Genet. Res., Camb. (1994), **64**, *pp.* 57–69 *With 5 text-figures*

Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle

MARK KIRKPATRICK*‡, WILLIAM G. HILL† AND ROBIN THOMPSON§

** Department of Zoology, University of Texas, Austin TX 78712, USA, † Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, U.K. § BBSRC Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, U.K.*

(*Received 20 October 1993 and in revised form 13 May 1994*)

Summary

Quantitative variation in traits that change with age is important to both evolutionary biologists and breeders. We present three new methods for estimating the phenotypic and additive genetic covariance functions of a trait that changes with age, and illustrate them using data on daily lactation records from British Holstein–Friesian dairy cattle. First, a new technique is developed to fit a continuous covariance function to a covariance matrix. Secondly, this technique is used to estimate and correct for a bias that inflates estimates of phenotypic variances. Thirdly, we offer a numerical method for estimating the eigenvalues and eigenfunctions of covariance functions. Although the algorithms are moderately complex, they have been implemented in a software package that is made freely available.

Analysis of lactation shows the advantages of the new methods over earlier ones. Results suggest that phenotypic variances are inflated by as much as 39% above the underlying covariance structure by measurement error and short term environmental effects. Analysis of additive genetic variation indicates that about 90% of the additive genetic variation for lactation during the first 10 months is associated with an eigenfunction that corresponds to increased (or decreased) production at all ages. Genetic tradeoffs between early and late milk yield are seen in the second eigenfunction, but it accounts for less than 8% of the additive variance. This illustrates that selection is expected to increase production throughout lactation.

1. Introduction

An individual's phenotype changes with age. A trait that changes with age can be represented as a trajectory, that is, a function of time. Because each character takes on a value at each of an infinite number of ages, and its value at each age can be considered as a distinct trait, such trajectories are referred to as 'infinite-dimensional' characters.

Many problems of interest to breeders and evolutionary biologists involve selection on this type of trait. The traditional way of analysing the quantitative genetics of infinite dimensional traits involves focusing on the phenotypic values at a small number of landmark ages, making discrete what is intrinsically a continuous process. Recently, the methods of quantitative genetics have been extended to infinite-dimensional traits to overcome this deficiency (Kirkpatrick & Heckman, 1989; Kirkpatrick *et al.* 1990; Kirkpatrick & Lofsvold, 1992; Gomulkiewicz & Kirkpatrick, 1992).

The infinite-dimensional approach can provide more accurate estimates of variation in the traits and improve estimates of their response to natural or artificial selection as compared to conventional methods. Improved estimates of phenotypic and genetic covariances can be realized using the fact that the measurements are ordered in time. The situation is analogous to the classical statistical problem of predicting the value of a dependent variable y as a function of an independent variable x. A standard approach is to regress observed values of y onto x. Then, given a value x^*, a prediction for the corresponding value y^* is determined by the regression equation. Alternatively, one might use the observed value of y corresponding to the observed value of x that is closest to x^*. In many situations the regression prediction will be superior because measurement error in y makes prediction from a single pair of observed x

‡ Corresponding author.

and y unreliable, while the regression approach gains power by using information from all the observations.

An estimate of the covariance between the values of a trait at two ages can likewise be improved by using information about the covariances at other ages. The classical approach of treating the value at each age as a discrete trait without regard for its place in the sequence of ages loses substantial information. In contrast, the infinite-dimensional approach seeks to retain this information by using, in effect, a regression of covariance on age. Given the notoriously large sampling errors inherent in estimates of covariances, any gain in the power of estimation is welcome.

This paper extends the recently-developed methods for the analysis of infinite-dimensional traits and demonstrates them using data on lactation records from British Holstein–Friesian dairy cattle. We begin by briefly reviewing the framework for estimating covariance functions that was introduced by Kirkpatrick *et al.* (1990). We then introduce three new methods within this framework. The first is a technique for estimating covariance functions referred to as the method of asymmetric coefficients. This method is illustrated with a simple worked example. The second is a technique for correcting the bias that appears along the diagonal of an estimated phenotypic covariance function or matrix. This bias arises because date-specific measurement errors inflate the phenotypic variances (the diagonal elements), but have no such bias on the phenotypic covariances (the off-diagonal elements) or any of the additive genetic parameters. Our strategy here is to use the unbiased off-diagonal elements to estimate the diagonal elements. The algorithm is demonstrated using a simple example. The third method is a numerical approach to calculate the eigenvalues and eigenfunctions of a covariance function which is useful to describe the patterns of variation. After these new methods are introduced, they are applied to lactation records from British Holstein–Friesian dairy cattle.

2. Estimating covariance functions

For any trait that changes in time, the phenotype of an individual at age t can be written $z(t)$. Variation in the population for this function is characterized by a covariance function. A covariance function is the infinite-dimensional analogue of a covariance matrix. The value of the phenotypic covariance function $\mathscr{P}(t_1, t_2)$ gives the phenotypic covariance between the value of the trait at ages t_1 and t_2. The phenotypic variance at age t_1 is written $\mathscr{P}(t_1, t_1)$. Likewise, the additive genetic covariance structure of a population is described by the additive genetic covariance function $\mathscr{G}$.

For any practical application, these covariance functions are estimated from breeding data. The approach advocated by Kirkpatrick *et al.* (1990) starts with measurements of individuals at each of n ages, denoted a_1 through a_n. Standard quantitative-genetic methods are used to obtain an estimate of the $n \times n$ covariance matrix for the measurements at these ages.

The goal now is to estimate the underlying covariance function from this matrix. In general, this is done by interpolating between the values of the covariance matrix, perhaps smoothing them in order to damp out the sampling error in the elements of the matrix. A variety of functions can be used for the interpolation. The approach we developed earlier is based on orthogonal functions (Kirkpatrick & Heckman, 1989; Kirkpatrick *et al.* 1990), and we will again use that method here.

We begin by briefly reviewing the approach, which is referred to as the method of 'symmetric coefficients' in this paper. It starts with the fact that any continuous covariance function can be represented as a weighted sum of orthogonal functions. That is, given a set of functions $\{\phi_i\}$, $i = 0, 1, \ldots$, that are orthogonal over the interval $[a_1, a_n]$, we can write the covariance function $\mathscr{P}$ as

$$\mathscr{P}(t_1, t_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} C_{ij} \phi_i(t_1) \phi_j(t_2), \tag{1}$$

where the C_{ij}s are constants. These constants form a symmetric matrix, $C_{ij} = C_{ji}$ (whence the term 'symmetric coefficients'), which guarantees that $\mathscr{P}$ is symmetric as required by the definition of a covariance function. The strategy developed by Kirkpatrick *et al.* (1990) is to use an estimated covariance matrix $\hat{\boldsymbol{P}}$ based on measurements taken at n ages to estimate a truncated set of the weighting coefficients C_{ij}. Our estimate of the covariance function $\mathscr{P}$, based on the first k orthogonal functions, is then

$$\hat{\mathscr{P}}(t_1, t_2) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \hat{C}_{ij} \phi_i(t_1) \phi_j(t_2), \tag{2}$$

where $k \leqslant n$. The statistical problem, then, is to estimate the matrix of coefficients $\boldsymbol{C}$ so that they can be substituted into eqn (2) to yield an estimate of the covariance function. As discussed by Kirkpatrick *et al.* (1990), we can obtain a 'full fit', in which $k = n$, such that the value of $\hat{\mathscr{P}}(t_1, t_2)$ exactly equals the corresponding value of $\hat{\boldsymbol{P}}$ when t_1 and t_2 equal two of the ages at which the data were taken. Alternatively, we can seek a 'reduced fit', in which $k < n$. Under a reduced fit, there will generally be discrepancies between $\hat{\mathscr{P}}$ and $\hat{\boldsymbol{P}}$. The rational for favouring a reduced fit is that the estimate $\hat{\boldsymbol{P}}$ includes sampling error, and we might prefer an estimate $\hat{\mathscr{P}}$ that smooths out the fluctuations that these errors introduce.

The methods for both the full and reduced estimates of a covariance function that were developed earlier lead to a symmetric coefficient matrix $\hat{\boldsymbol{C}}$. In the next section we introduce a new method that leads to an asymmetric coefficient matrix, and show its advantages.

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

3. The method of asymmetric coefficients

There are three reasons for developing the new method based on asymmetric coefficients. First, an estimate of a covariance function based on our earlier method has continuous first derivatives everywhere. This may be undesirable along the diagonal of the covariance function, where we might want to allow for the possibility of a crease, or discontinuous first derivative. A discontinuous first derivative along the diagonal is found in the covariance functions of several simple stochastic processes, including brownian motion, and so it seems desirable to allow for this possibility. The method of asymmetric coefficients makes no assumption about the continuity of first derivatives of the estimated covariance function along the diagonal.

Secondly, estimates of a covariance function based on asymmetric coefficients may be somewhat better behaved than those based on symmetric coefficients. The reason lies in the fact that estimates using symmetric coefficients involve the products of higher-order terms that result in functions that are less smooth than their asymmetric counterparts. The coefficient matrix $\hat{\boldsymbol{C}}$ derived using symmetric coefficients generally will have all non-zero elements. When substituted into eqn (2), this produces terms of order $\phi_{k-1}(t_1)\,\phi_{k-1}(t_2)$. With orthogonal polynomials as the ϕs, for example, this corresponds to the product of two $(k-1)^{\text{th}}$ order polynomials, which will often result in a quite 'wiggly' function. By contrast, the coefficient matrix $\hat{\boldsymbol{C}}$ derived using asymmetric coefficients has zeros in all elements C_{ij} for which $i+j \geqslant k$. Thus the terms of highest order to appear in eqn (2) are of the same order as ϕ_{k-1}. Hence asymmetric coefficients often lead to smoother estimates.

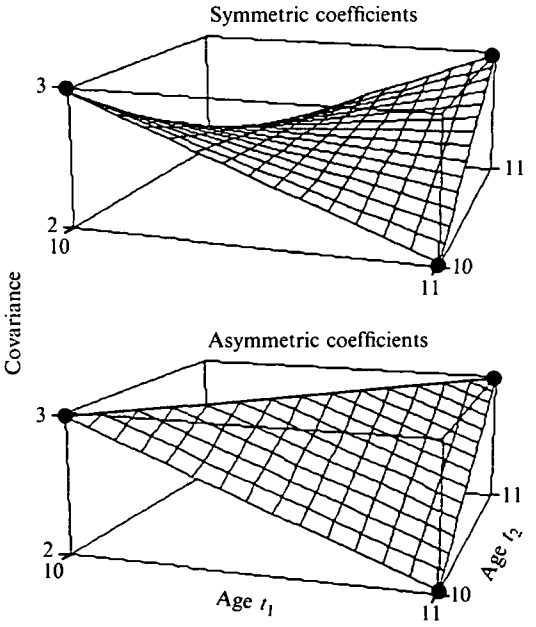

Fig. 1. Fits using the methods of symmetric coefficients (top) and asymmetric coefficients (bottom) with the example of eqn (5) discussed in the text. The solid circles show the original data points.

Thirdly, the asymmetric method can be used to correct for a bias in the diagonal elements of phenotypic covariance functions. We discuss this problem further in a later section ('Extrapolating to the diagonal').

These attractions of asymmetric coefficients are mitigated by the fact that some of the techniques developed earlier under the method of symmetric coefficients do not carry over to the new method. In particular, the algebraic technique for estimating the eigenfunctions and eigenvalues of the covariance matrix directly from the coefficients cannot be applied to asymmetric coefficients. It is still possible, however, to estimate these quantities by numerical methods using the methods we discuss in a later section ('Analysis of genetic variation').

The method of asymmetric coefficients seeks an estimate of the covariance function $\mathscr{P}$ that is of the form

$$\hat{\mathscr{P}}(t_1, t_2) = \begin{cases} \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \hat{C}_{ij}\,\phi_i(t_1)\,\phi_j(t_2), & t_1 \geqslant t_2 \\ \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \hat{C}_{ij}\,\phi_i(t_2)\,\phi_j(t_1), & t_1 < t_2. \end{cases} \tag{3}$$

Unlike the earlier method, there is no requirement that $\hat{C}_{ij} = \hat{C}_{ji}$ because the form of eqn (3) guarantees that $\hat{\mathscr{P}}$ will be symmetric, so we refer to this as the method of 'asymmetric coefficients'. The data matrix $\hat{\boldsymbol{P}}$ contains $n(n+1)/2$ parameters, and we can estimate no more than this number of coefficients. We choose to fit the coefficients $\hat{C}_{ij}$ with $i+j \leqslant k-1$, that is the lower left half of the matrix $\hat{\boldsymbol{C}}$. This choice tends to result in a smoother estimate than if higher-order coefficients were fitted, as discussed above.

The strategy we use to fit $\hat{\boldsymbol{C}}$ is to transform the problem into a standard least-squares formulation. By stacking the columns of the data matrix $\hat{\boldsymbol{P}}$ to form a vector $\boldsymbol{p}$, and similarly transforming the coefficient matrix $\hat{\boldsymbol{C}}$ into a vector $\boldsymbol{c}$, the statistical model can be written:

$$\boldsymbol{p} = \boldsymbol{X}\boldsymbol{c} + \boldsymbol{e}, \tag{4}$$

where $\boldsymbol{p}$ is a vector of observations (the estimated variances and covariances), $\boldsymbol{X}$ is a matrix defined by the values of the orthogonal functions evaluated at the measured ages, $\boldsymbol{c}$ is a vector of coefficients, and $\boldsymbol{e}$ is a vector of error terms. Our goal is to solve for the vector $\boldsymbol{c}$ that minimizes the error vector according to the weighted sum of squares criterion.

Algorithms for this calculation are described in Appendix A for the cases of both a full and a reduced fit. The method has been implemented as a computer program in a *Mathematica*® notebook (Wolfram, 1991). The program (which also performs other analyses and displays them graphically) is available from the senior author.

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

To illustrate our approach, consider the problem of fitting the covariance matrix

$$\hat{\boldsymbol{P}} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}, \tag{5}$$

based on measurements taken at the ages $\boldsymbol{a} = (10, 11)^{\mathrm{T}}$, as shown in Figure 1. We will find the full estimate of $\mathscr{P}$, and so $k = n = 2$. We choose to use normalized Legendre polynomials as the orthogonal functions. The first two of these polynomials are:

$$\phi_0(x) = 1/\sqrt{2} \quad \text{and} \quad \phi_1(x) = \sqrt{(3/2)}\,x \tag{6}$$

(see Kirkpatrick *et al.* 1990). Calculation of the coefficient matrix is described in detail for this example in Appendix A. It leads to the result

$$\hat{\boldsymbol{C}} = \begin{pmatrix} 6 & \sqrt{(3/3)} \\ -\sqrt{(3/3)} & 0 \end{pmatrix}. \tag{7}$$

Notice that the matrix $\hat{\boldsymbol{C}}$ is asymmetric, and that elements below the antidiagonal are zero. These two properties distinguish the asymmetric coefficient matrix from the symmetric matrix approach described by Kirkpatrick *et al.* (1990).

Substituting these coefficients into eqn (A 14) gives our estimate of the covariance function:

$$\hat{\mathscr{P}}(t_1, t_2) = \begin{cases} 3 + t_1 - t_2 & \text{for} \quad 10 \leqslant t_1 \leqslant t_2 \leqslant 11 \\ 3 + t_2 - t_1 & \text{for} \quad 10 \leqslant t_2 \leqslant t_1 \leqslant 11. \end{cases} \tag{8}$$

While the coefficient matrix from which it was calculated is asymmetric, the covariance function itself is symmetric (as required by the definition of a covariance function). Checking, we confirm that the entries in the original matrix $\hat{\boldsymbol{P}}$ are recovered when we substitute $t_1, t_2 = 10, 11$ into eqn (8). A perfect fit of the estimated covariance function to the data matrix results whenever a full fit is calculated.

The method of symmetric coefficients developed previously (Kirkpatrick *et al.* 1990) leads to somewhat different results. Using that method, the coefficient matrix for a full fit is

$$\hat{\boldsymbol{C}} = \begin{pmatrix} 5 & 0 \\ 0 & 1/3 \end{pmatrix}. \tag{9}$$

Unlike the coefficient matrix (7), this matrix is symmetric. (The off-diagonal coefficients are zero in this example, but that is not generally true.) The corresponding estimate of the covariance function is

$$\hat{\mathscr{P}}(t_1, t_2) = 223 - 21t_1 - 21t_2 + 2t_1 t_2. \tag{10}$$

As with the earlier estimate using asymmetric coefficients, the original data matrix of eqn (5) is recovered when we substitute $t_1, t_2 = 10, 11$ into this equation.

The symmetric and asymmetric expressions for $\hat{\mathscr{P}}$ are quite different. The differences are seen clearly in Fig. 1. A conspicuous and diagnostic discrepancy is that symmetric coefficient estimate of $\hat{\mathscr{P}}$ is smooth along the diagonal while the asymmetric coefficient estimate is not. The symmetric coefficient estimate also has more curvature.

4. Extrapolating to the diagonal

Estimates of phenotypic variances for the values of traits at specific ages are often inflated by factors that do not affect estimates of the covariances between ages. One source of this inflation is measurement error. A second source involves environmental factors that have effects over periods much shorter than the between-measurement intervals, such as weather, health, food quality, and hormonal state. This second type of factor tends to increased covariances close to the diagonal of the covariance function. For example, estimates of the phenotypic correlations of lactation test day records one day apart were 0·84, declining only to 0·82 for records five days apart (Pander *et al.* 1993). Thus we can view the diagonal elements of a phenotypic covariance matrix or function as being biased upwards, relative to a smoother underlying pattern that we expect on biological grounds. The upward bias appears as a ridge along the diagonal of estimated phenotypic covariance matrices and covariance functions. This bias distorts our picture of the covariance structure of the trait, and has practical implications in breeding programs that are based on age-specific variances.

We would therefore like to correct for the bias. Two strategies are available. A direct approach would be to estimate the measurement error directly, for example through repeated measures. A second, indirect approach is available when the characters of interest are age-specific measurements of the same trait through time. A familiar example is a growth trajectory, in which the data are measurements of the sizes of each individual at a series of ages. In this situation the basic phenotype of interest is a continuous function – the growth trajectory – that is an infinite-dimensional trait. Here we show how the phenotypic covariances for an infinite-dimensional trait can be used to estimate the variances (that is, the diagonal elements of the covariance matrix). These estimates are free of measurement error bias, and may lead to selection indices with increased efficiency.

Our strategy is as follows. On intuitive grounds, we expect covariance function for growth processes to be continuous. (This is a biological rather than mathematical argument, since there is nothing in the definition of a covariance function that requires it to be continuous.) Using the unbiased estimates for the phenotypic covariances (that is, $\mathscr{P}(t_1, t_2)$ where $t_1 \neq t_2$), we can extrapolate estimates of the phenotypic variances (that is, $\mathscr{P}(t_1, t_1)$). The algorithm begins with an estimated phenotypic covariance matrix of the sizes of individuals at the n ages a_i. We first estimate the phenotypic covariance function using only the $n(n-1)/2$ unbiased subdiagonal elements of $\hat{\boldsymbol{P}}$. The method of asymmetric coefficients described above

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

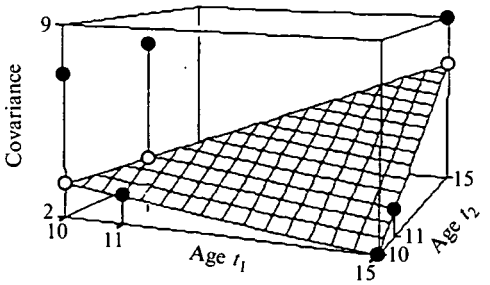

Fig. 2. Fit using the method of extrapolating to the diagonal using the example discussed in the text. The solid circles show the original data points; the open circles are the extrapolated values for the diagonal elements (the variances).

produces an estimate of the phenotypic covariance function in terms of a weighted sum of orthogonal functions. Because the diagonal elements of $\hat{\boldsymbol{P}}$ were omitted, this estimated interpolates the values of $\mathscr{P}(t_1, t_2)$ over the ranges $t_1 \in [a_2, a_n]$ and $t_2 \in [a_1, a_{n-1}]$, where $t_1 > t_2$. Secondly, the coefficients are used to extrapolate the estimated covariance function: the range of t_1 is extended downward from age a_2 to a_1 and the range of t_2 upward from age a_{n-1} to a_n, giving us the full range $t_1, t_2 \in [a_1, a_n]$. This extrapolation gives us an unbiased estimate of the diagonal of $\mathscr{P}$ along with the rest of the covariance function.

A detailed description of the algorithm is given in Appendix B. It has been implemented in a *Mathematica*™ notebook, which is available from the senior author. To illustrate, consider the estimated phenotypic covariance matrix

$$\hat{\boldsymbol{P}} = \begin{pmatrix} 7 & 3 & 2 \\ 3 & 8 & 3 \\ 2 & 3 & 9 \end{pmatrix} \tag{11}$$

based on measurements of some character taken at ages $\boldsymbol{a} = (10, 11, 15)^{\mathrm{T}}$, plotted in Fig. 2. This example will illustrate how the method naturally accommodates uneven intervals between the measured ages. The variances along the diagonal of $\hat{\boldsymbol{P}}$ have been inflated by the biases described earlier, and our aim is to obtain corrected estimates for them. In this approach the diagonal elements are not used in the estimate of the $\mathscr{P}$, and so a full fit uses $k = n-1 = 2$ orthogonal functions.

Appendix B shows that using the method of extrapolating to the diagonal, we obtain an estimated phenotypic covariance function

$$\hat{\mathscr{P}}(t_1, t_2) = \begin{cases} -17/4 + t_1 - t_2/4 & \text{for} \quad 10 \leqslant t_1 \leqslant t_2 \leqslant 15 \\ -17/4 + t_2 - t_1/4 & \text{for} \quad 10 \leqslant t_2 \leqslant t_1 \leqslant 15. \end{cases} \tag{12}$$

Evaluating this function at the measured ages ($t = 10$, 11, 15), we obtain the matrix

$$\hat{\boldsymbol{P}} = \begin{pmatrix} 3{\cdot}25 & 3 & 2 \\ 3 & 4 & 3 \\ 2 & 3 & 7 \end{pmatrix} \tag{13}$$

The results suggest that the variances shown along the diagonal of eqn (11) are overestimated by as much as 115% (7 *v.* 3·25 for the variance at age 10). These results are illustrated in Fig. 2.

5. Analysis of variation

The covariance function is an important descriptor of variation in the trajectory of a character that changes through time, and a substantial amount can be learned from its analysis. The spectrum, or eigenvalues and eigenfunctions, of a covariance function is particularly useful. The leading eigenvalues and eigenfunctions visualize major patterns of variation, and describe these patterns with many fewer parameters than the full covariance function. One important application involves the additive covariance function. Its leading eigenfunctions identify the types of evolutionary changes for which the population has substantial genetic variation available. Conversely, its spectrum also shows the types of changes for which there is not appreciable genetic variation, and therefore which will occur slowly if at all under selection.

The method of symmetric coefficients was specifically devised with this objective in mind. Calculations based on a symmetric coefficient matrix can be used to obtain estimates of eigenfunctions and eigenvalues directly (Kirkpatrick & Heckman, 1989; Kirkpatrick *et al.* 1990). The method of asymmetric coefficients, on the other hand, cannot be adapted to these calculations. We therefore propose an alternative using a numerical approach.

An estimate of a covariance function based on asymmetric coefficients can be evaluated on a square lattice of a moderate to large number of points. These values form a matrix whose spectrum (eigenvectors and eigenvalues) can then be calculated by standard methods. As the number of points on the lattice increases, the estimates of the eigenvalues will converge on those of the underlying covariance function (see Kirkpatrick & Heckman, 1989). The points of the eigenvectors can be interpolated to give estimates of the corresponding eigenfunctions (within a constant factor that is a function of the number of points in the lattice).

This may seem like a rather baroque method of estimating quantities that could be obtained much more directly by simply calculating the eigenvalues and eigenvectors of the original covariance matrix. The incentive for performing the less direct algorithm just described is that it is expected to give more accurate estimates (Kirkpatrick & Heckman, 1989). The reason for this seems to lie in the fact that simply calculating the spectrum of the original matrix discards all information about the ordering in time of the ages at which the measurements were taken. The methods for symmetric coefficients developed by Kirkpatrick *et al.* (1990) make use of this information; the indirect

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

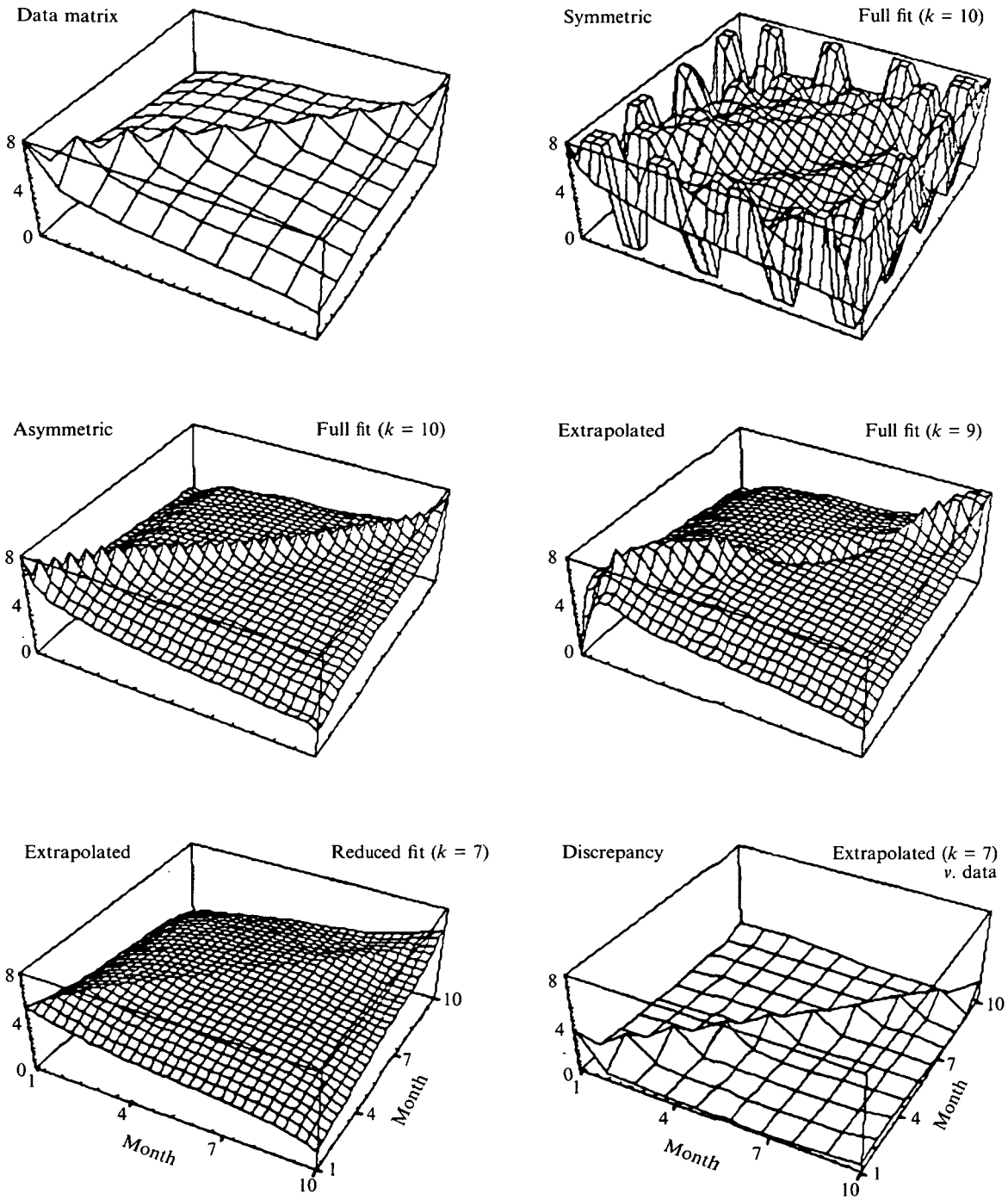

Fig. 3. Estimates of the phenotypic covariance function for lactation in British Holstein–Friesian dairy cattle. The original data (top left) show a ridge corresponding to upward biases in the diagonal elements (variances). The full fit using symmetric coefficients (top right) overfits the data; the plot has been truncated in the vertical dimension. The full fit with asymmetric coefficients (middle left) is much smoother, but reproduces the inflated diagonal. The extrapolated full fit (middle right) is poorly behaved along the diagonal. An extrapolated reduced fit with $k = 7$ (bottom left) is well behaved everywhere. The discrepancy between this estimate and the original data (bottom right) is substantial along the diagonal, corresponding to the bias, but very small elsewhere. Variances and covariances are in units of kg^2.

algorithm just described for use with asymmetric coefficients also does so.

6. Analysis of lactation in dairy cattle

We will illustrate the three methods developed above using lactation records of British Holstein–Friesian dairy cattle. The data are described in detail by Pander *et al.* (1992), and comprise their data set 2. Briefly, these were records of daily milk yield ('test day records') of 34029 heifers of known parents. The first record for each heifer was taken between day 5 and day 35 after the start of lactation, and successive records at monthly intervals for a total of 10 monthly records per individual. The data were analysed as if each measurement was taken at the midpoint of the

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

month interval. Additive genetic and phenotypic parameters were calculated from the sire component of covariances using restricted maximum likelihood (REML; see Patterson & Thompson, 1971; Meyer, 1985).

We analysed these data with the methods described above using Legendre polynomials. When reduced estimates are computed, the matrix $\boldsymbol{V}$ of the error covariances between the estimated covariances is required (Kirkpatrick *et al.* 1990). Since the REML program used does not estimate $\boldsymbol{V}$, we used the following approximation. The phenotypic and additive genetic covariances were viewed as if they had been estimated using a standard balanced half-sib breeding design with 700 random sires, each with 10 half-sib offspring. There was a total of 16000 residual degrees of freedom because daughters of a number of additional selected sires were used to increase connexions between herds. These sample sizes are a reasonable approximation to the more complex pedigree actually used to estimate the genetic parameters (see Pander *et al.* 1992). The $\boldsymbol{V}$ matrix was then estimated using the formulae in Appendix C of Kirkpatrick *et al.* (1990).

(i) *Estimating the phenotypic covariance function*

The phenotypic covariance matrix is plotted in three dimensions in Fig. 3 (top left). We begin by calculating full estimates ($k = n = 10$) of the phenotypic covariance function. The estimates based on symmetric coefficients and on asymmetric coefficients are compared in Fig. 3. Both show a conspicuous ridge running along the diagonal, corresponding to the date-specific measurement error described in the introduction. A secondary effect of the spike along the diagonal is to produce a series of parallel harmonic ridges in the symmetric coefficient estimate. These are not seen in the original data, but rather reflect side effects of how the polynomials used to construct $\hat{\mathscr{P}}$ accommodate the large elements along the diagonal. The estimate based on asymmetric coefficients is substantially smoother, as anticipated for the reasons discussed earlier, but still captures the diagonal ridge corresponding to the inflated variance estimates.

The method of extrapolating to the diagonal is used to eliminate the upward bias in the estimates of the diagonal of $\hat{\boldsymbol{P}}$. The full fit ($k = 9$ polynomials) gives an unsatisfactory estimate for $\mathscr{P}$ (Fig. 3, middle right). The extrapolation based on these high-order polynomials causes the estimated covariance function to take on extremely small values along the diagonal. In fact, this estimate is not positive semidefinite, and so does not qualify as a covariance function. The reduced fit with $k = 8$ suffers the same problem.

A reduced estimate with $k = 7$ (Fig. 3, bottom left), however, shows a covariance function that is both well-behaved and in keeping with our intuitive expectation based on the original data. The function rises smoothly to the diagonal. It fits the off-diagonal elements of $\hat{\boldsymbol{P}}$ very well: all of the differences are less than 2% in magnitude. In contrast, there is a large discrepancy between the extrapolated estimates of the diagonal elements of $\hat{\boldsymbol{P}}$ and those from the original matrix (Fig. 3, bottom right). The differences are, in fact, our estimates of the upward biases in the diagonal elements. They are substantial. Our extrapolated values differ by as much as 36% from the values of the diagonal elements of the original matrix $\hat{\boldsymbol{P}}$.

An even simpler estimate of the covariance function would be one that depended only on the difference between any pair of ages. Inspection of the data, however, shows for example that the phenotypic correlation between the first and second month of lactation is different from that between the seventh and eighth ($r_P = 0{\cdot}64$ *v.* $r_P = 0{\cdot}76$, respectively; Pander *et al.* 1992 Appendix Table 1), and so this alternative is not appropriate in this case.

(ii) *Estimation and analysis of the additive genetic covariance function*

To illustrate the method of asymmetric coefficients, we will again use the data of Pander *et al.* (1992). Their estimate of the 10×10 additive genetic covariance matrix $\hat{\boldsymbol{G}}$ and our estimates of the continuous covariance function are shown in Fig. 4. We first calculated full ($k = n = 10$) estimates $\hat{\mathscr{G}}$ of the covariance function using both symmetric and asymmetric coefficients. Both show severe fluctuations. The symmetric estimate takes on values that range from less than -3 to more than 7 kg^2 (Fig. 4, top right) even though the original matrix elements only span the range from 1·5 to 3·5 kg^2. The asymmetric estimate is again considerably better behaved, as expected, but it nevertheless takes on values as small as -0.12 kg^2 (Fig. 4, middle left).

We than calculated the symmetric and asymmetric estimates using reduced fits using $k = 9$ polynomials (Fig. 4, middle right and bottom left). Both are far better behaved than the full estimates. The goodness-of-fit tests give χ^2 (10 D.F.) = 36·4 for the symmetric fit and χ^2 (10 D.F.) = 30·8 for the asymmetric fits, indicating that the asymmetric fit is somewhat better. Both tests, however, show there are statistically significant discrepancies between the smoothed covariance function and the original data. We nevertheless prefer these reduced estimates because they are smoother and because the discrepancies between them and the original data matrix are small (less than 7% for both the symmetric and asymmetric fits).

We estimated the eigenvalues and eigenfunctions of $\hat{\mathscr{G}}$ using three different methods for comparison. First, we analysed the symmetric estimate with the algebraic method described by Kirkpatrick *et al.* (1990) using the reduced estimate with $k = 9$. Secondly, we carried

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

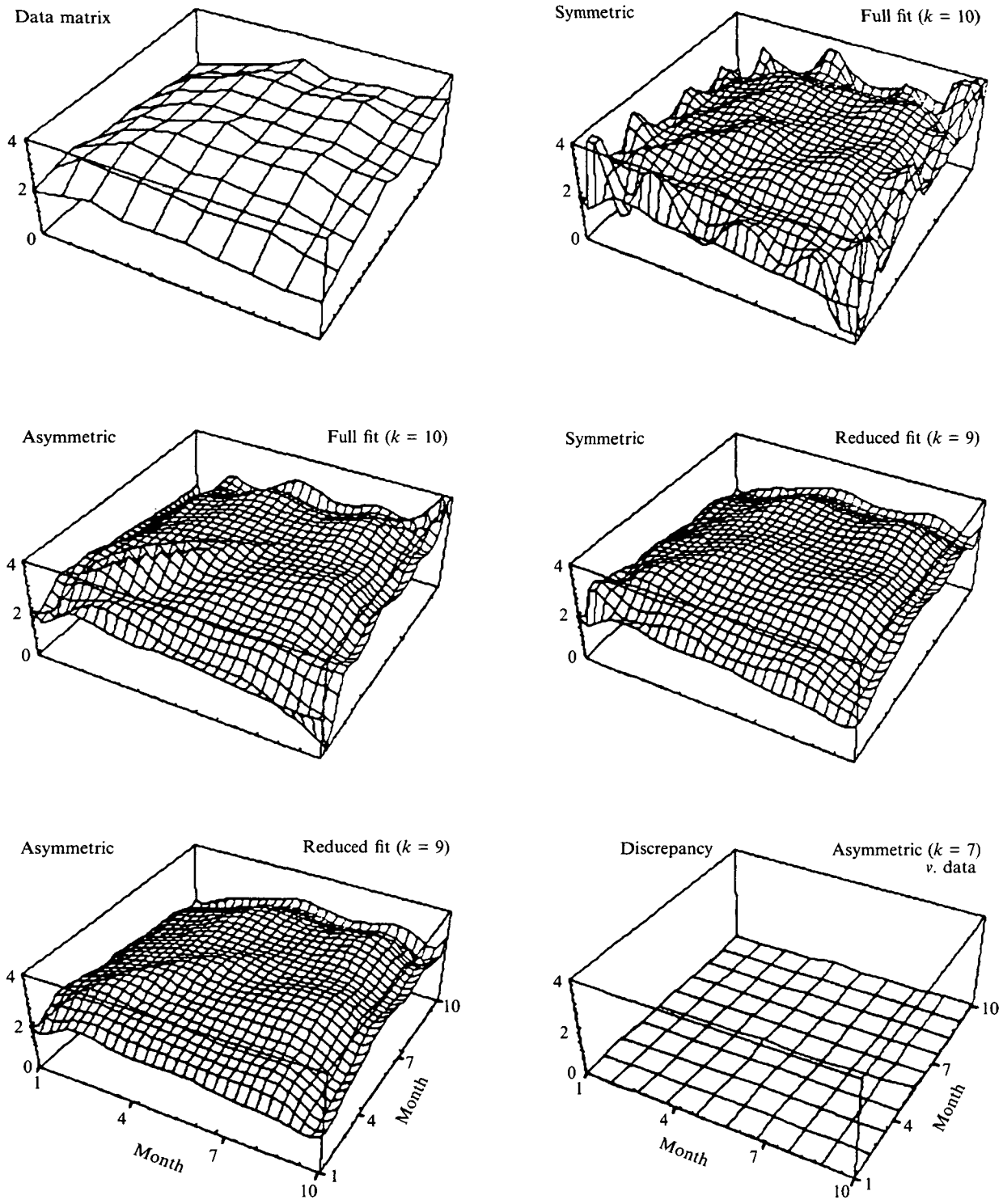

Fig. 4. Estimates of the additive genetic covariance function for lactation in British Holstein–Friesian dairy cattle. The original data are shown at top left. The full fit using symmetric coefficients (top right) overfits the data; the plot has been truncated in the vertical dimension. The full fit with asymmetric coefficients (middle left) is much smoother, but is poorly behaved in the off-diagonal corners. Reduced fits with $k = 9$ using symmetric coefficients (middle right) and asymmetric coefficients (bottom left) are much smoother. The discrepancy between the reduced asymmetric fit and the original data (bottom right) is very small everywhere. Variances and covariances are in units of kg^2.

out an analysis of the asymmetric estimate again with $k = 9$ using the numerical approach outlined above with a 31 × 31 matrix reconstructed from the estimated covariance function. Thirdly, we calculated the eigenvalues and eigenvectors of the original matrix $\hat{G}$. The eigenvalues from the last two methods were renormalized to make them comparable to those from the first method. (The eigenvalues of a covariance function are defined by an integration rather than a summation. Eigenvalues calculated by the second and third

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

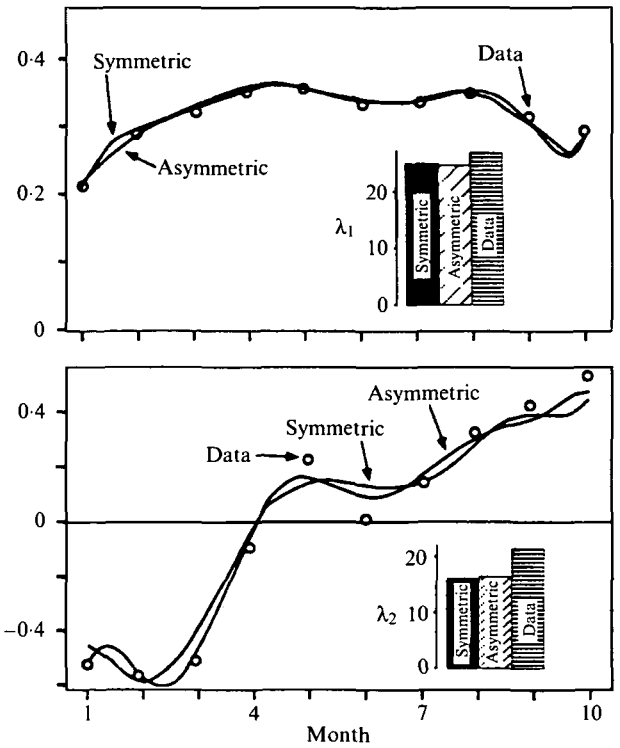

Fig. 5. Estimates of the first and second eigenfunctions (top and bottom panels, respectively) of additive genetic variation for lactation in British Holstein–Friesian dairy cattle. Each panel compares the estimates obtained via symmetric coefficients, asymmetric coefficients, and the corresponding eigenvector from the original additive genetic covariance matrix. Estimates using symmetric and asymmetric coefficients are reduced fits with $k = 9$ (see Fig. 4). Estimates of the corresponding eigenvalues are shown in the insets; note that estimates of λ_1 are about an order of magnitude greater than those for λ_2. Eigenvalues are in units of kg².

methods, in contrast, are defined by a matrix product that involves a summation whose value depends on the number of ages sampled. Renormalization accounts for the number of ages so that eigenvalues estimated by these different methods can be compared.)

The results are present in Fig. 5. The first eigenfunction is positive everywhere, showing that the principal axis of genetic variation corresponds to simultaneous increases or decreases in lactation at all ages. The leading eigenvalue shows that this eigenfunction accounts for about 90% of all additive genetic variation ($\lambda_1 = 228$ and the sum of all eigenvalues = 254, using the method of asymmetric coefficients). Thus there appears to be substantial genetic variation for enhanced milk production throughout the entire lactation period. A consequence of practical importance is that there are not strong tradeoffs between early and late lactation: genetic improvement at one age will tend to improve all ages. The second eigenfunction, which accounts for less than 8% of the genetic variation, shows a tradeoff between performance before *v.* after the fourth month of lactation.

All three methods give similar results in this example. Our experience with other examples, however, shows that this is not always so. It is likely that the agreement between the methods in this case results from the high precision of the parameter estimates in this very large data set and the relatively large number of measured ages. In other applications, we expect the infinite-dimensional methods will have substantially greater power than the conventional matrix-based ones (see Kirkpatrick & Heckman, 1989). Furthermore, the infinite-dimensional methods estimate the full eigenfunctions rather than a series of points along them.

7. Discussion

The methods developed here complement those developed earlier for estimating and analysing the structure of variation in traits that change with age. The technique of asymmetric coefficients may lead to smoother and more accurate estimates of the covariance and correlation functions whenever we are willing to allow there to be a crease (that is, a discontinuous first derivative) along its diagonal. The technique of extrapolating to the diagonal allows one to correct for biases that inflate the diagonal elements of a covariance matrix (the variances). The eigenvalues and eigenfunctions of covariance functions estimated by either method can be calculated numerically to reveal dominant components of variation and tradeoffs.

A question common to all of these methods is how to decide the appropriate degree of smoothing for the estimate of the covariance function. Kirkpatrick *et al.* (1990) developed a goodness-of-fit test, and suggested we choose the smoothest function (the one with the smallest number of orthogonal polynomials) that does not differ significantly from the original data matrix. Our analysis of phenotypic variation in lactation shows that this criterion is not always adequate. The function that it chooses may not be positive semi-definite, and therefore not qualify as a covariance function. Our preference in this case is for a smoother estimate of the covariance function that is better behaved (Fig. 3). Although it does differ significantly from the original data, the discrepancies are small.

An issue related to smoothing involves the positive definiteness of the covariance functions estimated by these methods. A covariance function, by definition, must be positive semidefinite. Estimates of covariance functions, like estimates of covariance matrices, can violate this requirement. Even if the matrix on which it is based is positive semi-definite, there is no certainty that the covariance function estimated by interpolating between the points of the matrix will be. The problem is illustrated in Fig. 3 by the estimates with $k = 9$. One might choose to use positive semi-definiteness as one of the criteria for choosing among estimates of the covariance function that differ in the

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

degree of smoothing. One might expect smoother fits generally to be less prone to violate positive semi-definiteness if the original data matrix does not.

Alternative methods for fitting functions might lead to better-behaved estimates of covariance functions, for example that are smoother, that conform better to the data, and that are less likely to violate the requirement of positive semidefiniteness. Polynomials, which are the basis for the estimates in this paper, are very often wiggly and can be poorly behaved when used for extrapolation. Other methods such as two dimensional splines are available (Lancaster & Salkauskas, 1986). They might lead to improved estimates.

One clear opportunity for alternative methods involves our method for extrapolating to the diagonal of the covariance function. The algorithm we developed produces a covariance function estimate that has a crease (a discontinuous first derivative) along its diagonal. Unfortunately, trajectories corresponding to covariance functions of this sort are not smooth: they are continuous, but do not have continuous derivatives (Soong, 1973, chapter 4). We usually expect on biological grounds that a growth process will be smooth. We therefore would prefer an estimate with continuous derivatives everywhere. It may be possible to extend the approach described here to cure this weakness in our method. In any event, we suspect that this extension typically would make only small changes to the quantitative results. The data analysed in Section 6 below suggest that further changes in the variance estimates produced by smoothing the diagonal crease will be small relative to the corrections made by the algorithm presented here.

Lactation in dairy cattle is an excellent candidate for infinite-dimensional analyses because changes in rate of production throughout lactation are of interest. We would like to maximize production over the whole lactation, and need to be able to predict lactation yield from a small number of records early in lactation, both at the phenotypic level so as to make early culling decisions, and at the genetic level to make early selection decisions. Previous analyses of lactation curves and yield prediction (reviewed by Danell, 1990) have not considered the underlying continuous covariance structure of the records. The deviation of an individual from the population mean at one or two early ages can be used to estimate its performance at any later age or set of ages by the standard methods of part-whole correlation (see, e.g. Falconer, 1989; VanRaden *et al.* 1991). Given the economic incentives, the relatively small amount of additional computation required by the infinite-dimensional method seems a small price to pay.

Our results show that allowance for inflation of the phenotypic variance by measurement error and date-specific effects (such as illness and weather) needs to be made in computing the underlying phenotypic covariance structure. Analyses of daily milk records have shown that almost all of the increase is associated with the variance of the daily record, but some residual effects span a few days. For example, the phenotypic correlations of records 1, 2, 10 and 30 days apart were 0·84, 0·82, 0·79 and 0·75 in a small data set (Pander *et al.* 1993). Test day records used in the present analysis were approximately 30 d apart, and the increase in diagonal elements of 30% or so (Fig. 3) correspond to these figures. In the repeatability model commonly used in the analysis of quantitative traits with multiple records (e.g. Falconer, 1989, chap. 8) it is assumed that $\mathscr{P}(t_1, t_2) = rV_P$ for all $t_1 \neq t_2$, and that $\mathscr{P}(t_1, t_1) = V_P$, where r is the repeatability. In our analysis we allow for this inflation of the variance but also for continuous changes in the covariance over the lactation.

The genetic analysis shows that, although there is substantial additive genetic variation, about 90% of it is associated with the first eigenfunction, which is positive at all ages (Fig. 5). Tradeoffs are seen in the second eigenfunction, which shows opposite effects on production before and after the fourth month of testing. This eigenfunctions, however, accounts for less than 7% of the genetic variation. The present analysis therefore formalizes what is known from examination of the genetic correlation matrix of test day records, which shows high positive values throughout (e.g. Pander *et al.* 1992), that selection on records from the first few months of production will have little negative consequences on performance in later lactation. A further development of the methods would involve defining joint covariance functions of yield of milk and, for example, of proportion of fat in the first lactation and (requiring more change in structure) of milk yield in the first and in later lactations.

Two directions for future work are suggested by this work. First, covariance functions would be better fitted assuming that the error structure of the estimated variances and covariances followed a multivariate Wishart distribution using likelihood to evaluate the fit. This approach requires numerical iteration. A covariance function estimate based on the methods from this paper (using least squares, assuming normally-distributed errors) would be a logical initial point for the iterations.

The ultimate extension would be a method in which the covariance function is estimated directly from the original observations, without the intermediate of a covariance matrix. In the analyses above, measurements records for individuals were grouped into 1-month categories for analysis. These pooled data were than analysed to give estimated covariance matrices, which in turn were analysed by the infinite-dimensional methods. A more direct approach would avoid the pooling entirely and instead treat each record according to the individual's actual age (or number of days since lactation began). Such a method might well increase the precision of covariance function estimates.

We thank Nick Barton and two anonymous referees for comments. This work was supported by NIH grant 45226-01 and NSF grant 9107140 to M.K.

Appendix A

The appendix describes the method of asymmetric coefficients. Programs for this analysis have been implemented in a *Mathematica*® notebook that is available on request from the senior author.

Here we will follow Kirkpatrick *et al.* (1990) by fitting orthogonal functions. These functions, denoted ϕ_i ($i = 0, 1, \ldots$), are orthogonal over the interval $[u, v]$. In the examples discussed in the text and below we use Legendre polynomials, in which case $u = -1$ and $v = 1$.

The method of asymmetric coefficients then proceeds according to the following seven steps.

(i) Form the vector $\boldsymbol{p}$ by stacking the successive columns of the lower left diagonal part of the phenotypic covariance matrix:

$$\boldsymbol{p} = (P_{11}, P_{21}, \ldots, P_{n1}, P_{22}, P_{32}, \ldots, P_{n2}, \ldots, P_{nn})^{\mathrm{T}}. \quad \text{(A 1)}$$

(ii) Form the coefficient vector $\boldsymbol{c}$ by stacking the successive columns of *upper left diagonal* part of the matrix $\boldsymbol{C}$:

$$\boldsymbol{c} = (C_{00}, C_{10}, \ldots, C_{k-1,0}, C_{01}, C_{11}, \ldots, C_{k-1,1}, \ldots, C_{0,k-1})^{\mathrm{T}}. \quad \text{(A 2)}$$

This contrasts with the method of symmetric coefficients, in which the coefficient vector is formed from the lower diagonal parts of the columns of $\boldsymbol{C}$, in the same way that $\boldsymbol{p}$ is formed according to eqn (A 1). Note that the subscripts run from 0 to $k-1$ rather from 1 to k in order to conform with the conventional numbering of the orthogonal functions.

(iii) The estimated phenotypic covariance matrix $\hat{\boldsymbol{P}}$ is based on measurements at n ages; these ages form the age vector $\boldsymbol{a}$. We use $\boldsymbol{a}$ to calculate the adjusted age vector $\boldsymbol{a}^*$:

$$a_i^* = u + \frac{(v-u)(a_i - a_1)}{a_n - a_1}, \quad \text{(A 3)}$$

$i = 1, 2, \ldots, n$. This operation rescales the range of the measured ages to the range of the orthogonal functions.

(iv) The next step is to form the matrix $\boldsymbol{X}$ from the orthogonal functions. The way in which the vectors $\boldsymbol{p}$ and $\boldsymbol{c}$ are formed makes the notation for specifying the elements of $\boldsymbol{X}$ somewhat awkward. We begin by defining four 'index functions', which are integer-valued functions that generate appropriate subscripts for the orthogonal functions and the adjusted age vector. The first two index functions are used to generate the subscripts for the matrix $\boldsymbol{C}$ as they appear on the right hand side of (A 2). The index function $\mathrm{I}_1(i,k)$ is based on the sequence

$$0, 1, \ldots, k-1, 0, 1, \ldots, k-2, \ldots, 0. \quad \text{(A 4)}$$

The value of $\mathrm{I}_1(i,k)$ is given by the ith element of (A 4). For example, $\mathrm{I}_1(5,3) = 1$. The function $\mathrm{I}_2(i,k)$ is based on the sequence

$$0, 0, \ldots, 0, 1, 1, \ldots, 1, \ldots, k-1, \quad \text{(A 5)}$$

in which there are k 0s, $(k-1)$ 1s, etc. The value of $\mathrm{I}_2(i,k)$ is given by the ith element of A 5; thus $\mathrm{I}_2(6,3) = 2$.

The third index function is $\mathrm{I}_3(i,k)$, and is based on the sequence

$$1, 2, \ldots, k, 2, 3, \ldots, k, \ldots, k. \quad \text{(A 6)}$$

The value of $\mathrm{I}_3(i,k)$ is given by the ith element of (A 6). For example, $\mathrm{I}_3(4,3) = 2$. The last two index functions are simply an incremented and decremented version of I_2 and I_3:

$$\mathrm{I}_4(i,k) = \mathrm{I}_2(i,k) + 1, \quad \text{(A 7)}$$

$$\mathrm{I}_5(j,k) = \mathrm{I}_3(j,k) - 1. \quad \text{(A 8)}$$

For all five index functions, the argument i can take on the values $i = 1, 2, \ldots, k(k+1)/2$.

With these definitions in hand, we now calculate the matrix $\boldsymbol{X}$:

$$X_{ij} = \phi_{\mathrm{I}_1(j,k)}(a^*_{\mathrm{I}_3(i,n)})\,\phi_{\mathrm{I}_2(j,k)}(a^*_{\mathrm{I}_4(i,n)}), \quad \text{(A 9)}$$

where $i = 1, 2, \ldots, n(n+1)/2$ and $j = 1, 2, \ldots, k(k+1)/2$. By way of comparison, the previous method of symmetric coefficients calls for $\boldsymbol{X}$ to be of the form

$$X_{ij} = \{\phi_{\mathrm{I}_5(j,k)}(a^*_{\mathrm{I}_3(i,n)})\,\phi_{\mathrm{I}_2(j,k)}(a^*_{\mathrm{I}_4(i,n)}) + \phi_{\mathrm{I}_2(j,k)}(a^*_{\mathrm{I}_3(i,n)})\,\phi_{\mathrm{I}_5(j,k)}(a^*_{\mathrm{I}_4(i,n)})\}/\{1 + \delta[\mathrm{I}_2(j,k), \mathrm{I}_5(j,k)]\}, \quad \text{(A 10)}$$

where $\delta[s,t] = 1$ if $s = t$ and is 0 otherwise.

(v) For a full estimate of $\mathscr{P}$, we solve for $\boldsymbol{c}$ using the relation

$$\boldsymbol{c} = \boldsymbol{X}^{-1}\boldsymbol{p} \quad \text{(A 11)}$$

Alternatively, we may be interested in a reduced estimate of $\mathscr{P}$, in which the number of orthogonal coefficients that are fit is smaller than the number of observations ($k < n$). To do so, we first calculate the matrix $\boldsymbol{V}$ that is the error covariance matrix corresponding to the vector $\boldsymbol{p}$; details are given in Kirkpatrick *et al.* (1990). We than calculate $\boldsymbol{c}$ using weighted least squares:

$$\boldsymbol{c} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{p}. \quad \text{(A 12)}$$

(vi) We now form the estimated coefficient matrix $\hat{\boldsymbol{C}}$ by unstacking the vector $\boldsymbol{c}$ (that is, performing the reverse of the operation described by eqn (A 2)). Using the notation of the index functions, the elements of $\hat{\boldsymbol{C}}$ are found by plugging successive values of i into the relation

$$\hat{C}_{[\mathrm{I}_1(i,k),\,\mathrm{I}_2(i,k)]} = c_i, \quad \text{(A 13)}$$

where $i = 1, 2, \ldots, k(k-1)/2$; all other elements of $\hat{\boldsymbol{C}}$ are 0.

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

(vii) The coefficient matrix generates our estimate for the covariance function:

$$\hat{\mathscr{P}}(t_1,t_2)=\begin{cases}\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}C_{ij}\,\phi_i(t_1^*)\,\phi_j(t_2^*), & t_1\geqslant t_2\\ \sum_{i=0}^{k-1}\sum_{j=0}^{k-1}C_{ij}\,\phi_i(t_2^*)\,\phi_j(t_1^*), & t_1<t_2,\end{cases} \quad \text{(A 14)}$$

where

$$t_i^*=u+\frac{(v-u)(t_i-a_1)}{a_n-a_1}. \quad \text{(A 15)}$$

The form of eqn (A 14) guarantees that $\hat{\mathscr{P}}$ is symmetric, as required by one of the defining criteria of covariance functions. In the case of a reduced fit ($k<n$), the consistency with the original data matrix can be tested for statistically significant deviation from the data using the goodness-of-fit test described in Kirkpatrick *et al.* (1990, Appendix C).

(i) *A worked example*

The method will be illustrated by the worked example presented by the covariance matrix of text eqn (5). We will use Legendre polynomials for the fit, which are defined over the interval $[-1,1]$, so $u=-1$ and $v=1$.

Following Step (i) above, stack successive columns of the subdiagonal part of $\hat{\boldsymbol{P}}$ to form the vector

$$\boldsymbol{p}=(3,2,3)^{\mathrm{T}}. \quad \text{(A 16)}$$

From Step (ii), the (unknown) vector of coefficients is

$$\boldsymbol{c}=(C_{00},C_{10},C_{01})^{\mathrm{T}}. \quad \text{(A 17)}$$

From Step (iii), the adjusted age vector is

$$\boldsymbol{a}^*=(-1,1)^{\mathrm{T}}. \quad \text{(A 18)}$$

We form the matrix $\boldsymbol{X}$ from the orthogonal functions as described in Step (iv):

$$\boldsymbol{X}\begin{pmatrix}[\phi_0(a_1^*)\ \ \phi_0(a_1^*)] & [\phi_1(a_1^*)\ \ \phi_0(a_1^*)] & [\phi_0(a_1^*)\ \ \phi_1(a_1^*)]\\ [\phi_0(a_2^*)\ \ \phi_0(a_1^*)] & [\phi_1(a_2^*)\ \ \phi_0(a_1^*)] & [\phi_0(a_2^*)\ \ \phi_1(a_1^*)]\\ [\phi_0(a_2^*)\ \ \phi_0(a_2^*)] & [\phi_1(a_2^*)\ \ \phi_0(a_2^*)] & [\phi_0(a_2^*)\ \ \phi_1(a_2^*)]\end{pmatrix}$$

$$=\begin{pmatrix}\frac{1}{2} & \left(-\frac{\sqrt{3}}{2}\right) & \left(-\frac{\sqrt{3}}{2}\right)\\ \frac{1}{2} & \frac{\sqrt{3}}{2} & \left(-\frac{\sqrt{3}}{2}\right)\\ \frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2}\end{pmatrix}. \quad \text{(A 19)}$$

The orthogonal coefficients for the full fit are calculated as described in Step (v):

$$\boldsymbol{c}=\boldsymbol{X}^{-1}\boldsymbol{p}=(6,-\sqrt{3}/3,\sqrt{3}/3)^{\mathrm{T}}. \quad \text{(A 20)}$$

By unstacking this vector according to Step (vi), we find the estimated coefficients matrix given by text eqn (7). Finally, using that result in Step (vii) we arrive at the estimated covariance function given by text eqn (8).

Appendix B

The technique of extrapolating to the diagonal makes use of the asymmetric coefficient fit described in Appendix A. Programs that run this analysis have been developed in a *Mathematica*® notebook that is available from the senior author on request.

As with the earlier methods, we are interested in fitting the data using k orthogonal functions. But because we are not using the diagonal elements of $\hat{\boldsymbol{P}}$ in the fit, we now require $k<n$ rather than $k\leqslant n$. The algorithm proceeds by the following steps.

(i) Form the vector $\boldsymbol{p}$ by stacking the *sub*diagonal parts of the columns of $\hat{\boldsymbol{P}}$:

$$\boldsymbol{p}=(P_{21},P_{31},\ldots,P_{n1},P_{32},P_{42},\ldots,P_{n2},\ldots,P_{n,n-1})^{\mathrm{T}}. \quad \text{(B 1)}$$

This vector is of length $n(n-1)/2$.

(ii) Form the coefficient vector $\boldsymbol{c}$ according to eqn (A 2).

(iii) We calculate two adjusted age vectors $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$:

$$a_i^*=u+\frac{(v-u)(a_{i+1}-a_2)}{a_n-a_2}, \quad \text{(B 2}a\text{)}$$

and

$$b_i^*=u+\frac{(v-u)(a_i-a_1)}{a_{n-1}-a_1}, \quad \text{(B 2}b\text{)}$$

$i=1,2,\ldots,n-1$.

(iv) Form the matrix $\boldsymbol{X}$:

$$X_{ij}=\phi_{\mathrm{I}_1(j,k)}(a^*_{\mathrm{I}_6(i,n)})\,\phi_{\mathrm{I}_2(j,k)}(b^*_{\mathrm{I}_4(i,n-1)}), \quad \text{(B 3)}$$

where

$$\mathrm{I}_6(i,n)=\mathrm{I}_3(i,n-1)+1, \quad \text{(B 4)}$$

$i=1,2,\ldots,n(n-1)/2$, and $j=1,2,\ldots,k(k+1)/2$.

(v) As in the previous section, the coefficients are calculated using eqn (A 11) for a full estimate, or eqn (A 12) for a reduced estimate. Note, however, that because only the off-diagonal elements of $\hat{\boldsymbol{P}}$ are being used, a full fit implies $k=n-1$ rather than $k=n$ polynomials (and likewise a reduced fit implies $k<n-1$). A reduced fit can be tested for consistency with the original data using the goodness-of-fit test described in Kirkpatrick *et al.* (1990, Appendix C). When extrapolating to the diagonal, however, the values along the diagonal are omitted from the test.

(vi) The coefficient matrix $\hat{\boldsymbol{C}}$ is formed from the vector $\boldsymbol{c}$ via eqn (A 13).

(vii) Finally, the estimated covariance function is obtained by substituting $\hat{\boldsymbol{C}}$ into eqn (A 14) using

$$t_1^*=u+\frac{(v-u)(t_1-a_2)}{a_n-a_2} \quad \text{(B 5}a\text{)}$$

https://doi.org/10.1017/S0016672300032559 Published online by Cambridge University Press

and

$$t_2^* = u + \frac{(v-u)(t_2-a_1)}{a_{n-1}-a_1}, \qquad \text{(B 5}b\text{)}$$

where t_1 and t_2 range over the interval $[a_1, a_2]$.

(i) *A worked example*

We will demonstrate the method of extrapolating to the diagonal using the covariance matrix given in text eqn (11), based on measurements taken at the ages

$$\boldsymbol{a} = (10, 11, 15)^{\mathrm{T}}.$$

(This age vector will illustrate how the infinite-dimensional method naturally accommodates unequal spacing of ages.) We will calculate a full estimate of $\mathscr{P}$ (that is $k = n-1 = 2$), again using Legendre polynomials.

The vectors $\boldsymbol{p}$, $\boldsymbol{c}$, $\boldsymbol{a}^*$, and $\boldsymbol{b}^*$ are

$$\boldsymbol{p} = (3, 2, 3)^{\mathrm{T}},$$

$$\boldsymbol{c} = (C_{00}, C_{10}, C_{01})^{\mathrm{T}},$$

and

$$\boldsymbol{a}^* = \boldsymbol{b}^* = (-1, 1)^{\mathrm{T}}.$$

Steps (iii)–(v) produce the same values for $\boldsymbol{X}$, $\boldsymbol{c}$, and $\hat{\boldsymbol{C}}$ found in the example of Section 2. Last, we follow Step (vi) to obtain the estimate of the covariance function given by text eqn (12).

References

Danell, B. (1990). Genetic aspects of different parts of the lactation. *Proceedings of the 4th World Congress on Genetic Applications to Livestock Production*, Edinburgh, **14**, 114–117.

Falconer, D. (1989). *Introduction to Quantitative Genetics*, 3rd edition. New York: Longman.

Gomulkiewicz, R. & Kirkpatrick, M. (1992). Quantitative genetics and the evolution of reaction norms. *Evolution* **46**, 390–411.

Kirkpatrick, M. & Heckman, N. (1989). A quantitative genetic model for growth, shape and other infinite-dimensional characters. *Journal of Mathematical Biology* **27**, 429–450.

Kirkpatrick, M., Lofsvold, D. & Bulmer, M. (1990). Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* **124**, 979–993.

Kirkpatrick, M. & Lofsvold, D. (1992). Measuring selection and constraint in the evolution of growth. *Evolution* **46**, 954–971.

Lancaster, P. & Salkauskas, K. (1986). *Curve and Surface Fitting: An Introduction*. London: Academic Press.

Meyer, K. (1985). Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics* **41**, 153–165.

Pander, B. L., Hill, W. G. & Thompson, R. (1992). Genetic parameters of test day records of British Holstein–Friesian heifers. *Animal Production* **55**, 11–21.

Pander, B. L., Thompson, R. & Hill, W. G. (1993). The effect of increasing the interval between recordings on genetic parameters of test day yields of British Holstein–Friesian heifers. *Animal Production* **56**, 159–164.

Pander, B. L., Thompson, R. & Hill, W. G. (1993). Phenotypic correlations among daily records of milk yields. *Indian Journal of Animal Sciences* **63**, 1282–1286.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545–554.

Soong, T. T. (1973). *Random Differential Equations in Science and Engineering*. New York: Academic Press.

VanRaden, P. M., Wiggans, G. R. & Ernst, C. A. (1991). Expansion of projected lactation yield to stabilize genetic variance. *Journal of Dairy Science* **74**, 4344–4349.

Wolfram, S. (1991). *Mathematica: A System for Doing Mathematics by Computer*, Second edition. Redwood City: Addison-Wesley.