Selection of locations of knots for linear splines in random regression test-day models

J. Jamrozik, J. Bohmanova & L.R. Schaeffer

Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, ON, Canada

Introduction

Functions with random coefficients in longitudinal models describe changes in covariances of the trait along a trajectory. Linear splines are one of many families of functions that can be used as random regressions in models for test-day (TD) data in dairy cattle (Bohmanova et al. 2008). These are piecewise functions consisting of independent linear segments, connected in selected points called ‘knots’ (De Boor 1978). Coefficients of linear splines are simple interpolation coefficients between the two knots flanking the record, and zeros between all other knots. Because at most only two coefficients are non-zero for a given observation, the resulting system of equations for the linear model with splines is sparser than that for a model with orthogonal polynomials or parametric functions (Misztal 2006). Other interesting properties of linear splines include their limited sensitivity to the data, because individual observations influence only a specific part of the function. This may result in a higher flexibility of fitting both average trajectories of the trait and its respective covariance structure. Each coefficient of

Keywords
Knots; linear splines; optimization algorithms; random regressions; test-day model.

Correspondence
J. Jamrozik, Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, ON, N1G 2W1, Canada.
Tel: +1 519 824 4120; Fax: +1 519 836 9873;
E-mail: jjamrozi@uoguelph.ca

Received: 26 January 2009; accepted: 20 April 2009

Summary

Using spline functions (segmented polynomials) in regression models requires the knowledge of the location of the knots. Knots are the points at which independent linear segments are connected. Optimal positions of knots for linear splines of different orders were determined in this study for different scenarios, using existing estimates of covariance functions and an optimization algorithm. The traits considered were test-day milk, fat and protein yields, and somatic cell score (SCS) in the first three lactations of Canadian Holsteins. Two ranges of days in milk (from 5 to 305 and from 5 to 365) were taken into account. In addition, four different populations of Holstein cows, from Australia, Canada, Italy and New Zealand, were examined with respect to first lactation (305 days) milk only. The estimates of genetic and permanent environmental covariance functions were based on single- and multiple-trait test-day models, with Legendre polynomials of order 4 as random regressions. A differential evolution algorithm was applied to find the best location of knots for splines of orders 4 to 7 and the criterion for optimization was the goodness-of-fit of the spline covariance function. Results indicated that the optimal position of knots for linear splines differed between genetic and permanent environmental effects, as well as between traits and lactations. Different populations also exhibited different patterns of optimal knot locations. With linear splines, different positions of knots should therefore be used for different effects and traits in random regression test-day models when analysing milk production traits.
the spline function affects only a part of the trajectory that could also give better numerical properties and a smaller number of estimation artefacts.

Bohmanova et al. (2008) compared random regression TD models with Legendre polynomials of order 4, with alternative models fitting linear splines with 4, 5 or 6 knots. The deviance information criterion favoured the spline model with 6 knots but overall differences between models, measured by residual variances, percentage of squared bias and correlations between predicted and observed records, were small. Models with splines gave smaller estimates of variances at extremes of lactation. Splines with 5 and 6 knots gave smaller error of prediction and a better stability of estimated breeding values than the model with Legendre polynomials.

One of the requirements for fitting linear splines is the selection of the number of knots and their positions along the trajectory. The number of knots corresponds to the number of linear segments. More knots would mean a smoother covariance surface, but at the cost of more parameters to estimate. Location of knots should correspond to the pattern of changes along the trajectory. Rapid changes require a larger number of knots in a given region whereas small changes would need a sparser distribution of knots. Unfortunately, no deterministic algorithm is known for optimal distribution of knots along a trajectory. Misztal (2006) described a heuristic approach for selection of knot positions when the data points on the trajectory are equally distributed. In the scenario when an accurate covariance function is known, he advocated that the extreme knots should bracket all the (time) points in the data. Additional knots should be placed so that correlations between adjacent knots would be in the range from 0.6 to 0.8. Two strategies are proposed when the parameters are to be estimated: forward or backward selection of knots with the criterion for adding (removing) knots based on the value of correlations between neighbouring knots.

Optimal number and placement of knots may be different for different random effects in the random regression model, as well as for different populations and traits. The objective of this research was to propose a method for the selection of knots for linear splines of a given order for an application in a random regression TD model under the assumption of a known covariance function. Additional objectives included comparisons of knot locations for different effects, traits and populations.

Material and Methods
Covariance function for linear splines
Let \( T = [T_1, \ldots, T_N] \) be a vector of \( N \) knots on a time scale (\( t \)). Linear splines are constructed by fitting a first-degree polynomial between each pair of knots. Covariables (\( z \)) of linear splines for a point \( t \) located between knots \( T_i \) and \( T_{i+1} \) can be described as

\[ z_i(t) = (t - T_i)/(T_{i+1} - T_i), \]

\[ z_{i+1}(t) = (T_{i+1} - t)/(T_{i+1} - T_i), \]

and

\[ z_j(t) = 0 \quad \text{for } j = 1, \ldots, i-1,i+2,\ldots,N. \]

For records located at a knot \( i \), covariables are defined as \( z_i(t) = 1 \) and \( z_j(t) = 0 \) for \( j = 1, \ldots, i-1, i+1, \ldots, N \). When \( t < T_i \) then \( z_i(t) = t/T_i \) and and \( z_j(t) = 0 \) for \( i \neq 1 \). Similarly, when \( t > T_N \) then \( z_N(t) = T_N/t \) and \( z_j(t) = 0 \) for \( i \neq N \). Vectors of covariables (\( z \)) have thus at most two non-zero elements. Let \( y_{it} \) be the observation of an animal taken at time \( t \) corresponding to the \( i \)th subclass of fixed effects. The random regression model can be presented in a simplified form as

\[ y_{it} = \text{Fixed}_i + a'z(t) + p'z(t) + e_{it}, \]

where Fixed, are the fixed effects of subclass \( i \), \( a \) and \( p \) are vectors of genetic and permanent environmental (PE) effects of an animal, respectively, \( z(t) \) is the vector of \( N \) covariables at time \( t \), and \( e_{it} \) is the error term. Covariance matrices for regression coefficients can be denoted as \( \text{var}(a) = G^a = [g_{ij}^a] \) and \( \text{var}(p) = G^p = [g_{ij}^p] \) for genetic and environmental effects, respectively.

To describe covariances along the trajectory it is sufficient, following Misztal (2006), to consider only two knots. Let \( T_1 = 0 \leq t \leq 1 = T_2 \) and let \( G = [g_{ij}] \) be a known \( N \times N \) covariance matrix for a given trait at the knots. The value of one level of an effect (e.g. one animal additive genetic effect) at point \( t \) is \( \alpha(t) = a'z(t) = (1-t)a_1 + ta_2 \), where \( a \) is a vector of \( N \) regression coefficients corresponding to this particular effect. Then \( \text{var}[\alpha(0)] = g_{11} = \text{var}(a_1), \text{var}[\alpha(1)] = g_{22} = \text{var}(a_2) \), and \( \text{cov}[\alpha(0), \alpha(1)] = g_{12} = \text{cov}(a_1, a_2) \). This means that the covariance structure between the linear spline coefficients can be derived in a simple way from a known covariance matrix for a trait at the knots (\( g_{ij} = g_{ij}^{\alpha} \)). Spline covariance function for any two points \( t_1 \) and \( t_2 \) can be presented as \( \Phi(t_1,t_2) = z(t_1)'Gz(t_2). \)
Selection algorithm

Differential evolution method (Price & Storn 1997) was used to obtain the best location of knots for linear splines of a given number of knots. This method is a multivariate optimization procedure from the general class of evolutionary algorithms. Assume $\mathbf{V}$ to be a known covariance matrix for a trait in a discrete time interval from $t_1$ to $t_M$ and let $N (N < M)$ be the number of knots of linear splines under consideration. Let $\mathbf{C}$ denote a correlation matrix derived from $\mathbf{V}$. The algorithms can be briefly described as follows:

1. Set up a starting population of $p$ individuals (solutions), each one with a genome of $M$ loci corresponding to number of points on the longitudinal scale ($t$). Assign genetic value for each locus as a random number from an assumed normal distribution with a known mean and variance.
2. Rank loci of each individual according to their genetic values and select the top $N$ loci for each. Positions of those loci on the genome would correspond to the knot locations on the time scale from $t_1$ to $t_M$.
3. Derive covariance matrix ($\mathbf{CO}$) between selected loci using the $\mathbf{V}$ matrix (as described in the previous paragraph) and a covariance matrix for the whole trajectory, for each individual.
4. Calculate an optimization criterion for each individual, defined as a sum-of-squared differences between elements of $\mathbf{C}$ and elements of a correlation matrix derived from $\mathbf{CO}$.
5. Generate a challenger for each individual as a progeny of current individuals using basic genetic concepts of mutation and recombination.
6. Perform steps in 2–4 for each challenger. If the value of optimization criterion for a challenger is better than for a current member of the population then replace a current member of the population by its challenger.
7. When the number of generations remains smaller than a predefined value, go to step 5.

This is an iterative process. After a finite and sufficiently large number of generations, an optimal (with respect to the optimization criterion) solution will be found.

Implementation

Covariance functions for milk, fat and protein yield and somatic cell score (SCS) in the first three lactations of Canadian Holsteins were derived from the estimates of genetic and PE covariance components of the multiple-trait random regression model with Legendre polynomials of order 4 (Bohmanova et al. 2002). Estimates for two scenarios, with two different ranges of days in milk (DIM), from 5 to 365 and from 5 to 305, were considered for each trait–lactation combination. In addition, genetic and PE covariance functions for DIM from 5 to 305 for milk yield in the first lactation of Australian (AUS), Canadian (CAN), Italian (ITA) and New Zealand (NZL) Holstein cows were derived using single-trait model estimates of genetic and PE covariance components presented by Jamrozik et al. (2002). The estimation model included Legendre polynomials of order 4 for random regressions.

For each covariance function (i.e. population–effect–lactation–trait–time scale combination) a differential evolution algorithm was run to determine optimal locations of knots for linear splines with the number of knots from 4 to 7. Population size for each optimization was set to 10 and the differential evolution was run for 2000 generations. The population size of 10 is typical for applications of the differential evolution algorithm. Larger values of $p$ require more computing time per generation. Smaller populations may have more problems with finding the optimum because of the reduced variation in the population. We tested several other values for population size and 10 gave the best overall performance of the algorithm. Genetic values at each locus were generated from a normal distribution with mean equal to $-100$ and variance equal to 200. The genetics terms that we are using here (i.e. locus, population, generation) should be considered only as a genetic algorithm’s nomenclature and with no particular references to the actual animal breeding and genetics context.

Results

Optimal locations of spline knots will be presented for selected population–random effect–lactation–trait combinations, illustrating the general patterns in distribution of knots for different scenarios. Goodness-of-fit of the spline covariance function (SSD) was calculated for one of the triangles of the symmetric correlation matrices. This allowed comparison of this characteristic between different traits.

Optimal locations of knots and goodness-of-fit for genetic and PE covariance functions for milk, fat, protein and SCS on DIM from 5 to 365 in the first lactation of Canadian Holsteins are presented in Table 1. Goodness-of-fit for spline covariance functions improved with the number of knots for all effects and traits. PE effect for milk, fat and protein
yields was characterized with substantially larger values of the goodness-of-fit criterion compared with the genetic effect. Similar trends were shown by Pool et al. (2000) when orthogonal polynomials were used as random regressions. Splines with larger number of knots for the PE effect should be used to obtain the same accuracy of fit for both genetic and PE covariance functions. Milk yield had the best goodness-of-fit followed by protein yield. SCS was the trait with the poorest goodness-of-fit. The first (last) knots were located very close to the beginning (end) of the time scale although not necessary at 5 or 365 DIM. Distributions of knots were not symmetric, in general, at least for splines with smaller numbers of knots. The first two knots corresponded to DIM < 100 which was the part of the time scale with more rapid changes of covariances. Differences in goodness-of-fit for covariance functions decreased with increasing number of knots.

Location of knot positions differed between traits. Sum of absolute differences between corresponding knot locations (DIST) was used to quantify those differences. Milk and protein yield had the most similar locations of knots (average DIST over different orders of splines = 6.5), values of DIST between milk and fat and milk and SCS were 13.6 and 12.0, respectively. Locations of knots for the PE covariance functions were different than those corresponding to the genetic effects. The average value of DIST between genetic and PE effects for milk yield was equal to 17.6, distances between knots for genetic effect of milk yield and PE effect for other traits were of a similar order.

Table 2 gives positions of knots and goodness-of-fit for milk yield on DIM from 5 to 365 in the first three lactations of Canadian Holsteins. The general trend of both goodness-of-fit and locations of knots between lactations followed the trend described for changes in goodness-of-fit for genetic and PE covariance functions for milk yield on DIM from 5 to 365 in the first lactation of Canadian Holsteins, by number of knots.

<table>
<thead>
<tr>
<th>Covariance function</th>
<th>Lactation</th>
<th>Number of knots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic</td>
<td>First</td>
<td>6,74,275,362</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>7,80,254,354</td>
</tr>
<tr>
<td></td>
<td>Third</td>
<td>6,84,257,365</td>
</tr>
<tr>
<td>PE</td>
<td>First</td>
<td>6,74,192,361</td>
</tr>
<tr>
<td></td>
<td>Second</td>
<td>23,87,196,347</td>
</tr>
<tr>
<td></td>
<td>Third</td>
<td>5,153,293,357</td>
</tr>
<tr>
<td>SSD</td>
<td></td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23.4</td>
</tr>
</tbody>
</table>
of those parameters between traits for the first lactation. In addition, goodness-of-fit decreased with lactation number for both effects.

Table 3 compares locations of knots for genetic and PE covariance functions for milk yield of Canadian Holsteins for different lactation lengths (5–305 versus 5–305). The longer time interval resulted obviously in a poorer goodness-of-fit. Distribution of knots varied between different time scales with no systematic patterns of changes.

Different populations exhibited different patterns of optimal knot locations and goodness-of-fit for covariance functions (Table 4). Knot locations from the single-trait covariance function for the Canadian population differed from the results based on the multiple-trait estimates (Table 1). This can be attributed to differences in the shape of Legendre polynomials covariance from single- and multiple-trait models. CAN and ITA exhibited similar locations of knots for splines with more than 4 knots and results for AUS and NZL followed similar trends.

Discussion

Optimal locations of knots for linear splines were not the same for different populations, random effects, lactations and traits. Thus, different spline functions should be used for optimal modelling of covariance functions for different traits. Only random regressions describing the changes of covariances along the trajectory were taken into account in this research. The best spline function for averages across time (i.e. fixed regressions) would likely require spline functions of yet another order and locations of knots. Sensitivity of random regression genetic evaluation models with linear splines for changes in knot locations should be examined further.

Only splines with number of knots between 4 and 7 were considered in this study. Linear splines with <4 segments are of no practical use in random regression TD models because they will provide insufficient fit to the covariance structure. On the
other hand, more than 7 regression coefficients per individual will cause computation problems and a possible over-parameterization of the model.

The underlying assumption in this study was the known covariance function. This covariance function could be based on estimates using different models like orthogonal polynomials or parametric functions. Covariances among regression coefficients for linear splines with a fixed location of knots can then be derived in a very simple manner. The method proposed could serve the purpose of preliminary selection of the number and location of knots for linear splines. Nevertheless, a re-estimation of covariance components for the model with random spline coefficients should follow.

The differential evolution algorithm was found to be fast and efficient in the multidimensional optimization problem attempted in this research. Computing times ranged from 3 to 5 min of CPU for the 305 DIM with 4 knots and 365 DIM with 7 knots models, respectively (Opteron 848 processor running at 2.2 GHz). Selection of optimal locations of knots for splines could have also been done in a simpler albeit more resource-consuming manner by examining all possible location of knots on the time scale. For example, selection of 5 knots from 360 possible candidates would require performing 49 002 354 071 comparisons. Another application of evolutionary algorithms to the problem of selection of knots for linear splines could be to perform an optimization for both number of knots and their locations simultaneously. The optimization criterion would take into account goodness-of-fit and impose a penalty for the number of knots. Finally, a multi-trait (or multi-lactation) covariance function could be considered with the simultaneous optimization process for many traits (lactation).

Knots have been assigned in an arbitrary way in most applications of random regression models with linear splines. Usually, the flanking knots have been set to the most extreme points of the trajectory. Our results indicated, however, that this is not necessarily the best location of the first and last knots.

Conclusions

Covariance matrix for coefficients of linear splines can be derived in a simple way from existing estimates of covariance functions. Optimal number and location of knots for linear splines in random regression TD models are not the same for different populations, random effects, lactations and traits. TD models for dairy production traits with linear splines as random regressions should account for these differences.

Acknowledgements

We are grateful to Dr Cedric Gondro and Dr Brian Kinghorn for introducing us to the world of evolutionary algorithms. The use of their differential evolution software is acknowledged.

References


