Rothamsted Repository Download

A - Papers appearing in refereed journals

GENETICS AND BREEDING

Genetic and Environmental Smoothing of Lactation Curves with Cubic Splines

I.M.S. WHITE,* R. THOMPSON,† and S. BROTHERSTONE*
*Institute of Cell Animal and Population Biology,
University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, Scotland
†Rothamsted Experimental Station,
AFRC Institute of Arable Crops Research,
Harpenden, Herts AL5 2JQ, England and
Roslin Institute (Edinbugh),
Roslin, Midlothian EH25 9PS, Scotland

ABSTRACT

Most approaches to modeling lactation curves involve parametric curves with fixed or random coefficients. In either case, the resulting models require the specification of an underlying parametric curve. The fitting of splines represents a semiparametric approach to the problem. In the context of animal breeding, cubic smoothing splines are particularly convenient because they can be incorporated into a suitably constructed mixed model. The potential for the use of splines in modeling lactation curves is explored with a simple example, and the results are compared with those using a random regression model. The spline model provides greater flexibility at the cost of additional computation. Splines are shown to be capable of picking up features of the lactation curve that are missed by the random regression model.

(**Key words**: lactation curves, smoothing, cubic splines, genetic effects)

Abbreviation key: **RR** = random regression.

INTRODUCTION

Data for test day production of dairy cows provide an example of longitudinal data or repeated measures, the essential feature of which is the presence of correlations between tests on the same animal. Both genetic and environmental covariances need to be taken into account. Various methods have been proposed for analyzing such data, ranging from simple curve fitting to a full multivariate analysis. The latter amounts to treating the measurements at successive times as separate but correlated traits. Most methods of analysis for longitudinal data can be regarded as being derived from a model in which the traits have a patterned covariance matrix. Different methods make more or less stringent assumptions about the structure of the matrix.

Two curve-fitting methods, random regressions and spline fitting, use patterned covariance matrices in the analysis of longitudinal data. The random regression method, with polynomial or other simple functions as covariates, is being used increasingly by animal breeders (3, 5), but, as yet, splines have not achieved the same popularity. However, for modeling growth curves, Verbyla et al. (7) recommended spline curves, which are flexible and can be fitted into the familiar mixed model framework. The spline terms play a role similar to that of the quadratic and higher order terms in a polynomial random regression model or to the exponential term of the lactation curve of Wilmink (8). Woolliams and Waddington (10) used splines to model lactation curves at the phenotypic level and reported a twofold increase in precision relative to the use of the lactation curve of Wood; no evidence of bias was found.

This paper shows that fitting splines to test day records is computationally feasible and leads to curves for estimated breeding value that are more flexible than those derived from other methods. First, we outline the basic mathematical theory of splines, and then we show how splines can be fitted by BLUP and REML methods, which are familiar to animal breeders.

MATERIALS AND METHODS

Cubic Splines

A cubic spline is a piecewise cubic function that is constrained so that the function and its first two derivatives are continuous at the breakpoints (knots) between one cubic segment and the next. Given q knots $x_1 < x_2 < \ldots < x_q$, the piecewise cubic curve

represented on the interval $x_j \le t \le x_{j+1}$ by

$$g(t) = \frac{(t - x_j)g_{j+1} + (x_{j+1} - t)g_j}{x_{j+1} - x_j}$$
$$- \frac{1}{6}(t - x_j)(x_{j+1} - t)$$
$$\left[\left(1 + \frac{t - x_j}{x_{j+1} - x_j}\right)v_{j+1} + \left(1 + \frac{x_{j+1} - t}{x_{j+1} - x_j}\right)v_j\right]$$

has value g_j and second derivative v_j at x_j, $j = 1 \ldots q$. The first term is a broken-stick curve produced by linear interpolation between the knots, and the second term, which vanishes at the knots, represents a cubic deviation. Between x_1 and x_q, the curve is continuous and has a continuous second derivative. The parameters (g_i, v_i), $i = 1 \ldots q$ of a cubic spline satisfy, in addition, the $q - 2$ constraints

$$\frac{g_{j+1} - g_j}{x_{j+1} - x_j} - \frac{g_j - g_{j-1}}{x_j - x_{j-1}} = \frac{1}{6}(x_j - x_{j-1})v_{j-1}$$
$$+ \frac{1}{3}(x_{j+1} - x_{j-1})v_j + \frac{1}{6}(x_{j+1} - x_j)v_{j+1}$$

for $j = 2 \ldots q - 1$, which ensure the continuity of the first derivative (see Appendix). Natural cubic splines are defined to be linear outside the interval (x_1, x_q) and, therefore, have $v_1 = v_q = 0$.

Given data $(t_1, y_1), \ldots, (t_n, y_n)$, the smoothing spline is the natural cubic spline that minimizes

$$\sum (y_i - g(t_i))^2 + \alpha \int_{x_1}^{x_q} \left(\frac{d^2 g}{dt^2}\right)^2 dt.$$

The integral is a roughness penalty, and the parameter α determines the trade-off between fidelity to the data and smoothness. Usually α is a parameter to be estimated, but, occasionally, it is specified a priori as a number ν ($1 \le \nu \le q$) of equivalent degrees of freedom. The case $\nu = 2$ ($\alpha = \infty$) corresponds to fitting a straight line. The case $\nu = q$ ($\alpha = 0$) produces a spline of best fit with no smoothing. When $q = n$ and the knots coincide with $t_1, \ldots, t_n$, this spline interpolates the data.

The term smoothing spline is usually restricted to the case in which the knot sequence consists of the distinct values in $t_1 \ldots t_n$. Here, it is assumed that $3 \le q \le n$, but no assumption is made about the position of the knots, which are chosen independently of $t_1 \ldots t_n$.

A good general introduction to splines and their statistical application is given by Green and Silverman (2).

Mixed Model Formulation

Natural cubic splines can be incorporated into the standard mixed model. The following demonstration extends some results of Verbyla et al. (7) to the case in which the knots are chosen independently of the data locations $t_1 \ldots t_n$.

Let $\mathbf{g}^T = (g_1 \ldots g_q)$ and $\mathbf{v}^T = (v_2 \ldots v_{q-1})$. Later, $\mathbf{a(t)}^T$ is used to denote $(1,t)$, $\mathbf{X}$ ($q \times 2$) has row i $\mathbf{a(x_i)}^T$ and $\mathbf{T}$ ($n \times 2$) has row i $\mathbf{a(t_i)}^T$. In matrix notation, the spline equation becomes

$$g(t) = \zeta(t)^T \mathbf{g} + \lambda(t)^T\mathbf{v},$$

and the constraints become $\mathbf{Q}^T\mathbf{g} = \mathbf{Rv}$. Here $\zeta(t)^T = (\zeta_1 \ldots \zeta_q)$ and $\lambda(t)^T = (\lambda_2 \ldots \lambda_{q-1})$, and each component of $\zeta(t)$ and $\lambda(t)$ is a function of t. It is easy to check that, for any fixed value of t, the vectors $\zeta(t)$ and $\lambda(t)$ have at most two nonzero components and that $\zeta(t)^TX = \mathbf{a(t)}^T$. For the constraint equations, $\mathbf{Q}^T$ is ($q - 2 \times q$), $\mathbf{R}$ is symmetric of dimension $q - 2$, and each matrix has at most three nonzero components in each row. Also $\mathbf{Q}^TX = \mathbf{0}$. These results are discussed in more detail in the Appendix.

A general solution to the constraints can be written

$$\mathbf{g} = \mathbf{Xb} + \mathbf{Q(Q^TQ)}^{-1}\mathbf{Rv} \qquad [1]$$

where the components of $\mathbf{b}$ are the intercept and slope of the regression of $g_1 \ldots g_q$ on $x_1 \ldots x_q$. Substitution for $\mathbf{g}$ and the use of $\zeta(t)^TX = \mathbf{a(t)}^T$ leads to an expression for the spline function in terms of $\mathbf{b}$ and $\mathbf{v}$:

$$g(t) = \mathbf{a(t)}^T\mathbf{b} + z(t)^T\mathbf{v} \qquad [2]$$

where

$$z(t)^T = \zeta(t)^T\mathbf{Q(Q^TQ)}^{-1}\mathbf{R} + \lambda(t)^T. \qquad [3]$$

The vector of spline values at the data points $t_1 \ldots t_n$ is $\mathbf{Tb} + \mathbf{Zv}$, where $\mathbf{Z}$ is the ($n \times q - 2$) matrix with row i $z(t_i)^T$. According to Green and Silverman (2), the roughness penalty is $\mathbf{v}^T\mathbf{Rv}$ so that the quantity minimized by the smoothing spline is

$$(\mathbf{y} - \mathbf{Tb} - \mathbf{Zv})^T(\mathbf{y} - \mathbf{Tb} - \mathbf{Zv}) + \alpha\mathbf{v}^T\mathbf{Rv}.$$

Differentiation of this expression with respect to **b** and **v** and setting derivatives to 0 leads to equations for a mixed model with fixed effects **Tb** (linear trend) and random effects **Zv**, with $cov(v) = \sigma_s^2 \mathbf{R}^{-1}$, where $\sigma_s^2 = \alpha^{-1}\sigma_e^2$. This connection between mixed model equations and a minimization problem was noted by Robinson (6). A convenient way to choose an appropriate degree of smoothing is by REML estimation of the variance component σ_s^2.

Families of Splines

In suffix notation, the mixed model for a single spline is

$$y_j = b_0 + b_1 t_j + \sum_{k=2}^{q-1} v_k z_k(t_j) + e_j \quad j = 1 \ldots n$$

where element k of the row matrix $\mathbf{z}(t)$ (Equation [3]) has been written $z_k(t_j)$.

From here, y_j = milk yield, and t_j = lactation stage (DIM) for a particular animal at test j. Suppose there are m animals and that animal i enters test j at t_{ij} DIM. Use y_{ij} for the test j measurement for animal i ($j = 1, \ldots n_i$). Then, a natural extension of the spline model is

$$\begin{aligned} y_{ij} = \ & b_0 + b_1 t_{ij} + b_{i0} + b_{i1} t_{ij} \\ & + \sum_{k=2}^{q-1} v_k z_k(t_{ij}) \\ & + \sum_{k=2}^{q-1} v_{ik} z_k(t_{ij}) + e_{ij} \end{aligned}$$

where the first two terms represent an overall linear regression, the third and fourth terms (animal and animal × linear) describe the deviation from the overall regression for animal i, and the fifth and sixth terms (spline and animal × spline) represent, respectively, a mean spline deviation and the deviation from the mean spline for animal i. The final term e_{ij} is the residual error with variance σ_e^2. The parentheses are ANOVA expressions for the individual terms in which spline is shorthand for the joint effect of the covariates $z_2(t), \ldots, z_{q-1}(t)$.

Inspection of the equation shows that the structure of the model is identical to that of a random regression (**RR**) model with covariates t and $z_j(t), j = 2 \ldots q - 1$. However, instead of representing polynomials of increasing degree or a set of unrelated functions of

DIM (such as t^2, log t), $z_2(t) \ldots z_{q-1}(t)$ represent a sequence of shifted curves. Each curve is approximately parabolic and centered on an internal knot.

In the absence of genetic effects, the covariance structure for the values of v in the case of a single spline is assumed to be true independently for each animal (i.e., the overall covariance matrix is block diagonal) with a block $\sigma_s^2 \mathbf{R}^{-1}$ for each animal. The linear terms are usually treated as fixed for a single spline, and the intercept and slope terms are usually treated as random when fitting a family of splines. The covariance matrix for the values of b is assumed block diagonal with a block Φ (2×2) for each animal. With this covariance structure, the model fits a separate spline with the same degree of smoothing for each animal. Both linear and deviation components are shrunk toward a common value.

Genetic effects are included either as an animal model (fitting the interaction terms with an associated numerator relationship matrix) or a sire model (fitting additional interaction terms sire, sire × linear, and sire × spline). In either case, the directly fitted animal iteractions are also included and represent permanent environmental effects. The smoothing parameter σ_s^2 and covariance matrix Φ are estimated separately at each level of deviation (genetic, permanent, and environmental). In addition, there is a smoothing parameter determining the smoothness of the mean spline.

The model can be further extended to include additional fixed or random effects. In general, any factor F, which would be included as an F × test interaction in an analysis treating tests as different traits, can be treated as an F × spline term, representing a family of spline curves, one for each level of F. The spline method has the advantage that factor F effects are constrained to change smoothly through the lactation, which is usually appropriate.

Estimated Covariance Matrices

The values of a spline at the knots depend on the parameters **b** and **v** from Equation [1], assuming that **b** and **v** independently generate a covariance matrix

$$cov(\mathbf{g}) = \mathbf{X}\Phi\mathbf{X}^T + \sigma_s^2 \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{R}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T.$$

Estimates of genetic and environmental covariances are obtained by substituting appropriate values of σ_s^2 and Φ in this equation. The covariance function c(s, t) = cov(g(s),g(t)) for values of s and t between the knots is derived in the same way from Equation [2] as $c(s, t) = \mathbf{a}(t)^T\mathbf{b}(s) + \mathbf{z}(t)^T\mathbf{v}(s)$, where $\mathbf{b}(s) = \Phi\mathbf{a}(s)$

and $\mathbf{v}(s) = \sigma_s^2 \mathbf{R}^{-1}\mathbf{z}(s)$. Considered as a function of t for fixed s, $c(s, t)$ is a natural spline with knots $x_1 \ldots x_q$ and parameters $\mathbf{b}(s)$, $\mathbf{v}(s)$.

Application

Lactation curves were fitted to test day records of milk production for 2885 progeny of 30 Holstein-Friesian sires in 503 herds using ASREML (1). Each curve modeled daily milk yield as a spline function of DIM. The ASREML program fits conventional smoothing splines with densely distributed knots (i.e., one knot for each value of DIM occurring in the data set). However, we fixed the number of knots in advance and fitted smoothing splines with 10 knots using the method described. Ten knots kept computational requirements to a reasonable level and allowed sufficiently accurate modeling of the lactation curves. Knots were placed at the mean number of DIM for each test. The lactation stage of animals entering the first test varied between 4 and 40 d, and successive tests were conducted at approximate 30-d intervals.

The hierarchical model had linear regression and spline terms for the general mean, sires, and cows within sires. The regression coefficients (intercept and slope) were treated as correlated random variables. The following additional fixed effects were included in the model: age at calving, the percentage of Holstein genes, and herd-test-month. The residual variance was assumed to be constant for each test but was allowed to vary between tests.

As well as the full model with 10 knots, reduced models with 3, 4, 5, 6, and 8 knots were fitted. Results were compared with those obtained using the exponential curve of Wilmink (8). To emphasize the similarity to the spline equation, this equation can be written

$$g(t) = b_0 + b_1 t + vz(t)$$

where $z(t) = \exp(-Dt)$ and D is a constant. The spline model with 10 knots was fitted twice: once using a sire model and again using an animal model.

RESULTS

Of the fixed effects in the model, the breed difference (Holstein – Friesian) was estimated at 1.60 kg (SE = 0.45), the effect of age at calving as 0.17 kg/mo (SE = 0.02), and the mean decline in production as 0.86 kg/mo (SE = 0.005). Table 1 gives variance estimates for each test. Genetic and environmental variances show the same pattern of variation over tests. Variances are constant in the middle of the lactation and rise at each end. The heritability rises from 0.23 to a fairly constant 0.30 from tests 5 to 9 before dropping slightly to 0.27 at test 10. Most of the genetic and environmental variance derived from the linear part of the spline. The spline deviation component accounted for, at most, 12% of the total. Estimated genetic and phenotypic correlations are given in Table 2. The genetic correlations never fall below 0.75, but the phenotypic correlations drop to 0.40. Correlations with a midlactation test fall away symmetrically and increasingly rapidly as the gap between the tests increases. In contrast, correlations with the first test fall away rapidly at first and remain almost constant between tests 6 to 9 before falling again at test 10.

Figure 1 shows the mean spline curve and the estimated breeding values for the 30 sires as deviations from the mean curve (with exaggerated vertical scale). For the most part, the resultant curves (not shown) are similar in shape, rising rapidly to a maximum after about 50 d, followed by a steady decline. The main source of variation in the curves was in the vertical position, reflecting overall performance, but,

TABLE 1. Mean DIM, variance estimates, and heritabilities by test.

| Test | DIM | Variance | | | | |
		G[1]	E	R	P	h^2
				(kg^2)		
1	18	3.40	9.10	5.17	15.11	0.23
2	48	3.06	7.47	4.07	12.30	0.25
3	78	2.94	6.79	3.57	11.09	0.27
4	109	2.93	6.62	3.08	10.43	0.28
5	139	2.95	6.60	2.25	9.59	0.31
6	169	2.96	6.57	2.60	9.91	0.30
7	199	2.95	6.52	2.37	9.62	0.31
8	229	2.96	6.62	2.32	9.68	0.31
9	259	3.09	7.23	2.06	10.07	0.31
10	290	3.45	8.81	3.04	12.70	0.27

[1]G = Genetic, E = environmental, R = residual, and P = phenotypic.

TABLE 2. Estimated genetic correlations (below diagonal) and phenotypic correlations (above diagonal) for tests 1 to 10.

Test	Test									
	1	2	3	4	5	6	7	8	9	10
1	. . .	0.64	0.60	0.55	0.52	0.48	0.47	0.46	0.45	0.40
2	0.98	. . .	0.66	0.63	0.61	0.57	0.55	0.53	0.51	0.44
3	0.93	0.98	. . .	0.68	0.68	0.64	0.62	0.58	0.55	0.46
4	0.88	0.95	0.99	. . .	0.72	0.69	0.67	0.63	0.58	0.48
5	0.85	0.92	0.97	0.99	. . .	0.74	0.73	0.68	0.63	0.52
6	0.82	0.90	0.95	0.98	0.99	. . .	0.73	0.70	0.65	0.54
7	0.82	0.89	0.93	0.96	0.98	0.99	. . .	0.74	0.71	0.60
8	0.82	0.88	0.91	0.93	0.95	0.97	0.99	. . .	0.76	0.68
9	0.81	0.85	0.88	0.89	0.90	0.92	0.95	0.98	. . .	0.75
10	0.77	0.81	0.82	0.82	0.83	0.85	0.88	0.93	0.98	. . .

as Figure 1 shows, there is also considerable variation in shape. The shape of the deviation curve does not seem to be related to the level.

The exponential curve of Wilmink (8) was fitted as a three-parameter random regression model to the same data, and the results were compared with those using the spline fit. An attempt was made to optimize exponent (D) of the exponential term. Several runs were conducted using different values and choosing D to maximize the likelihood. Some values of D produced negative genetic correlations between early and late parts of the lactation, but the correlations were positive for the optimum value D = 0.068. Agreement was good between the overall breeding values estimated, and the equivalent of Figure 1 for the exponential curves showed the same general shape. However, the variations in slope after the peak were necessarily absent.

Reduction of the number of knots had a large effect on the fit as measured by the negative log-likelihood. Taking the value for the 10-knot spline as 0, the values for 3-, 4-, 5-, 6-, and 8-knot splines were 312, 267, 184, 133, and 41. On the same scale, the value for the exponential model was 39 and was roughly equivalent to the spline model with 8 knots.

Using a sire model, fitting the spline model with 10 knots took approximately 60 min (four iterations) and using the exponential model, about 30 min (eight iterations). The spline fit was repeated using an animal model, which took about 180 min.

DISCUSSION

The mean lactation curve for a large population of animals may have a shape that is sufficiently regular to be fitted by a simple parametric function, but individual curves at sire and animal level show considerable variation in shape. Thus, successful modeling of lactation curves is likely to require a model with random coefficients such as the RR or spline model. Both RR and splines fit into the mixed model framework; unbalanced data can easily be dealt with, and genetic and environmental effects form a natural part of the model.

Because the two types of models are similar in structure, the mixed model equation arrays for a spline model with q knots and an RR model of order q are of the same size and equal sparsity. The models differ in the nature of the covariates used and the associated covariance structure. The RR model typically includes polynomial terms or a set of unrelated functions of DIM. The covariates of the spline model comprise a sequence of similar curves centered on the knots. In general, the RR model requires a full set of $1/2q$ (q + 1) variances and covariances to be estimated at each level (genetic and permanent environmental). The structure in the covariates of the spline model is such that a patterned covariance matrix, known within a constant, can be assigned a priori. The constant reflects the smoothness of the spline curve or family of spline curves. Thus, at each level, only four parameters need to be estimated for the spline model (σ_s^2 and the three elements of Φ).

A particular advantage of the spline approach is the extra flexibility it offers in the shape of fitted curves. Polynomials, even of high degree, are unlikely to produce such a good fit because of their inbuilt constraints. For example, a cubic curve has two turning points and one point of inflection, but the inflection point is necessarily midway between the turning points.

Preselection of knots allows the computational effort required to fit the model to be tailored to the computational power available. Once the number of knots has been determined, prior knowledge of the curves can be used to guide placement of the knots. For fitting a single spline, Wold (9) recommended that there should be no more than one extreme and one inflection point between knots, and that extremes should be centered in intervals between knots and

inflection points located near knots. When a family of related splines is being fitted, this advice can be followed only insofar as the curves share the same extremes and inflection points. We have ignored one part of this advice by placing the second knot at about 50 d, at which the curves tend to peak.

With preselected knots, smoothing takes place at two stages. Limitation of the number of knots to q has a smoothing effect; to invoke the roughness penalty produces further smoothing. This result is in contrast to regression splines, which rely on the first source of smoothing alone (i.e., there is no roughness penalty). Regression splines can be fitted by standard least squares calculations but show undesirable local behavior if there are too few knots [Chapter 9.3; (4)].

A comparison of the result of fitting the Wilmink curve and fitting a sequence of spline models with different numbers of knots shows that the Wilmink curve does better than the spline models with fewer than eight knots. It seems unlikely that this is due to the form of the Wilmink curve. A more likely explanation is that the better fit is due to the two extra covariance parameters allowed in the Wilmink model. Because the Wilmink model has the same number of terms as a three-knot spline, and five knots must be added to achieve a fit for the spline model equal to that for the Wilmink model, each extra parameter is roughly equivalent to two or three extra knots, which suggests that the spline model could be improved by allowing covariances between the spline (v) and regression (b) coefficients. There seems to be no compelling statistical or biological reason for the conventional assumption that these are independently distributed.
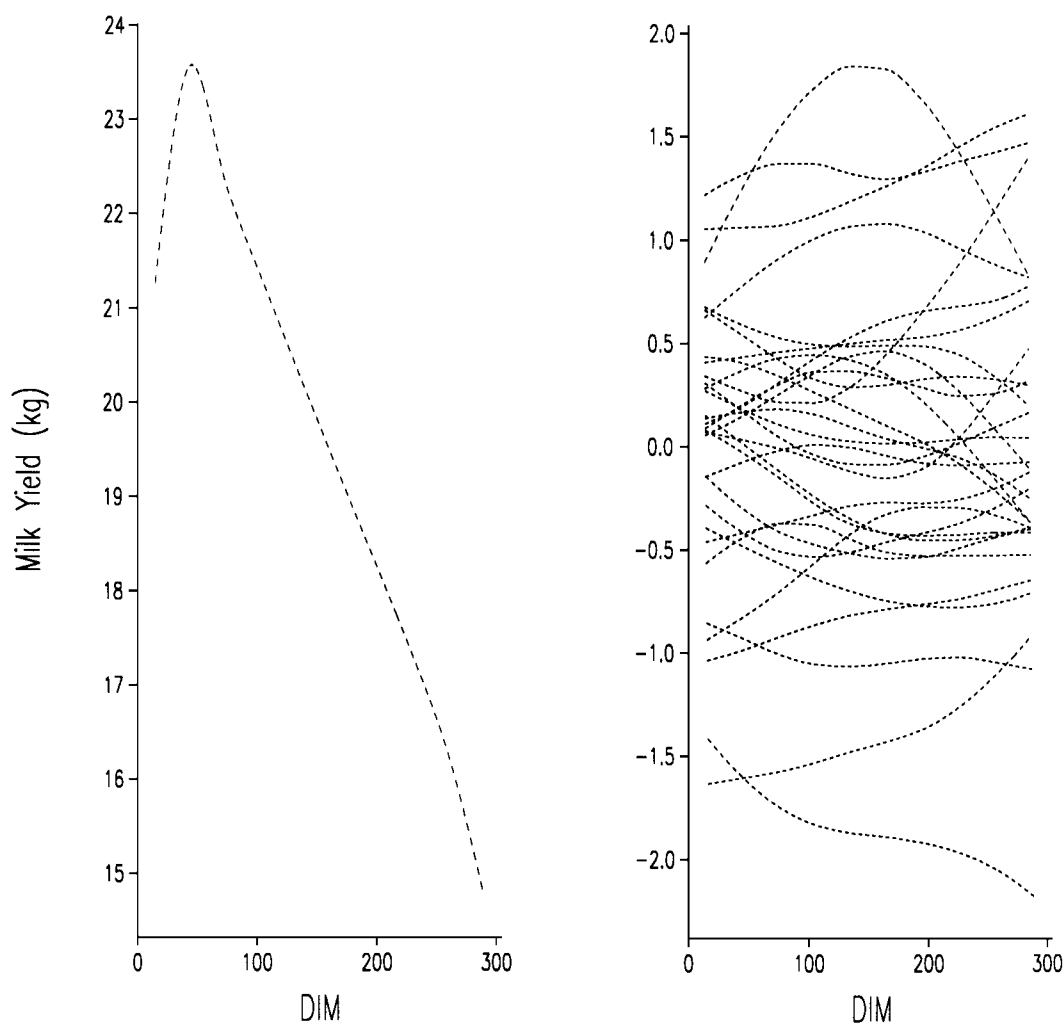

Figure 1. Fitted mean spline curve (left) and BLUP estimates (right) of breeding values for 30 sires as deviations from the mean curve.

CONCLUSIONS

Splines provide significant, extra flexibility in fitting lactation curves compared with RR models containing polynomial or simple functions of DIM as covariates. This flexibility extends to the covariance function on which estimates of genetic and environmental correlations are based. Fixing the number and position of knots makes the procedure computationally feasible. Figure 1 demonstrates that splines can capture interesting features of curves of EBV that are missed by simpler functions. With present methods, fitting a spline model requires more computation than does a simple RR model, but there is scope for improving present methods by exploiting special structure in the spline matrix equations.

REFERENCES

1 Gilmour, A. R., R. Thompson, B. R. Cullis, and S. J. Welham. 1997. ASREML Manual. New South Wales Dep. Agric., Orange, 2800, Australia.
2 Green, P. J., and B. W. Silverman. 1994. Nonparametric Regression and Generalized Linear Models. Chapman & Hall, London, England.
3 Jamrozik, J., and L. R. Schaeffer. 1997. Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. J. Dairy Sci. 80:762–770.
4 Hastie, T. J., and R. J. Tibshirani. 1990. Generalized Additive Models. Chapman & Hall, London, England.
5 Meyer, K., and W. G. Hill. 1997. Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated' records by restricted maximum likelihood. Livest. Prod. Sci. 47:185–200.
6 Robinson, G. K. 1991. That BLUP is a good thing: the estimation of random effects. Stat. Sci. 6:15–51.
7 Verbyla, A. P., B. R. Cullis, M. G. Kenward, and S. J. Welham. 1997. Smoothing splines in the analysis of designed experiments and longitudinal data. Res. Rep. 97/4, Dep. Stat., Univ. Adelaide, Adelaide, South Australia, Australia.
8 Wilmink, J.B.M. 1987. Adjustment of test-day milk, fat and protein yields for age, season and stage of lactation. Livest. Prod. Sci. 16:335–348.
9 Wold, S. 1974. Spline functions in data analysis. Technometrics 16:1–11.
10 Woolliams, J. A., and D. Waddington. 1998. The effectiveness of mixed-model smoothing for prediction of lactation yields. Page 10 in Proc. Br. Soc. Anim. Sci., Scarborough, England. Br. Soc. Anim. Sci., Midlothian, England.

APPENDIX

For $j = 1 \ldots q - 1$, let $h_j = x_{j+1} - x_j$. After rearrangement, the constraint equations can be written

$$\frac{g_{j+1} - g_j}{h_j} - \frac{1}{6} h_{j-1}(2v_j + v_{j+1}) = \frac{g_j - g_{j-1}}{h_{j-1}}$$
$$+ \frac{1}{6}h_{j-1}(v_{j-1} + 2v_j).$$

In this form, the equations declare equality between the left and right derivatives of the piecewise cubic curve at each internal knot x_j, $j = 2 \ldots q - 1$.

The components of $\zeta(t)$ and $\lambda(t)$ are derived from the coefficients of $g_1 \ldots g_q$, $v_2 \ldots v_{q-1}$ as given in the nonmatrix form of the spline equation. For $x_j \leq t \leq x_{j+1}$, $\zeta_j = (x_{j+1} - t)/h_j$, $\zeta_{j+1} = (t - x_j)/h_j$, and all other components are 0. Also, $\lambda_u = -\frac{1}{6}(t - x_j)(x_{j+1} - t)(1 + \zeta_u)$ for $u = j, j+1$ and is 0 for all other values of u. When $t \leq x_1$, ζ_1 and ζ_2 are calculated as for $x_1 < t < x_2$, and $\lambda_2 = -\frac{1}{6}h_1(t - x_1)$ is the only nonzero component of λ. Similarly, when $t > x_q$, ζ_{q-1} and ζ_q are calculated as for $x_{q-1} < t < x_q$, and $\lambda_{q-1} = -\frac{1}{6}h_{q-1}(x_q - t)$ is the only nonzero component of λ. In all cases, λ_1 and λ_q are identically 0.

That $\zeta(t)^T X = (1, t)$ can be obtained directly from these results or by noting that $\zeta(t)^T g$ is the value at t of the broken-stick graph interpolating between $(x_1, g_1), \ldots, (x_q, g_q)$ and by considering the two special cases $g_i = 1$ (case 1) and $g_i = x_i$, $i = 1 \ldots q$ (case 2).

The matrices Q and R are derived by inspection of the nonmatrix form of the constraint equations. For $j = 1 \ldots q - 2$, $q_{j,j} = h_j^{-1}$, $q_{j+1,j} = -h_j^{-1} - h_{j+1}^{-1}$, $q_{j+2,j} = h_{j+1}^{-1}$, and $r_{j,j} = \frac{1}{3}(h_j + h_{j+1})$. For $j = 1 \ldots q - 3$, $r_{j,j+1} = r_{j+1,j} = \frac{1}{6}h_{j+1}$. All other elements of Q and R are 0. That $Q^T X = 0$ follows by direct evaluation or by noting that the columns of X correspond to the special cases 1 and 2, for which $v = 0$.