Chapter 19

Analysis of longitudinal data

-Random Regression Analysis

Julius van der Werf

1 Introduction

In univariate analysis the basic assumption is that a single measurement arises from a single unit (experimental unit). In multivariate analysis, not one measurement but a number of different characteristics are measured from each experimental design, e.g. milk yield, body weight and feed intake of a cow. These measurements are assumed to have a correlation structure among them. When the same physical quantity is measured sequentially over time on each experimental unit, we call them repeated measurements, which can be seen as a special form of a multivariate case. Repeated measurements deserve a special statistical treatment in the sense that their covariance pattern, which has to be taken into account, is often very structured. Repeated measurements on the same animal are generally more correlated than two measurements on different animals, and the correlation between repeated measurements may decrease as the time between them increases. Modeling the covariance structure of repeated measurements correctly is of importance for drawing correct inference from such data.

Measurements that are taken along a trajectory can often be modeled as a function of the parameters that define that trajectory. The most common example of a trajectory is time (longitudinal), and repeated measurements are taken on a trajectory of time. However, there could be other variables defining the trajectory, e.g. by location (spatial), environmental determinants (moisture, temperature, humidity) or physiological determinants (fat versus a trajectory of weight).

The term 'repeated measurement' can be taken literally in the sense that the measurements can be thought of as being repeatedly influenced by identical effects, and it is only random noise that causes variation between them. However, repeatedly measuring a certain characteristic may give information about the change over time of that characteristic. The function that describes such a change over time may be of interest since it may help us to understand or explain, or to manipulate how the characteristic changes over time. Common examples in animal production are growth curves and lactation curves.

Generally, we have therefore two main arguments to take special care when dealing with repeated measurements. The first is to achieve statistically correct models that allow correct inferences from the data. The second argument is to obtain information on a trait that changes gradually over time.

Experiments are often set up with repeated measurements to exploit these two features. The prime advantage of longitudinal studies (i.e. with repeated measurements over time) is its effectiveness for studying change. Notice that the interpretation of change may be very different if it is obtained from data across individuals (cross sectional study) or on repeated measures on the same individuals. An example is given by Diggle et al. (1994) where the relationship between reading ability and age is plotted. A first glance at the data suggests a negative relationship, because older people in the data set tended to have had less education. However, repeated observations on individuals showed a clear improvement of reading ability over time.

The other advantage of longitudinal studies is that it often increases statistical power. The influence of variability across experimental units is canceled if experimental units can serve as their own control.

Both arguments are very important in animal production as well. A good example is the estimation of a growth curve. When weight would be regressed on time on data *across* animals, not only would the resulting growth curve be more inaccurate, but also the resulting parameters might be very biased if differences between animals and animals' environments were not taken into account.

Models that deal with repeated measurements have been often used in animal production. In dairy cattle, the analysis of multiple lactation records is often considered using a 'repeatability model'. The typical feature of such a model from the genetic point of view is that repeated records are thought of expressions of the same trait, that is, the genetic correlation between repeated lactation is considered to be equal to unity. Models that include data on individual test days have often used the same assumption. Typically, genetic evaluation models that use measures of growth do often consider repeated measurements as genetically different (but correlated) traits. Weaning weight and yearling weight in beef cattle are usually analyzed in a multiple trait model.

Repeatability models are often used because of simplicity. With several measurements per animal, they require much less computational effort and less parameters than a multiple trait model. A multiple trait model would often seem more correct, since they allow genetic correlations to differ between different measurements. However, covariance matrix for measurements at very many different ages would be highly overparameterised. Also, an unstructured covariance matrix may not be the most desirable for repeated measurements that are recorded along a trajectory. As the mean of measurements is a function of time, so also may their covariance structure be. A model to allow the covariance between measurements to change gradually over time, and with the change dependent upon differences between times, can make use of a *covariance function*.

As was stated earlier, repeated measurements can often be used to generate knowledge about the change of a trait over time. Whole families of models have been especially designed to describe such changes as regression on time, e.g. lactation curves and growth curves. The analysis may reveal causes of variation that influence this change. Parameters that describe typical features of such change, e.g. the slope of a growth curve, are regressions that may be influenced by feeding levels, environment, or breeds. There may also be additive genetic variation within breeds for such parameters. One option is then to estimate curve parameters for each animal and determine components of variation for such parameters. Another option is use a model for analysis that allows regression coefficients to vary from animal to animal. Such regression coefficients are then not fixed, but are allowed to vary according to a distribution that can be assigned to them, therefore indicated as *random regression coefficients*.

We will present models that use random regression in animal breeding. Typical applications are for traits that are measured repeatedly along a trajectory, e.g. time. Different random regression models will be presented and compared. The features of random regression models and estimation of their parameters will be discussed.

Alternative approaches to deal with repeatedly measured traits along a trajectory are the use of covariance functions, and use of multiple trait models. These approaches have much in common, since they all deal with changing covariances along a trajectory. Different models that allow the study of genetic aspects of changes of traits along a trajectory will be presented and discussed.

2 Exploring correlation patterns in repeated measurements

There are several ways to explore the correlation structure in repeated measurements Diggle et al. (1994) . Graphical displays can be very useful to identify patterns relevant to certain questions, e.g. the relationship between response and explanatory variables. Figures 2-1, 2-2, and 2-3 (adapted from Diggle et al. 1994) illustrate this by showing graphs for body weight in 5 pigs as a function of time. Figure 2-1 shows the lines connecting the weights on an individual pig in consecutive weeks. This graph shows that (1) pigs gain weight over time, 2) pigs that are largest at the beginning tend to be largest at the end, and 3) the variation among weights is lower at the beginning than at the end.

The second observation is important in relation to correlation structure, and has important biological implications. Figure 2-2 gives a clearer picture of the second point. By plotting deviations from the mean, the graph is magnified. In Figure 2-2, we observe that lines do cross quite often, and rankings do change for different times on the axis. Measurements tend to cross less in the later part of the experiment, i.e. correlations might be higher in later part of the trajectory. With many individuals, it is more difficult to interpret such graphs.

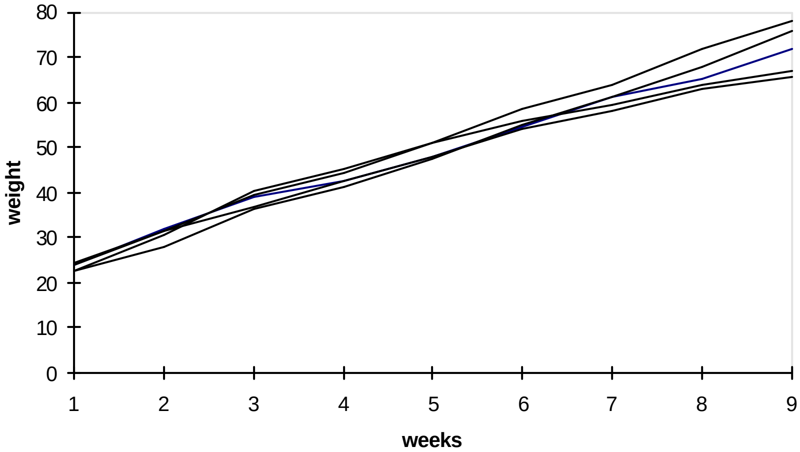

Figure 2.1. Body weight for 5 pigs measured at 9 consecutive weeks

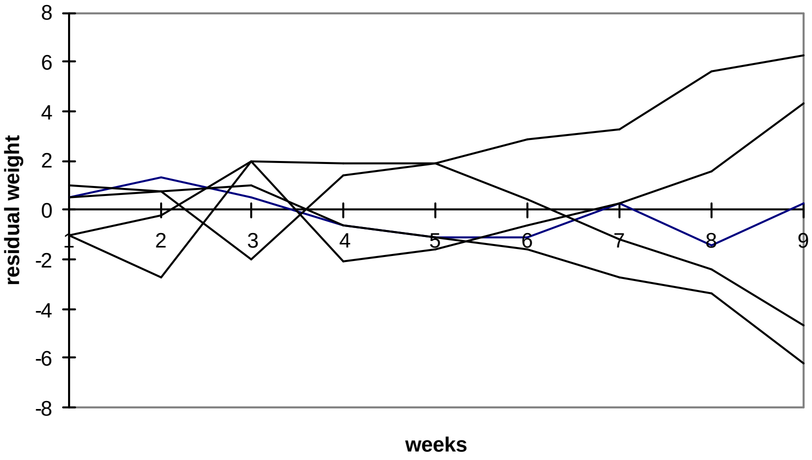

Figure 2-2. Residual body weight (deviation from week mean) for 5 pigs measured at 9 consecutive weeks

Exact values such as correlations between measurements at different time points can not be obtained from graphs like in Figure 2-1. When observations are made at equally spaced times, associations between repeated measurements at two fixed times are easily plotted and measured in terms of correlations. With unequally spaced observations, this is less evident. Diggle et al. (1994) suggest to use a variogram. This is a function that describes the association among repeated values and is easily estimated with irregular observation times. A variogram is defined as

$$g\,(u) = \tfrac{1}{2}[\{E(\,y(t) - y(t-u)\}^2\,], u \geq 0$$

where $g\,(u)$ describes the squared differences between measurements that are u time units apart. The variogram is calculated from observed half-squared differences between pairs of residuals,

$$v_{ijk} = \tfrac{1}{2}(r_{ij} - r_{ik})^2$$

and the corresponding time differences

$$u_{ijk} = t_{ij} - t_{ik}$$

where y_{ij} is the j^{th} observation on animal i, and residuals are $r_{ij} = y_{ij} - E(\,y_{ij}\,)$, i.e. they can be calculated as deviations from contemporary means. If the times are regular, $\hat{g}\,(u)$ is estimated as the average of all v_{ijk} corresponding to the particular u. With irregular sampling times, the variogram can be estimated from the data pairs (u_{ijk}, v_{ijk}) by fitting a curve A variogram for the example of Figure 2-2 is given in Figure 2-3. As an example, the point for u=8 is obtained as the average of the half- squared differences between the residual for the first and the 9th observation on the five pigs:

$$\hat{g}(u=8) = \sum_{i=1}^{5} \tfrac{1}{2}\{(y_{i1} - \bar{y}_{.1}) - (y_{i9} - \bar{y}_{.9})\}^2 / 5,$$

where $\bar{y}_{.j}$ is the average of the jth observation.

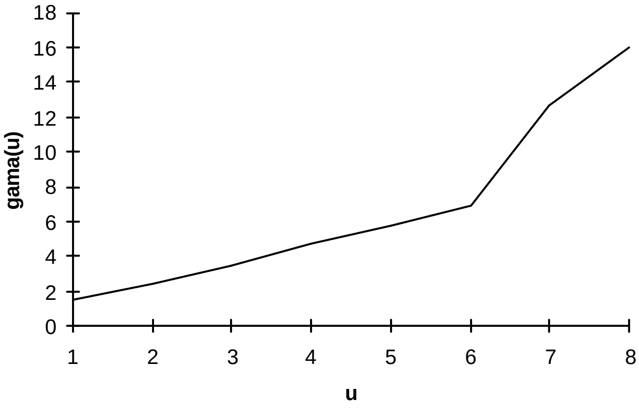

Figure 2-3. Variogram for the pig example, showing $g\ (u)$ (gama(u)) for u=1,…8 (u= distance between measurements in weeks)

A structure often used in repeated measurements to describe the correlation matrix is the *autocorrelation structure*. We can define autocorrelation as the correlation between two measurements as a function of the distance (in time) between the measurements.
The autocorrelation function can be estimated from the variogram as $\hat{r}(u) = 1 - \hat{g}(u) / \hat{s}^2$, where $\hat{s}^2$ is the 'process variance', which is calculated as the average of all half squared differences $\tfrac{1}{2}(y_{ij} - y_{lk})^2$ with i$\neq$l.

There exist several types of correlation models. In a *uniform correlation model*, we assume that there is a positive correlation r between any two measurements on the same individual (independent of time). In matrix terms the correlation matrix between observation on the same animal is written as

$$V_0 = (1-r)I + rJ$$

where ***I*** is the identity matrix and ***J*** is a matrix with all elements equal to 1. The uniform correlation model is used in what is generally called a 'repeatability model' in animal breeding.

In the *exponential correlation model* , correlations between two observations at the same animal at times j and k are

$$v_{jk} = s^2 \exp(-f|t_j - t_k|).$$

In this model, the correlation between a pair of measurements on the same individual decreases towards zero as the time separation between measurements increases. The rate of decay is faster for larger values of f. If the observation times are equally spaced, than the correlation between the jth and the kth measurements can be expressed as $v_{jk} = s^2 r^{|j-k|}$, where $r = e^{-f}$. Sometimes the correlation decreases slow initially, and then decreases sharply towards zero. Such behaviour may be better described by a *Gaussian correlation function*:

$$v_{jk} = s^2 \exp(-f(t_j - t_k)^2).$$

The exponential correlation model is sometimes called the first order autoregressive model. In such a model, the random part of an observation depends only on the random part of the previous observation: $e_j = re_{j-1} + z_j$, where z_j is an independent random variable. Models where random variables (e.g. errors) depend on previous observations are called ante-dependence models, and a when random variable depend on p previous variables we have a pth order Markov model.

In general we can distinguish three different sources of random variation that play a role in repeated measurements (Diggle et al., 1994):

- Random effects. When the units on which we have repeated measurements are sampled at random from a population, we can observe variation between units. Some units are on average higher responders than other units. In animal breeding, an obvious example of such effects are animal effects, or more specific, the (additive) genetic effects of animals.
- Serial correlation. This refers to part of the measurement that is part of a time varying stochastic effect. Such an effect causes correlation between observations within a short time interval, but common effects are less correlated if measurements are further away.
- Measurement error. This is an error component of an observation which effect is each time independent of any other observations.

If a model is build that accommodates these three random effects, the variance structure of each of the effects needs to be described. Diggle et al (1990) give a general formula for the variance of observations on one experimental unit as

$$\text{var}(e) = v^2 J + s^2 H + t^2 I \qquad [2.1]$$

where v^2, s^2 and t^2 are variance components for the three random effects, J is a matrix with ones, and H is specified by a correlation function.

This model given in [2.1] often used in analysis of longitudinal data. In fact, model [2.1] is not at all general. The random effects are assumed to be constant over all measurements within a unit. If we think of this effect as a the genetic effect of an animal we can imagine very well these to vary between ages (over time), and this may even bear our special interest. Therefore, J should be replaced by a correlation function. The serial correlation effect may be seen as the temporary environmental effects often used in animal breeding data. For both the random and the serial correlation effect, the question is how the correlation (c.q. covariance-) function should be defined.

The patterns as described in this section, and as often use in the statistical literature, show smooth functions that seem natural for many stochastic processes. However, the additive genetic effect on a trait over time maybe more irregular. For example, some genes could be mainly active during the first 4 months of growth of a pig, with high correlations between measurements within this period, but other genes may take over during the last month. Also the permanent environmental effect does not necessarily follow the pattern of the correlation structures shown in this section. In dairy cattle, the permanent environmental effect might be explained by differences between raising of animals before first calving, possibly having a large effect on the first part of lactation, and only then gradually decreasing. We could therefore require a method used to describe the change of covariances over time to be flexible, and not relying on predefined structures.

A flexible approach is to define a function for the covariance structure that is based on regression. The next section will describe the development of covariance functions based on regression on orthogonal polynomials. Like polynomial regression is suitable and flexible for fitting linear function of the means, it can be used to fit covariance structures. Alternatively, models to fit covariance structures over time could be based on time variables defined based upon a biological model (e.g. growth and lactation curves). Such models will be presented in a later section, when random regression will we discussed.

Random Regression Models

Random regression models can typically be used when a trait is expressed repeatedly, e.g. over time or in different environments. In that case, the effect changes gradually along a trajectory of time, or of some other continuous variable (temperature, elevation, rainfall. For simplicity, we think of the expression of body weight as a function of time. If the random effects are modeled as a function of time, then both the variance as the covariance between expression at different times are modeled as a continuous function. Note that previously we often modeled repeated measures of weight as multiple traits, e.g. weaning weight, post-weaning weight, yearling weight. The advantage of random regression is that traits can be measured at any point along a trajectory, i.e. at any age, and we do not have to chop this up in distinct traits.

In linear models we are used to fitting weight as a regression of age. This is often a fixed regression, indicating that for each animal that is a certain amount of time younger or older than an average age there will be a weight correction. This correction is the same for all animals, hence a fixed regression. In random regression models, we estimate a different regression coefficient for each animal. Hence, each animal has his/her own slope (some grow faster than others) and we estimate the variance of all slope parameters. An animal individual's slope is estimated as a BLUP, depending on the variance of slopes (like the breeding value is derived from the variance of breeding values.

Hence, each animal may have 3 breeding values for weight, if we fit a three order regression. The first is an intercept, the 'average weight', i.e. how much the animal deviates from the population mean over all ages, the second is a slope, 'the growth', i.e.

how much the animals deviates more/less at the beginning/end of the trajectory from other animals' weight, and the third is a quadratic term (harder to interpret)
The regression coefficients are not the same for each animal, but they are drawn from a population of regression coefficients. In other words, regression coefficients in ***a*** and ***p*** are *random regression coefficients* with var(***a***)= $\boldsymbol{K}_a$ and var(***p***)= $\boldsymbol{K}_p$, where a is additive genetic effect and p is permanent environmental effect. If the second term of **a** represents a random 'slope', and the variation in this term gives an indication of variation in growth curve. If most of the variation is in the first term, there is no genetic variation in growth curve, and individual growth curves would be pretty much parallel to each other. A lot of variation in 'slope' means that we can select on 'growth curve benders'. As a geneticist, think of genes for early vs. genes for late growth. Note the similarity of this interpretation with multivariate models (the higher genetic correlations between weight at different ages, the smaller the prospect of selection for growth curve benders), and with principal components analysis (different variables indicating different aspects of a process – the different random regression variables could be made orthogonal by decomposing the K-matrices, e.g. see Van der Werf, 2002)

In fact, we have rewritten a multivariate mixed model to a mixed model in a format of a univariate random regression model, with each random effect having *k* random regression coefficients. A model for n observations on q animals can then be written as

$$\mathbf{y}= \mathbf{Xb}+ \sum_{j=0}^{k-1} \mathbf{Z}_j\mathbf{a}_j + \sum_{i=0}^{k-1} \mathbf{Z}_j\mathbf{p}_j + \mathrm{e}, \qquad [4\text{-}6]$$

where $\mathbf{Z}_j$ are n by q matrices for the ith polynomial, and $\mathbf{a}_j$ and $\mathbf{p}_j$ are vectors with random regression coefficients for all animals for additive genetic and permanent environmental effects. The matrix **Z** contains the regression variables, i.e. the coefficients are those of the polynomials in F **(i.e. rather than a 1's, Z contains 1, x, x^2, etc.**. We can order the data vector by sorting records by animal, and we can stack the $\mathbf{a}_j$ and $\mathbf{p}_j$ vectors and sort them by animal, each animal having *k* coefficients in **a** and *k* coefficients in **p** (to simplify notation, we assume equal order of fit for CF's for both random effects, therefore having equal incidence matrices). We can then write $\mathbf{Z}^*$ as a block diagonal matrix of order *n by k*q,* with for each animal *i* block $\mathbf{Z}_i^*$= F_i.
The mixed model can be written as

$$\mathbf{y}= \mathbf{Xb}+ \mathbf{Z}^*\mathbf{a} + \mathbf{Z}^*\mathbf{p} + \mathrm{e},$$

with $\mathbf{a}'= \{\mathbf{a}_1' ,...\mathbf{a}_q'\}$ and $\mathbf{p}'= \{\mathbf{p}_1' ,...\mathbf{p}_q'\}$, with $\mathbf{a}_i$ and $\mathbf{p}_i$ being the sets of random regression coefficients for animal i for the additive genetic and the permanent environmental effects, respectively. If all animals have measurements on the same age points, all $\mathbf{Z}_i^*$ are equal and $\mathbf{Z}^*= \mathbf{I_q}\otimes \mathrm{F}$;

The variances and covariances of the random effects for the model can be written as:

$$\mathrm{var}\ (\mathbf{a})= \mathbf{A}\otimes\mathbf{K}_a$$

$$\text{var}(\mathbf{p}) = \mathbf{I} \otimes \mathbf{K}_p$$

and $\text{cov}(\mathbf{a},\mathbf{p})=0.$

where $\mathbf{K_a}$ and $\mathbf{K}_p$ are the random regression variances for a additive genetic and permanent environmental effects, respectively. The mixed model equations for the random regression model with covariance functions (RR -model) have a similar structure as a repeatability model, except that more coefficients are generated through the polynomic regression variables from Φ which are incorporated in **Z**. In the additive genetic effects part of the equations there is for each animal a diagonal block $\Phi_i'\Phi_i$ + $a^{ii}\sigma_\varepsilon^2 K_a^{-1}$, and there are off diagonal blocks $a^{ij}\sigma_\varepsilon^2 K_a^{-1}$ with a^{ij} the (i,j)th element of the inverse of the numerator relationships matrix ($\mathbf{A}^{-1}$). The part for the permanent environmental effects is block diagonal with diagonal blocks equal to $\Phi_i'\Phi_i + \sigma_\varepsilon^2 K_p^{-1}$. Schematically, the mixed model equations will be like

$$\begin{bmatrix} X_i'X_i & \cdots & X_i'\Phi_i & \cdots & X_i'\Phi_i & \cdots \\ \vdots & . & \vdots & . & \vdots & . \\ \Phi_i'X_i & \cdots & \Phi_i'\Phi_i + a^{ii}s_e^2K_a^{-1} & \cdots & \Phi_i'\Phi_i & \cdots \\ \vdots & . & \vdots & . & \vdots & . \\ \Phi_i'X_i & . & \Phi_i'\Phi_i & . & \Phi_i'\Phi_i + s_e^2K_p^{-1} & . \\ \vdots & . & \vdots & . & \vdots & . \end{bmatrix} \begin{bmatrix} b \\ \vdots \\ a_i \\ \vdots \\ p_i \\ \vdots \end{bmatrix} = \begin{bmatrix} X_i'y_i \\ \vdots \\ \Phi_i'y \\ \vdots \\ \Phi_i'y \\ \vdots \end{bmatrix}$$

where the subscript i refers to those part of the equations for animal i. For the earlier example, we a 3-order CF with measurements at standardized ages [-1 0 1], Φ'Φ is

The ASREML package can be used for random regression analysis. The latter package requires the user to define a regression model (e.g. a 3rd order polynomial regression on 'days in milk = dim', and random regression is achieved by defining a random interaction term between animal and this polynomial regression term.

weight = herd poly(dim,4) !r poly(dim,3).animal

The first term is a 4th order polynomial regression of milk on days in milk (*dim*) as a fixed effect. This basically fits an average lactation curve equal for all animals. The random term indicates individual animal variation around this mean curve, interpreted as a dim by animal interaction (each animal another regression coefficient on dim) Alternatively, in ASREML, the regression coefficients (e.g. the Legendre regression on age as in the F matrix for each animal) can be constructed 'by hand' based on the age of the measurement and provided in a data file. ASREML allows estimation of variances and covariance components between these regression coefficients when they are taken as random. This covariance matrix should be equal to the **K**-matrix.

Covariance functions

A covariance function can be defined as " a continuous function to give the variance and covariance of traits measured at different points on a trajectory", see e.g. Kirkpatrick et al (1990). Covariance functions are automatically estimated in random regression models; a covariance function is defined for each of the random effects that explain variation. Covariance functions (and RR models) can be defined based on many regression models, e.g. polynomials, Legendre polynomials, splines, etc). In mathematical terms, a covariance function (CF), e.g. for the covariance between breeding values u_l and u_m on an animal for traits measured at ages x_l and x_m is:

$$\mathrm{cov}(u_l, u_m) = (f(x_l, x_m) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \phi_i(x_l)\phi_j(x_m) k_{ij} \qquad [3\text{-}1]$$

where ϕ_i is the ith (i=0,..,k-1) polynomial for a k-order of fit, x is a standardized age ($-1 \le x \le 1$) and k_{ij} are the coefficients of the CF. The ages can be standardized by defining a_{min} and a_{max} as the first and the latest time point on the trajectory considered, and standardizing age a_i to $x_i = [2(a_i-a_{min}) /(a_{max}-a_{min})]-1$.

The CF can be written in matrix notation as

$$\hat{G} = \mathrm{F}\,\mathbf{K}\mathrm{F}'$$

where $\hat{G}$ is the genetic covariance matrix of order *t* for breeding values at t given ages, F is a *t by k* matrix with orthogonal polynomials. When using Legendre polynomials, the matrix F can be written as $\mathbf{M}\mathrm{L}$, with $\mathbf{M}$ being a *t by k* matrix with elements $m_{ij} = a_i^{(j-1)}$ (i=1,..t; j=1,..k), and L being a matrix of order k with Legendre polynomial coefficients.

Besides estimating CF directly from data via Random Regression, one can also estimate them from a covariance matrix $\hat{G}$ that was previously estimated. For example, a genetic covariance matrix among siz 50-day lactation periods (1-50; 51-100; etc) is given in Table 4. A CF of order 3 is estimated, approximating these coefficients via best fit (for details, see Kirkpatrick et al., 1990). Note that a CF of order 6 would have given a full fit. In addition, estimates for a 3rd order RR model obtained from the same data via REMl are given.

Table 4 and Figure 1 show that for a 3 order fit genetic variances estimated from random regression were higher at the periphery of the trajectory. Also covariances between ages most far apart were more extreme in the CF estimated from RR model. Genetic correlation between the first and the last month of lactation was near zero with the CF from the RR model, whereas it was near 0.7 in a bivariate analysis (Table 4-4).

From this comparison, it appears that estimating CF parameters from a random regression model may not always give reliable parameters. In our example, particularly genetic variance was overpredicted near the edges. This example shows a typical property of random regression models, that variances and covariances tend to be exaggerated near the edges. This is very much a property of polynomials in general, although the extend would depends somewhat on the availability of data near the extremes of the trajectory. Often, there is little data, and a lower order of fit might give a quite good likelihood of the whole data, in spite of its bad fit near the edges. A higher order fit does not always resolve these problems, and the use of splines is usually advocated as a remedy (Gilmour, 2006) .

TABLE 4. Pre-estimated genetic covariance matrix ($\tilde{G}$) and fitted matrices using CF coefficient from REML in a random regression model, and from fitting $\tilde{G}$ (variances on diagonal, correlations on off-diagonals)

$\tilde{G}$: pre-estimated covariance matrix						
	2.7576	0.9250	0.8940	0.8419	0.7318	0.7290
	0.9250	1.7367	0.9634	0.9240	0.7905	0.6268
	0.8940	0.9634	2.4029	0.8250	0.8160	0.7074
	0.8419	0.9240	0.8250	1.6709	0.8512	0.6395
	0.7318	0.7905	0.8160	0.8512	2.4600	0.8951
	0.7290	0.6268	0.7074	0.6395	0.8951	2.4519
G estimated from CF from REML random regression model (order 3)						
	5.4653	0.9522	0.7786	0.4982	0.1795	-0.0759
	0.9522	3.7534	0.9306	0.7198	0.4041	0.0877
	0.7786	0.9306	2.8598	0.9181	0.6745	0.3501
	0.4982	0.7198	0.9181	2.5851	0.9051	0.6600
	0.1795	0.4041	0.6745	0.9051	2.9545	0.9128
-	0.0759	0.0877	0.3501	0.6600	0.9128	4.4768
G estimated from $\tilde{G}$ (3-order fit Legendre polynomials						
	2.4413	0.9643	0.8962	0.8528	0.8110	0.6856
	0.9643	2.0702	0.9799	0.9444	0.8612	0.6526
	0.8962	0.9799	1.8981	0.9843	0.8955	0.6553
	0.8528	0.9444	0.9843	1.7387	0.9528	0.7483
	0.8110	0.8612	0.8955	0.9528	1.7097	0.9121
	0.6856	0.6526	0.6553	0.7483	0.9121	2.3127

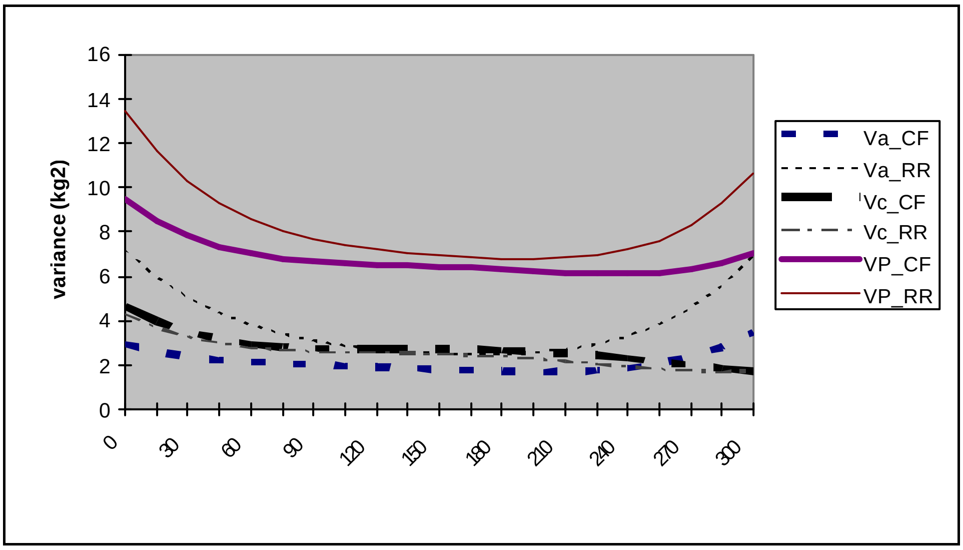

Figure 1 Phenotypic, additive genetic and permanent environmental variances over lactation estimated by covariance function (3-order Legendre polynomials) from multiple trait variance-covariance matrices (CF) and by REML directly from data (RR)

The conclusion is that although it is theoretically most appealing to estimate CF parameters directly from a random regression model, this method may not always give the most reliable estimates. Other regression techniques than polynomials (e.g. the use of splines), and other statistical models (e.g. varying the temporary environmental variance along the trajectory) an often help.

Analyzing patterns of variation

Kirkpatrick and Heckman (1989) and Kirkpatrick et al (1990) show that covariance functions can be used to analyze 'patterns of inheritance' in the covariance matrix $\tilde{G}$. For this purpose they determined eigenvalues and eigenfunctions from the coefficient matrix for a given covariance function.

In a way, this is a similar approach as principal component analysis. If we consider the covariance structure among 25 type traits in dairy, we might be able to say that one main eigenvalue is due to some kind of linear combination of all type traits related to udder scores. We would find this if this is a group of traits highly correlated among each other, but not highly correlated to other traits. In the canonical decomposition of covariance functions, determining such major components has a special meaning, because it shows at which ages the observed variables are correlated, and where they are not. In other words, it shows how independent variables act on the trait along the trajectory. For example we may determine that a first major eigenvalue is related to a linear combination of test days in the first part of lactation (the combination being defined by the eigenvector attached to that eigenvalue), whereas another eigenvalue may be a combination of test day variables in later lactation. If this was found for the genetic covariance matrix, the interpretation could be that two main and independent components could be distinguished in milk production, each acting on different parts of lactation, and those two components could be related to different genes, possible on two different parts of the genome. The last would be of interest in QTL analysis: one canonical variable could be linked to one marker, whereas another is linked to another marker. Other examples include: analysis of growth over time (different genetic effects for different part of the growth curve) and genotype by environment interaction (variation in environmental sensitivity)

In contrast to multiple traits, the variables in repeated measurement can be ordered along a trajectory. In that case, The transformation of variables described by each eigenvector can be written as a continuous function of age. This is indicated as eigenfunction (Kirkpatrick and Heckman, 1989). Eigenfunctions are calculated as follows:

Consider the covariance function

$$\hat{G} = \mathrm{F}\,\mathbf{K}\mathrm{F}'$$

for a set of ages in age vector **a**, where the age coefficients are build in the regression coefficient in F . The matrix **K** is decomposed into eigenvalues **D** and eigenvectors **E** as **K** =**EDE**', and we can then evaluate eigenfunctions for a give set of ages as F**E**

Taking the earlier example:

$$\hat{K} = \begin{vmatrix} 1348 & 66.5 & -111.7 \\ 66.5 & 24.3 & -14.0 \\ -111.7 & -14.0 & 14.5 \end{vmatrix}$$

$$D = \begin{vmatrix} 1361 & 0 & 0 \\ 0 & 24.5 & 0 \\ 0 & 0 & 1.5 \end{vmatrix} \qquad E = \begin{vmatrix} -0.995 & -0.079 & 0.056 \\ -0.050 & 0.915 & 0.400 \\ 0.083 & -0.395 & 0.915 \end{vmatrix}$$

and $\Phi E = \begin{vmatrix} -0.511 & -1.802 & 0.997 \\ -0.769 & 0.256 & -0.684 \\ -0.634 & 0.441 & 1.976 \end{vmatrix}$

The columns of F**E** represent eigenfunctions, and each has an eigenvalue attached to it. The rows refer to each of the ages, i.e. -1, 0 and 1 for row 1,2 and 3, respectively.

To obtain the eigenfunction coefficients, we have to use L**E** which is:

$$\Lambda E = \begin{vmatrix} -0.769 & 0.256 & -0.684 \\ -0.062 & 1.121 & 0.489 \\ 0.1971 & -0.937 & 2.170 \end{vmatrix}$$

And the first eigenfunction could be written as

$$y_1(x) = -0.769 - 0.062x + 0.197x^2$$

Figure 4 shows the three eigenfunction plotted for the example of Kirkpatrick et al. (1990). It should be noticed that the sign a the evaluated values between eigenfunctions is irrelevant (for example, the firs eigenfunction has only positive values in the Genetics paper of Kirkpatrick et al. What matter is how the values of the eigenfunctions change over the trajectory. In this example, the main eigenfunction is almost constant for all ages. Since it has the largest eigenvalues attached to it, the interpretation is that the major part of the genetic variance is explained by a factor that is constant for all ages. Selection on this factor will increase weight for all ages. Since this eigenvalue is very dominant, selection for weight at any age will improve weight on all ages. In multiple trait terms, weight at different ages is highly correlated (from the G used in the example, we can calculate correlations of 0.88 between weight at 2 and 3 weeks; 0.86 between weight at 2 and 4 weeks ,and 0.99 between weight at 3 and 4 weeks.

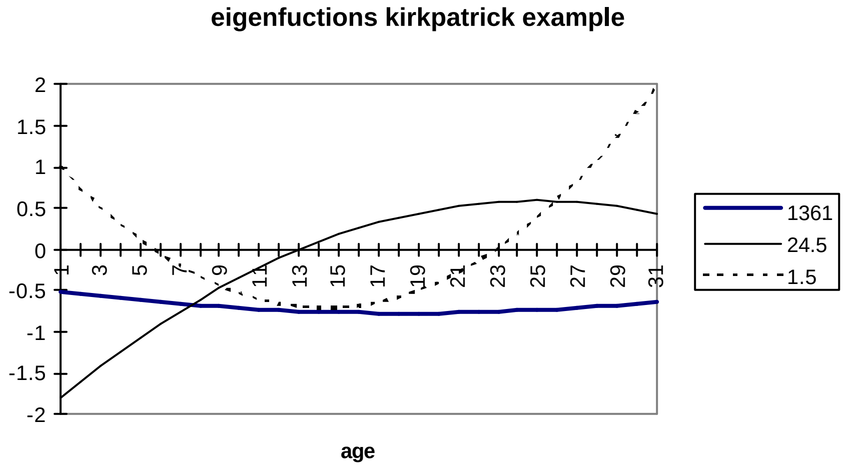

Figure 4 Eigenfunctions for the example of Kirkpatrick et al (1990)

A very interesting pattern is shown by the second eigenfunction. Selection on this variable decreases weight at early age and increases weight at later ages. Selection on the variable represented by this eigenfunction could therefore be used to change the growth curve, e.g. select for lower weight at start and higher weight at the end of a trajectory considered. In this example, the 2nd and 3rd eigenvalues are not very large (relative the first eigenvalue, and the possibilities to change the growth curve may be limited.

Notice that we could have drawn the same conclusion from inspection of the high genetic correlations represented in G. However, with considering more ages along the trajectory would make such interpretations more difficult. We could also have calculated eigenvalues of G directly, being 1714, 82 and 6, i.e. not a very different pattern than the eigenvalues from K. However, it is important to see that the ages used in this example were symmetrically chosen. In multiple trait evaluation this is not necessarily the case. Therefore, the most important difference between eigenvalue decomposition of a multivariate covariance matrix and a eigenvalue decomposition of a covariance function is that the last takes the ordering measurement along a trajectory into account.

Summarizing Discussion

We have presented various ways to analyze repeated measurements where interest is in a model that uses correct variance covariance structures between the observations, and that can make use and enable inferences on the gradual change of the measurements over time. A random regression model seems the most appropriate for modeling such data, and such models more or less implicitly use covariance functions. Canonical transformation can be used to simplify large scale genetic evaluations, and to reduce the rank of the covariance matrices used for each random effect. Eigenvalue decomposition possibly reveal patterns in the covariance structure, and might be of help to implement selection rules that aim for a change of the curves. Such analysis might also be useful when detecting more specifically the mode of action of Quantitative Trait Loci in specific parts of the genome, determined by genetic markers, or to identify parts of genetic variation that are specifically correlated with third traits of interest. Such analysis has similarities with principal component analysis, but the extra dimension is added by considering principal components as a function of time (eigenfunctions).

Examples of the use of random regression or covariance functions are: analysis of test day models, analysis of weight and growth data, feed intake, etc (note, this could include component traits, e.g. analysis of fatness traits as a function of weight. Trait measurements can be modeled as a function of time, but also as a function of a continuous environmental variable (herd production level, ambient temperature, etc). For example genetic variation in susceptibility to heat can be modeled by regression production on a heat stress index (a function of temperature and humidity). Variation in susceptibility to disease can be measures as a regression of parasite infection level on an environmental variable that measures environmental risk to disease.

Many studies have considered genetic aspects of growth by first estimating parameters for growth curves, and subsequently estimating variance components for the growth curve parameters. Such analysis could be improved upon by the use of random regression models. Main differences between these approaches is that the first (two-step) approach maybe less able to estimate curves for animals with missing data, and more general, does not use information from relatives. The values of such information is well known to animal breeders, not only in gaining accuracy, but also to account for directional selection. Varona et al (1997) presented random regression models in a Bayesian manner, and give a good discussion on the merits of such models over the two-step procedure with first estimating curve parameters and subsequently estimating their variance components.

It is often of interest not only to analyze the behavior of a repeatedly measured trait over time as such, but also to study correlations of certain curve parameters with ‘third’ traits. An example is to study growth curves by random regression models for weight data, and to correlate CF parameters with meat quality traits such as fat and muscle. Animals that tend to grow faster in the last phase before slaughter may have also a different pattern for onset of body fat, different mature weight and a different maturity rate (age at first calving!). Multivariate random regression analyses are required here. Such analyses will form a computational challenge (with a need to explore robustness of the estimation) but will be the basis of a very interesting biological debate on how to improve such dynamic traits of growth and development such that animals will have improved performance in the prevailing production system.

Fitting a curve with many parameters will generally give an accurate fit of the covariance structure. However, there is often an interest in fewer numbers of parameters. Models based on ‘biological curves’ may appeal because certain parameters have a ‘biological meaning’. The biological meaning from polynomials can be determined by plotting eigenfunctions from polynomials. Lindsey (1993) argues that it is generally preferable to choose a model that describes the mechanism that generates the data. Herewith, we can refer to modeling certain residual covariance matrices, which may have autocorrelation structures. Lindsey also discusses growth curves and refer to Sandland and McGilchrist (1979) who provide a number of reasons why polynomials are unattractive for growth models:

1) growth processes can undergo changes of phase which cannot be accommodated by polynomials
2) the stochastic structure of the model will be distorted if the polynomial is inappropriate
3) polynomials cannot easily represent asymptotic behavior of a growth curve.

Anderson and Pedersen (1996) argue that many growth curves are non-linear functions (e.g. Gompertz, logistic regression) for which is more difficult to introduce random effects. They also argue that average curves of the exponential form (e.g. y= a exp(-bx-c/x) are not of the same form if the parameters a, b and c vary from animal to animal. Sometimes, transformations to linear models are possible, although transformations to stabilize between animal variation may destabilize within animal variation (see Anderson and Pedersen (1997) for an example).

The need to avoid polynomials depends on the trajectory considered. In certain instances, it may be more important to accurately account for asymptotes, in which case polynomials are less appropriate. The use of splines is often advocated as being a robust technique in regression analysis and should probably be considered as very useful in random regression analysis as well. Also, the behavior of different random regression models in relation to data structure needs more study. For some traits, there may be many more data point at the younger ages, and there may be sequential selection. In general, estimating covariance matrices between certain ages of the trajectory can be useful as a reference for checking parameter estimates for covariance functions, as was also demonstrated in these notes.

In general, arguments for fitting mean growth curves for populations, or subpopulations, can also be used for random regression models. The same holds true for the number of parameters that should be used to fit regression models. However, a practical argument for analyzing (large size) animal breeding data is that more random regression coefficients rapidly increase computing demands, and that for predicting breeding values, accurately fitting first moments (means) is usually more important than accurately fitting second moments (variances)

References

Anderson, S, and Pedersen, B. 1996. Growth curve and food intake curves for group housed gilts and castrated male pigs. Animal Sci. 63:457-464.

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. 1994. Analysis of longitudinal data. Clarendon Press, Oxford.

Ducrocq, V.P., and Besbes. 1993. Solution of multiple trait animal models with missing data on some traits. J. Anim. Breed. Genet. 110:81-92.

Hayes, J.F. and Hill, W.G., 1981. Modification of estimates of parameters in the construction of genetic selection indices ('bending') Biometrics 37:483-493.

Jamrozik, J. and L.R. Schaeffer. 1997. Estimates of genetic parameters for a test day model with random regressions for production of first lactation Holsteins. J. Dairy Sci. (in press).

Johnson, D.L. and Thompson, R. 1995. Restricted maximum Likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. J. Dairy Sci. 78_449-456.

Kirkpatrick, M., Lofsvold, D., and Bulmer, M. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. Genetics 124:979-993.

Kirkpatrick, M., Thompson, R., and Hill, W.G. 1994. Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. Genetical Research 64:57-69.

Lindsey, J.K. 1993. Models for repeated measurements. Clarendon Press, Oxford.

Meyer, K, and Hill, W.G. 1997. Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated' records by Restricted Maximum Likelihood. Livest. Prod. Sci. 47:185-200.

Meyer, K. 1997. Estimates of covariance functions for mature weight of beef cows in the Wokalup selection experiment. Proc. Assoc. Advmt. Anim. Breed. Genet.

Meyer, K. 1997. Estimation of genetic and phenotypic covariance functions for longitudinal data. Proc. Assoc. Advmt. Anim. Breed. Genet. 12: 534-537.

Meyer, K. DFREML and covariance functions with random regression. In preparation

Ptak, E., and L.R. Schaeffer. 1993. Use of test day yields for genetic evaluation in dairy sires and cows. Livest. Prod. Sci. 34:23

Sandland, R.L. and McGilchrist, C.A. 1979. Stochastic growth curve analysis. Biometrics 35:255-271.

Varona, L., Moreno, C, Carcia Cortes, L.A., and Altarriba, J. 1997. Multiple trait genetic analysis of underlying biological variables of production functions. Livest. Prod. Sci. 47: 201-209.

Van der Werf, J.H.J., Goddard, M.E., and Meyer, K. 1997. The use of covariance functions and random regression for genetic evaluation of milk production based on test day records.

Veerkamp, R.F. and M.E. Goddard. 1997. Covariance functions across herd production levels for test day records on milk fat and protein yield. J. Dairy Sci. (abstract).

Visscher, P.M. 1994. Bias in genetic R^2 from half sib designs. Proc. 5th WCGALP Guelph.

Wiggans, G. and Goddard, M.E. 1997. Test day model with 30 traits and genetic covariance matrix of reduced rank. J. Dairy Sci. In Press