

## Advances in methodology for random regression analyses

*K. Meyer*

Animal Genetics and Breeding Unit (a joint venture between NSW Department of Primary Industries and the University of New England), University of New England, Armidale, NSW 2351, Australia.  
Email: kmeyer@didgeridoo.une.edu.au

*Abstract.* Random regression analyses have become standard methodology for the analysis of traits with repeated records that are thought of as representing points on a trajectory. Modelling curves as a regression on functions of a continuous covariable, such as time, for each individual, random regression models are readily implemented in standard, linear mixed model analyses. Early applications have made extensive use of regressions on orthogonal polynomials. Recently, spline functions have been considered as an alternative. The use of a particular type of spline function, the so-called B-splines, as basis functions for random regression analyses is outlined, emphasising the local influence of individual observations and low degree of polynomials employed. While such analyses are likely to involve more regression coefficients than polynomial models, it is demonstrated that reduced rank estimation via the leading principal components is feasible and likely to yield more parsimonious models and more stable estimates than full rank analyses. The combined application of B-spline basis function and reduced rank estimation is illustrated for a small set of data for beef cattle.

*Additional keywords:* B-spline functions, reduced rank estimation, principal components.

### Introduction

Over the last decade, analyses fitting the so-called ‘random regression’ (RR) model have become a standard procedure for the analyses of data from livestock recording schemes, where the traits of interest are recorded repeatedly and, along with their covariances, are changing along some continuous scale. Typical examples are records for test-day production taken over the course of lactation of dairy cows, and weights of animals, taken at several ages. Generally, the resulting lactation or growth curves are modelled through higher order polynomials. While being conceptually appealing and straightforward, practical applications of RR analyses have been plagued by problems associated with large numbers of parameters to be estimated, poor polynomial approximation, implausible estimates at the extremes, and high computational requirements.

This paper advocates the use of B-spline functions as a more robust alternative to polynomials for modelling curves in RR analyses. Furthermore, a reparameterisation of RR models to estimate the leading principal components of curves directly and to obtain reduced rank estimates of the covariance matrices for RR coefficients is proposed. This is shown to be advantageous, yielding more parsimonious models and stable estimates and requiring considerably fewer computations than full rank analyses. The combined application of these techniques is illustrated for a set of beef cattle data.

### Background

Records for traits that are measured ‘repeatedly’ per individual along some continuous scale can be thought of as representing points on a curve. Most commonly this continuous covariable, henceforth called the ‘control variable’, is time or age, but scenarios involving other kinds of continuous scale, such as size, weight or spatial measurements or environmental variables, for example temperature, are readily conceived. The underlying curve is then modelled by a mathematical function, and such traits are therefore commonly referred to as ‘function-valued’ (FV) traits.

In principle, many types of functions are suitable to model these curves, including ‘standard’ growth curves such as the Gompertz curve, which may be exponential. However, by restricting the functions considered to the class represented by a regression on functions of the control variable, the resulting curves can be modelled within the linear, mixed model framework commonly employed in quantitative genetic analyses, and standard methods of estimation, such as restricted maximum likelihood (REML), can be employed. While the curves are generally not ‘linear’ (i.e. not straight lines), they are linear functions of the regression coefficients (i.e. linear in the parameters to be estimated). The shape of the curves is then determined by the shape of the functions of the control variable, which are fitted as covariables.

The analysis of FV traits is thus a straightforward extension of the standard analysis of repeated measures data, which consider traits to be ‘points’ rather than curves. In standard analyses, we model genetic and other random effects by regressing observations on indicator variables, which have values of unity or zero for individuals that do or do not have a record, respectively, for the trait. For RR analyses, the indicator variables are replaced by functions of the control variable, and we fit a set of RR coefficients for each individual and source of variation, rather than individual effects. This yields subject-specific estimates of trajectories, such as growth or lactation curves. Similarly, changes in mean can be accommodated by fitting corresponding or higher order regression curves as fixed effects.

The functions of the control variable, fitted as covariables in RR analyses, are commonly referred to as basis functions. The underlying idea is that any trajectory can be represented as a linear combination of the basis functions. Theoretically, this may involve infinitely many terms. In practice, however, a truncated expansion involving relatively few terms generally suffices. There are many suitable basis functions. Kirkpatrick and Heckman (1989) suggested the use of orthogonal polynomials, preferring them over trigonometric series. In particular, Legendre polynomials, as employed by Kirkpatrick *et al.* (1990), have been the most common choice for RR analyses of data from livestock improvement programs.

RR models allow for continual changes in both mean and covariances of traits with changes of the control variable. Covariances between measurements at any two points along the trajectory are given by a ‘covariance function’ (CF). This involves the basis functions, evaluated for the 2 points of interest, and some coefficients. Like the trajectories, CFs have conceptually infinitely many terms, but good estimates can generally be obtained by an approximation involving a relatively small number of basis functions (Kirkpatrick and Heckman 1989). Conveniently, the necessary coefficients are the covariances among regression coefficients, which can be estimated in a RR analysis (Meyer 1998). Further details are outlined in a recent review of the quantitative genetics of FV traits by Meyer and Kirkpatrick (2005b).

RR models provide an appealing and conceptually simple, albeit powerful framework for the analysis of ‘repeated records’ from livestock improvement programs with desirable properties. Each record is contributing information at the value of the control variable at which it is measured, alleviating the need for, sometimes arbitrary or inappropriate, corrections for differences in the latter. Any information implicit in the order and spacing of records is utilised. Moreover, covariances for individual points are modelled more appropriately. In the context of genetic evaluation, this implies that data are used more efficiently, and that estimates of breeding values have higher accuracies. In terms of estimation of variance components, RR models facilitate parsimonious description of changing and potentially complex covariance structures.

### B-splines to model trajectories

To date, practical applications of RR models have commonly fitted orthogonal (Legendre) polynomials of the control variable as covariables. Higher order polynomials are flexible and have been found to be capable of modelling changes in means and variances along a continuous scale well. However, polynomials often put a high emphasis on observations at the extremes, and are known to be potentially problematic for high orders of fit. ‘Runge’s phenomenon’ describes the observation that the error of polynomial approximation of a curve increased with the order of polynomial fit, with errors predominantly due to oscillations at the extremes of the curve (e.g. de Boor 2001). Such behaviour may have contributed to problems, evident through implausible estimates of variance components at ends of the range of the control variable, observed in a number of instances. In particular, this occurred often for data with substantially different numbers of observations at the 2 extremes of the control variable, and higher orders of polynomial fit.

An alternative to high degree polynomials are ‘piece-wise polynomials’, i.e. curves constructed from pieces of lower degree polynomials, so-called splines, joined smoothly at selected points, the so-called knots. The name spline originates from the long thin, flexible strip of wood traditionally used in drawing curves — when fixed at given points (‘knots’), the wood takes the shape which minimises the energy required for bending, thus yielding the smoothest shape possible. Splines are readily fitted within mixed model analyses (Ruppert *et al.* 2003; Verbyla *et al.* 1999). A RR analysis of test day records of dairy cows using splines has been presented by White *et al.* (1999).

A particular type of spline curve is the so-called B-spline (de Boor 2001). B-splines are often preferred to other types of splines due to their good numerical properties (Ruppert *et al.* 2003). Rice and Wu (2001) suggested the use of the B-spline basis functions to model random effects curves in a mixed model analysis, and demonstrated their efficacy in estimating covariance functions. Other applications, also at the phenotypic level, have been described by Shi *et al.* (1996) and James *et al.* (2000). Torres and Quaas (2001) performed a genetic RR analysis using B-splines to model test day records of dairy cows, and Meyer (2005a, 2005b) presented applications to the RR analysis of growth records of beef cattle.

#### *Splines in general*

Consider the simple scenario of a spline function consisting of linear segments. This is commonly described as a ‘broken-stick’ curve, and is a straightforward extension of the parametric, linear regression. Let  $y_i$  denote observations recorded at values  $t_i$  of the control variable, and assume we have knots  $T_k$ . In fitting a linear spline, we then assume a non-parametric regression model:

$$y_i = \beta_0 + \beta_1 t_i + \sum_k \beta_{1k} (t_i - T_k)_+ + \varepsilon_i \quad (1)$$

for  $y_i$ , with  $\beta_0$  and  $\beta_{1k}$  denoting the intercept and linear regression coefficients, respectively,  $\varepsilon_i$  is the residual error pertaining to  $y_i$ , and  $(x)_+ = \min(0, x)$  equal to  $x$  if  $x$  is positive and equal to 0 otherwise (Ruppert *et al.* 2003). This gives a slope of  $\beta_1$  for the first segment, with  $t_i < T_1$ , a slope of  $\beta_1 + \beta_{11}$  for the second segment with  $T_1 \leq t_i < T_2$ , and a slope of  $\beta_1 + \sum_{k=1}^m \beta_{1k}$  for the segment bordered by  $T_m$  and  $T_{m+1}$ . For splines consisting of polynomial segments of degree  $p$ , (equation 1) is expanded to:

$$y_i = \beta_0 + \beta_1 t_i + \dots + \beta_p t_i^p + \sum_k \beta_{pk} (t_i - T_k)_+^p + \varepsilon_i \quad (2)$$

where  $\beta_p$  are the  $p$ -th degree regression coefficient. Terms  $(x)_+^p$  are known as truncated power functions, i.e. equation 2 gives a spline with truncated power basis functions of degree  $p$ .

Fitting spline functions as shown above can yield ‘wiggly’ estimates of curves, especially if  $p$  is low and if there are many knots, and the choice of knots can be more influential than desired. The ‘roughness’ of the estimate can be reduced by imposing a penalty, which reduces the influence of the regression coefficients, and thus yields a smoother curve. A common choice for ‘smoothing splines’ involve penalising the second derivatives of the spline function (Green and Silverman 1994). As a simpler alternative, Ruppert *et al.* (2003) suggested to constrain the sum of squared regression coefficients  $\beta_{pk}$  in (eqn 2), imposing a penalty of  $\lambda \sum_k \beta_{pk}^2$  added to the criterion to be minimised in least-squares or maximum likelihood estimation, to smooth the resulting curve. The smoothing parameter,  $\lambda$ , governs the degree of smoothing, small values yielding curves close to the unconstrained fit, and large values resulting in estimates more similar to the corresponding parametric regression. With smoothing, choice of knots is less crucial than for  $\lambda = 0$ .

Such penalised splines are readily accommodated in a mixed model framework. In essence, imposing a roughness penalty yields a system of equations where the penalised regression coefficients are estimated as random effects, i.e. shrinking them towards their mean, while the unpenalised coefficients are treated as fixed effects. The shrinkage is determined by  $\lambda$ , which is equivalent to the ratio of error variance and variance due to the regression coefficients. This implies that a suitable value of  $\lambda$  can be estimated from the data in a REML analysis. A widely used spline in mixed model analyses is the cubic smoothing spline, i.e.  $p = 3$ , considered in depth by Green and Silverman (1994) and Verbyla *et al.* (1999).

Spline curves can be integrated with other fixed and random effects in a mixed model. For longitudinal data, Ruppert *et al.* (2003) and Durbán *et al.* (2005) considered a model comprising subject-specific curves, fitting a linear, penalised spline for the overall curve and corresponding, random curves for the deviations for each individual. Similarly, Verbyla *et al.* (1999) proposed a model of analysis which included 3 subject- and time-specific terms for each

individual, namely the intercept and linear term and the random spline deviation. In an application to dairy data, White *et al.* (1999) showed that a mixed model analysis fitting splines for each animal is equivalent to a RR analysis. However, the matrix of covariances among RR coefficients is assumed to be highly structured. While intercept and slope are assumed correlated and have different variance, all random spline terms are assumed to have the same variance and to be independently distributed. This results in a total of only 4 variance components due to the RR to be estimated.

### B-splines

Alternative sets of basis functions to the truncated power base considered above exist. A common choice are B-spline functions, which yield equivalent fits, but have better numerical properties (Eilers and Marx 2005; Ruppert *et al.* 2003). Here, the ‘B’ stands for basis (de Boor 2001). The basis functions of B-splines comprise a set of overlapping, smooth and non-negative functions, which are unimodal and sum to unity. Like the truncated power basis, they can have different degree  $p$ .

B-spline functions can be defined recursively. Basis functions of degree  $p = 0$  have values of unity for all points in a given interval, and zero otherwise. For the  $k$ -th interval given by knots  $T_k$  and  $T_{k+1}$  with  $T_k < T_{k+1}$ ,

$$B_{k,0}(t) = \begin{cases} 1 & \text{if } T_k \leq t < T_{k+1} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Higher degree basis functions,  $B_{k,p}$  for  $p > 0$ , are then determined from the values of the lower degree basis functions and the width of the adjoining intervals between knots. The general relationship is:

$$B_{k,p}(t) = \frac{t - T_k}{T_{k+p} - T_k} B_{k,p-1}(t) + \frac{T_{k+p+1} - t}{T_{k+p+1} - T_{k+1}} B_{k+1,p-1}(t) \quad (4)$$

There are  $n + p$  basis functions for a B-spline of degree  $p$  with  $n$  intervals. For each  $B_{k,p}(t)$  in (equation 4), there are a limited number of non-zero functions for degree  $p - 1$  to be considered. Efficient strategies to evaluate the basis functions recursively exist and are described in the relevant literature (e.g. de Boor 2001). Alternatively, for equally spaced knots, B-spline functions can be obtained as the difference between splines with a basis of truncated power functions; see Eilers and Marx (2005) for details. These functions (equation 4) can be used as basis functions in a RR analysis, yielding estimates of subject-specific trajectories which are B-spline curves and corresponding estimates of CF which are tensor products of splines (Rice and Wu 2001).

Figure 1 shows the values of spline basis functions for a continuous scale divided into 5 equal intervals, i.e. involving 6 knots (4 interior at 0.2, ..., 0.8 and 2 exterior at 0 and 1), together with the first 6 Legendre polynomials (not normalised). Legendre polynomials are essentially non-zero

for the whole range, i.e. when fitting a polynomial regression each point contributes information to the complete curve to be estimated. In other words, individual observations have a ‘global’ influence. As shown in Figure 1, cubic, quartic and quintic Legendre polynomials have relatively large values close to the upper limit of the range of the control variable, which may offer some explanation for implausibly high estimates of variances at the upper extremes encountered in RR analyses fitting cubic or higher order Legendre polynomials.

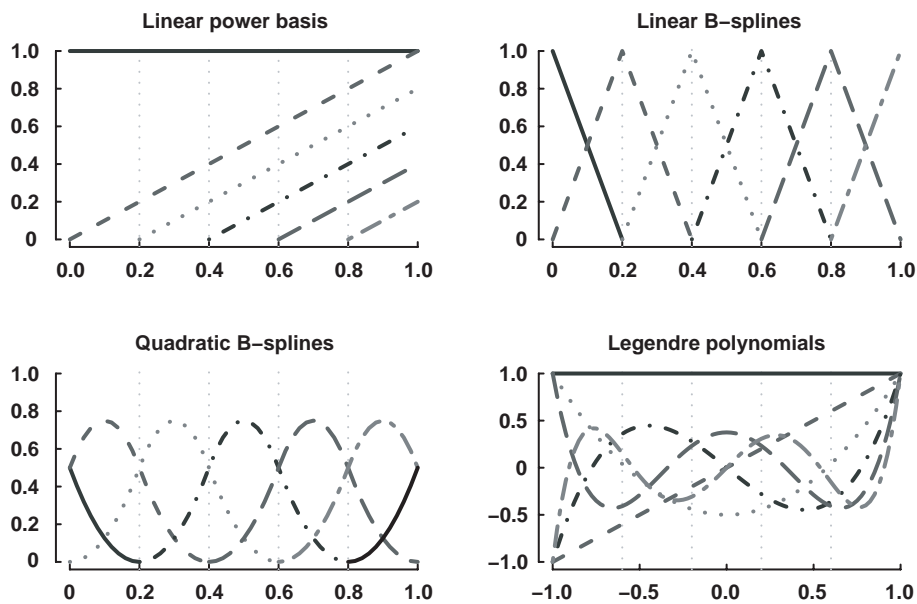
In contrast, B-splines of degree  $p$  are non-zero for at most  $p + 1$  adjacent intervals, resulting in a ‘local’ influence of individual observations only. For  $p = 0$  (not shown), B-spline functions correspond to weights in a ‘standard’ multivariate analysis, considering observations in each interval to be a different trait. Hence, RR analyses fitting B-splines of higher degree can be thought of as derived from a multivariate scenario, but with ‘blending’ of information across intervals or traits, where the degree of blending increases with  $p$ . Moreover, for equal intervals, all functions are of the same shape, resulting in even distribution of importance of individual points.

Calculation of the B-spline functions via equations 3 and 4 requires  $2p$  additional knots to be specified. Values shown in Figure 1 were obtained by adding  $p$  equally spaced knots either end of the observed range (i.e. at  $-0.2$  and  $1.2$  for  $p = 1$  and at  $-0.4, -0.2, 1.2$  and  $1.4$  for  $p = 2$ ). This resulted in individual functions simply being horizontally shifted copies of each other. Alternatively, we might have defined the external knots (0 and 1) to have multiplicity  $p + 1$ , resulting in somewhat different shapes of functions spanning the external knots (Eilers and Marx 2005).

### B-splines in mixed models

Random regression analyses can employ B-splines as basis functions. Fitting a B-spline of degree  $p$  with  $n$  intervals for a random effect requires  $n + p$  RR coefficients to be fitted. However, as each basis function contributes only to (at most)  $p + 1$  intervals, the respective design matrices in a mixed model analysis are considerably sparser (having  $\leq p + 1$  non-zero elements per row) than for corresponding RR analyses fitting an equal number of coefficients but using polynomial basis functions. Thus, while RR analyses fitting B-spline basis functions generally involve more RR coefficients than analyses fitting a polynomial regression, additional computational requirements owing to a higher order of fit are, in part at least, off-set by a greater sparsity of the mixed model equations.

Like splines with a truncated power basis, B-splines can be penalised to obtain smoother curves. The penalty for the former, described above, resulted in a formulation analogous to a ridge regression. Corresponding penalties for a B-spline basis, suggested by Eilers and Marx (1996), are based on differences between adjacent B-spline coefficients. The degree of fit and order of difference for the penalty can be chosen independently. Eilers and Marx (2005) showed that penalties based on second order differences are a good approximation to the more commonly used roughness penalties on second derivatives used in fitting smoothing splines (Green and Silverman 1994). The resulting functions are referred to as ‘P-splines’. Nomenclature here is not unequivocal though, other authors simply use P-splines to denote any kind of penalised splines. Again, estimation of penalised B-splines is readily integrated in a mixed model



**Figure 1.** Values of covariables (non-zero values only shown) for random regression analyses fitting splines with 5 equal intervals, and fitting Legendre polynomials.

framework of analysis, treating the penalised spline coefficients as random effects Eilers (1999). Interpreting this as a RR analysis, the definition of the penalty again imposes a fairly rigid structure on the covariance matrix among RR coefficients.

Penalised splines have been conceived with smoothing of fixed trajectories in mind, and perform best with a certain degree of overfitting, i.e. a comparatively large number of knots. Ruppert *et al.* (2003) recommended a number of knots equal to about a quarter of the unique values of the control variable, and amounting to at least 35. In that case, or for small datasets, the structure imposed by the penalties on the covariance matrix among regression coefficients may be adequate. In contrast, RR analyses employed in animal breeding are generally restricted to relatively few knots and thus regression coefficients, with few assumptions about the structure of the covariance matrix of RR coefficients. A logical alternative to P-splines then is a RR analysis fitting B-spline basis functions where each coefficient is simply treated as a random effect without restrictions on variances and covariances as suggested by Rice and Wu (2001), resulting in an unstructured covariance matrix among RR coefficients. The same approach has been taken by Shi *et al.* (1996) and James *et al.* (2000), who combined it with reduced rank estimation to avoid overparameterisation.

Few explicit guidelines for the choice of knots are available. Fewer knots tend to yield smoother curves but a less detailed local fit. For splines curves fitted as fixed effects, an obvious strategy is to place knots where curves are expected to change. With penalised splines, the choice of knots is less critical. Often, knots are chosen to be equally spaced over the range of the control variable  $t$ . This is simple and easy to report. Demonstrating good interpolation even when  $t$  was unevenly distributed over its range, Eilers and Marx (2005) argued that this is adequate whenever penalised splines were fitted. An alternative recommendation is to place knots so that the resulting intervals span equal quantiles of  $t$  (Ruppert *et al.* 2003). A review of the properties of B-splines and P-splines is given by Eilers and Marx (2005).

Residual error variances in RR analyses are generally assumed to be independently distributed, but are often considered to change with the control variable  $t$ . Commonly, changes in error variance, or its logarithmic value, are modelled through a polynomial variance function. An alternative to the latter is a penalised spline function, as described above (eqn 2). More generally, Ruppert *et al.* (2003) suggested to fit a 'double mixed model', where heteroskedastic error variances are modelled as a function of  $t$  and, potentially, any other systematic effects in the mixed model.

RR analyses require mean trends to be modelled as fixed effects. Often there are multiple curves, nested within levels of a fixed effect. Generally, mean trajectories are modelled

by corresponding functions to those employed in fitting random, subject-specific curves. Using B-splines to model random curves, Rice and Wu (2001) considered a simple, fixed regression on B-spline basis functions to model mean trajectories. Alternatively, we might want to fit a penalised spline for this purpose. In the simplest scenario, this might involve a constant smoothing factor  $\lambda$  for all values of the control variable  $t$ . In other cases, we might want to allow  $\lambda$  to vary over the range of  $t$  (Ruppert and Carroll 2000). This could be achieved by modelling changes in the variance of the penalised spline coefficients as a function of  $t$ . As for residual errors, a suitable form of variance function might be a low degree spline of the logarithmic values of the variance (Crainiceau *et al.* 2004). Further research is required to evaluate the scope for such models in RR analyses.

### Reduced rank estimation

To date, covariance matrices among RR coefficients have, by and large, been considered unstructured. While, assuming normality, matrices are required to be positive (semi-) definite, this implies that for  $k$  coefficients fitted, there are  $k(k+1)/2$  coefficients to be estimated, i.e. that the number of parameters increases quadratically with the order of fit. Notable exceptions are RR analyses fitting smoothing splines (e.g. White *et al.* 1999), as described above. Analyses of FV traits assuming a parametric correlation structure involve few parameters [see Jaffrézic and Pletcher (2000), for a review], but impose this structure directly on the covariances among observations at different points of the trajectory, i.e. are not RR analyses. However, when applied to the within-subject covariances only, this can be combined with a RR model for the other random effects fitted (Foulley *et al.* 2000; Meyer 2001).

Thus RR analyses can involve a substantial number of parameters to be estimated, and large datasets are required to estimate these accurately. With the number of mixed model equations to be handled proportional to the orders of fit and numbers of random effects levels, computational demands can be, to say the least, impressive if not prohibitive, especially for REML estimation of covariance functions. This has motivated the use of a principal components parameterisation, which allows estimation of the most important components only, for the analysis of FV traits (Kirkpatrick and Meyer 2004).

Consider the eigendecomposition of a covariance matrix  $\mathbf{K}$ . This gives the matrix as the product of the matrix of eigenvectors  $\mathbf{E}$ , and the diagonal matrix of eigenvalues  $\mathbf{\Lambda}$ .

$$\mathbf{K} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}' \quad (5)$$

$\mathbf{K}$  is orthogonal, i.e.  $\mathbf{E}\mathbf{E}' = \mathbf{I}$ . The columns of  $\mathbf{E}$ ,  $\mathbf{e}_i$ , are the eigenvectors. These define linear transformations of the original variables to new variables, the so-called 'principal components' (PC), which are uncorrelated and have variances as given by the corresponding eigenvalues.

Eigenvectors and eigenvalues are usually ordered in descending order of the latter. The  $i$ -th PC then explains maximum variation given PCs 1, ...  $i - 1$  (Jolliffe 1986). Thus, if we consider the first  $m$  principal components only, we minimise the error of truncation for this number of terms. Often, the bulk of variation is explained by the first few PCs, and  $m$  can be considerably smaller than the size of  $\mathbf{K}$  with small or negligible loss of information. This is the principle underlying the use of PC analysis as a ‘dimension reduction’ technique.

*Principal components for function-valued traits*

The FV equivalent to eigenvectors are eigenfunctions (EF). As shown by Kirkpatrick *et al.* (1990), EFs can be useful in analysing and visualising patterns of variation of FV traits. In particular, genetic EFs can be used to predict the deformation in genetic trajectories due to selection. Like covariance functions, EFs, theoretically have infinite dimensions, but are, in practice, approximated as the weighted sum of a limited number of basis functions.

Estimates of EF and eigenvalues of CFs can be obtained from the coefficient matrix of the CF, i.e. the covariance matrix of RR coefficients. For orthogonal basis functions, such as Legendre polynomials, estimates are given directly by the eigendecomposition of the latter matrix (Kirkpatrick and Heckman 1989). Let  $\mathbf{K}$  of size  $k \times k$  be the coefficient matrix of a CF. Estimates of the EFs are then:

$$\hat{\kappa}_i(t) = \sum_{r=1}^k \phi_r(t) e_{ri} \tag{6}$$

where  $e_{ri}$  are the elements of the  $i$ -th eigenvector of  $\mathbf{K}$ , and  $\phi_r(t)$  is the  $r$ -th basis function. Evaluating (equation 6) for all values of  $t$  yields a curve, the eigenfunction. For non-orthogonal basis functions, such as B-splines, we need to adjust estimates for the basis used to obtain the eigenfunctions and eigenvalues of the corresponding CF (James *et al.* 2000). Alternatively, eigenfunctions and eigenvalues can be extracted numerically, by evaluating the CF for a fine grid of values of  $t$ , and calculating the eigendecomposition of the resulting covariance matrix (Rice and Wu 2001).

For simplicity of illustration, consider observations  $y_i$  for a FV trait, recorded at  $t_i$ , which are affected by a single random effect only. A RR model fitting  $k$  terms is then:

$$y_i = \sum_{r=1}^k \phi_r(t_i) \tau_r + \varepsilon_i \tag{7}$$

with  $\tau = \{\tau_r\}$  the vector of RR coefficients with variance  $V(\tau) = \mathbf{K}$ , and  $\varepsilon_i$  the residual error. Instead of modelling the trajectory for  $y_i$  as the weighted sum of basis functions  $\phi_r(\cdot)$ , we can rewrite (equation 7), expressing the trajectory as a weighted sum of eigenfunctions. This is often referred to as the ‘Karhunen-Loève’ expansion. Using (equation 6) and for  $m = k$ ,

$$y_i = \sum_{r=1}^m \kappa_r(t_i) \tau_r^* + \varepsilon_i = \sum_{r=1}^m \left[ \sum_{s=1}^k \phi_r(t_i) e_{sr} \right] \tau_r^* + \varepsilon_i \tag{8}$$

with  $\tau^* = \mathbf{E}'\tau$ . If the trajectory modelled represented additive genetic effects of the individual, values  $\tau_r^*$  would be the breeding values for the respective principal components.

Dimension reduction is achieved by considering  $m < k$ , truncating (equation 8) to the  $m$  leading, most variable PCs. Note that there are now 2 levels of truncation: first, the trajectory is approximated by its first  $m$  EFs, assuming these are infinite-dimensional; second, the EFs are estimated as the weighted sum of a limited number,  $k \geq m$ , of functions  $\phi_r(t)$  such as orthogonal polynomials or B-splines. Further details are discussed by Kirkpatrick and Meyer (2004).

This reduces the number of RR coefficients to be estimated to describe a trajectory from  $k$  to  $m$ , with a corresponding reduction in the size of the mixed model equations, and thus computational requirements. Yet, there are still  $k$  covariables  $\phi_r(t)$  to be fitted. The resulting covariance matrix among RR coefficients then has size  $k \times k$  as before, but reduced rank  $m$ , and  $m(2k - m + 1)/2$  parameters to be estimated. While there are  $m$  eigenvalues and the first  $m$  columns of  $\mathbf{E}$  have  $km$  elements, the number of parameters is less than  $(k + 1)m$ , as  $\mathbf{E}$  is required to be orthogonal and have columns with norm of unity, which reduces the number of ‘free’ elements in the columns of  $\mathbf{E}$  accordingly.

*Direct estimation of eigenfunctions*

PCs for random effects have generally been estimated by first obtaining full rank estimates of the covariance matrix, and then carrying out an eigenvalue decomposition of the matrix. A better approach is to estimate the PCs directly, and, at the same time, to restrict estimation to the most important components only (Kirkpatrick and Meyer 2004). This can readily be done within the mixed model framework, requiring only a simple reparameterisation. Let

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{9}$$

describe the usual RR model with  $\mathbf{y}$ ,  $\mathbf{b}$  and  $\boldsymbol{\varepsilon}$  the vectors of observations, fixed effects and residuals,  $\boldsymbol{\alpha}$  and  $\boldsymbol{\gamma}$  the vectors of RR coefficients modelling animals’ additive genetic and permanent environmental effects, and  $\mathbf{X}$ ,  $\mathbf{Z}$  and  $\mathbf{W}$  the respective incidence matrices. Further, let  $V(\boldsymbol{\alpha}) = \mathbf{K}_\alpha \otimes \mathbf{A}$  and  $V(\boldsymbol{\gamma}) = \mathbf{K}_\gamma \otimes \mathbf{I}$ , with  $\mathbf{A}$  the numerator relationship matrix between animals.

For full rank estimation under the usual model, elements of  $\mathbf{Z}$  and  $\mathbf{W}$  are basis functions  $\phi_r(\cdot)$  evaluated for values of the control variable pertaining to  $y$ . Reparameterising to the Karhunen-Loève form as in (equation 8), gives

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}^*\boldsymbol{\alpha}^* + \mathbf{W}^*\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon} \tag{10}$$

with  $\mathbf{Z}^* = \mathbf{Z}(\mathbf{I}_{N_A} \otimes \mathbf{Q}_\alpha)$  and  $\mathbf{W}^* = \mathbf{W}(\mathbf{I}_N \otimes \mathbf{Q}_\gamma)$ ,  $N_A$  denoting the total number of animals in the analysis, including any parents without records, and  $N$  the number of animals in the data. Let  $\mathbf{K}_\alpha = \mathbf{E}_\alpha \boldsymbol{\Lambda}_\alpha \mathbf{E}'_\alpha$  be the eigendecompositions of the covariance matrices among RR coefficients for a single animal. To estimate the leading  $m$  PCs only,  $\mathbf{Q}_x$  could be

chosen as above, to be equal to the first  $m$  columns of the respective  $\mathbf{E}_x$ .

Alternatively, we could choose to scale each column by the square root of its eigenvalue,  $\mathbf{Q}_x^m = (\mathbf{E}_x \mathbf{\Lambda}_x^{1/2})^m$ , (superscript  $m$  denoting the submatrix consisting of the first  $m$  columns), so that the vectors of RR coefficient,  $\alpha^*$  and  $\gamma^*$  have variances proportional to identity matrices. Eigenvalues are then simply the norms of the columns of  $\mathbf{Q}_x^m$ . This formulation allows the values for the elements of  $\mathbf{Q}_x^m$  determined by the orthogonality constraint on  $\mathbf{Q}_x^m$  to be obtained by solving a small system of linear equations (Meyer and Kirkpatrick 2005a). It is equivalent to a factor-analytic representation of  $\mathbf{K}_x$  where all specific variances are zero (Jennrich and Schluchter 1986; Thompson *et al.* 2003). As often done for such applications, we can apply an orthogonal transformation of the parameters. Let  $\mathbf{K}_x = \mathbf{L}_x \mathbf{L}'_x$  with  $\mathbf{L}_x$  the Cholesky factor of  $\mathbf{K}_x$ . The left singular vectors of  $\mathbf{L}_x$  are equal to the eigenvectors of  $\mathbf{K}_x$  and the corresponding singular values are the square root of the eigenvalues of  $\mathbf{K}_x$  (Harville 1997, page 232). Hence, choosing the matrix of right singular vectors of  $\mathbf{L}_x$  as transformation yields  $\mathbf{Q}_x = \mathbf{L}_x$  (Smith *et al.* 2001) i.e. estimating the non-zero elements of the first  $m$  columns of the Cholesky factor of  $\mathbf{K}_x$  is equivalent to estimating the first  $m$  PCs of  $\mathbf{K}_x$ .

This puts the ‘variance components’ to be estimated into the incidence matrices  $\mathbf{Z}^*$  and  $\mathbf{W}^*$ . Additional parameters are the variances of residuals which determine  $\mathbf{R} = \mathbf{V}(\varepsilon^*)$ . REML estimates can be obtained in the same fashion as for standard multivariate and RR models. The REML likelihood is:

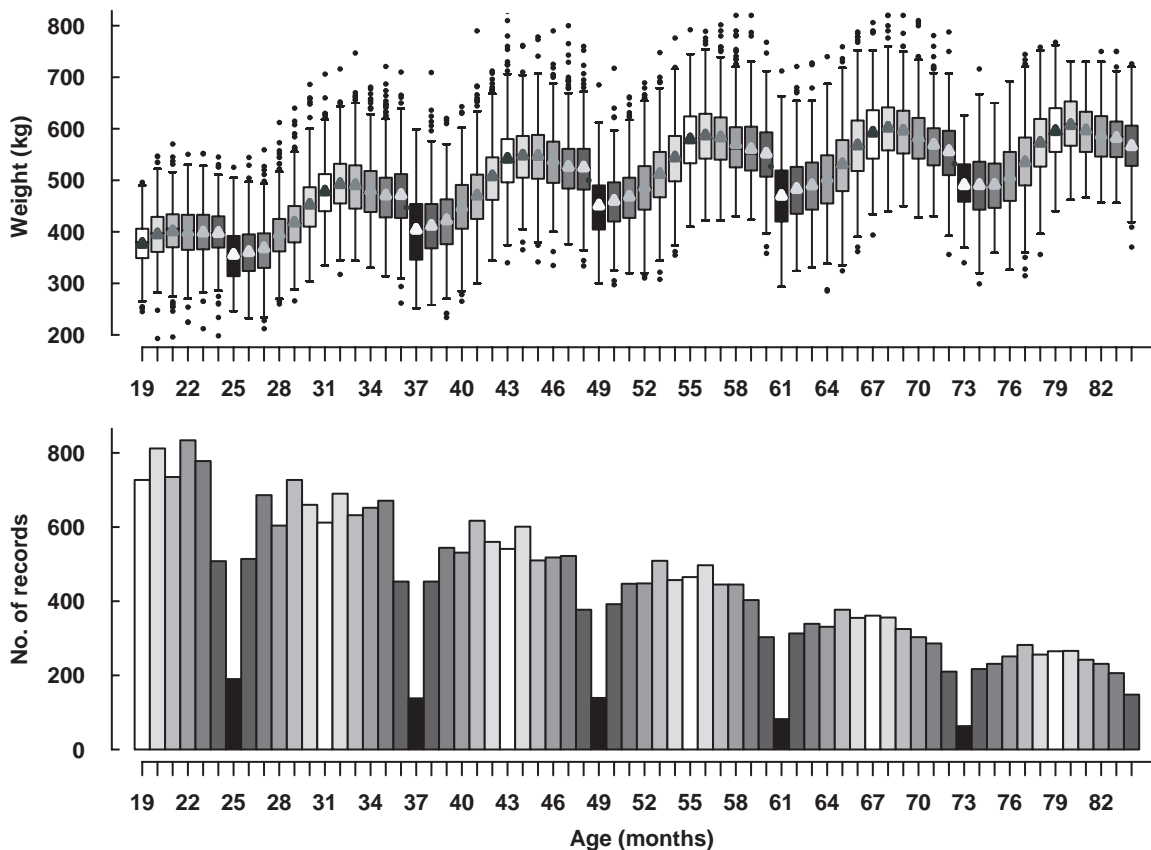
$$-2 \log L = \text{constant} + \log |\mathbf{A}| + \log |\mathbf{R}| + \log |\mathbf{C}| + \mathbf{y}'\mathbf{P}\mathbf{y} \quad (11)$$

with  $\mathbf{C}$  the coefficient matrix in the mixed model equations pertaining to (equation 10), and  $\mathbf{y}'\mathbf{P}\mathbf{y}$  a weighted sum of squares of residuals; see Meyer and Kirkpatrick (2005a) for further details.

The likelihood ( $\log L$ ) can be maximised using common optimisation techniques. In particular, the so-called average information (AI) algorithm (Gilmour *et al.* 1995) is widely used in animal breeding applications. Meyer and Kirkpatrick (2005a) describe an AI-REML procedure for reduced rank estimation via the leading PCs, and Thompson *et al.* (2003) outline a closely related algorithm for analyses fitting factor-analytic models.

**Application**

The combined use of B-spline basis functions and reduced rank estimation considering the leading eigenfunctions only is illustrated for a small set of mature



**Figure 2.** Distribution of records over ages (bottom), together with means (▲) and quartiles (top).

cow weight records from a selection experiment at the Wokalup research station in Western Australia.

### Data

Records consisted of weights of Polled Hereford cows, born between 1974 and 1988, collected during the Wokalup selection experiment. This comprised a herd of about 300 cows weighed on a monthly basis, except during the calving season, mostly April and May each year. Details of the experiment are described by Meyer *et al.* (1993). Records for cows aged 1.5–7 years at weighing, taken between mid-1976 and the end of 1990, were selected. After edits, this resulted in 28643 records on 908 cows, with up to 63 records per animal. While cows born after 1981 did not have a chance to have records up to 7 years of age, 75 and 61% of cows had at least 13 and 24 records, respectively, themselves and, considering records on parents or sibs, 97% of animals had information at the genetic level on at least 24 ages at recording available. Furthermore, almost all of the younger cows had dams with weights at the later ages for which they did not have a chance to have records themselves.

The research station has a Mediterranean climate with winter rains and summer droughts, resulting in a large seasonal variation in pasture availability. With short mating and thus calving periods of about 2 months, ages of cows and months of recording are strongly correlated. Figure 2 shows the distribution of records over ages at recording (in months).

### Analyses

Estimates of covariance functions for genetic and permanent environmental effects of the cow were obtained using REML, fitting a random regression on quadratic B-spline functions of age at recording in days. The seasonal pattern of feed availability and fluctuations in weights of cows suggested intervals that were fractions of a year. Earlier analyses (Meyer 2000), using quadratic segmented polynomials with a truncated power basis at the phenotypic level only, had shown annual intervals to be too coarse, but

half-yearly intervals to allow the pattern of variation to be modelled adequately. Subdividing the range of ages into 11 equal intervals of 6 months each resulted in  $k = 13$  RR coefficients to be estimated. Decreasing intervals to 3 months each increased the number of spline coefficients to  $k = 24$ . Analyses were carried out estimating up to  $m = 8$  PCs for  $k = 13$ , and up to  $m = 5$  components for  $k = 24$ . For simplicity, the same order of fit and number of principal components were considered for both random effects, except for one analysis. In addition, a full rank model ( $m = 13$ ) was fitted for  $k = 13$ . As suggested by Meyer (2000), variances due to temporary environmental effects were modelled through a cyclic step function with 12 classes corresponding to month of age at recording. With a strong confounding of age and month of the year, this was, in essence, equivalent to modelling temporary environmental effects as a function of the month of recording.

Fixed effects fitted included a mean weight trajectory, modelled through a regression on a quadratic B-spline with 11 equal intervals, and contemporary groups, defined as paddock-date of weighing-year of birth of cow subclasses, with 3949 levels. Results from different models of analysis were compared using the REML form of the Akaike Information criterion (AIC), corrected for small sample size,  $-2\log L + 2p + 2p(p+1)/(N-p-1)$  (Burnham and Anderson 2004), and the Bayesian Information criterion (BIC), calculated as  $-2\log L + (N-r(\mathbf{X}))p$  (Wolfinger 1993), with  $\log L$  the REML maximum likelihood,  $N$  the number of records,  $r(\mathbf{X})$  the rank of the design matrix for fixed effects, and  $p$  the number of parameters estimated. For these analyses, the ‘penalty factor’,  $[N-r(\mathbf{X})]$ , for the number of parameters in the BIC amounted to 10.114.

### Results

As shown in Figure 2, weights of cows fluctuated greatly over a 12 months period, reflecting extreme seasonal climatic differences. Cows were usually heaviest in summer after weaning, with a mean January weight of 512.3 kg, and lightest

**Table 1. Maximum log likelihood (logL), Akaike (AIC) and Bayesian (BIC) Information Criterion for different analyses**

Rank <sup>A</sup>	Neqns <sup>B</sup>	p <sup>C</sup>	6-months intervals			p <sup>C</sup>	3-months intervals		
			LogL <sup>D</sup>	-0.5 AIC <sup>D</sup>	-0.5 BIC <sup>D</sup>		logL <sup>D</sup>	-0.5 AIC <sup>D</sup>	-0.5 BIC <sup>D</sup>
3/3	10131	84	-1810.84	-1895.09	-2235.63	150	-1464.59	-1615.38	-2223.14
4/4	12189	104	-829.63	-934.01	-1355.56	192	-290.23	-483.53	-1261.18
5/5	14247	122	-279.72	-402.25	-896.68	232	390.98	157.08	-782.25
5/7	16063	137	-68.32	-205.99	-761.14				
6/6	16305	138	-86.41	-225.09	-784.28				
7/7	18363	152	-56.24	-209.06	-824.91				
8/8	20421	164	-36.68	-201.63	-866.04				
13/13	30711	194	-18.88	-214.21	-999.94				

<sup>A</sup>Number of principal components fitted; genetic/permanent environmental.

<sup>B</sup>Number of equations in mixed model matrix.

<sup>C</sup>Number of parameters to be estimated.

<sup>D</sup>+85000.

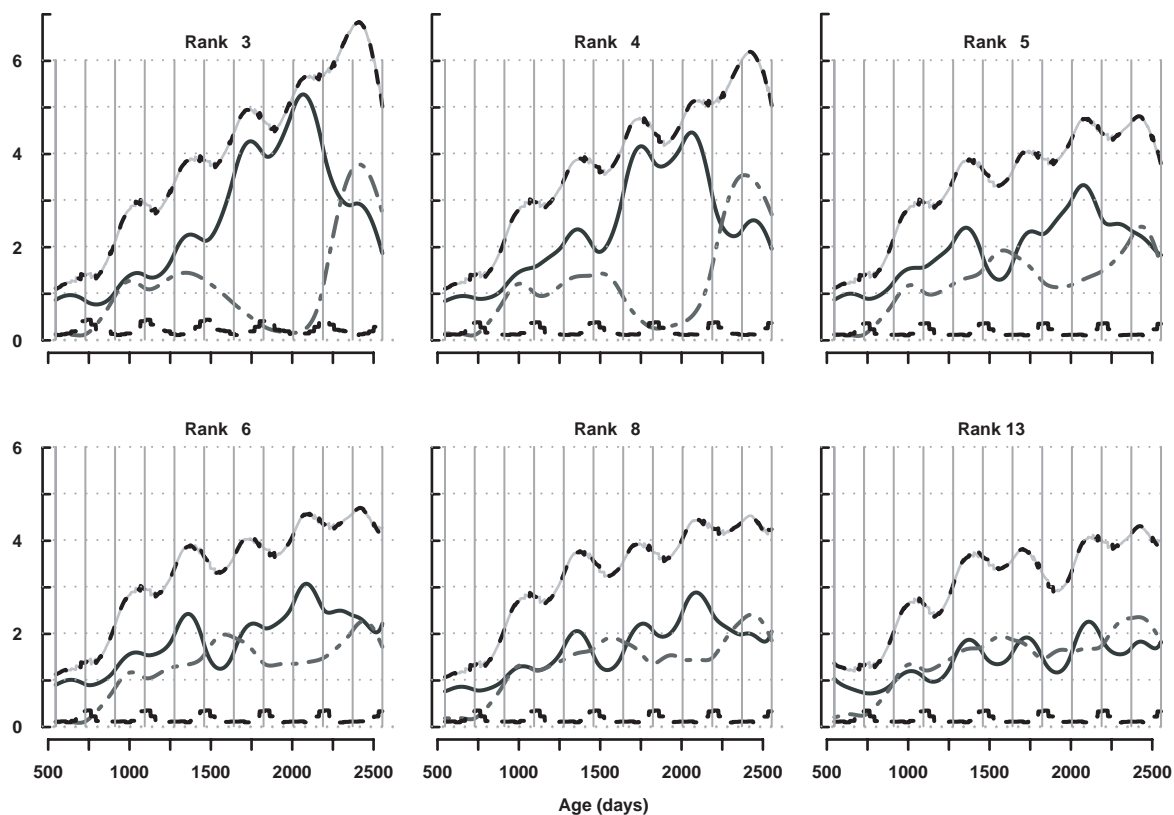
just after calving, with a mean June weight of 424.4 kg. Records were also highly variable, with an overall coefficient of variation of 19%, and, as shown in Figure 2, numerous observations outside the range given by 150% of the inter-quartile distance spanned by the second and third quartile, especially at the upper side. Within month standard deviations ranged from 75.4 kg in July to 99.9 kg in December.

Characteristics of the different analyses carried out are summarised in Table 1. Numbers of equations to be handled and thus computational requirements per likelihood evaluation were determined by the number of PCs fitted. As demonstrated by Meyer (2005c) for a multi-trait analysis involving 8 traits, increasing the number of PCs fitted by 1 can increase the number of operations required by a factor close to 2. A similar relationship was observed for our analyses. A full rank analysis involved 30711 equations and 194 and 612 parameters to be estimated for models fitting 13 and 24 RR coefficients, respectively.

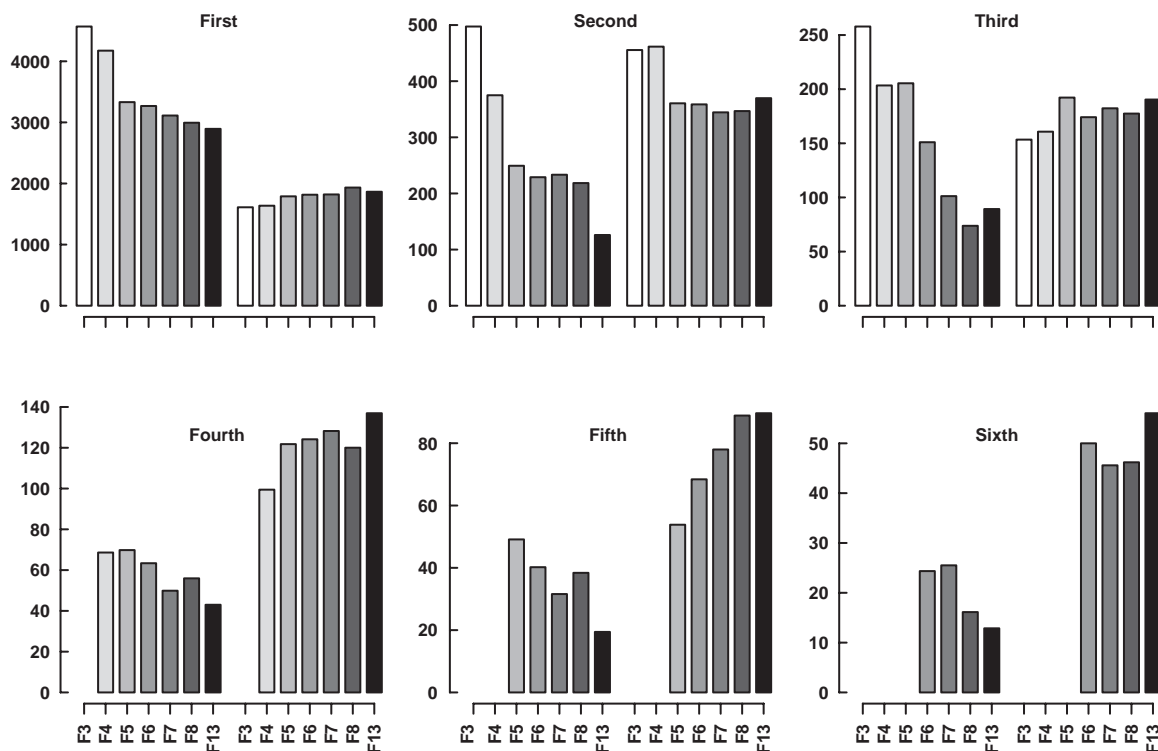
Likelihoods increased significantly with the number of PCs considered. However, likelihood ratio tests are known to favour the more detailed models. Similarly, the AIC was lowest for the model fitting the most parameters, fitting 5 PCs with  $k = 24$ . For  $k = 13$ , the AIC decreased until  $m = 8$ , the highest number of PCs considered for a reduced rank fit.

In contrast, with a stringent penalty for the number of parameters estimated, the BIC for analyses considering the same number of PCs for both random effects, was lowest fitting the first 6 PCs only with  $k = 13$ , indicating that there was little advantage to fitting more PCs or increasing the number of intervals. Inspection of the eigenvalues (see below) suggested that the last PCs for permanent environmental effects explained more variation than those for genetic effects. A model fitting 5 genetic and 7 permanent environmental PCs (for  $k = 13$ ) involved almost the same number of parameters as  $m = 6$  for both components, but provided a better fit to the data.

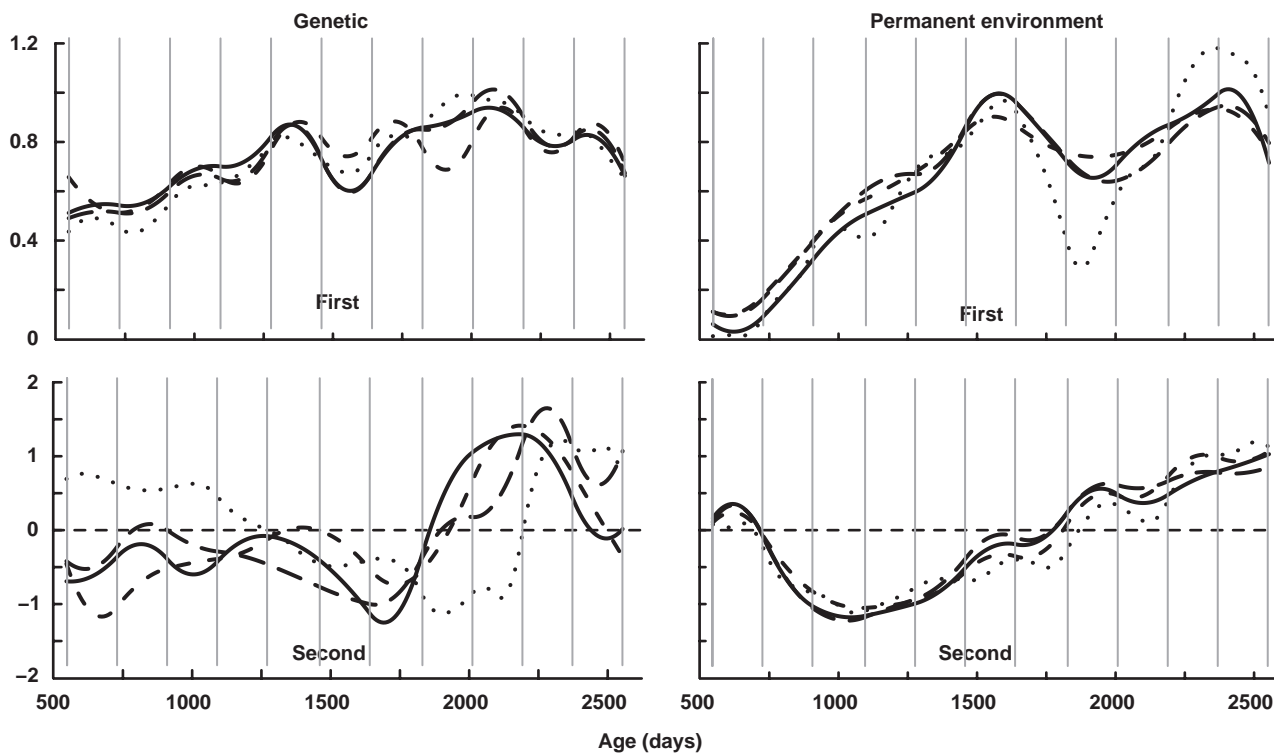
Estimates of variance components for the ages in the data are shown in Figure 3 for analyses fitting 13 coefficients, with vertical lines indicating the position of the knots. As suggested by the corresponding likelihood values, fewer than 5 or 6 PCs were not sufficient to model the variation in the data adequately, resulting in implausibly high estimates of variance components after about 1500 days of age. Estimates of phenotypic variances ( $\sigma_P^2$ ) were essentially the same for all analyses fitting 5 or more PCs, with only a slight tendency of values at the highest ages to decrease with increasing order of fit. While  $\sigma_P^2$  followed a consistent, cyclic pattern corresponding to changes in means, genetic ( $\sigma_A^2$ ) and



**Figure 3.** Estimates of genetic (solid line), permanent environmental (dot-dashed line), temporary environmental (long-dashed line) and phenotypic (dashed line) variances (in 1000 kg<sup>2</sup>) from analyses fitting B-splines with 6-months intervals and considering the first 3 to 8 and all (Rank 3, ..., Rank 8, Rank 13) principal components.



**Figure 4.** Estimates of the first six genetic (left) and permanent environmental (right) eigenvalues from analyses fitting B-splines with 6-months intervals and considering the first 3 to 8 and all (F1, ..., F8, F13) principal components.



**Figure 5.** Estimates of the first 2 eigenfunction for genetic (left) and permanent environmental (right) effects from analyses fitting B-splines with 6-months (solid line: Rank 5, long-dashed line: Rank 7, dashed line: Rank 13) and 3-months (dotted line: Rank 5) intervals.

permanent environmental ( $\sigma_R^2$ ) variances exhibited less regular changes over time. With  $\sigma_p^2$  remaining more or less the same, curves for  $\sigma_A^2$  and  $\sigma_R^2$  tended to become less divergent with increasing numbers of PCs, indicating a strong sampling correlation. With only 908 animals in the data and well over 100 parameters to be estimated, this is hardly surprising.

Corresponding estimates of the first 6 eigenvalues are summarised in Figure 4. As for most growth data, the first genetic PC explained the bulk of genetic variation, increasing from 85.3% for  $m = 5$  to 90.7% for  $m = 13$ . Estimates of genetic eigenvalues tended to decrease with increasing number of PCs considered. Conversely, estimates of permanent environmental eigenvalues tended to increase, again emphasising strong sampling correlations and some repartitioning of estimates between sources of variation. From the fourth PC onwards, permanent environmental PCs explained consistently more variation than their genetic counterpart. This suggested that more permanent environmental than genetic PCs should be considered. As shown above, an analysis fitting 5 genetic and 7 environmental PCs indeed provided a better fit to the data than any of the other analyses. Further work is required to determine the best model to be fitted.

Estimates of the first 2 EFs are presented in Figure 5. Estimates fitting  $k = 24$  spline coefficients showed considerably more fluctuations than those obtained fitting  $k = 13$  coefficients, indicating some overparameterisation in the former analyses. Again, estimates clearly reflected the cyclic, seasonal changes in variation in the data. As commonly found for CFs of growth, the first genetic EF was positive throughout, i.e. selection for increased size at any age is likely to increase size at all ages (e.g. Kirkpatrick *et al.* 1990).

## Conclusions

Reduced rank estimation via the leading principal components is a powerful tool for high-dimensional, genetic analyses. It is of particular interest for scenarios like RR analyses, where a subset of the PCs explains almost all variation. While the number of parameters to be estimated increases with the number of basis functions required to approximate the PCs, computational requirements for estimation are largely determined by the number of PCs. Moreover, sampling variances increase with the number of parameters estimated. The leading PCs tend to be estimated most accurately (Kirkpatrick and Meyer 2004; Meyer 2005*d*). Hence, reduced rank estimation can provide estimates with essentially the same mean square errors as full rank analyses. This is of special relevance for small datasets and models involving many parameters, such as the application shown above. James *et al.* (2000), Kirkpatrick and Meyer (2004) and Meyer and Kirkpatrick (2005*b*) presented small simulation studies examining the sampling behaviour of reduced rank estimates in a RR context.

RR analyses fitting splines instead of (orthogonal) polynomials are appealing. However, fitting many knots, as is common practice in semi-parametric smoothing or penalised spline estimation, can result in many effects to be estimated and thus be computationally demanding, in particular for genetic evaluation. Although, as outlined above, the number of variance components for such models tends to be small, RR models fitting a smaller number of knots and allowing for unstructured covariances among the regression coefficients may be preferable, in particular for modelling covariance functions. However, even with relatively few knots, RR models fitting B-spline basis functions tend to involve more coefficients than corresponding analyses with polynomial basis functions.

Fortunately, few PCs generally suffice to explain the bulk of variation for FV traits, especially if observations are sparse. Hence, combined with reduced rank estimation, RR analyses fitting B-splines do not need to be any more complex than corresponding polynomial analyses. As shown for the example above, such analyses are well capable of modelling FV traits with large changes in variation along the trajectory, and are well suited to the analysis of 'repeated records' from livestock improvement schemes.

## References

- Burnham KP, Anderson DR (2004) Multimodel inference: Understanding {AIC} and {BIC} in model selection. *Sociological Methods and Research* **33**, 261–304. doi:10.1177/0049124104268644
- Crainiceau CM, Ruppert C, Carroll RJ (2004) Spatially adaptive Bayesian P-splines with heteroscedastic errors. Working paper no. 61, John Hopkins University, Department of Biostatistics. Available online at: [www.bepress.com/jhbiostat/paper61](http://www.bepress.com/jhbiostat/paper61) (verified 3 August 2005).
- de Boor C (2001) 'A practical guide to splines.' 2nd edn. (Springer Verlag: New York)
- Durbán M, Harezlak J, Wand MP, Carroll RJ (2005) Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine* **24**, 1153–1167. doi:10.1002/sim.1991
- Eilers PHC (1999) Discussion on paper by Verbyla, Cullis, Kenward and Welham. *Applied Statistics* **48**, 307–308.
- Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science* **11**, 89–121. doi:10.1214/ss/1038425655
- Eilers PHC, Marx BD (2005) Splines, knots, and penalties. Available online at: [http://www.stat.lsu.edu/faculty/bmarx/splines\\_knots\\_penalties.pdf](http://www.stat.lsu.edu/faculty/bmarx/splines_knots_penalties.pdf) (verified 3 August 2005).
- Foulley JL, Jaffrézic F, Robert-Granié C (2000) EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis. *Genetics Selection Evolution* **32**, 129–141. doi:10.1051/gse:2000110
- Gilmour AR, Thompson R, Cullis BR (1995) Average Information REML, an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450.
- Green PJ, Silverman BW (1994) 'Nonparametric regression and generalized linear models. A roughness penalty approach.' Monographs in statistics and applied probability. Vol. 58. (Chapman & Hall: London)
- Harville DA (1997) 'Matrix algebra from a statistician's perspective.' (Springer Verlag)

- Jaffrézic F, Pletcher SD (2000) Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. *Genetics* **156**, 913–922.
- James GM, Hastie TJ, Sugar CA (2000) Principal component models for sparse functional data. *Biometrika* **87**, 587–602. doi:10.1093/biomet/87.3.587
- Jennrich RI, Schluchter MD (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.
- Jolliffe IT (1986) 'Principal component analysis.' (Springer Verlag: New York)
- Kirkpatrick M, Heckman N (1989) A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27**, 429–450. doi:10.1007/BF00290638
- Kirkpatrick M, Lofsvold D, Bulmer M (1990) Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* **124**, 979–993.
- Kirkpatrick M, Meyer K (2004) Simplified analysis of complex phenotypes: direct estimation of genetic principal components. *Genetics* **168**, 2295–2306. doi:10.1534/genetics.104.029181
- Meyer K (1998) Estimating covariance functions for longitudinal data using a random regression model. *Genetics, Selection, Evolution* **30**, 221–240.
- Meyer K (2000) Random regressions to model phenotypic variation in monthly weights of Australian beef cows. *Livestock Production Science* **65**, 19–38. doi:10.1016/S0301-6226(99)00183-9
- Meyer K (2001) Estimating genetic covariance functions assuming a parametric correlation structure for environmental effects. *Genetics, Selection, Evolution* **33**, 557–585. doi:10.1051/gse:2001102
- Meyer K (2005a) Estimates of covariance functions for growth of Angus cattle from random regression analyses fitting B-spline functions. *Proceedings of the Association for Advancement of Animal Breeding Genetics* **16**, 52–55.
- Meyer K (2005b) Random regression analyses using B-splines to model growth of Australian Angus cattle. *Genetics, Selection, Evolution* **37**, 473–500.
- Meyer K (2005c) Reduced rank estimates of the genetic covariance matrix for live ultra-sound scan traits. *Proceedings of the Association for Advancement of Animal Breeding Genetics* **16**, 56–59.
- Meyer K (2005d) Sampling behaviour of reduced rank estimates of genetic covariance functions. *Proceedings of the Association for Advancement of Animal Breeding Genetics* **16**, 286–289.
- Meyer K, Carrick MJ, Donnelly BJP (1993) Genetic parameters for growth traits of Australian beef cattle from a multi-breed selection experiment. *Journal of Animal Science* **71**, 2614–2622.
- Meyer K, Kirkpatrick M (2005a) Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. *Genetics, Selection, Evolution*. **37**, 1–30. doi:10.1051/gse:20040304
- Meyer K, Kirkpatrick M (2005b) Up hill, down dale: quantitative genetics of curvaceous traits. *Philosophical Transactions of the Royal Society B* **360**, 1443–1455. doi:10.1098/rstb.2005.1681
- Rice JA, Wu CO (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259. doi:10.1111/j.0006-341X.2001.00253.x
- Ruppert D, Carroll RJ (2000) Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**, 205–223. doi:10.1111/1467-842X.00119
- Ruppert D, Wand MP, Carroll RJ (2003) 'Semiparametric regression.' (Cambridge University Press: New York)
- Shi M, Weiss RE, Taylor JMG (1996) An analysis of pediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics* **45**, 151–164.
- Smith AB, Cullis BR, Thompson R (2001) Analysing variety by environment data using multiplicative mixed models and adjustments for spatial field trends. *Biometrics* **57**, 1138–1147. doi:10.1111/j.0006-341X.2001.01138.x
- Thompson R, Cullis BR, Smith AB, Gilmour AR (2003) A sparse implementation of the Average Information algorithm for factor analytic and reduced rank variance models. *Australian and New Zealand Journal of Statistics* **45**, 445–459. doi:10.1111/1467-842X.00297
- Torres RAA, Quaas RL (2001) Determination of covariance functions for lactation traits on dairy cattle using random-coefficient regressions on B-splines. *Journal of Animal Science* **79**(Suppl.1), 112. [Abstract.]
- Verbyla AR, Cullis BR, Kenward MG, Welham SJ (1999) The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistics* **48**, 269–311.
- White IMS, Thompson R, Brotherstone S (1999) Genetic and environmental smoothing of lactation curves with cubic splines. *Journal of Dairy Science* **82**, 632–638.
- Wolfinger RD (1993) Covariance structure selection in general mixed models. *Communications in Statistics — Simulation and Computing* **22**, 1079–1106.

Received 14 February 2005, accepted 29 April 2005