parameters are estimated with these polynomials conversions are necessary to find out whether they are realistic or not. Finally, there is problem with numerical stability because, even though the polynomials are orthogonal on a uniform scale, they are far from orthogonal with real-data distributions. Stability may be improved by reparameterization to diagonal variances, however, this increase the complexity. Rank reduction (i.e., elimination of dimensions with very small eigenvalues) that is usually performed with the diagonalization may result in large changes for some points on the trajectory; especially there are large changes in variances along the trajectory. Successful estimation of variances with Legendre polynomials requires large data sets, even distribution of data points on the trajectory and careful modeling of other effects (Druet et al., 2003).

Several alternatives exist to Legendre polynomials. White et al. (1999) used cubic polynomials and Torres and Quaas (2001) used B-splines with 10 knots. One coefficient of splines affects only parts of the trajectory resulting in possibly better numerical properties and fewer estimation artifacts. Foulley and Robert-Granié (2002) advocated the use of fractional polynomials where also roots and logs are implemented. Subsequently, the properties at the extremes can be improved and some artifacts eliminated. However, both approaches still result in cryptic parameters.

Recently, several studies at the University of Georgia look at the application of linear splines. When knots are at points corresponding to a multiple trait model (MTM), the (co)variances of splines and multiple trait models are the same for all effects other than

permanent environment and the residuals. This greatly simplifies preparation of parameter files as in many cases the literature and common-sense information may be sufficient.

The purpose of this paper is to present properties and then report applications of the linear splines in random regression models (RRMS).

Materials and Methods

Random Regression Models (RRMs)

A random regression model for growth in beef can be defined as:

$$y_{ijkt} = \sum_{m=0}^{1} CG_{im}(a_t)^m + \sum_{l=0}^{n} d_{jl}\Phi_l(a_t) + \sum_{l=0}^{n} p_{jl}\Phi_l(a_t) + \sum_{l=0}^{n} m_{kl}\Phi_l(a_t) + \sum_{l=0}^{n} mp_{kl}\Phi_l(a_t) + e_{ijkt}$$

where y_{ijkt} is t^{th} observation of animal j of dam k, CG_{mi} m^{th} fixed regression coefficient of contemporary group i; d_{jl} and p_{jl} are l^{th} random regression coefficients of direct genetic and permanent environmental effects of animal j; m_{kl} and mp_{kl} are l^{th} random regression coefficients of maternal genetic and permanent environmental effects of dam k; and e_{ijkt} is the random measurement error; $\Phi(a_t)$ represents a vector of covariables at age a_t. For linear splines, a vector of spline coefficients (Φ) at age t (a_t) for knots q_1, q_2 and q_3 can be defined as:

for $q_1 < a_t \le q_2$ $\Phi(a_t) = [(1-\alpha)\ \alpha\ 0]$ where $\alpha = \frac{a_t - q_1}{q_2 - q_1}$,

for $q_2 < a_t \le q_3$ $\Phi(a_t) = [0\ (1-\alpha)\ \alpha]$ where $\alpha = \frac{a_t - q_2}{q_3 - q_2}$, and

for $a_t > q_3$ $\Phi(a_t) = [0\ 0\ \alpha]$ where $\alpha = \frac{a_t}{q_3}$.

The choice of linear splines was due to two factors. First, each spline coefficient has localized effects (Green and Silverman, 1994; Wold, 1974) and thus would result in fewer artifacts than polynomials. Second, parameters for models with linear splines are very easy to derive from parameters of MTM. This can be illustrated by listing the RRMS for the direct effect only:

$$y_{ijkt} = \ldots + d_{j1}\Phi_1(a_t) + d_{j2}\Phi_2(a_t) + d_{j3}\Phi_3(a_t) + \ldots$$

The spline coefficients and the model for specific weights corresponding to standard weights in MTM are:

Birth weight: $\Phi_1(1) = 1;\ \Phi_2(1) = 0;\ \Phi_3(1) = 0;$ $y_{ijkt} = \ldots + d_{j1} + \ldots$

Weaning weight: $\Phi_1(205) = 0;\ \Phi_2(205) = 1;\ \Phi_3(205) = 0;$ $y_{ijkt} = \ldots + d_{j2} + \ldots$

Yearling weight: $\Phi_1(365) = 0;\ \Phi_2(365) = 0;\ \Phi_3(365) = 1;$ $y_{ijkt} = \ldots + d_{j3} + \ldots$

Thus, the direct effects in RRMS for standard weights are the same as in MTM, and subsequently the variances are identical. Generalizing, when the knots in RRMS correspond to traits in MTM, the variances in the corresponding effects except the residual are the same. However, the residual effect in MTM is split into the permanent environment plus the residual effect.

Data

Three data sets were used for comparisons. The first data set was simulated using covariances matrices as constructed by Legarra et al. (2004) and transformed to RRM with cubic Legendre polynomials. The simulation involved a total of 29,400 animals in three generations. Four data sets were simulated. The first data set (3EXACT) consisted of three records per animal at exactly 1, 205 and 365 days of age. With this data set, properly designed RRM should be in perfect agreement with MTM. The second data set (3SPREAD) contained three records per

animal. These records were located in 45 days interval around 1d, 205d and 365 days of age. The distribution of the spread in this and later cases was uniform. With the second data set, RRM would be expected to maintain accuracy because it accounts for changes in variances while accuracy of MTM would be expected to be lower. The third (5EXACT) data set was formed by adding records at 100 and 300 days of age to the first data set. The fourth data set (5SPREAD) was created by including two extra records in 45 days interval around 100 days and 25 days interval around 300 days of age to the second data set. With extra records, RRM was expected to be more accurate than MTM, and that accuracy should be similar for the both data sets. Comparisons involved 3 models: RRM with cubic Legendre polynomials (RRML), RRMS and MTM (for 3 traits only). The first data set was used to compare accuracies obtained with all the models for all the data sets, and to compare computing costs. Computing costs for RRMs included the original model and models after diagonalization.

The second data set contained about 540,000 Gelbvieh animals, of which 90% had weaning weights, 80% had birth weights and 30% had yearling weights. Comparisons involved the same models as above and were used to obtained correlations among EPDs obtained from those models as well as computing times.

The third data contained a subset of the data above, with weight records on 18,900 Gelbviehs, of which 100, 75 and 17% had birth (BWT), weaning (WWT) and yearling (YWT) weights, respectively. This data set was used for comparing estimates of parameters obtained with RRMS and MTM.

Results and Discussion

Table 1 presents accuracies computed as correlation between the true (simulated) and predicted breeding values. The accuracies for all the three methods using 3EXACT were the same. The accuracies with 3SPREAD were essentially the same for RRML and RRMS but lower for MTM, as expected. With 5EXACT and 5SPREAD, the accuracies of RRML and RRMS increased, also as expected. However, the increase in RRMS was slightly smaller than in RRML indicating differences between these methods.

Table 1: Accuracies (%) of breeding values in multiple trait model (MTM), random regression model with Legendre polynomials (RRML) and random regression model with splines (RRMS), and four datasets

	3EXACT			3SPREAD			5EXACT		5SPREAD	
Age	**MTM**	**RRML**	**RRMS**	**MTM**	**RRML**	**RRMS**	**RRML**	**RRMS**	**RRML**	**RRMS**
1	56.6	56.6	56.6	55.1	56.6	56.6	56.9	56.9	57.0	56.9
205	53.5	53.5	53.5	52.0	53.5	53.5	55.9	55.8	56.0	55.9
365	52.8	52.8	52.8	51.4	52.8	52.8	53.9	53.7	54.0	53.8

(Co)variances of RRML and RRMS are equivalent at standard days but are different at other days. Figure 1 shows the direct variance as the function of age for MTM, RRML and RRMS. The variance for RRMS is concave in between the knots; the concavity increases with decrease of genetic correlation between the adjacent knots. Figure 2 shows genetic correlations for the direct effect between birth weight and other days. The correlations with RRMS are

inflated especially around 100 days of age. The inflated correlation resulted in too large contribution of records especially around 100 d to prediction of birth weight.

Figure 1. Direct genetic variance of random regression models with Legendre polynomials (RRML), splines with knots located at 1, 205 and 365 d (RRMS 3 knots), splines with knots located at 1, 100, 205 and 365 d (RRMS 4 knots), and multitrait model (MTM)

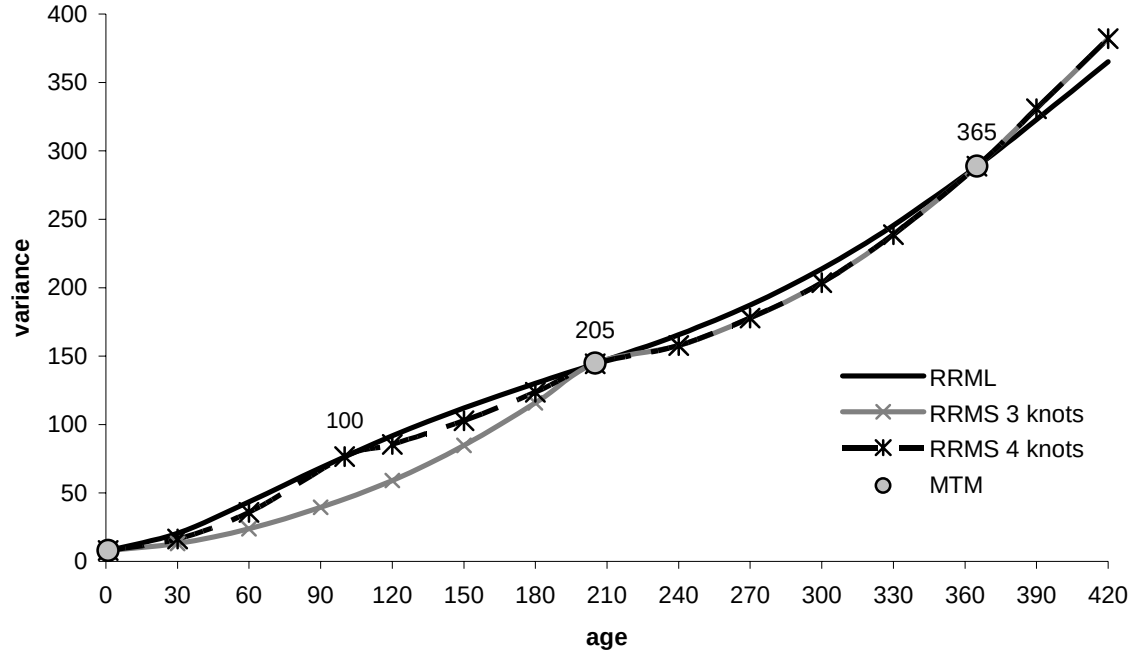
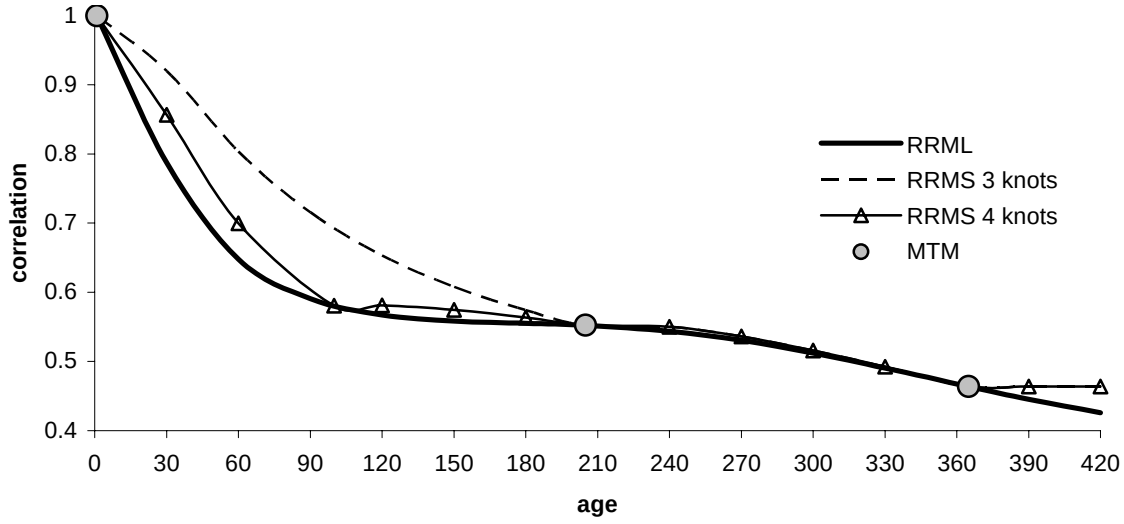


Figure 2. Direct genetic correlations between weight at birth and other ages in RRML and RRMS with 3 and 4 knots and MTM



Figures 1 and 2 also contain graphs of RRMS obtained when an extra knot was added at 100 d. In this case, the variances of RRML and RRMS are very similar. It is worth noting that

despite the differences in variances, the accuracies of RRML and RRMS were very similar and higher than those with MTM. This agrees with the opinion of C. R. Henderson indicating that mixed model equations are robust with respect to slightly inaccurate parameters. A selection of number and location of knots will be a topic of a separate study.

The rank of RRM was reduced by dropping random regression coefficients with eigenvalues that explained less then 1 % of variance. Although the computation costs were reduced, this affected accuracy. Correlations between the rank reduced and non-reduced RRML predictions were 0.89, 0.91 and 0.95 for the direct genetic effect and 0.80, 0.98 and 0.97 for the maternal genetic effect at 1, 205 and 365 days of age, respectively. This was because even though the eigenvalue corresponding to the eliminated direct genetic variance component accounted for only 0.102 % of the total variance, it explained a large portion of the variance at birth (Figure 3). The elimination of this eigenvalue resulted in decrease of direct genetic variance at birth by 5.13 (65.2 %). The change in variance due to the rank reduction was close to zero after 150 days of age. Similarly, reduction of the maternal effect caused decrease of variance at early ages and almost no change at late ages. Foulley and Robert-Granié (2002) mentioned that the rank of RRM should be reduced with caution.

Figure 3: Differences in direct and maternal genetic variance between original and rank reduced RRML

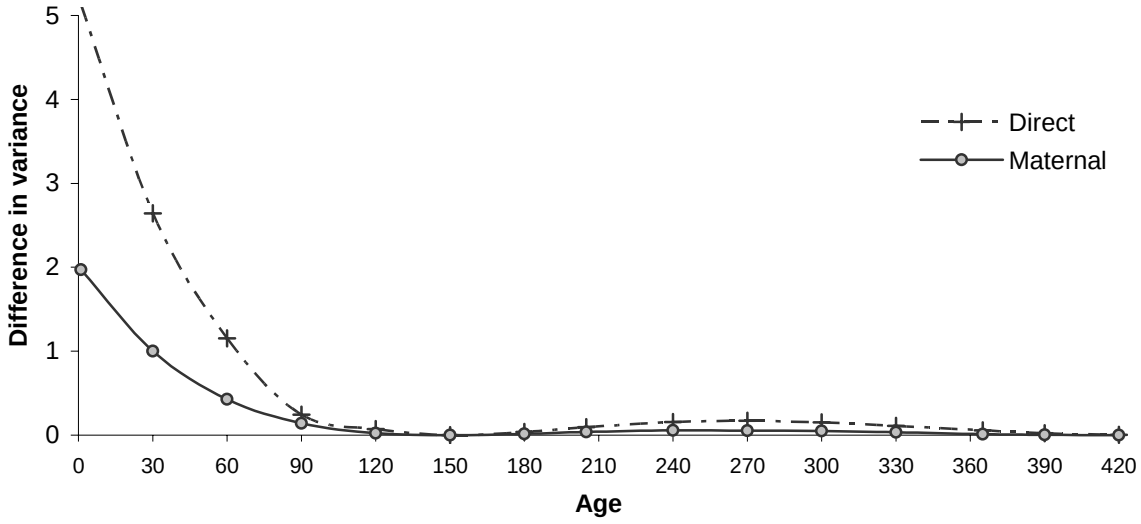


Table 2 shows the number of rounds and computing time for the original, diagonalized and reduced models. Solution method was preconditioned conjugate gradient with a diagonal preconditioner. After diagonalization, the number of rounds required to converge decreased from 571 to 101 in RRML, and from 184 to 81 in RRMS. The RRMs with splines were the fastest due not only to showing higher convergence but also to having only three covariables per effect rather than four in RRML.

Table 2: Computing cost of models with the 3EXACT data set[a]

	MTM		RRML		RRMS	
	Rounds	**Time**	**Rounds**	**Time**	**Rounds**	**Time**
Original	288	2m31s	571	20m40s	184	3m40s
Diagonalization			101	4m4s	81	1m47s

[a] 2.8 GHz processor

Table 3 presents the number of round till convergence obtained with the national Gelbvieh data set. The model with Legendre polynomials did not converge until the diagonalization was implemented.

Table 3: Number of rounds until convergence with the field data set[a]

	MTM	**RRML**	**RRMS**
Original	678	> 2000	254
Diagonalization	-	274	-

[a] 2.8 GHz processor

Table 4 present estimates of variance components for a subset of the Gelbvieh data set obtained with MTM and RRMS. It is clear that the estimates of RRMS are at the same scale as MTM, and that both sets are similar. The estimates of genetic correlations between the direct and maternal effects seem inflated in MTM but are lower in RRMS. This could be due to ability of RRMS to account for changes in variances for records with a spread while the preadjustment in MTM is just for the fixed effects

Conclusions

Random regression model using linear splines offers several advantages over models using other functions. First, its parameters are on the scale of multiple trait models and thus are easy to create and to evaluate. Second, this model is numerically more stable than RRML. One issue is determining the number of knots. One rule is to have such knots so that the correlations among adjacent point are high but not too high, e.g., 0.5 to 0.8. Too low correlations diminish the modeling capacity, and too high correlations result in too many knots and subsequently more numerical problems. However, small imperfections in modeling variances with RRMS seem to reduce its accuracy very little.

Table 4. Estimates and their posterior standard deviations of (co)variance components for direct additive genetic, maternal additive genetic, maternal permanent environmental and residual effects using multi-trait model and random regression model with a linear spline function

Trait[a]	Multi-trait model			Random regression model		
	BWT	WWT	YWT	BWT	WWT	YWT
			Direct additive genetic variance			
BWT	8.1±1.0			8.1±8		
WWT	18.1±4.0	195.4±37.6		20.8±4.0	225.5±29.5	
YWT	40.5±5.9	155.4±60.7	754.9±17.0	41.4±8.9	260.0±42.3	855.1±178.8
			Maternal additive genetic variance			
BWT	1.0±0.3			1.2±0.4		
WWT	-2.4±1.4	76.3±17.8		-3.5±1.6	79.7±18.6	
YWT	-6.6±2.8	-22.4±23.3	148.4±48.6	-2.3±3.5	5.7±17.7	185.1±84.7
			Maternal permanent environmental variance			
BWT	1.4±0.3			1.4±0.2		
WWT	5.3±1.6	96.4±14.0		6.4±1.1	78.2±13.8	
YWT	8.1±3.3	57.9±22.7	111.5±48.1	6.2±2.7	76.6±18.8	146.7±35.6
			Residual variance[b]			
BWT	7.7±0.6			7.7±0.5		
WWT	10.8±2.5	407.4±22.3		9.4±2.4	362.9±18.6	
YWT	10.0±4.0	343.1±38.1	859.7±81.2	9.2±5.8	283.3±29.3	726.2±107.7

[a] BWT: birth weight, WWT: weaning weight, YWT: yearling weight.
[b] Estimates for random regression model are sums of estimated direct permanent environmental and residual variances.

Literature Cited

Druet, T., F. Jaffrézic, D. Boichard, and V. Ducrocq. 2003. Modeling lactation curves and estimation of genetic parameters for first lactation test-day records of French Holstein cows. J. Dairy Sci. 86: 2480-2490.

Foulley, J.L., and C. Robert-Granié. 2002. Basic statistical methods for longitudinal data. Montpellier, France.

Green, P. J., and B. W. Silverman. 1994. Nonparametric regression and generalized linear models. Chapman & Hall, London

Legarra, A., I. Misztal and J.K. Bertrand. 2004. Constructing covariance functions for random regression models for growth in Gelbvieh beef cattle. J. Anim. Sci. 82:1564-1571.

Torres, R. A., and R. L. Quaas. 2001. Determination of covariance functions for lactation traits on dairy cattle using random-coefficient regressions on B-splines. Int. Anim. Agric. Food Sci. Abstracts. 112.

White, I. M., R. Thompson, S. Brotherstone, 1999. Genetic and environmental smoothing of lactation curves with cubic splines. J. Dairy Sci. 82: 632-638.

Wold, S. 1974. Spline functions in data analysis. Technometrics. 16: 1-11.