Statistical Models for the Genetic Analysis of Longitudinal Data

Florence Jaffrézic

Doctor of Philosophy

University of Edinburgh

2001

Table of Contents

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

Florence Jaffrézic
September, 2001

Acknowledgements

I would like to thank especially Bill Hill who has always believed in me and in my ideas. He always was of great support and gave me extremely useful encouragements. It was a very important motivation.

Very special thanks also to Ian White who has always been here for my every day questions. Thank you so much for his patience, his good mood and his very useful help.

I wish to express my gratitude to Robin Thompson for always showing interest in my work and for his very valuable ideas. Thank you also for the implementation with Arthur Gilmour of the OWN function in ASREML, which has proved to be extremely useful.

Many thanks to Peter Visscher and Vincent Ducrocq for interesting discussions and comments, and to Sue Brotherstone for always being very kind and helpful for all the computer problems.

I am especially grateful to Scott Pletcher for extremely useful exchanges of ideas and work about model comparisons. It has been an invaluable help for my PhD. I will never thank him enough !

I acknowledge with appreciation the financial support received from the INRA (Station de Génétique Quantitative et Appliquée, Jouy-en-Josas).

I wish to thank all the people I have shared an office with, especially Heli (for her very useful help with Fortran and Latex !), Stuart and Victor for nice discussions, as well as Lutz for entertaining coffee breaks. Very special thanks also to all my friends in Edinburgh who made my stay so pleasant !

Je voudrais dédier cette thèse à mes parents sans qui je ne serais jamais arrivée jusqu'ici. Merci de tout mon coeur ! Merci aussi à Pascal pour son soutien et sa patience. Merci d'avoir toujours été là, même dans les moments difficiles.

List of publications

Statistical models for estimating the genetic basis of repeated measures and other function-values traits.
(F. Jaffrezic, S.D. Pletcher)
Published in Genetics 156 (2000) 913-922.

A link function approach to model heterogeneity of residual variances over time in lactation curve analysis.
(F. Jaffrezic, I.M.S. White, R. Thompson, W.G. Hill)
Published in Journal of Dairy Science 83 (2000) 1089-1093

Nonparametric estimation of covariance structure for genetic analysis of repeated measures and other function-values traits.
(F. Jaffrezic, S.D. Pletcher, W.G. Hill)
Submitted to Genetical Research (2001)

Contrasting models for lactation curve analysis.
(F. Jaffrezic, I.M.S. White, R. Thompson, P.M. Visscher)
In press in Journal of Dairy Science (2001)

Modelling and analysis of incomplete and short lactations.
(F. Jaffrezic, P. Minini)
Submitted to Journal of Dairy Science (2001)

Generalized character process models: Estimating the genetic basis of traits that cannot be observed and that change with age or environmental conditions.
(S.D. Pletcher, F. Jaffrezic)
In press in Biometrics (2001)

Abstract

Animal breeders and evolutionary geneticists are often faced with the problem of analysing traits that change as a function of age or some other independent and continuous variable. Three main approaches have been put forward to deal with this kind of data: random regression models, that are the most commonly used at present, character processes (CP) that have recently been proposed and focus on a parametric modelling of the covariance structure, and structured antedependence (SAD) models suggested in the statistical literature.

The first objective of this work was to compare and contrast these different methodologies for genetic analysis. As the range of all possible models can be very large in practice, it is advisable to have a preliminary idea of the covariance structure of the data, and a non-parametric approach based on the variogram was proposed. It is especially adapted for exploratory analysis when a large number of observations is available per subject over time and was applied to the analysis of daily records for milk production in dairy cattle. Model comparisons in the univariate case showed that character processes were generally better able to fit the covariance structure than random regression with fewer parameters. However, CP models do not allow a straightforward extension to the multivariate case. Further research showed that structured antedependence models offer similar advantages to character processes compared to random regression while allowing an extension to multi-trait analyses. SAD models were even able to capture the highly non-stationary correlation pattern in the application to lactation curve analysis. For genetic evaluation of dairy cattle, longitudinal models can easily provide estimation of individual cumulative milk productions as well as genetic values at 305 days. However, these predictions do not take into account the drying-off process and can be highly overestimated for short lactations. A methodology to correct them was suggested. All these analyses were performed in the case of normally distributed longitudinal data. An extension to the genetic analysis of non-normally repeated measures was considered. Estimation procedure becomes much more complicated and requires the use of Markov Chain Monte Carlo methods.

In this study antedependence models appeared to be the most appropriate for genetic analysis of longitudinal data. In their traditional specification, however, times of measurement were assumed to be on a discrete scale and equally spaced. This can be quite a stringent assumption in practice and a continuous extension of these models was proposed.

Chapter 1

General Introduction

Animal breeders and evolutionary geneticists are often faced with the problem of analysing traits that change as a function of age or some other independent and continuous variable. This work was initially motivated by the genetic analysis of lactation curve for dairy cattle. Advantages of using individual test day information instead of summary measure of 305-day yield to evaluate the genetic merit for milk production are widely accepted. Firstly, more environmental variation can be removed from the phenotypic observations by considering the effects acting on the repeated measures that cannot be taken into account when modelling 305-day yields. Secondly, larger accuracy of cows' genetic evaluations may be achieved because of the use of more data per animal. Thirdly, selection tools to improve lactation shape characteristics such as persistency may be obtained. Genetic analysis of repeated measures is also a more general issue and can be applied to a wide range of areas such as growth curve analysis of laboratory and agricultural species, or the study of age-specific fitness components such as reproductive output.

Detailed descriptions of the extension of classical quantitative genetics to the analysis of function-valued traits is given by Kirkpatrick and Heckman (1989) and Pletcher and Geyer (1999). In short, the method assumes the observed character is best described by a function (or stochastic process) of some independent and continuous variable. Although any continuous variable is acceptable (e.g., the level of some environmental factor), the most common is age, and all of the examples will focus on characters that change with age. Further, it is assumed that the character values at each age constitute a multivariate normal distribution on some scale. This assumption is for most practical cases a good approximation and is very convenient for calculations. It was however relaxed by Pletcher and Jaffrezic (2001, appended to this thesis).

As with traditional quantitative genetics, it is assumed that the observed phenotypic trajectory of the character is random and influenced by one or more unobservable factors. In the simplest case one might consider the additive contribution of many genes along with unpredictable environmental effects. More complicated models involving interactions among different genes or specific environmental effects (e.g., maternal effects) are straightforward, although computational difficulties will likely arise. For the additive model, we assume the observed phenotype can be decomposed as

$$X(t) = \mu(t) + g(t) + e(t) + \epsilon, \tag{1.1}$$

where $\mu(t)$ is a nonrandom function, the genotypic mean function of $X(t)$, and $g(t)$ and $e(t)$ are Gaussian random functions, which are independent of one another and have an expected value of zero at each age (Kirkpatrick and Heckman, 1989 ; Pletcher and Geyer, 1999). They represent the age-dependent genetic and environmental deviations, respectively. In this context, $e(t)$ is often referred to as the permanent environmental effect and ϵ is the residual variation—ϵ is assumed normally distributed with constant and unknown variance over time.

The goal of the analysis is to decompose the observed variation in $X(t)$ into its genetic and environmental contributions by estimating *covariance functions* for $g(t)$ and $e(t)$. A covariance function, $r(s,t)$, is a bivariate continuous function that describes the covariance between any two ages, $r(s,t) = \text{Cov}\{X(s), X(t)\}$. By the independence of $g(t)$ and $e(t)$, the phenotypic covariance function of $X(t)$

is given by $P(s,t)$ as

$$P(s,t) = G(s,t) + E(s,t). \tag{1.2}$$

where $G(s,t)$ is the genetic covariance function, and $E(s,t)$ the environmental covariance function, which also includes the residual variance. These functions are estimable via maximum likelihood (ML) or restricted maximum likelihood (REML) when there are data on individuals of various relatedness (Lynch and Walsh, 1998 ; Pletcher and Geyer, 1999).

There have been at least four different methods suggested for estimating the desired covariance functions: orthogonal polynomials (Kirkpatrick and Heckman, 1989), random regression (Diggle et al., 1994 ; Jamrozik et al., 1997 ; Meyer, 1998), character processes (Pletcher and Geyer, 1999), and structured antedependence models (Nunez-Anton and Zimmerman, 2000). All four methods are based on likelihood estimation—although the orthogonal polynomial approach was originally published as a least squares estimation (Kirkpatrick et al., 1990).

Random Regression (RR): Random regression models employ parametric forms for the unobserved functions in (1.1). Although traditionally a parametric mean curve is often used to estimate $\mu(t)$, this is not essential. However, the individual deviations from this curve (i.e., the $g(t)$ and $e(t)$) are assumed to be parametric functions of time, and polynomials are often used. For example, the age-dependent deviations from the population mean due to an individual's genotype might be linear in time, such that

$$g(t) = a_1 + a_2 t.$$

where the a_i are random genetic regression coefficients. The regression coefficients are unobservable, random effects; they have a specific value for each individual; and they are assumed to be multivariate normally distributed. The environmental deviations, $e(t)$, are assumed independent of the genetic effects, and they are modelled similarly.

Genetic and environmental covariances as a function of age are determined by the variances and covariances among the regression coefficients. Following the example presented above, the genetic covariance between ages s and t is

$$\begin{aligned} G(s,t) &= \mathrm{Cov}(g(s), g(t)) \\ &= \mathrm{Cov}(a_1 + a_2 s, a_1 + a_2 t) \\ &= \mathrm{Var}(a_1) + (s+t)\mathrm{Cov}(a_1, a_2) + st\mathrm{Var}(a_2). \end{aligned}$$

The primary objective in these models is to choose the most appropriate parametric functions for the genetic and the permanent environmental deviations. In many cases the parametric functions are nested and likelihood ratio testing can be used. Since this involves testing the significance of parameters on the boundary of their feasible parameter space, the test statistics are often mixtures of Chi-square distributions (Stram and Lee, 1994).

Character process model (CP): In contrast to the RR models, the character process model does not attempt to model the forms of the $g(t)$ or $e(t)$ functions. Instead, parametric models for the covariance functions themselves (i.e., $G(s,t)$ and $E(s,t)$ in equation (1.2)) are the target of analysis (Pletcher and Geyer, 1999).

Again taking the genetic covariance function as an example, the covariance function can be decomposed into

$$G(s,t) = v_G(s)v_G(t)\rho_G(|s-t|) \tag{1.3}$$

where $v_G(t)^2$ describes how the genetic variance changes with age and $\rho_G(|s-t|)$ describes the genetic correlation between two ages. There are no restrictions on the form of $v_G(\cdot)$, and it is often modelled using simple polynomials (linear, quadratic, etc.). As presented in Pletcher and Geyer (1999) the character process model assumes correlation-stationarity, i.e. the correlation between two ages is assumed to be a function only of the time distance ($|s-t|$) between them. Although strictly speaking this assumption is almost surely wrong, experience suggests that it is expected to provide a reasonable approximation in most cases (Pletcher and Geyer, 1999). The benefit of correlation stationarity is that it allows numerous choices for $\rho(\cdot)$, all of which satisfy several theoretical requirements (Pletcher and Geyer, 1999). Chapter 2 presents a possible non-stationary extension of these models using a non-linear transformation upon the time axis. When the data are collected at equally spaced intervals, CP models with a constant variance and an absolute exponential correlation ($\rho(s,t) = \theta_c^{|s-t|}$) function are equivalent to an autoregressive model of order 1.

Orthogonal Polynomials (OP): Kirkpatrick and Heckman (1989) originally presented the use of orthogonal polynomials as a non-parametric way of "smoothing" previously estimated covariance matrices. This was the first attempt to formalize the estimation of covariance functions in a genetic context. As with the CP model, the shapes of the individual age-dependent deviations were not considered, and models for the structure of the variance-covariance matrix itself were the focus of attention. Kirkpatrick and Heckman (1989) suggest that the genetic covariance function be represented as

$$G(s,t) = \sum_{i=0}^{m}\sum_{j=0}^{m} \phi_i(s)\phi_j(t)k_{ij} \tag{1.4}$$

where m determines the number of polynomial terms used in the model, k_{ij} are the $m(m+1)/2$ unknown parameters to be estimated (the coefficients of the linear combination), and ϕ_i is the i^{th} Legendre polynomial (Kirkpatrick et al., 1990). The environmental covariance function is modelled similarly. Meyer and Hill (1997) present a method for estimating covariance functions such as (1.4) directly from the data using REML.

As originally presented, the orthogonal polynomial approach is similar in spirit to the CP model, and both differ in principle from the RR approach. In the RR methods, the primary model development occurs at the level of individual deviations (equation (1.1)). The analyst begins by considering the behaviour of individual age-specific deviations. The resulting covariance structure is a consequence of these deviations. For the CP and OP models, the situation is reversed. The analyst begins by considering the structure of the covariance matrix (equation (1.2)), and the shapes of the individual deviations are a consequence of this structure. Although RR and OP methodologies were originally proposed as two different concepts, Meyer (1998) showed that in most cases they are equivalent.

Structured antedependence models: The concept of this methodology is again different from the previous ones. The idea of antedependence models, as originally proposed by Gabriel (1962), is that an observation at time t can be explained by the previous ones. An antedependence structure of order r is defined by the fact that the ith observation ($i > r$) given the r preceding ones is independent of all further observations (Gabriel, 1962). Generalizing this concept to genetic analysis, a second order structured antedependence model for the genetic part $g(t)$ can be written as:

$$g(t_0) = \epsilon_g(t_0) \tag{1.5}$$

$$g(t_1) = \phi_1 \; g(t_0) + \epsilon_g(t_1) \tag{1.6}$$

$$g(t_j) = \phi_1 \; g(t_{j-1}) + \phi_2 \; g(t_{j-2}) + \epsilon_g(t_j) \tag{1.7}$$

for $j \geq 2$. Here, ϕ_1 and ϕ_2 are regression parameters, and $\epsilon_g(t)$ is assumed to be normally distributed, with mean zero and variance $\sigma_g(t)$ that can change with time. This corresponds to a generalization of simple autoregressive models that assume constant variances. In structured antedependence (SAD) models, Nunez-Anton and Zimmerman (2000) propose to use a parametric function for variances $\sigma_g(t)$ using for example a polynomial of time. SAD models require very few parameters for the covariance structure, and increasing the order of antedependence only involves one extra parameter at each step. The same model can be written for environmental effects $e(t)$. At their first order, SAD models are closely related to character processes with an exponential correlation and a quadratic variance. The CP approach is more general in the sense that many different correlation functions can be considered. However, SAD models of order s allow more flexibility for the correlation function as $2s$ parameters are included whereas only 2 parameters were included for the CP models.

The first objective of this work was to compare and contrast these different methodologies and evaluate their performance. In Chapter 2, a variety of simulated data sets was explored and types of covariance structures (genetic and environmental) accommodated by each method are described. Empirical data on age-specific mortality and reproductive output in fruit fly, *Drosophila melanogaster*, and growth in beef cattle were considered. Ability of each model to adequately fit empirical data was evaluated.

The range of all possible models can be very large in practice, especially for the character process methodology where it is possible to combine different functions of variance and correlation for both genetic and environmental parts. It is in general not possible to investigate all the possible combinations, and it would therefore be extremely useful to have a preliminary idea of the covariance structure in order to choose the most appropriate model. The object of Chapter 3 is to propose a non-parametric approach based on the variogram (Diggle and Verbyla, 1998) especially adapted for exploratory analysis when a large number of observations is available per subject over time and data are unbalanced. The methodology is illustrated using both simulated data sets and actual data on age-specific fertility in *Drosophila* and daily records for milk production in dairy cattle.

Models were compared in Chapter 2 in the case of univariate genetic analysis. Character process models were generally better able to fit the covariance structure than random regression with fewer parameters. However, CP models do not have a straightforward extension to the multivariate case. We focus in Chapter 4 on structured antedependence models, that have similar advantages to character processes and can easily be extended to the multivariate case. Their performance was compared to random regression models. Bivariate phenotypic and genetic analysis of fertility and mortality in *Drosophila*, and of milk, fat and protein yields in dairy cattle are presented.

An application and comparison of the different models in the case of lactation curve analysis is presented in Chapter 5 in order to help deciding which would be the most appropriate for genetic evaluation of dairy cattle based on test-day records. The most commonly used at present are random regression models, but very few comparisons with other methodologies have been done. A methodology to model residual variances that change with time using a structural model is proposed in Chapter 6.

Genetic evaluation of dairy cattle for milk production requires prediction of individual genetic values at 305 days. Longitudinal models can easily provide these predictions as well as individual cumulative milk productions which are much more accurately estimated than with extrapolation procedures previously used. However, classical models (random regression, character process, antedependence models) ignore completely the drying off process and predictions obtained rely on the assumption that cows are never made dry. This can be a problem especially for cows with shorter lactations as it will induce an overestimation of the predicted productions, and eventually of the genetic values. A methodology that corrects predictions obtained with longitudinal models for the probability of each cow to be dried off at each time is proposed in Chapter 7.

All these analyses were performed in the case of normally distributed longitudinal data. An extension of character process models to the genetic analysis of non-normally distributed repeated measures is appended to this thesis. Estimation procedure becomes much more complicated and Markov Chain Monte Carlo methods were used. This approach was investigated using simulated data and applied to a large data set measuring mortality rates in *Drosophila* (Pletcher and Jaffrezic, 2001).

Chapter 2

Statistical models for estimating the genetic basis of repeated measures and other function-valued traits

2.1 Introduction

A simple and efficient procedure for the genetic analysis of characters that change as a function of age (or some other independent and continuous variable) is desirable for researchers in several fields of biology and genetics. Plant and animal breeders are often faced with the genetic analysis of "repeated measures" data, such as lactation in dairy cows or growth rates in important agricultural species. Biologists interested in the evolution of life histories study the genetic basis of age-specific fitness components, such as survival or reproductive output; while evolutionary ecologists often examine the genetic relationship between values of a single character expressed over a continuous range of environmental variables.

Recent conceptual and computational advances have made the genetic analysis of such *function-valued* traits readily accessible. Four methods have been advanced in the literature. First, random regression (RR) models have been widely used for the analysis of longitudinal data in the traditional statistical literature (Diggle et al., 1994 ; Verbeke and Molenberghs, 1998) and recently have been applied in the animal breeding context (Jamrozik et al., 1997). Second, the use of orthogonal polynomials (OP) to approximate covariance matrices was initially suggested by Kirkpatrick and Heckman (1989) and is closely related to random regression models (Meyer and Hill, 1997 ; Meyer, 1998). Third, the character process (CP) model was recently proposed by Pletcher and Geyer (1999) and is based on stochastic process theory. Fourth, structured antedependence (SAD) models that correspond to a generalization of autoregressive models and have been proposed by Nunez-Anton and Zimmerman (2000). We develop and consider a general extension of the process model to take advantage of new methods for estimating complicated correlation structures. Each of these methods has been implemented in relatively easy to use computer software packages which are freely available.

The aim of this chapter is to compare and contrast the four approaches and evaluate their performance. We explore a variety of simulated data sets and describe the types of covariance structures (genetic, environmental, and otherwise) accommodated by each method. Using empirical data on age-specific mortality and reproductive output in the fruit fly, *Drosophila melanogaster*, and on growth in beef cattle, we evaluate the ability of each model to adequately fit empirical data.

2.2 Examples and Analyses

2.2.1 Estimation procedures

Models considered have been described in the Introductory Chapter. We propose to relax the stationarity assumption for character process models using a method proposed by Nunez-Anton (1998) and Nunez-Anton and Zimmerman (2000). The idea is to implement a non-linear transformation upon the time axis, $f(t)$, such that correlation stationarity holds on the transformed scale—on the original scale the correlation is non-stationary. The correlation function is then defined as $\rho(s,t) = \rho(|f(s) - f(t)|)$, and the functions suggested by Pletcher and Geyer (1999) remain valid. Ideally the transformation function should contain a small number of parameters with interpretable effects.

Nunez-Anton and Zimmerman (2000) suggest a Box-Cox power transformation such that

$$\begin{aligned} f(t,\lambda) &= (t^\lambda - 1)/\lambda \quad \text{if } \lambda \neq 0 \\ &= \text{Log } t \qquad\quad \text{if } \lambda = 0 \end{aligned}$$

where λ is a parameter to be estimated. For an absolute exponential correlation function: $\rho(s,t) = \theta^{|f(s)-f(t)|}$, the correlations on the sub-diagonals are monotone increasing if $\lambda < 1$ or monotone decreasing if $\lambda > 1$. If $\lambda = 1$ the non-stationary model reduces to a stationary one. Thus, a likelihood ratio test of the null hypothesis $H_0 : \lambda = 1.0$ can be used to quantitatively examine the extent of non-stationarity in the data. Additional flexibility in the non-stationary pattern might be achieved by considering more than one parameter λ. For example, one might incorporate distinct λ_i for different values of $|s - t|$, which is equivalent to a separate λ_i for each sub-diagonal of the covariance structure.

All covariance parameters were estimated using restricted maximum likelihood (REML). In all cases a non-parametric mean function was used (i.e., a separate mean was fitted for each distinct age in the data), which ensures a consistent estimate of the covariance structure (Diggle et al., 1994). Comparison among models was based on the Bayesian Information Criterion (BIC) (Schwarz, 1978), which provides for likelihood based comparison among non-nested models. It penalizes the likelihood for the number of parameters involved in the model and takes into account the number of observations (which is not the case for criterion such as AIC (Akaike, 1974)). BIC is

$$\text{Loglikelihood} - \frac{1}{2} \times \text{number of parameters in the model} \times \log n^*$$

where $n^* = n - p$ when using REML with n the number of observations in the data set and p the number of fixed effects. The model selected is the one that maximizes the criterion. Other selection criteria could also have been considered: minimizing the standard error of genetic value predictions (for beef cattle data), or using the score test to check the goodness-of-fit of covariance structures.

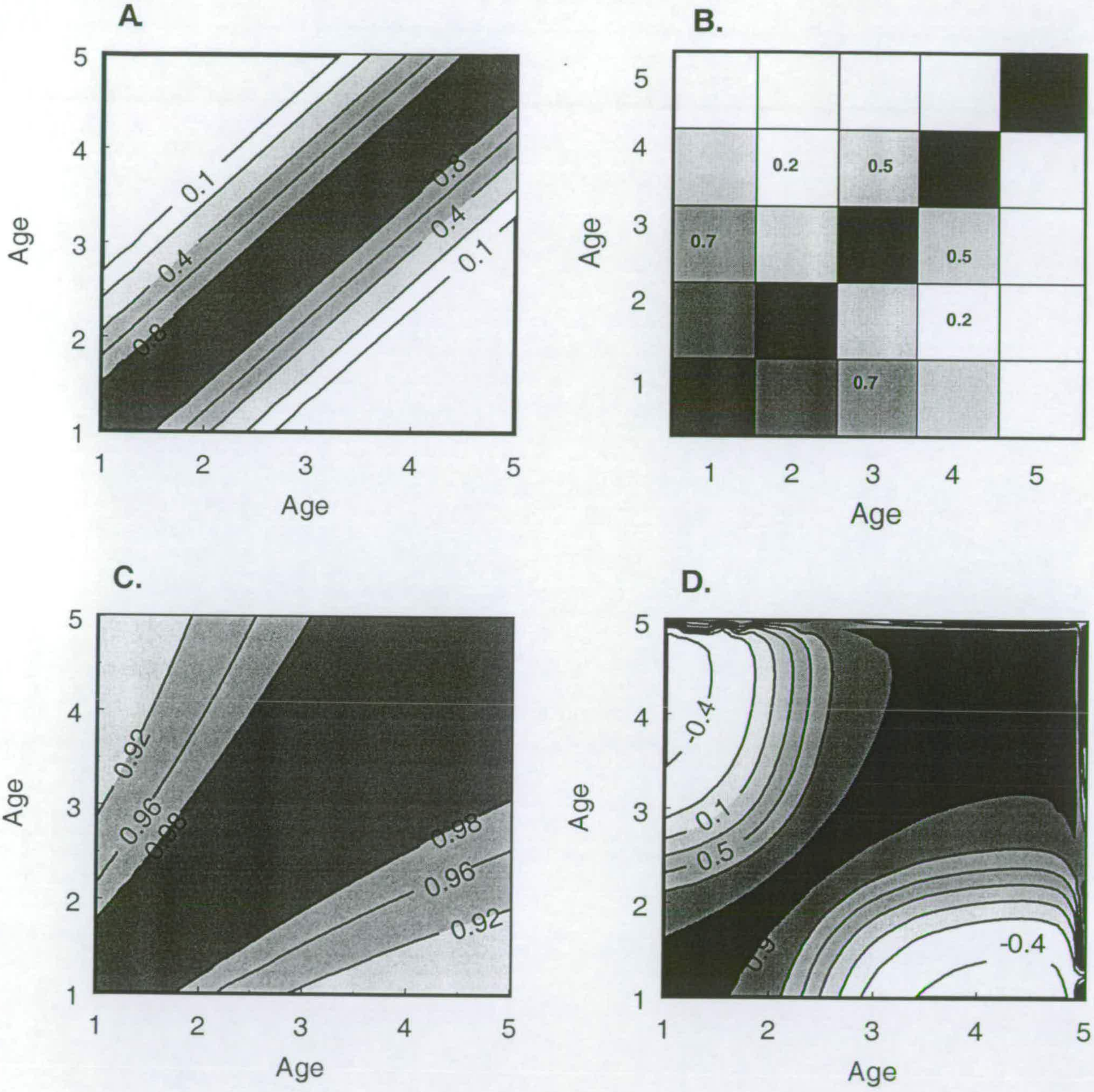
To determine the best fitting model under each technique, a large number of models were fitted to each data set. For the character process method, over 100 different models (i.e., different combinations of polynomial variance functions and stationary and non-stationary correlation functions) were investigated, and the best model was chosen according to the BIC criterion. We chose to examine a large number of CP models for reasons of thoroughness. The CP models are relatively new, and the behaviour of these models is not well-known. In practice, such an exhaustive search is not required, as standard model selection procedures (e.g., sequential addition of polynomial terms to the variance function) result in identical conclusions (results not presented). For both random regression and orthogonal polynomial methods, the appropriate polynomials of increasing degree were fit until an increase in degree no longer resulted in a significant increase in the log-likelihood at the $\alpha = 0.05$ level (Meyer and Hill, 1997). We find that a reasonable approach to model selection requires of the order of 5–10 model fits for each method. For SAD models, the order of antedependence was increased until the added correlation coefficient was close to zero. Quadratic variances were first considered and the polynomial order was reduced when appropriate. A more detailed description of these models is given in Chapter 4.

Estimates of the covariance structure based on random regression, orthogonal polynomials and antedependence models were obtained using the software package ASREML (Gilmour et al., 2000), while estimates of the character process model (and certain orthogonal polynomial models) were obtained using computer software developed by S. Pletcher (personal communication, C code and executable files freely available). A series of exploratory analyses were conducted to ensure the two software packages produced comparable log-likelihoods. A small number of covariance structures could be fitted by both packages (models of constant variance and correlation across ages, and small orthogonal polynomial models) and these structures were fitted to several data sets. In all cases, identical log-likelihoods were reported by each package.

2.2.2 Simulated Data

Many data sets were simulated according to various covariance structures. All were built assuming a standard sire design (i.e., groups of half-sibs) in which 12 offspring from each of 70 sires were measured at five different ages. Under such a design, the estimated between-sire covariance function is directly proportional to the genetic covariance function. The environmental covariance function and residual error are estimated based on the within-sire and the within-animal variation. We present the results of four representative data sets. Because the magnitude of the variance and covariances were different among the simulations, we set the residual variance for all simulations to approximately 10% of the total variance at age 0.

Figure 2.1: Contour plots of the simulated genetic covariance structures for: A–data generated according to a stationary character process (CP) model, B–data simulated according to a CP model with arbitrary and non-stationary correlation (this is a discrete valued matrix rather than a continuous function), C–data generated under a random regression (RR) model with linear deviations, and D–data simulated assuming an orthogonal polynomial (OP) model of degree two.


A.
B.
C.
D.
Age
Age
1
2
3
4
5
0.1
0.4
0.8
0.8
0.4
0.1
0.2
0.5
0.7
0.5
0.2
0.7
0.92
0.96
0.98
0.98
0.96
0.92
-0.4
0.1
0.5
0.9
-0.4

The first data set was simulated according to a stationary CP covariance structure, the purpose of which was to assess the behaviour of SAD, RR and OP models when the genetic correlation decreases to zero within the range of the data. The genetic covariance function was composed of a quadratic variance (i.e., a quadratic $v^2(\cdot)$ from equation 1.3) and "normal" correlation ($\rho(t_i, t_j) = exp(-0.8(t_i - t_j)^2)$ (Figure 2.1a). The environmental covariance function was composed of a linear variance and "Cauchy" correlation function ($\rho(t_i, t_j) = 1/(1 + 0.05(t_i - t_j)^2)$ (Pletcher and Geyer, 1999). We refer to this data set as the stationary CP data.

To examine a well-behaved covariance function with a somewhat non-stationary correlation, we simulated data with genetic variance function identical to that in the stationary CP data, but with an arbitrary non-stationary correlation structure (Figure 2.1b). The environmental covariance was assumed identical to that in the stationary CP data. This data set is the non-stationary CP data.

The third data set was simulated according to a random regression model with linear deviations for both the genetic and environmental parts. The chosen parameter values resulted in genetic and environmental correlations that remained quite high over all ages in the data (Figure 2.1c).

The last data set that we present was simulated according to an OP model, with quadratic Legendre polynomials for the genetic and environmental parts (i.e., $m = 2$ in equation 1.4). The shapes of the covariance functions were rather undulating, as is expected from functions based on orthogonal polynomials. Parameter values were chosen such that the environmental correlation remained quite high over time while the genetic correlation was highly non-stationary (Figure 2.1d).

To compare the fit of the models we calculated goodness-of-fit statistics for the estimated variance and correlation functions under each model with respect to the simulated structure. Goodness-of-fit was quantified by the concordance correlation coefficient, r_c, described by Vonesh et al. (1996) (see appendix). The possible values of r_c are in the range : $-1 \leq r_c \leq 1$, with a perfect fit corresponding to a value of 1 and a lack of fit to values ≤ 0. This coefficient allowed separate evaluation of the fit for variance and correlation for the genetic and environmental parts, which was not possible with likelihood based criteria that only provide one measure for the overall fit of the model.

2.2.3 Empirical Data

***Drosophila* reproduction and mortality:** Age-specific measurements of reproduction and mortality rates were obtained from 56 different recombinant inbred (RI) lines of *Drosophila melanogaster*, which are expected to exhibit genetically based variation in longevity and reproduction (J.W. Curtsinger and A.A. Khazaeli, unpublished results). Age-specific measures of mortality and average female reproductive output were collected simultaneously from two replicate cohorts for each of 56 RI lines. Live/dead observations were made every day, while egg counts were made every other day. For both mortality and reproduction the data were pooled into 11 5-day intervals for analysis. Mortality rates were log-transformed and reproductive measures were square-root transformed to ensure the age-specific measures were normally distributed.

Growth in beef cattle: These data come from the Wokalup selection experiment in Western Australia and correspond to January weights of 436 beef cows, from 77 sires. Weights were recorded between 19 and 82 months of age, with up to 6 records per cow. Analyses are carried out within 83 contemporary groups (year-paddock-age of weighing subclasses), fitted as fixed effects. Additional information, along with access to the data, can be obtained from Dr. Karin Meyer's web page at the Animal Genetics unit of the University of New England, Australia (http://agbu.une.edu.au/∼meyer).

The aim of experimental data such as reproduction and mortality in *Drosophila* is to study the genetic variation and correlations of such fitness components over time. On the other hand, data such as growth in beef cattle, or milk production of dairy cattle as analysed later, are used to calculate genetic values for breeding programs. Although these two goals are very different, they both require the most appropriate modelling of the covariance structure, either to be studied directly or to provide more accurate breeding value predictions.

2.3 Results

2.3.1 Simulations

For the stationary CP data, the best random regression model according to the BIC criterion was characterized by quadratic and linear deviations for the genetic and environmental parts, respectively. Higher order polynomials did not converge to a maximum and could not be considered. The best OP model contained a cubic polynomial for the genetic covariance and a quadratic for the environmental part. As expected, the simulated structure was accurately recovered by the stationary character process model. Concordance coefficients r_c describing the goodness-of-fit for the variance and correlation functions are given in Table 2.1. For the RR and OP models, the environmental covariance structure (including both the variance and correlation) was very well fitted ($r_c \approx 1$). The genetic variance was also well modelled, but both models had trouble dealing with the rapidly decreasing genetic correlation function. Although the OP model could better estimate the genetic correlation (r_c=0.61 for OP compared to 0.36 for RR), it contains significantly more parameters than the regression model (17 vs. 10), and both models exhibit similar behaviour. The polynomial structures are unable to handle correlation patterns that decrease asymptotically to zero within the range of the data, and the correlation obtained by both models goes negative (Figure 2.2).

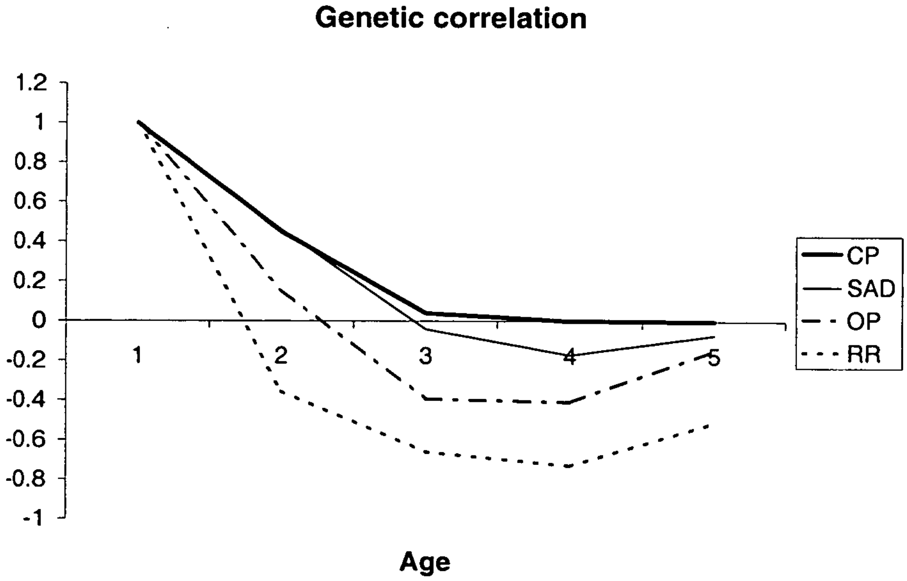
Table 2.1: Goodness-of-fit values for covariance functions estimated from three different methods on simulated data.

Simulated Covariance Structure	Model	VarG	CorrG	VarE	CorrE	BIC
Stationary CP						
	CP	0.98	1.0	1.0	1.0	-4591
	SAD	0.99	0.98	0.35	0.99	-5652
	RR	0.96	0.36	0.93	0.87	-7414
	OP	0.98	0.61	0.98	0.98	-6605
Non-stationary CP						
	CP	0.91	0.03	0.99	1.0	-4454
	SAD	0.90	0.75	0.44	0.99	-5479
	RR	0.95	0.10	0.94	0.81	-7397
	OP	0.84	0.70	0.98	0.97	-6628
Random Regression						
	CP†	1.0	0.93	0.96	0.93	-3817
	SAD	0.99	1.0	0.83	0.96	-3965
	RR	1.0	0.94	0.99	1.0	-3803
	OP	1.0	0.94	0.99	1.0	-3803
Orthogonal Polynomial						
	CP†	0.86	0.10	0.69	0.94	-14334
	SAD	0.98	0.30	0.74	0.99	-14276
	RR	0.30	0.15	0.94	0.90	-14371
	OP	0.99	0.83	0.99	1.0	-14272

† The best fitting correlation function was a non-stationary CP model.

The best SAD model was of second order antedependence for both genetic and environmental parts with a linear 'variance' for the genetic part and quadratic for the environmental one. Regarding BIC criterion, they performed better than either RR or OP models, with fewer parameters (only 8). They could deal better with the asymptotic correlation pattern (r_c=0.98) as shown on Figure 2.2.

Figure 2.2: Genetic correlations between age 1 and other for the simulated stationary character process data and fitted genetic correlations obtained from the random regression model with linear deviations, orthogonal polynomial of degree three and second order structured antedependence model with linear innovation variance.



The aim of the second simulated data set was to investigate the behaviour of these models in the case of a rather simple non-stationary genetic correlation structure. The best RR and OP models were the same as for the stationary CP data detailed in the previous paragraph. The RR model dealt very poorly with the non-stationary pattern of the genetic correlation (r_c=0.10); the correlation was estimated to be very high over all ages. Again, the greater number of parameters in the best fitting OP model over the regression model provided a better fit to the correlation structure (r_c=0.70). Surprisingly, the CP model failed to accurately estimate the non-stationary correlation pattern (Table 2.1). Our non-stationary extension did not significantly improve the goodness-of-fit (BIC=−4454 and −4456 for stationary and non-stationary models, respectively; P=0.052 for a likelihood ratio test of $\lambda = 1.0$). However, the goodness-of-fit of the fitted non-stationary correlation ($r_c = 0.55$) is substantially better than that of the stationary model ($r_c = 0.03$), which provides an interesting commentary on model selection criteria. In retrospect, the non-stationarity in this data set was predominantly between extreme ages (ages 1 and 5). It is possible that more observations per individual are needed to detect small to moderate levels of non-stationarity (see fly reproduction data). The best SAD model was of first order antedependence with linear variance for the genetic part and second order antedependence with quadratic variance for the environmental part. It proved to be the best model to capture the non-stationary correlation pattern (r_c=0.75). The BIC criterion was again higher than for either RR or OP models.

All methods did a reasonable job of estimating the genetic and environmental covariance structures generated according to a random regression model with linear deviations. Under this model the correlations (both genetic and environmental) remained quite high over time. Our non-stationary extension of the CP model was successful in providing a good fit to the data. The genetic covariance structure was described by a quadratic variance and non-stationary correlation given by the characteristic function of the Uniform distribution (Pletcher and Geyer, 1999), and the environmental variance function was linear with a Cauchy correlation. The goodness-of-fit for the genetic correlation structure was improved substantially over a stationary model (r_c=0.74, BIC=-3819 and r_c=0.93, BIC=-3817 for the stationary and non-stationary CP models, respectively).

The data set simulated with an OP structure might be considered pathological in that the genetic covariance structure is highly irregular. In fact, the genetic correlation is negative between early ages but highly positive between late ages (Figure 2.1d). This pattern may not be found very often in practical cases, but it is, however, typical for OP models (Kirkpatrick et al., 1994). Convergence problems hindered our ability to obtain estimates of high dimensional random regression models, and the best RR model was not able to accommodate either the simulated genetic variance or correlation ($r_c = 0.30$ and $r_c = 0.15$, respectively). Both the genetic and environmental covariance structure was described by a quadratic variance and non-stationary correlation given by the characteristic function of the Uniform distribution. When compared to random regression, the CP model is much better at estimating the genetic variance function but is slightly worse at approximating the correlation structure (Table 2.1). The environmental covariance is better behaved and much less of a problem. As seen with the random regression simulations, the strong positive correlation across all ages is well fitted by all the methods. SAD model with a third order antedependence for the genetic part and second order for the environmental part with quadratic variances proved to be better able to deal with this covariance structure than either CP or RR models and had a higher BIC value.

2.3.2 Empirical

Drosophila reproduction and mortality: For age-specific mortality and reproduction in *Drosophila* both SAD and CP models provided a significantly better fit, according to the BIC criterion, than either the orthogonal polynomial or random regression methods (Table 2.2). In fact, they achieved higher likelihoods despite containing fewer parameters than the OP or RR models.

Table 2.2: Results of covariance function estimation on empirical data. NPCov: number of parameters in the covariance structure. N: total number of observations.

	Method	Genetic	Environmental	NPCov	Log L	BIC
Fly Mortality						
(N=955)						
11 fixed effects	SAD	ante(3)-quad	ante(1)-quad	10	-162.4	-234.3
	CP	Quad-Cauchy	Lin-Cauchy	7	-186.0	-247.7
	OP	Cubic	Quadratic	17	-242.1	-338.0
	RR	Quadratic	Quadratic	13	-298.2	-380.4
Fly Reproduction						
(N=1109)						
11 fixed effects	CP	Const-Exp†	Quad-Cauchy†	8	494.1	427.5
	SAD	ante(2)-quad	ante(2)-const	8	461.8	395.3
	OP	Cubic	Quadratic	17	451.4	353.4
	RR	Quadratic	Linear	10	374.0	300.5
Beef Cattle Growth						
(N=1626)						
24 fixed effects	CP	Lin-Exp	Lin-Exp	7	-6895.6	-7010.0
	RR	Constant	Linear	6	-6910.7	-7021.4
	OP	Linear	Linear	8	-6908.3	-7026.4

† The best fitting correlation function was a non-stationary CP model.

Table 2.3: Character process model estimates of genetic and environmental covariance functions for empirical data (standard errors are given in brackets). θ_0, θ_1 and θ_2: parameters of the variance function such that a quadratic variance is represented as $v^2(t) = \theta_0 + \theta_1\, t + \theta_2\, t^2$. θ_C and λ: parameters of the correlation function.

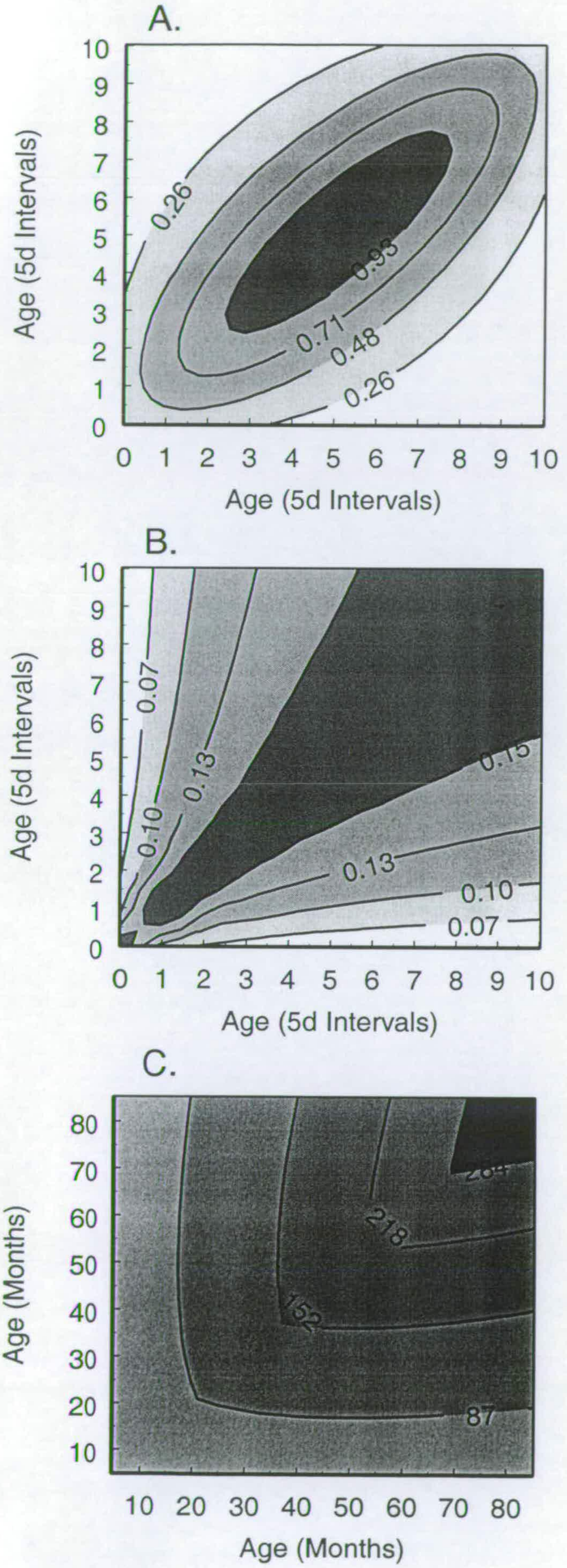
	Parameters	Genetic	Environmental	Residual
Fly Mortality				
	θ_0	0.28(0.12)	0.53(0.05)	None
	θ_1	0.35(0.08)	-0.03(0.007)	
	θ_2	-0.03(0.007)	—	
	θ_C	0.10(0.02)	1.76(0.29)	
Fly Reproduction				
	θ_0	0.18(0.03)	0.10(0.02)	None
	θ_1	—	-0.01(0.01)	
	θ_2	—	-0.002(0.001)	
	θ_C	0.26(0.15)	4.0(2.0)	
	λ	-0.63(0.30)	0.51(0.13)	
Beef Cattle Growth				
	θ_0	0.0001*(186.3)	0.0001*(257.8)	1000.8 (85.35)
	θ_1	4.12(6.95)	38.94 (7.77)	
	θ_C	0.99(0.02)	0.99 (0.003)	

* Parameter estimate is at the lower boundary and asymptotic standard errors may not be reliable.

For age-specific mortality, the best CP model for the genetic covariance was a quadratic variance with a Cauchy correlation function ($\rho_G(t_i, t_j) = 1/(1 + \theta(t_i - t_j)^2)$). The BIC criterion was slightly higher for SAD model with antedependence of order 3 for the genetic part and order 1 for the environmental part with quadratic variances. For fly reproduction the best character process model was a constant variance at all ages coupled with a non-stationary correlation function described by the absolute exponential, $\rho_G(t_i, t_j) = \theta^{|f(t_i)-f(t_j)|}$. Parameter estimates and their standard errors for the CP model are presented in Table 2.3, and the fitted genetic covariance structures are presented in Figure 2.3(a and b).

The simplicity of the character process model allows quantitative statements about the predominant attributes of the genetic covariance function. Genetic variance for *Drosophila* mortality declines significantly with age, while genetic variance is constant at all ages for reproductive output. For mortality, the parameter in the genetic correlation function was significantly different from zero ($p < 0.0001$) suggesting that mortality rates become less genetically correlated as ages become further separated in time. This is true for reproductive output as well, and the significant non-stationarity parameter in the genetic correlation provides evidence for an increase in the correlation between two equidistant ages with increasing age.

Figure 2.3: Contour plots of genetic covariance functions fitted by the character process model. A–age-specific mortality in the fruit fly, *Drosophila melanogaster*, B–age-specific reproduction in *D. melanogaster*, C–age-specific growth in beef cattle.


A.
Age (5d Intervals)
Age (5d Intervals)
0
1
2
3
4
5
6
7
8
9
10
0.26
0.48
0.71
0.93
0.26
B.
Age (5d Intervals)
Age (5d Intervals)
0.07
0.10
0.13
0.15
0.13
0.10
0.07
C.
Age (Months)
Age (Months)
10
20
30
40
50
60
70
80
87
152
218
284

Beef cattle: Although differences in fit among the methods are less dramatic for beef cattle than for *Drosophila*, the character process model again provides a significantly better fit (as determined by the BIC criterion) than either random regression or orthogonal polynomial methods (Table 2.2). The best fitting model for the genetic part was a linear variance (increasing with age) and an absolute exponential correlation ($\rho_G(t_i, t_j) = \theta^{|t_i - t_j|}$). There was no evidence for non-stationarity in the data. Parameter estimates and their standard errors for the CP model are presented in Table 2.3, and the fitted genetic covariance structure is shown in Figure 2.3c. SAD models were not fitted to this data set because it was too unbalanced: at most 6 measures were available per animal whereas 24 different times of measurement were possible.

2.4 Discussion

The quantitative genetic analysis of repeated measures and other function-valued traits requires the estimation of continuous covariance functions for each source of variation in a particular statistical model. Traditionally, statistical geneticists interested in characters that change gradually along some continuous scale have had to settle for models that are either overparameterized (i.e., standard multivariate methods) or oversimplified (e.g., composite character analysis) (Meyer, 1998 ; Pletcher and Geyer, 1999). In recent years, however, the introduction and development of random regression models, orthogonal polynomial models, and models based on stochastic process theory (i.e., the character process model) have provided important alternatives. Other types of random regression models (e.g., non-linear models as suggested by Lindstrom and Bates (1990) and Davidian and Giltinan (1995)) may prove useful, but they are currently difficult to implement.

Through extensive investigation of a variety of simulated covariance structures and empirical data, we find that under most conditions structured antedependence (SAD) and character process (CP) models provide the best description of the underlying covariance structure. It is clear from the simulation results that they can adequately capture a correlation that declines rapidly to zero as character values become further separated in time, whereas both random regression models and orthogonal polynomials have noticeable problems approximating such a structure (Table 2.1; stationary CP data and Figure 2.2). Polynomials do not have asymptotes, and the rapid decline in correlation tends to force both methods to estimate correlations that are strongly negative within the range of the data. Although the characteristics of covariance functions for natural organisms remain generally unknown, this is a serious limitation as asymptotic behaviour in covariances/correlations are to be expected (Pletcher and Geyer, 1999). Other parameterizations of the RR models (e.g., using orthogonal polynomials in the regression) may prove more useful in this regard. On the other hand, RR and OP models deal quite well when the correlation structure remains high over time (see environmental correlation in CP simulated data; Table 2.1).

A further advantage of the CP models appears to be the ability to model the variance and correlation separately. As mentioned previously, for random regression models the entire covariance structure is implicitly determined by the shapes of the regression polynomials, and covariance surfaces described by orthogonal polynomials have a fixed relationship between variance and correlation. This limitation is exemplified in the analysis of growth in beef cattle. For the genetic deviation, the best fitting RR model included only a random intercept. This implies not only that the variance is considered constant over time, but also that the correlation is constant and equal to 1 across all ages, which is probably not appropriate (Figure 2.3c). Applying the same argument to the fertility data in *Drosophila*, the best fitting CP model for the genetic part was a constant variance with a rather rapid decline in correlation between increasingly separated ages (Table 2.3). Such a combination is simply not possible under the RR or OP methods. It is also likely that the separation of variance and correlation was a major factor contributing to the ability of the CP model to reasonably estimate the genetic variation with a much smaller number of parameters (four parameters) than random regression (10 parameters) or orthogonal polynomial (17 parameters) models (Table 2.2).

The data sets we examined were small in comparison to those commonly analyzed in agricultural and breeding contexts. Using extremely large data sets, complicated covariance and correlation models may be of greater use, and the random regression and orthogonal polynomial methods may begin to show an advantage. Large data sets would also relieve the convergence problems we experienced with high order random regression and orthogonal polynomial models. Unfortunately, most quantitative genetic studies of natural and experimental populations are extremely labor intensive, and sample sizes will often be similar to those reported here. For these situations, the properties of the character process models (e.g., easy hypothesis testing, few and interpretable parameters) make it a useful option.

Despite their apparent success in this study, there are several important limitations of the process models that suggest avenues for further development. First, additional ways of relaxing the stationarity assumption (Pletcher and Geyer, 1999) without greatly increasing the number of parameters are needed. Although not appropriate in all situations, a promising direction proposed by Nunez-Anton and Zimmerman (2000) has been studied here and seems to offer reasonable flexibility in practice. As shown in the simulation study, SAD models offer a higher degree of flexibility than character processes in capturing non-stationary correlation patterns. Second, CP models require the manipulation (inversion, factorization, etc.) of matrices whose dimensions are proportional to the number of ages in the data set, regardless of the size of the model itself (Meyer, 1998). A method of reparameterization, similar to that used for RR and OP models (Meyer, 1998), would be useful. Third, a method for estimating the eigenfunctions of covariance functions used by the process models would provide insight into patterns of genetic constraints across ages (Kirkpatrick et al., 1990 ; Kirkpatrick and Lofsvold, 1992).

Lastly, the genetic analysis of two or more function-valued traits is an important goal. Generalization of regression models to multi-trait analyses is straightforward and has already been used, for instance, to analyze age-dependent milk production, fat, and protein content in dairy cattle (Jamrozik et al., 1997). Bivariate character process models might be implemented by defining a parametric cross-covariance function between the two traits, but appropriate forms for this function are yet to be discovered. A promising way forward for multivariate analysis seems to be offered by antedependence models as shown in Chapter 4.

Appendix

Goodness-of-fit of the covariance structure

The concordance correlation coefficient r_c described by Vonesh et al. (1996) was used in the simulation study to evaluate the goodness-of-fit for both the variance and correlation functions estimated by the models when compared to the simulated structure. For the correlation structure, for instance, we consider:

$$r_c = 1 - \frac{\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}(y_{ij} - \hat{y}_{ij})^2}{\sum_{i,j}(y_{ij} - \bar{y})^2 + \sum_{i,j}(\hat{y}_{ij} - \hat{y})^2 + T(T-1)(\bar{y} - \hat{y})^2/2} \tag{2.1}$$

where $\hat{y}_{ij}$ represents the estimated correlation between times t_i and t_j given by the model, and y_{ij} is the correlation between times t_i and t_j in the simulated data. T represents the total number of times at which measurements were taken. $\bar{y}$ and $\hat{y}$ are the mean of the correlation values for the simulated data and for the model, respectively. The concordance coefficient for the variance estimate is much simpler and given by

$$r_c = 1 - \frac{\sum_{i=1}^{T}(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2 + \sum_i(\hat{y}_i - \hat{y})^2 + T(\bar{y} - \hat{y})^2} \tag{2.2}$$

where the y now refer to the actual and estimated variances rather than correlations.

The coefficient r_c is directly interpretable as a concordance coefficient between observed and predicted values. It directly measures the level of agreement (concordance) between y_{ij} and $\hat{y}_{ij}$, and its value is reflected in how well a scatter plot y_{ij} versus $\hat{y}_{ij}$ falls about the line identity. The possible values of r_c are in the range : $-1 \leq r_c \leq 1$, with a perfect fit corresponding to a value of 1 and a lack of fit to values ≤ 0.

Chapter 3

Non-parametric estimation

3.1 Introduction

Animal breeders and evolutionary geneticists are often faced with the problem of analysing traits that change as a function of age or some other independent and continuous variable. This is the case for example for lactation curve analysis in dairy cattle, growth curve analysis of laboratory and agricultural species, or the study of age-specific fitness components such as reproductive output. Many techniques have already been proposed to deal with this kind of data. The most commonly used at present are random regression models (Diggle et al., 1994). Another approach, called 'character process models', has recently been proposed by Pletcher and Geyer (1999), and corresponds to a parametric modelling of the covariance structure. An overview of these techniques is presented in the introductory chapter.

These methods require an a priori formulation of a parametric model, however, and so the main difficulty is to choose the most appropriate model. In fact, the number of possible models can be very large in practice, especially for the character process methodology where it is possible to combine different functions of variance and correlation for both the genetic and environmental parts. It is in general not possible to investigate all the possible combinations. It would therefore be extremely useful in practice to have an idea of the covariance structure in order to choose the most appropriate parametric model.

When a small number of measures with common times of measurement is available for each subject, it is possible to estimate an unstructured covariance matrix with standard software. However, this is in general not feasible when the number of measurements per subject is large and when data are unbalanced, which can be the case for example for daily records for milk production in dairy cattle. The aim of this chapter is to propose a non-parametric procedure that deals with this kind of data, and requires no a priori assumption about the model. This methodology is based on the 'variogram' (Diggle and Verbyla, 1998), and will be illustrated using simulated and actual data sets.

3.2 Variogram approach

We focus here on the analysis of repeated measures over time, but this approach can also be applied to traits that change as a function of another independent and continuous variable. In order to present the variogram methodology, we first consider the case of a phenotypic analysis, and then propose a way to extend it to genetic analysis.

3.2.1 Phenotypic analysis

Let t_j $(j = 1, ..., J)$ be the times of measurement, and y_{ij} the measure on individual i $(i = 1, ..., I)$ taken at time t_j. It is not necessary for individuals to have measures at all times. It is assumed that y_{ij} is the realization of a random variable $Y_i(t_j)$, where $Y_i(t)$ are a set of I mutually independent Gaussian processes with mean value functions $\mu_i(t) = E(Y_i(t))$ and common covariance function $P(s,t) = cov(Y_i(s), Y_i(t))$.

For a general Gaussian process $Y(t)$ with mean value $\mu(t)$ and covariance function $P(s,t)$ we define the residual process to be the zero-mean process $Z(t) = Y(t) - \mu(t)$. Then, as presented by Diggle and Verbyla (1998), the variogram of $Z(t)$ is the function:

$$\gamma(s,t) = \frac{1}{2}E[(Z(s) - Z(t))^2] \text{ for } s \neq t \tag{3.1}$$

As, $E(Z(s)) = E(Z(t)) = 0$, it follows that:

$$\gamma(s,t) = \frac{1}{2}[P(s,s) + P(t,t) - 2P(s,t)] \tag{3.2}$$

where $P(s,t)$ is the phenotypic covariance function. This description of the variogram does not assume stationarity, i.e. it is not assumed that $\gamma(s,t) = \gamma(s-t)$ as in classical definitions.

For a set of longitudinal data (y_{ij}, t_j) with known mean value function $\mu_i(t)$, the variogram cloud is the set of points $((t_j, t_k, v_{ijk})$, for $i = 1, ..., I$, $j = 1, ..., J$ and $k > j)$ in three-dimensional space, where:

$$v_{ijk} = \frac{1}{2}[(y_{ij} - \mu_i(t_j)) - (y_{ik} - \mu_i(t_k))]^2 \tag{3.3}$$

If the data contain replicated pairs (t_j, t_k) across subjects, the sample variogram $\bar{v}(t_j, t_k)$ is defined as the average of such pairs across subjects. Let $r(t_j, t_k)$ be the number of subjects contributing to $\bar{v}(t_j, t_k)$. When all the $r(t_j, t_k)$ are large, the sample variogram may be an adequate estimator for $\gamma(t_j, t_k)$. When $r(t_j, t_k)$ are small, a smoother estimator for $\gamma(t_j, t_k)$ is desirable. Note that when the data are balanced in the sense that the observation times are common to all I subjects, $r(t_j, t_k) = I$ for all (t_j, t_k).

If the mean value structure is known, then the squared residuals, $z_{ij}^2 = (y_{ij} - \mu_i(t_j))^2$ are unbiased for the variance function $v(t_j)$. As for the variogram, if replicated values of z_{ij}^2 at each time t_j are available from different subjects, the sample means of these sets of replicated values provide adequate non-parametric estimates of the variance function. In other cases, a smoother estimator for $v(t_j)$ is again desirable.

In most applications, $\mu_i(t_j)$ is unknown and will then have to be replaced by an appropriate estimate $\hat{\mu}_i(t_j)$. In practice, we propose to pre-correct data y_{ij} for fixed effects using a simple regression model, and to fit a non-parametric mean-curve in the variogram with: $\hat{\mu}_i(t_j) = \bar{y}_{.j}$. Diggle et al. (1994), in Chapter 4, provide a discussion about fixed effects estimation. As it is to be used for exploratory purposes, the aim of this estimation procedure is to be simple and computationally fast rather than statistically efficient.

3.2.2 Genetic analysis

It is assumed that the observed phenotypic process $Y(t)$ is a Gaussian process and can be decomposed as:

$$Y(t) = \mu(t) + g(t) + e(t) \tag{3.4}$$

where $\mu(t)$ are the fixed effects, $g(t)$ and $e(t)$ the genetic and environmental effects, which are assumed to be mean zero Gaussian processes, independent of each other, and with covariance functions $G(s,t)$ and $E(s,t)$, respectively.

In the case of a one-way classification, data are assumed to be divided into groups (eg. half-sib families, clones, etc.). The idea is to consider simple ANOVA on group means for each time independently that will provide variance estimates, and to combine these with the variogram approach in order to obtain covariance estimates.

The linear mixed model can be written as:

$$y_{sij} = \mu_j + u_{sj} + e_{sij} \tag{3.5}$$

where y_{sij} is the observation at time t_j for individual i from group s ($j = 1, ..., J$, $i = 1, ..., n_s$ and $s = 1, ..., S$), u_{sj} is the group effect and e_{sij} the residual term at time t_j. When considering each time t_j independently, u and e are assumed to be independent and normally distributed with variances $v_G(t_j)$ and $v_E(t_j)$, respectively. If the groups are half-sib families, for example, $v_G(t_j)$ is equal to a quarter of the additive genetic variance at time t_j.

Variance functions

Let us assume first a balanced setting, i.e. all groups have the same number n_s of subjects and individuals have observations at all times t_j. Observations y_{sij} are assumed to have been corrected previously for fixed effects. μ_j represents the mean curve in the population and can be approximated by the average $\bar{y}_{..j}$ at each time t_j. Using a simple ANOVA on group means, the variance cloud $v_{1sj} = (\bar{y}_{s.j} - \bar{y}_{..j})^2$ provides an estimate for $\gamma_1(t_j) = (1 - (1/S))(v_G(t_j) + (1/n_s)v_E(t_j))$, and $v_{2sij} = (y_{sij} - \bar{y}_{s.j})^2$ for $\gamma_2(t_j) = (1 - (1/n_s))\ v_E(t_j)$.

Variogram cloud

Extending results for single times, two variogram clouds can be defined:

$$v_{1sjk} = \frac{1}{2}[(\bar{y}_{s.j} - \bar{y}_{..j}) - (\bar{y}_{s.k} - \bar{y}_{..k})]^2 \tag{3.6}$$

and

$$v_{2sijk} = \frac{1}{2}[(y_{sij} - \bar{y}_{s.j}) - (y_{sik} - \bar{y}_{s.k})]^2 \tag{3.7}$$

Extending the ANOVA result and the variogram approach, the first variogram cloud provides estimates for:

$$\gamma_1(t_j, t_k) = \frac{S-1}{2S}[(v_G(t_j) + v_G(t_k) - 2G_{jk}) + \frac{1}{n_s}(v_E(t_j) + v_E(t_k) - 2E_{jk})] \tag{3.8}$$

and the second provides estimates for:

$$\gamma_2(t_j, t_k) = \frac{n_s - 1}{2n_s}(v_E(t_j) + v_E(t_k) - 2E_{jk}) \tag{3.9}$$

where G_{jk} and E_{jk} represent the group and environmental covariances between times t_j and t_k, respectively.

Extension to the unbalanced case is given in the appendix.

3.3 Simulation study

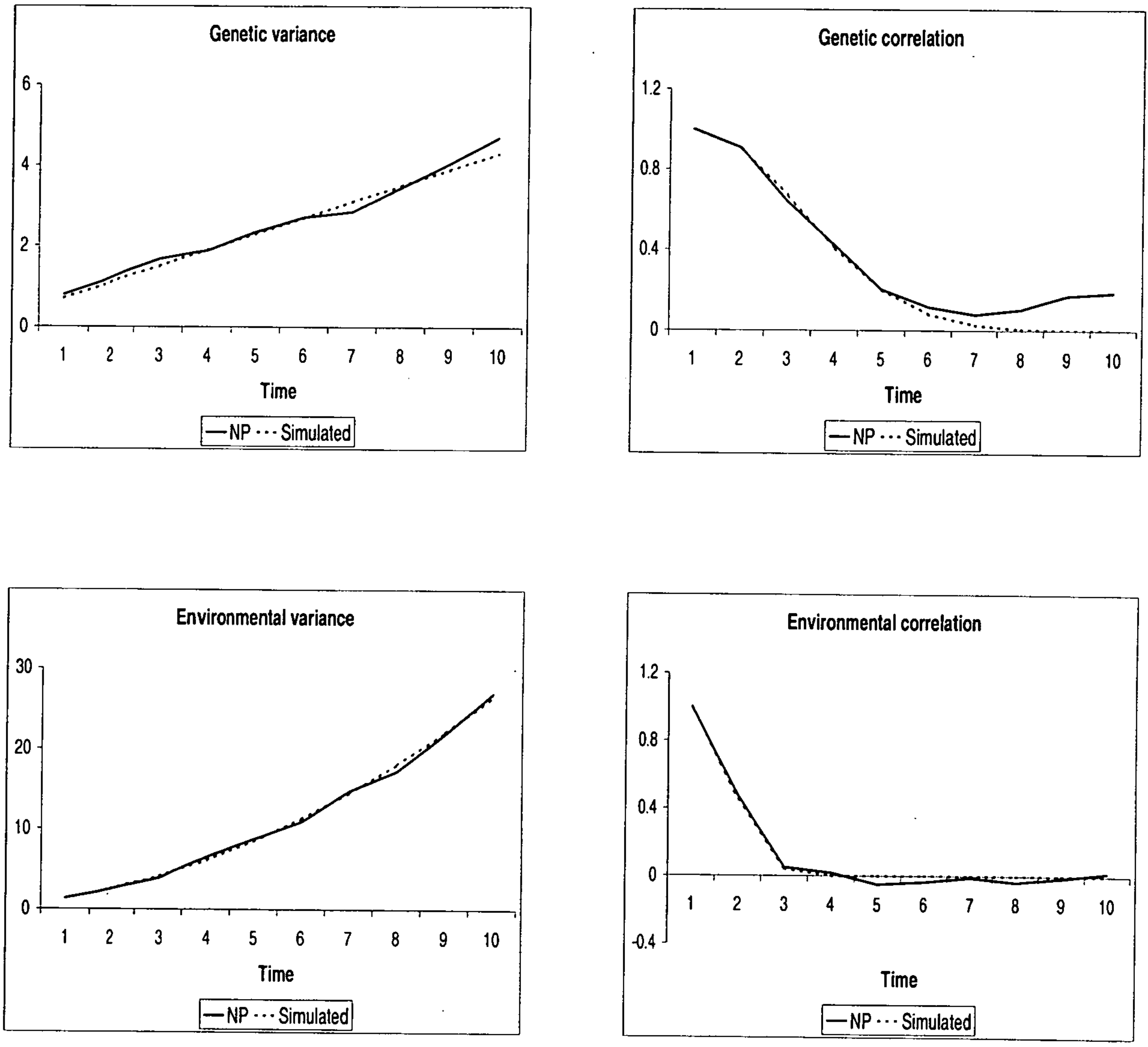
Implementation of the variogram genetical analysis was easy and consequently, calculations were fast. In order to check the behaviour of this estimation procedure, different data sets were investigated.

3.3.1 Stationary correlation

A balanced design was considered, with 100 sires, 20 progeny per sire, and 10 measures per progeny. A stationary character process model was considered with a linear variance ($\sigma_S^2(t) = Var_G(t) = 0.3 + 0.4t$) and Gaussian correlation ($\rho_G(t,s) = \exp(-0.1(t-s)^2)$) for the genetic part, and quadratic variance ($Var_E(t) = 0.5 + 0.6t + 0.2t^2$) and Gaussian correlation ($\rho_E(t,s) = \exp(-0.8(t-s)^2)$) for the environmental part.

The time (t) values were discrete. In order to check the simulated covariance structure, REML estimates for a character process model were calculated on the simulated data set using ASREML (Gilmour et al., 2000). Figure 3.1 presents the estimated genetic and environmental variances and correlations using the non-parametric estimation procedure presented above. The non-parametric estimation procedure provided very good estimates for both the genetic and environmental covariance structures. For 10 repeated measurements, the non-parametric estimates were also relatively smooth, and no additional smoothing seemed to be required. The slight discrepancy observed for the genetic correlation for large lag times is probably due to a lack of information. More simulated replicates should be studied in order to make sure this discrepancy is not significant.

Figure 3.1: Variance and correlation functions for a stationary process. Balanced design: 100 sires with 20 progeny per sire and 10 measures per subject. NP: Non-parametric estimates.



3.3.2 Non-stationary correlation

The proposed non-parametric estimation procedure makes no assumption about stationarity of the covariance or correlation structure. In order to check its behaviour for a non-stationary correlation pattern, the non-stationary extension of the character process models proposed by Jaffrézic and Pletcher (2000) - Chapter 2 - was used to simulate a data set with the same balanced sire design as previously and 10 measures per subject. The parameter of non-stationarity λ was chosen to be 0.5 with exponential correlations for both the genetic and environmental parts ($\rho_G(t,s) = \exp(-0.1(|\frac{t^\lambda - s^\lambda}{\lambda}|))$, $\rho_E(t,s) = \exp(-0.8(|\frac{t^\lambda - s^\lambda}{\lambda}|))$).

Table 3.1: Non-stationary environmental correlation (exponential with $\lambda = 0.5$). Balanced design: 100 sires with 20 progeny per sire and 10 measures per subject. Simulated (above diagonal) and non-parametric estimated (below diagonal) correlation.

	1	2	3	4	5	6	7	8	9	10
1	1	0.52	0.31	0.20	0.14	0.10	0.07	0.05	0.04	0.03
2	0.54	1	0.60	0.39	0.27	0.19	0.14	0.10	0.08	0.06
3	0.33	0.60	1	0.65	0.45	0.32	0.23	0.17	0.13	0.10
4	0.22	0.37	0.63	1	0.69	0.49	0.36	0.27	0.20	0.16
5	0.10	0.20	0.40	0.68	1	0.71	0.52	0.39	0.29	0.23
6	0.07	0.14	0.28	0.50	0.72	1	0.73	0.55	0.41	0.32
7	0.05	0.10	0.21	0.36	0.52	0.72	1	0.75	0.57	0.44
8	0.01	0.05	0.13	0.26	0.38	0.54	0.75	1	0.76	0.59
9	0.01	0.02	0.07	0.19	0.28	0.40	0.55	0.75	1	0.77
10	0.02	0.01	0.05	0.13	0.19	0.28	0.41	0.55	0.75	1

The simulated genetic correlation remained quite high over time, whereas the simulated environmental correlation rapidly decreased as ages were further apart and was highly non-stationary. Table 3.1 gives the simulated and estimated environmental correlation matrices. It appeared that the non-parametric estimation procedure was able to capture the non-stationary pattern of the correlation function, and the estimates provided were very close to the actual values. This non-parametric estimation procedure can therefore be useful in practice to check the stationary assumption for the correlation function.

3.4 Application

3.4.1 Daily records in dairy cattle

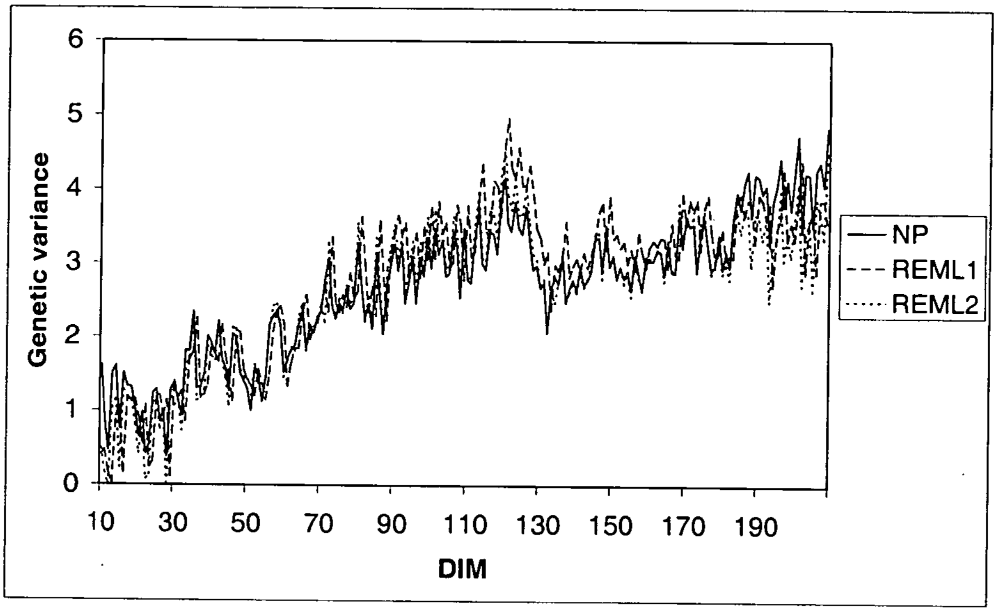
Daily records for milk production for first lactation were analysed using this non-parametric procedure. Data came from the Langhill experimental farm (Edinburgh, UK), and comprised 438 cows from 50 sires. The number of daughters per sire varied from 1 to 22, with 9 on average. Using a simple regression model, data were previously corrected for fixed effects: age at calving, percentage of Holstein genes, line (selected or control), diet (forage or concentrates). Estimation for the mean curve is included in the definition of the variogram: a non-parametric curve is considered, fitting one mean at each time. In order to have enough observations per sire at each time, we considered only data from day 10 to day 240. The total number of observations was 83634, with a maximum of 230 records for cows with complete measures.

Figure 3.2 shows the estimates of genetic and environmental variances. In order to check the non-parametric estimates, as well as their ability to deal with unbalanced data and fixed effects estimation, REML estimates for the variances were also calculated using ASREML and considering each time independently.

REML1 represents estimates obtained while estimating fixed effects at the same time, and REML2 are estimates obtained on the data set previously corrected for fixed effects. It can be seen that variance estimates obtained here with the three methodologies were extremely close. A similar analysis was performed for the covariance estimates. Unstructured covariance matrices for both the genetic and environmental parts were obtained using the package REMLPK (Meyer, 1985). However, as it cannot provide estimates for unstructured covariance matrices of size 230 by 230, this analysis was performed for only a few given times. It can be seen that covariance estimates obtained with the non-parametric approach and with REML were also very similar.

As completely unstructured covariance matrices cannot be obtained with standard software for all the observed ages, this non-parametric methodology should prove to be extremely useful to study the covariance and correlation structure for these daily records. Figure 3.3 shows estimates of genetic and environmental correlations for days in milk 10, 80 and 210. As expected from previous analyses (White et al., 1999), the genetic correlation is quite high for all pairs of ages (about 0.8), except for the early stage of lactation. For example, the correlation between day 210 and day 10, as well as between day 80 and day 10, is about 0.2. For all stages of lactation, the environmental correlation is high for days in milk close in time (for example, correlation of 0.8 between day 210 and 190), and decreases steadily as days become further apart (correlation of 0.6 between day 210 and day 130, and of 0.2 between day 210 and day 10).

Figure 3.2: Genetic and environmental variances for daily records for milk production in dairy cattle, given in kg^2 (DIM: Days in milk). NP: Non-parametric estimates. REML1: REML with fixed effects estimated at the same time. REML2: REML on the data set pre-corrected for fixed effects.



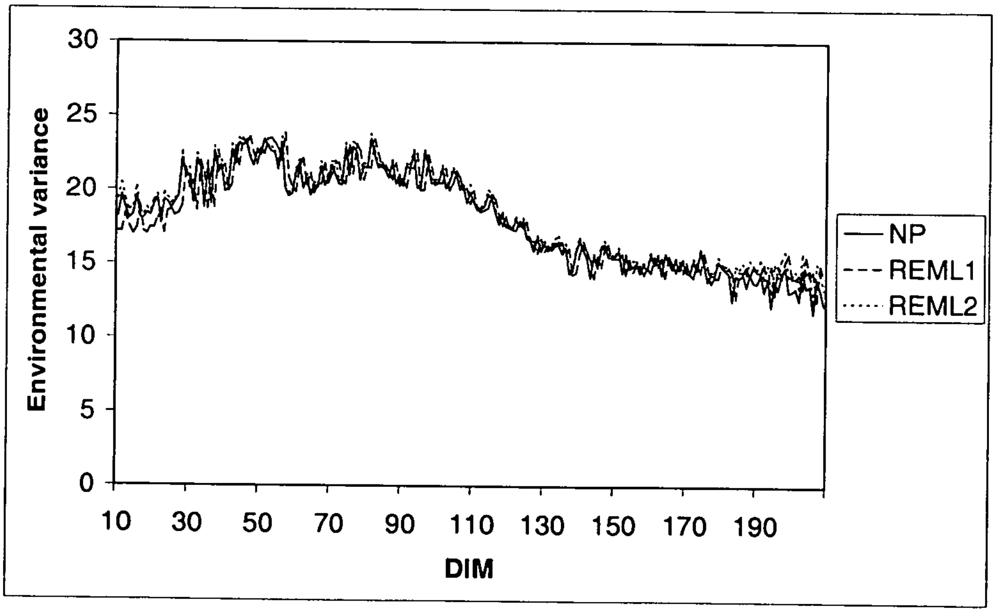
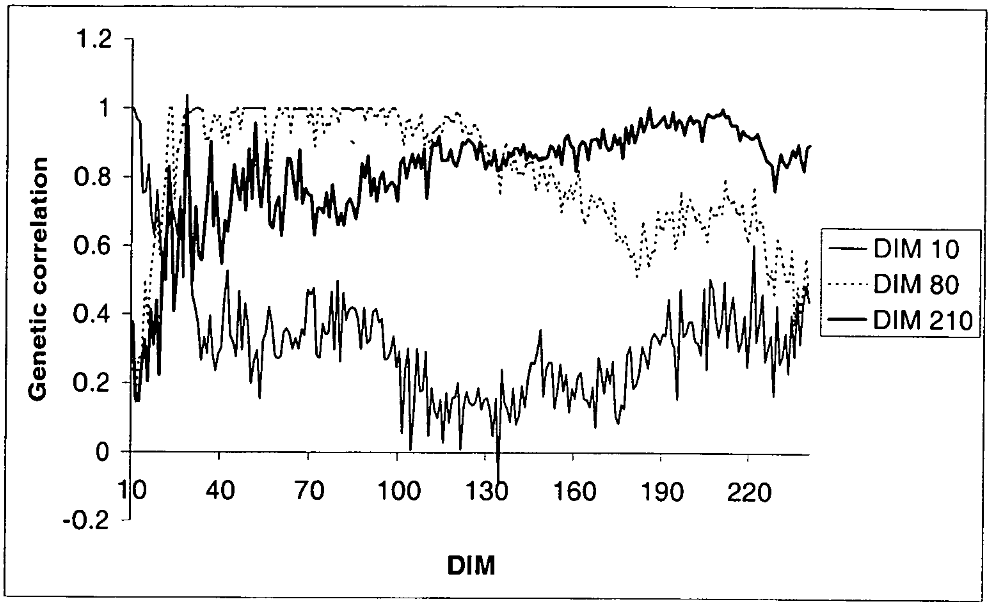
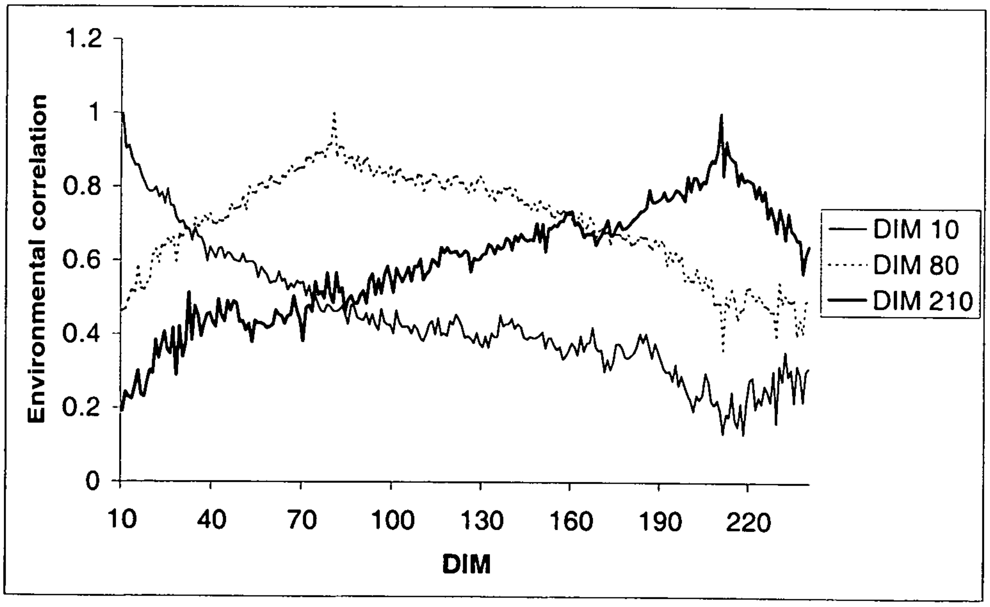


Figure 3.3: Genetic and environmental correlations for daily records for milk production in dairy cattle. DIM 10 : Correlation between day in milk 10 and others. DIM 80 : Correlation between day in milk 80 and others. DIM 210 : Correlation between day in milk 210 and others.





3.4.2 Fertility data in Drosophila

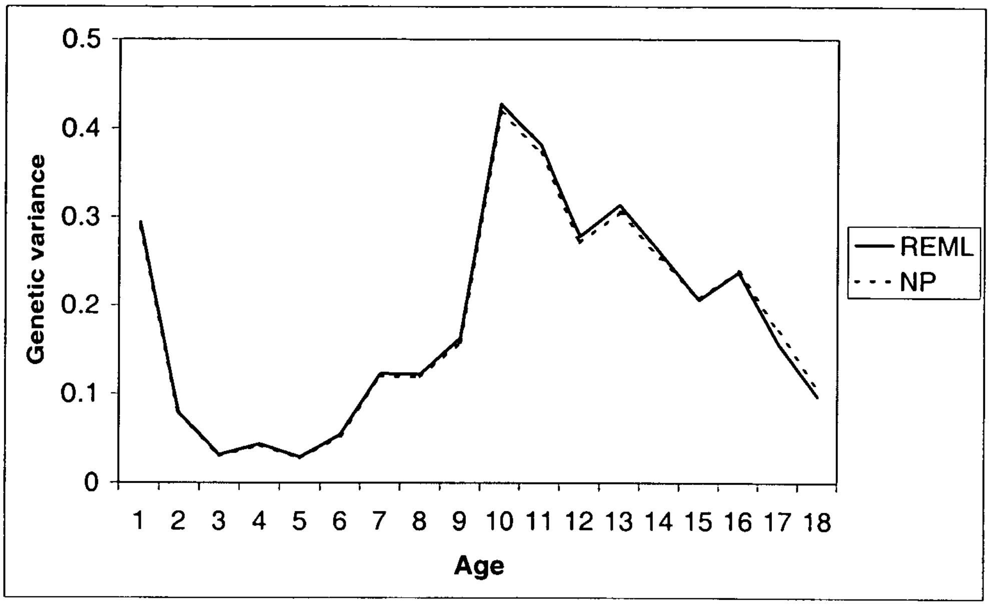
Age-specific measurements of reproduction were obtained from 56 different recombinant inbred (RI) lines of *D. melanogaster*, which are expected to exhibit genetically based variation. Age-specific measures for average female reproductive output were collected from two replicate cohorts for each of the lines. Egg counts were made every other day, and observations were square-root transformed so that the age-specific measures were approximately normally distributed. In order to have enough observations for each line, only the 18 first ages (out of 34) were considered.

Figure 3.4 shows estimates of genetic and environmental variances using both the non-parametric procedure presented above and a REML analysis performed with the software ASREML. The procedures showed very similar results for both genetic and environmental parts. If a parametric model were to be chosen, a quadratic function would probably be appropriate for the environmental variance.

For the genetic variance, however, the choice of a parametric function may be more difficult. In fact, in a previous study (see Chapter 2 ; Jaffrezic and Pletcher (2000)), data were pooled into 5-day intervals, and it was found that the best parametric model for the genetic variance, using a likelihood based criterion, was a constant function estimated at 0.18. However, the variation observed here for the genetic variance with both the non-parametric and REML methodologies may be worthwhile to study. The genetic variance seems to drop quickly for early ages, then increases rapidly at about age 10, and decreases thereafter. The causes of these large changes may therefore be worth investigating.

Table 3.2 gives non-parametric estimates for the correlation matrices. It appears that both the genetic and environmental correlations seem to be non-stationary, as was also found in Chapter 2.

Figure 3.4: Genetic and environmental variances for fertility data in Drosophila. Each age corresponds to a 2-day interval. NP: Non-parametric estimation.



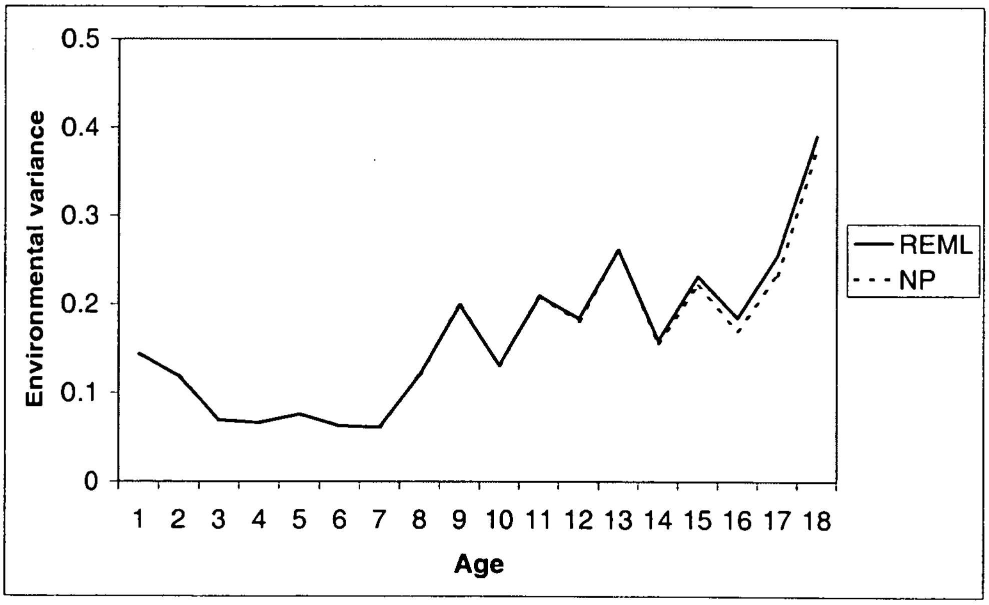


Table 3.2: Non-parametric estimates for genetic (above diagonal) and environmental (below diagonal) correlation for fertility data in Drosophila (table gives correlation for every 4-day interval).

	1	2	3	4	5	6	7	8	9
1	1	0.34	0.50	0.37	0.34	0.47	0.44	0.34	0.56
2	0.03	1	1.0	0.44	0.54	0.32	0.20	0.33	0.22
3	0.30	0.16	1	0.87	0.68	0.80	0.67	0.50	0.73
4	0.17	0.31	0.32	1	0.76	0.77	0.58	0.46	0.45
5	0.33	0.04	0.48	0.35	1	1.0	0.90	0.87	0.91
6	-0.04	0.22	0.20	0.17	0.07	1	0.92	1.0	0.93
7	0.16	0.09	0.35	-0.03	0.30	0.37	1	0.95	1.0
8	0.05	-0.20	0.14	0.08	0.22	0.12	0.37	1	1.0
9	-0.05	0.03	-0.13	0.03	0.00	0.24	-0.07	0.17	1

3.5 Discussion

In the analysis of repeated measurements, before assuming a parametric model it is advisable to have an idea of the shape of the variance and correlation functions for both the genetic and environmental parts. When a small number of observations is available for each subject at a fixed set of times, it is possible to estimate unstructured covariance matrices with standard software. However, this is not feasible when the number of observations over time is large, and when data are unbalanced. In this case, the proposed non-parametric procedure would prove to be extremely useful.

As shown in the above analyses, this methodology presents several positive aspects. It is, first of all, easy to implement as it involves mainly sum and average calculations. Moreover, the computing time required is small even for a large data set such as the daily records for milk production, especially because it is a non-iterative procedure. Secondly, it is able to provide estimates close to REML even for a non-stationary correlation structure, as was shown in the simulation study, or for unbalanced data sets, as the Langhill data. Finally, it is able to deal with a large number of observations over time, and provides estimates for covariances and correlations between all ages, which was not possible with usual softwares.

It should however be used mainly for exploratory purposes as it does not always provide statistically efficient estimates. As pointed out by Diggle and Verbyla (1998), one of the difficulties of this approach can be fixed effects estimation. Nevertheless, when only a few fixed effects are considered, as was the case for the Langhill data, it was shown that the non-parametric analysis on pre-corrected data performs well compared to the REML which estimates fixed effects at the same time. Another point that needs to be further investigated concerns extension to an animal model, that would take into account the relationship matrix. This does not seem to be straightforward, and requires further study.

The extension of this non-parametric approach to multiple trait analysis is obvious as formulae given in this chapter can also be used to estimate cross-covariance and cross-correlation functions between different traits. This could for example be useful for the joint analysis of milk, fat and protein in dairy cattle, and could also help generalizing the character process methodology to multivariate analyses.

Appendix

Unbalanced analysis

In the case of an unbalanced design, let n_{sj} be the number of individuals in group s with measures at time t_j. The ANOVA variance estimate at time t_j is:

$$\bar{v}_{1j} = \frac{1}{\sum_{s=1}^{S} n_{sj}} \sum_{s=1}^{S} n_{sj}(\bar{y}_{s.j} - \bar{y}_{..j})^2 \tag{3.10}$$

Let n_{dj} the average number of daughters per sire with measures at time t_j. The previous variance cloud will provide estimates for:

$$\gamma_1(t_j) = (1 - \frac{1}{S})(v_G(t_j) + \frac{1}{n_{dj}} v_E(t_j)) \tag{3.11}$$

A straightforward extension of this result to covariance estimates is:

$$\bar{v}_{1jk} = \frac{1}{\sum_{s=1}^{S} n_{sjk}} \sum_{s=1}^{S} \frac{n_{sjk}}{2} [(\bar{y}_{s.j} - \bar{y}_{..j}) - (\bar{y}_{s.k} - \bar{y}_{..k})]^2 \tag{3.12}$$

where n_{sjk} is the number of individuals in group s with measures for both time t_j and t_k.

This variogram cloud will give estimates for:

$$\gamma_1(t_j, t_k) = \frac{S-1}{2S}[(v_G(t_j)+v_G(t_k)-2G_{jk})+\frac{1}{\sum_{s=1}^{S} n_{sjk}} \sum_{s=1}^{S}(\frac{n_{sjk}}{n_{sj}} v_E(t_j)+\frac{n_{sjk}}{n_{sk}} v_E(t_k)-2\frac{n_{sjk}^2}{n_{sj}n_{sk}} E_{jk}) \tag{3.13}$$

Provided that n_{sjk} is not too different from n_{sj} and n_{sk}, this variogram cloud will give estimates for:

$$\gamma_1(t_j, t_k) = \frac{S-1}{2S}[(v_G(t_j) + v_G(t_k) - 2G_{jk}) + \frac{1}{n_{djk}}(v_E(t_j) + v_E(t_k) - 2E_{jk})] \tag{3.14}$$

where n_{djk} is the average number of subjects per group with measures at times t_j and t_k. Other weights could also be used, such as those proposed by Robertson (1962).

Chapter 4

Genetic analysis of multivariate repeated measures

4.1 Introduction

The need for a rigorous method of analysis for biological characters that are best considered as functions of some independent and continuous variable is rapidly expanding. Important examples of such function-valued traits include growth curves, age-specific fitness components such as survival or reproductive output, lactation curves in dairy cattle, and gene expression profiles across age or environmental treatments.

Several techniques have been proposed for single trait (univariate) analyses. These include random regression models, which are based on a parametric modelling of individual curves (Diggle et al., 1994); character process models, which focus on parametric modelling of the covariance structure (Pletcher and Geyer, 1999); and orthogonal polynomials (Kirkpatrick and Heckman, 1989), which can be interpreted in terms of either individual curves or the covariance structure (Meyer, 1998). A comparison among these methods revealed that, in most cases, character process models provided a better fit to the covariance structure (genetic and non-genetic) than either random regression or orthogonal polynomials, and they did so with fewer parameters (see Chapter 2 ; Jaffrezic and Pletcher, 2000).

Unfortunately, extension of the character process models to the simultaneous analysis of two or more function-valued traits (i.e., a multivariate function-valued analysis) is not straightforward. The primary hinderance is the development of reasonable parametric forms for the cross-covariance functions, which describe the genetic and non-genetic covariance between the two traits. Athough a multivariate extension of random regression models is straightforward, their poor performance in the univariate case argues strongly against their use in a multivariate setting. Moreover, the nature of the parameterization results in a dramatic increase in the number of parameters required to describe the covariance structure (for example, for a quartic random regression model, a genetic univariate analysis requires 30 parameters, whereas a bivariate analysis would require 110 parameters).

The aim of this chapter is to develop techniques that maintain the spirit of the character process models, which make reasonable assumptions about the covariance structure in the data in order to greatly reduce the number of parameters in the model, and at the same time allow for a straightforward extension to the multivariate case. Structured antependence (SAD) models (Zimmerman and Nunez-Anton, 1997 ; Nunez-Anton and Zimmerman, 2000) provide an ideal framework for this development, and we extend the SAD models to study the relationship between two function-valued traits. The performance of these models is compared to character process and random regression models, and several examples, including phenotypic and genetic analysis of age-specific fertility and mortality in *Drosophila* and of milk, fat and protein yields through lactation for dairy cattle, are presented.

4.2 Materials and Methods

4.2.1 Structured antedependence models

Univariate

As in classical quantitative genetics for the analysis of function-valued traits (Kirkpatrick and Heckman, 1989 ; Pletcher and Geyer, 1999), it is assumed that the observed phenotypic trait $X(t)$ changes continuously over time or some other independent variable and that its trajectory can be decomposed as follows:

$$X(t) = \mu(t) + g(t) + e(t) \tag{4.1}$$

In the simplest case, $\mu(t)$ represents the mean value at each time for the population, for example the average number of eggs in the case of fertility analysis ; $g(t)$ and $e(t)$ represent individual deviations from this mean value due to the genetic and environmental effects, respectively. Traditionally, the genetic effect is assumed to be the additive contribution of a very large number of genes. It is assumed that $g(t)$ and $e(t)$ are Gaussian variables, independent of each other, with mean zero and covariance functions $G(s,t)$ and $E(s,t)$, respectively. $G(s,t) = \mathrm{Cov}(g(s), g(t))$ represents the covariance for the genetic effects between any two times. The aim of the analysis is therefore to be able to estimate these genetic and environmental covariance functions.

Several methodologies have already been proposed for this purpose. The most commonly used are random regression and character process models. Jaffrezic and Pletcher (2000) as well as the Introductory Chapter present a description of these two approaches. They rely on very different concepts: random regression models focus on modelling individual deviations $g(t)$ and $e(t)$, and the covariance structure is a consequence of these deviations. Character processes directly model the covariance structures $G(s,t)$ and $E(s,t)$ by assuming parametric functions for variances and correlations. Shapes on the individual deviations are a consequence of this structure. In previous analyses (see Chapter 2 ; Jaffrezic and Pletcher, 2000), character process models have proved to be generally able to fit better the covariance structure than random regression models with fewer parameters. However, their extension to the multivariate case is not straightforward as parametric cross-covariance functions between different traits are yet to be discovered.

As genetic analysis of two or more function-valued traits is an important goal, we considered another kind of models that have been proposed in the statistical literature by Zimmerman and Nunez-Anton (1997), and seem to offer similar advantages to character processes to model the covariance structure adequately with few parameters. The concept of this methodology is again different from the two previous ones. The idea of antedependence models, as originally proposed by Gabriel (1962), is that observation at time t can be explained by the previous ones. For example, a first order antedependence model will assume that observation at time t depends only on observation at time $(t-1)$. Generalizing this concept to genetic analysis, a second order structured antedependence model for the genetic part $g(t)$ can be written as:

$$g(t) = \phi_1 \; g(t-1) + \phi_2 \; g(t-2) + \epsilon_g(t) \tag{4.2}$$

where ϕ_1 and ϕ_2 are correlation parameters, and $\epsilon_g(t)$ is assumed to be normally distributed, with mean zero and variance $\sigma_g(t)$ that can change with time. In structured antedependence (SAD) models, Nunez-Anton and Zimmerman (2000) propose to consider a parametric function for these variances, for example a polynomial of time. This parametrization requires very few parameters to model the covariance structure, and increasing the order of antedependence only involves one extra parameter at each step. The same model can be written for environmental effects $e(t)$. It would also be possible to consider a dependence for non-successive lags, for example lags $(t-1)$, $(t-2)$ and $(t-6)$ without the inbetween coefficients. This pattern has previously been observed in some practical cases.

Structured antedependence models correspond to a non-stationary extension of autoregressive processes, and their first order is also closely related to character process models with an exponential correlation and a quadratic variance. The CP approach is more general in the sense that many different correlation functions can be considered. However, SAD models of order s allow more flexibility for the correlation function as $2s$ parameters are included whereas only 2 parameters were included for the CP models.

Let $\boldsymbol{G}$ be the genetic covariance matrix. It is of dimension $J \times J$, where J is the number of measurement times, and has components $G(s,t)$. Parametric specification of the genetic covariance function with structured antedependence models is not straightforward. However, using a result presented by Pourahmadi (1999), it is easy to obtain the genetic covariance matrix. Based on a Cholesky decomposition of the inverse of the covariance matrix, it can be shown that $\boldsymbol{G}^{-1}$ can be written as:

$$\boldsymbol{G}^{-1} = \boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L} \tag{4.3}$$

where $\boldsymbol{L}$ is a lower triangular matrix with 1's on the diagonal and the negatives of the correlation coefficients ϕ_j as below-diagonal entries. $\boldsymbol{D}$ is a diagonal matrix with variances $\sigma_g(t_j)$ as components. The parametric specification of antedependence models of order s is equivalent to the last $J-s-1$ subdiagonals of $\boldsymbol{G}^{-1}$ are zero. Correlation and variance parameters are estimated by REML procedures.

Bivariate

If two variables y_{1t} and y_{2t} are considered, it is easy to extend structured antedependence models to study the relationship between the two variables. For example, considering data ordered as $\boldsymbol{y} = (\boldsymbol{y_1}', \boldsymbol{y_2}')'$, it is possible to study the influence of y_1 on y_2.

If a first order antedependence model is considered, the model can be written as:

$$\begin{aligned} y_{1t} &= \theta_1 \, y_{1,t-1} + e_{1t} \\ y_{2t} &= \theta_2 \, y_{2,t-1} + \phi_1 \, y_{1,t} + \phi_2 \, y_{1,t-1} + e_{2t} \end{aligned}$$

The error terms e_1 and e_2 are assumed to have zero means and to be uncorrelated over time. It is possible to generalize Pourahmadi's (1999) covariance matrix parametrization, in this bivariate case, considering matrix $\boldsymbol{L}$ as:

$$\begin{pmatrix} 1 & & & & & & & \\ -\theta_1 & \ddots & & & & & & \\ & \ddots & \ddots & & & & & \\ (0) & & -\theta_1 & 1 & & & & \\ -\phi_1 & & & (0) & 1 & & & \\ -\phi_2 & \ddots & \ddots & & -\theta_2 & \ddots & & \\ & \ddots & \ddots & & & \ddots & \ddots & \\ (0) & & -\phi_2 & -\phi_1 & (0) & & -\theta_2 & 1 \end{pmatrix}$$

When allowing variances to change over time, for example as a linear function, the diagonal matrix $\boldsymbol{D}$ can be written: $\boldsymbol{D} = \text{Diag}\{\exp(a_1+b_1t_j)\}$, for $j = 1, ..., n_1$, and $\text{Diag}\{\exp(a_2+b_2t_j)\}$, for $j = n_1+1, ..., n_1+n_2$ where n_1 and n_2 are the number of times of measurement for the first and second trait, respectively.

It is straightforward to extend this covariance parametrization to higher order antedependence models, and other relationships between the two traits.

4.2.2 Bivariate random regression models

The extension of random regression to the multivariate case is straightforward. For example in a bivariate analysis with linear random deviations and ignoring fixed effects for simplicity, the model can be written as:

$$y_{1t_i} = a_1 + b_1 t_i + e_{1i} \qquad (4.4)$$

$$y_{2t_j} = a_2 + b_2 t_j + e_{2j} \tag{4.5}$$

where e_1 and e_2 are independent error terms. Therefore the cross-covariance function between the two traits can be calculated as:

$$\begin{aligned} Cov(y_{1t_i}, y_{2t_j}) &= Cov(a_1 + b_1 t_i, a_2 + b_2 t_j) \\ &= Cov(a_1, a_2) + Cov(a_1, b_2)t_j + Cov(b_1, a_2)t_i + Cov(b_1, b_2)t_i t_j \end{aligned}$$

4.3 Examples

Two real examples were considered to illustrate these methodologies: firstly, the bivariate analysis of fertility and mortality rate in *Drosophila*, and secondly the multivariate analysis of milk, fat and protein yields for dairy cattle. Calculations were performed using ASREML (Gilmour et al., 2000).

4.3.1 Data sets

***Drosophila* reproduction and mortality:** Age-specific measurements of reproduction and mortality rates were obtained from 56 different recombinant inbred (RI) lines of *Drosophila melanogaster*, which are expected to exhibit genetically based variation in longevity and reproduction. Age-specific measures of mortality and average female reproductive output were collected simultaneously from two replicate cohorts for each of 56 RI lines. Live/dead observations were made every day, while egg counts were made every other day. For both mortality and reproduction, the data were pooled into 11 5-day intervals for analysis. Mortality rates were log-transformed and reproductive measures were square-root transformed so that the age-specific measures were approximately normally distributed.

Milk, fat and protein yields for dairy cattle: These data comprised records on 9277 cows in first lactation, daughters of 464 Holstein-Friesian sires. The lactation stage of animals at first test varied between 4 and 40 days, with successive tests at approximately 30 day intervals. Records on milk yield, as well as fat and protein yields were available, with 10 measurements per cow for each of these three variables. Fixed effects considered were the age at calving, the percentage of North American Holstein genes, and herd-test-month. For the mean curve, a non-parametric curve was considered, fitting one mean at each test.

4.4 Results

4.4.1 Fly data

Univariate phenotypic analysis: Preliminary univariate analyses were performed in order to select the most appropriate structured antedependence model (SAD) for both variables. Models of order r ($r = 1, 2, ..., R$) were considered until the correlation coefficient ϕ_{R+1} was close to zero. For all SAD models, a quadratic function was considered to model variances: Log $\sigma_i^2 = a + b\ t_i + c\ t_i^2$. These models were compared to a character process with quadratic variance and exponential correlation (CP) as well as to a quadratic random regression model (RR2).

Table 4.1 shows that for both variables a first order structured antedependence model (SAD(1)) can be considered. Improvement obtained with a second order was not significant. SAD(1) fitted much better than a quadratic random regression model with fewer parameters, and almost as well as the character process model.

Table 4.1: Univariate phenotypic analysis for fertility and mortality rate in *Drosophila*. CP: character process model with quadratic variance and exponential correlation. RR2: quadratic random regression model.

		Fertility			Mortality		
Model	NPCov	Log L	Parameters		Log L	Parameters	
			ϕ_1	ϕ_2		ϕ_1	ϕ_2
SAD(1)	4	390.14	0.75		-256.59	0.73	
SAD(2)	5	390.59	0.73	0.03	-255.89	0.76	-0.04
CP	4	405.59			-259.39		
RR2	6	339.64			-342.04		

4.4.2 Bivariate analysis

Structured antedependence models

Influence of fertility on mortality rate (M1): Data were ordered considering fertility first and mortality after in order to be able to study the effect of fertility on mortality rate using an antedependence model. Variances were modelled with quadratic polynomials. Let Fert(t) and Mort(t) be the fertility and mortality variables at time t, respectively. As chosen previously, a first order SAD model was considered for fertility. For mortality rate, models with increasing antedependence order were considered until the added correlation coefficient was close to 0. The chosen model had a likelihood equal to 160.8 and can be written as:

$$\text{Fert}(t) = 0.75\ \text{Fert}(t-1) + \epsilon_t \tag{4.6}$$

$$\text{Mort}(t) = 0.67\ \text{Mort}(t-1) - 0.33\ \text{Fert}(t) + e_t \tag{4.7}$$

This analysis shows that mortality rate at time t is strongly positively correlated with mortality rate at time $(t-1)$, but is also negatively correlated with fertility at time t. This could in fact be expected as a high fertility level proves good health conditions and therefore low mortality rate.

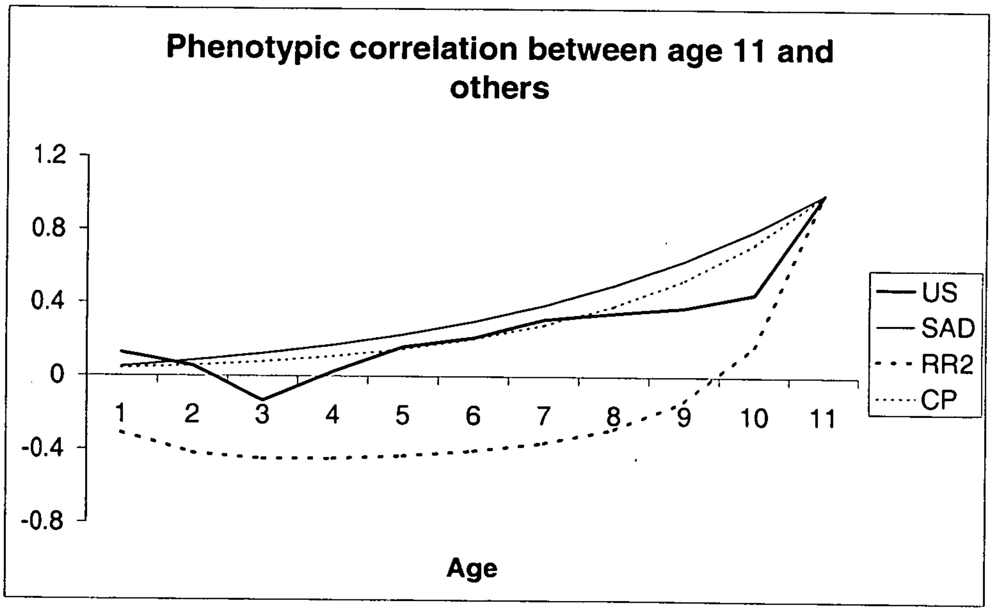
Influence of mortality rate on fertility (M2): The chosen model had a likelihood of 176.4 and can be written as:

$$\text{Mort}(t) = 0.73\ \text{Mort}(t-1) + e_t \tag{4.8}$$

$$\text{Fert}(t) = 0.66\ \text{Fert}(t-1) - 0.11\ \text{Mort}(t) + \epsilon_t \tag{4.9}$$

As this model (M2) has a higher likelihood than the previously considered structured antedependence model (M1), it will be chosen for the bivariate analysis.

Figure 4.1: Phenotypic correlation for mortality between age 11 and others. US: Unstructured model, SAD: chosen structured antedependence model (M2), RR2: quadratic random regression model, CP: character process model with exponential correlation.



Random regression model: A quadratic random regression model for both fertility and mortality variables involved 21 parameters, and likelihood was equal to 67.74. This was much lower than for structured antedependence models despite the larger number of parameters (only 9 parameters for SAD model). This likelihood difference seemed to be mainly due to the poor ability of random regression models to deal with asymptotic correlation patterns, as illustrated by Figure 4.1. In fact, the correlation for mortality rate between early and late ages, for example, was close to 0. However, instead of decreasing asymptotically to zero, as polynomials do not have asymptotes, the random regression correlation went negative. This problem has already been pointed out in Chapter 2 and Jaffrezic and Pletcher (2000).

4.4.3 Bivariate genetic analysis

Table 4.2 gives likelihood and parameter estimates for the two structured antedependence models (M1 and M2) as well as for the quadratic random regression, considering the same model for both genetic and environmental parts. In the genetic analysis, difference in the number of parameters between SAD and random regression models was even larger than in the phenotypic analysis: 18 compared to 42. In spite of this difference, the likelihood was much higher for SAD model than for the quadratic random regression model.

The genetic cross-correlation matrix for the chosen structured antedependence model (M2) is given in Table 4.3. Figures 4.2 and 4.3 show genetic cross-correlation patterns for some given ages for fertility and mortality. The genetic cross-correlation matrix was found to be asymmetrical with very low genetic correlation for fertility at early ages and mortality rate. Correlations were strongly negative between fertility and mortality at late ages.

Table 4.2: Bivariate genetic analysis for fertility and mortality rate in *Drosophila* (M1 and M2: structured antedependence models. RR2: Quadratic random regression model. ϕ_1: cross-correlation parameter).

Model	NPCov	LogL	Parameter ϕ_1	
			Genetic	Environmental
M1	18	287.2	-0.13	-0.34
M2	18	302.3	-0.12	-0.06
RR2	42	134.7		

Table 4.3: Genetic cross-correlation with the chosen structured antedependence model (M2).

		Fertility					
		1	3	5	7	9	11
	1	-0.12	-0.20	-0.22	-0.20	-0.17	-0.13
	3	-0.04	-0.33	-0.50	-0.51	-0.46	-0.38
Mortality	5	-0.02	-0.18	-0.55	-0.69	-0.68	-0.59
	7	-0.02	-0.15	-0.44	-0.68	-0.73	-0.66
	9	-0.02	-0.14	-0.42	-0.65	-0.72	-0.66
	11	-0.02	-0.14	-0.41	-0.64	-0.72	-0.66

Figure 4.2: Genetic cross-correlation for fertility at ages 1, 5 and 11 and mortality at all ages with the chosen structured antedependence model.

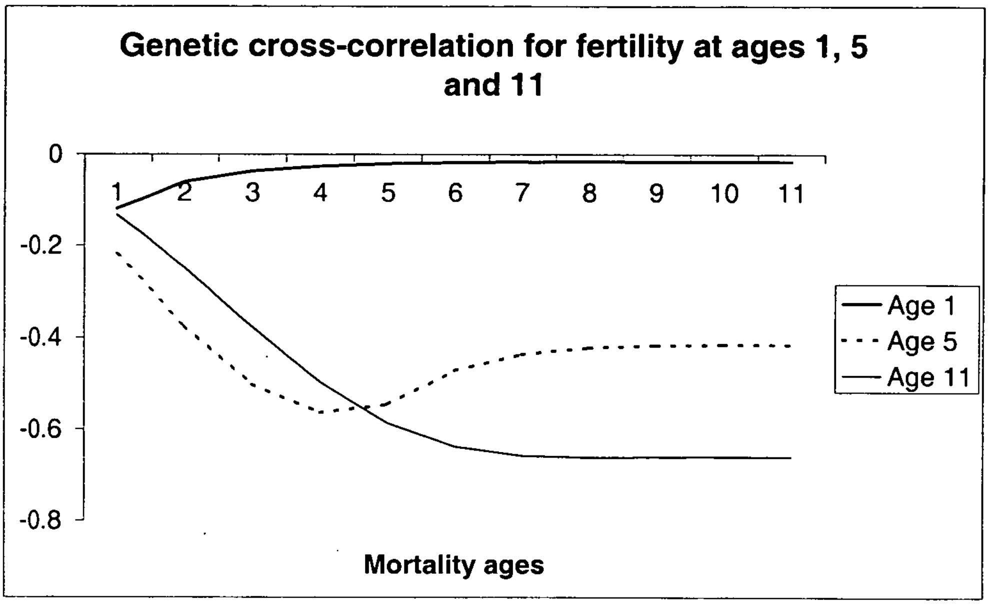
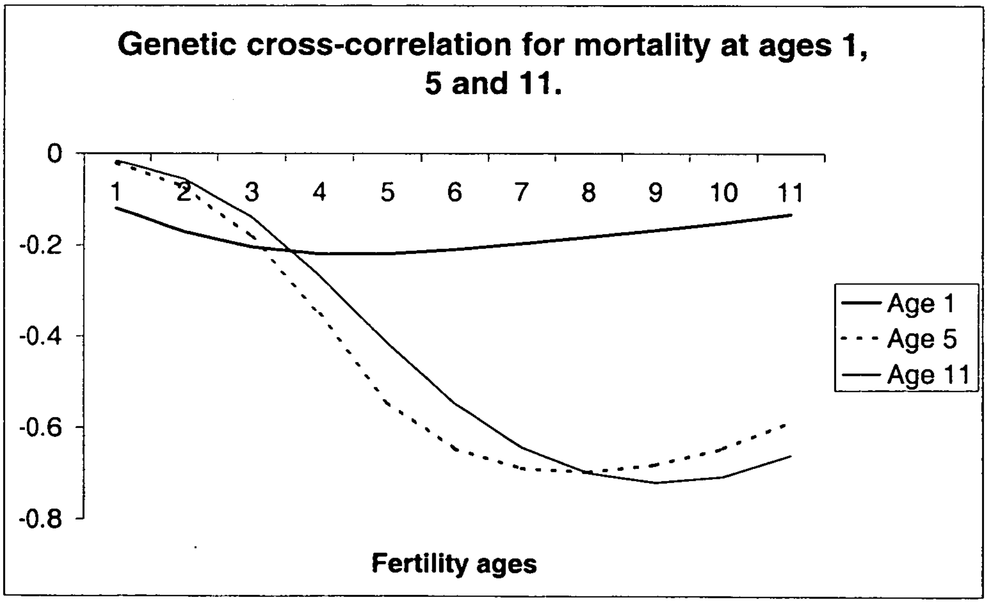


Figure 4.3: Genetic cross-correlation for mortality at ages 1, 5 and 11 and fertility at all ages with the chosen structured antedependence model.



4.4.4 Dairy cow data

Univariate phenotypic analysis

The same steps as presented previously were followed in order to choose the most appropriate antedependence model. Different orders of structured antedependence models (SAD) were compared to character process models with quadratic variance and exponential correlation (CP), or non-stationary correlation (CPNS), as well as to quadratic (RR2), cubic (RR3) and quartic (RR4) random regression models. Fat and protein yields were multiplied by 10 in order to have variances of about the same order as for milk. Likelihoods for these different models are given in Table 4.4.

Parameter estimates for the fourth order structured antedependence model (SAD(4)) were:

$$\begin{aligned}
\text{Milk}(t) &= 0.50\ \text{Milk}(t-1) + 0.22\ \text{Milk}(t-2) + 0.10\ \text{Milk}(t-3) + 0.05\ \text{Milk}(t-4) + e_1(t \\
\text{Fat}(t) &= 0.37\ \text{Fat}(t-1) + 0.21\ \text{Fat}(t-2) + 0.14\ \text{Fat}(t-3) + 0.11\ \text{Fat}(t-4) + e_2(t) \\
\text{Prot}(t) &= 0.51\ \text{Prot}(t-1) + 0.22\ \text{Prot}(t-2) + 0.10\ \text{Prot}(t-3) + 0.06\ \text{Prot}(t-4) + e_3(t
\end{aligned}$$

Univariate analysis for milk, fat and protein yields showed that structured antedependence models of order 1 were about equivalent to character process models (see Table 4.4). Increasing orders of SAD allowed more flexibility and had a higher likelihood than CP models. SAD models of order 3 or 4 performed better than cubic random regression models, and even better than a quartic random regression model in the milk yield analysis, while requiring far fewer parameters: 7 parameters for a 4th order SAD model, 15 parameters for a quartic random regression model. This difference in the number of parameters would be even larger in a bivariate anlaysis as 55 parameters would be required for a bivariate quartic regression, but only 18 in a bivariate SAD(4) model.

Table 4.4: Univariate phenotypic analysis for milk, fat and protein yields in dairy cattle (SAD : structured antedependence models up to order 4. CP and CPNS: character process model with quadratic variance and exponential correlation stationary and non-stationary, respectively. RR: quadratic, cubic and quartic random regression models. NPCov: number of parameters in the covariance structure).

			Log L	
Model	NPCov	Milk	Fat	Protein
SAD(1)	4	-1731	346	1334
SAD(2)	5	1587	3674	4581
SAD(3)	6	2155	4798	5238
SAD(4)	7	2253	5217	5371
CP	4	-1874	604	1852
CPNS	5	-1505	1175	2593
RR2	6	677	4230	1948
RR3	10	1564	4943	4564
RR4	15	2046	5365	6163

Phenotypic bivariate analysis for Milk and Protein yields

In all SAD models, a quadratic function of time was used to model variances. As before, the model was chosen by progressively increasing the order of the structured antedependence model until the added correlation coefficient was close to 0. The chosen model was:

$$\text{Milk}(t) = 0.63 \text{ Milk}(t-1) + 0.24 \text{ Milk}(t-2) + \epsilon_t \tag{4.10}$$

$$\text{Prot}(t) = 0.65 \text{ Prot}(t-1) + 0.31 \text{ Milk}(t) - 0.21 \text{ Milk}(t-1) + e_t \tag{4.11}$$

where Milk(t) and Prot(t) represent milk and protein yields at time t, respectively. For this model, likelihood was equal 1294.4. This model was compared to a bivariate quadratic random regression model (Log L = -284.1). The likelihood was therefore higher for structured antedependence model (11 parameters) than for quadratic random regression (21 parameters).

Table 4.5 gives the phenotypic cross-correlation between milk and protein yields for the structured antedependence model. It appeared that the correlation between milk and protein yields was quite high for all tests, with a cross-correlation of about 0.95 along the diagonal. The cross-correlation matrix was nearly symmetrical.

Table 4.5: Phenotypic cross-correlation between milk and protein, and milk and fat yields with the chosen structured antedependence model.

		Milk					
		1	2	4	6	8	10
	1	0.94	0.58	0.50	0.43	0.35	0.26
	2	0.56	0.93	0.67	0.56	0.46	0.34
Protein	4	0.50	0.67	0.95	0.76	0.62	0.46
	6	0.42	0.56	0.75	0.97	0.76	0.57
	8	0.35	0.46	0.61	0.74	0.97	0.69
	10	0.26	0.34	0.45	0.54	0.67	0.97
	1	0.73	0.44	0.38	0.33	0.27	0.20
	2	0.35	0.70	0.49	0.40	0.33	0.25
Fat	4	0.40	0.56	0.76	0.60	0.48	0.36
	6	0.37	0.49	0.65	0.80	0.62	0.46
	8	0.32	0.42	0.55	0.66	0.83	0.58
	10	0.25	0.32	0.43	0.52	0.62	0.85

Phenotypic bivariate analysis for Milk and Fat

The chosen model was:

$$\text{Milk}(t) = 0.60\ \text{Milk}(t-1) + 0.26\ \text{Milk}(t-2) + \epsilon_t \tag{4.12}$$

$$\text{Fat}(t) = 0.44\ \text{Fat}(t-1) + 0.14\ \text{Fat}(t-2) + 0.31\ \text{Milk}(t) - 0.18\ \text{Milk}(t-1) + e_t \tag{4.13}$$

For this model, likelihood was equal 1091, with 12 parameters to model the covariance structure. Bivariate quadratic random regression model had a likelihood equal to -333, and 21 parameters. This likelihood was again lower than for the structured antedependence model in spite of a larger number of parameters. Table 4.5 gives the phenotypic cross-correlation between milk and fat yields obtained with the structured antedependence model. The correlation was lower than between milk and protein, and was equal to about 0.75 along the diagonal. The cross-correlation matrix was again nearly symmetrical.

4.5 Discussion

Structured antedependence models were considered for the genetic analysis of bivariate repeated measurements. These models require few parameters to model variance and correlation structures. Univariate analyses showed that structured antedependence models of order 1 (SAD(1)) are about equivalent to character process (CP) models. Increasing orders of SAD models allow more flexibility to model the correlation structure than CP models, and therefore significantly improve the goodness-of-fit for the covariance structure. These analyses also showed that SAD models of order 3 or 4 performed better than cubic random regression and even, in some cases, than a quartic random regression model, with far fewer parameters.

Multivariate extension of random regression models requires a very large number of parameters. For example, a bivariate genetic analysis considering only a quadratic model for both genetic and environmental parts requires 45 parameters. When increasing to the cubic order for both parts, the number of parameters jumps to 75 ! In contrast, increasing the order of a structured antedependence model only adds 2 parameters at each step. Moreover, the examples analysed showed that bivariate SAD models generally perform better than random regression, with fewer parameters, and offer a high degree of flexibility to model the cross-covariance structure.

The parametrization considered here for the covariance matrix in the bivariate analysis allows only an asymmetrical modelling of the dependence between the two traits. Further study should therefore be undertaken in order to model simultaneously the dependence on trait 1 over trait 2 as well as trait 2 over trait 1. However, the chosen structured antedependence models proved to be able to deal with a symmetric cross-correlation pattern as shown in the cow data analysis as well as an asymmetric pattern as for the fly data.

Chapter 5

Contrasting models for lactation curve analysis

5.1 Introduction

Several methodologies have already been proposed for genetic evaluation of production traits for dairy cattle based on test-day-records. Currently, the most commonly used are random regression models (Diggle et al., 1994 ; Jamrozik and Schaeffer, 1997). The idea of these models is to consider a mean curve in the population, which can be either parametric or non-parametric, and to model individual deviations from this mean curve for each animal. These deviations are usually modelled with polynomial functions and more specifically orthogonal polynomials that have desirable numerical properties. Estimates of genetic values at each time are directly obtained from these individual curves. Another approach, called character process models, has recently been proposed by Pletcher and Geyer (1999) and concentrates on the modelling of the covariance structure. If a completely unstructured matrix were considered, which corresponds to a multivariate analysis, the number of parameters to be evaluated would be very large.

The character process approach aims at reducing the number of parameters in the covariance structure by considering appropriate parametric functions for the variance and correlation. Structured antedependence models (Zimmerman and Nunez-Anton, 1997 ; Nunez-Anton and Zimmerman, 2000) have also been proposed in the statistical literature. They seem to offer the same advantages as character processes, ie. flexibility in modelling the covariance structure, with few parameters. They correspond to a generalization of autoregressive models, allowing the variance to change with time. The aim of this chapter is to investigate and compare the behaviour of these different approaches for lactation curve analysis.

5.2 Materials and methods

Models considered were described in the Introductory Chapter. All calculations were performed using the program ASREML (Gilmour et al., 2000). An average information algorithm (Gilmour et al., 1995) was used for covariance parameter estimations.

In order to implement the structured antependence models, a Cholesky decomposition of the inverse of the covariance matrix as presented in Chapter 4 and by Pourahmadi (1999) was used.

Data set

These methodologies were applied to the genetic evaluation of first lactation milk production for dairy cattle. Lactation curves were fitted to test day records for 9277 progeny of 464 Holstein-Friesian sires, assumed unrelated. Observations were made over two years (1993 and 1994). The lactation stage of animals at first test varied between 4 and 40 days, with successive tests at approximately 30 day intervals. All cows had 10 measurements. The fixed effects considered were the age at calving, the percentage of North American Holstein genes, and herd-test-month. An exponential curve (Wilmink, 1987) was fitted as a fixed regression model for the general curve of the population:

$$g(t) = \alpha_0 + \alpha_1 t + \alpha_2 \exp(-Dt) \tag{5.1}$$

where t stands for days in milk and parameter D was assumed to be known and equal to 0.068, chosen based on previous studies (White et al., 1999).

Model comparisons

The aim was to compare the performance of the three approaches in modelling genetic and permanent environmental parts for lactation curve analysis. Many different combinations of variance (polynomials up to quadratic) and correlation functions (exponential: $\theta^{|t_i - t_j|}$, Gaussian: $\exp(-\theta(t_i - t_j)^2)$, Cauchy: $1/(1 + \theta(t_i - t_j)^2))$, stationary or non-stationary, were considered for the character process approach. Polynomials up to the quartic order were fitted for the random regression models. Antedependence up to order 4 was considered for SAD models.

5.3 Results

Phenotypic analysis

As shown in Chapter 4 and in Table 5.1, likelihood was higher for a structured antedependence model of order 3 compared to a quartic random regression, although the number of parameters was much smaller (7 compared to 15). Parameter estimates for the fourth order structured antedependence model were:

$$\text{Milk}(t) = 0.50\,\text{Milk}(t-1) + 0.22\,\text{Milk}(t-2) + 0.10\text{Milk}(t-3) + 0.05\,\text{Milk}(t-4) + e(t) \tag{5.2}$$

The fourth correlation parameter was quite small, therefore a third order SAD model may be enough for the phenotypic analysis of milk production. Likelihood for character process models was about the same as a first order antedependence model. CP models had trouble dealing with the highly non-stationary phenotypic correlation structure.

Table 5.1: Model comparison for the phenotypic analysis of milk production (NPCov: number of parameters in the covariance structure, CP and CPNS: character process with quadratic variance and exponential stationary or non-stationary correlation, RR: random regression models up to the quartic order).

Model	NPCov	Log L
SAD(1)	4	-1731
SAD(2)	5	1587
SAD(3)	6	2155
SAD(4)	7	2253
CP	4	-1874
CPNS	5	-1505
RR2	6	677
RR3	10	1564
RR4	15	2046

Genetic analysis

Based on likelihood (Table 5.2), antedependence models seem to offer a high degree of flexibility for the covariance structure with few parameters. In fact, likelihood was in general much higher for these models than most random regression or character processes. The genetic part was well fitted by a simple correlation structure, and an antedependence of order 1 was appropriate. Increasing the order of antedependence beyond this did not provide a significant improvement. On the other hand, the environmental part had a much more complex covariance structure and an antedependence of order 3 was necessary.

Table 5.2: Model comparison for the genetic analysis of milk production (NPCov: number of parameters in the covariance structure).

Model	Genetic	Environmental	NPCov	Log L
	Unstructured			
1	US	US	110	4126
2	SAD(1)	US	59	4109
	Structured antedependence model			
3	SAD(1)	SAD(3)	11	3845
4	SAD(2)	SAD(3)	12	3852
5	SAD(3)	SAD(3)	13	3854
6	SAD(2)	SAD(2)	11	3796
7	SAD(1)	SAD(1)	9	3580
	Character process			
8	Quad-ExpNS	Quad-ExpNS	11	3266
9	Quad-Exp	Quad-ExpNS	10	3259
10	Lin-Exp	Quad-ExpNS	9	3244
11	Quad-Exp	Quad-Exp	9	2759
12	Lin-Exp	Quad-Exp	8	2733
	Random regression			
13	Quartic	Quartic	31	3623
14	Quad	Quartic	22	3607
15	Cubic	Cubic	21	3336
16	Quad	Quad	13	2767

The chosen SAD model provided a likelihood even higher than a quartic-quartic random regression, that has many more parameters (31 instead of 11). Parameter estimates were:

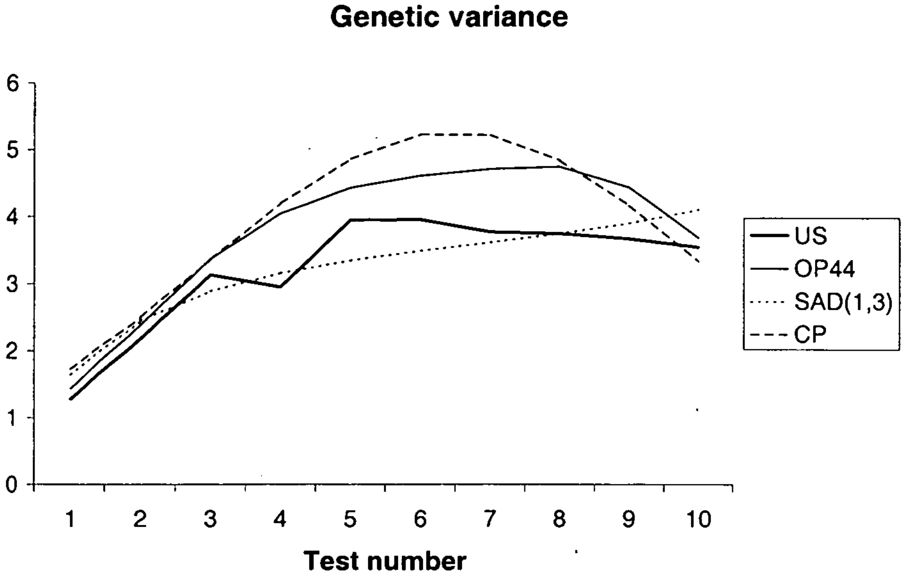
$$\text{Gen}(t) = 0.997\ \text{Gen}(t-1) + \epsilon(t) \tag{5.3}$$

$$\text{Env}(t) = 0.67\ \text{Env}(t-1) + 0.16\ \text{Env}(t-2) + 0.06\ \text{Env}(t-3) + e(t) \tag{5.4}$$

Table 5.3 gives the estimated genetic and environmental correlation matrices for the chosen antedependence model. The genetic correlation remains quite high over time and therefore can be well modelled with a simple correlation structure. On the other hand, the environmental correlation is highly non-stationary and requires a much more complex correlation structure. The non-stationary extension of character process models could not deal well with this pattern as correlations on the sub-diagonals were not monotone either increasing or decreasing. Other more appropriate parametric forms of correlation functions need to be discovered.

Figure 5.1 shows the genetic and environmental variance estimates for the different models. It can clearly be seen that although the innovation variance considered for the structured antedependence model was quadratic, the fit was considerably improved compared to character process models with a quadratic variance, and estimates were much closer to the unstructured ones. The difficulty in modelling the environmental variance seems to be mainly due to the last test. Without this variance increase at the end of the lactation, much simpler polynomial functions would fit adequately.

Figure 5.1: Genetic and environmental variances estimated under the four different models (US: Model 1 in Table 5.2. OP44: Model 13. SAD(1,3): Model 3. CP: Model 9).



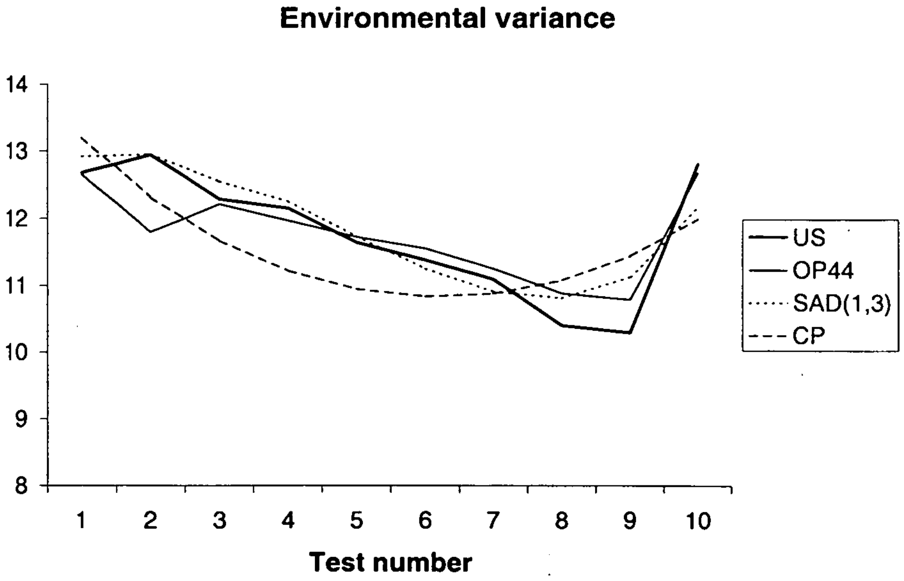


Table 5.3: Genetic correlation for SAD(1) (above diagonal) and environmental correlation for SAD(3) (below diagonal).

	1	2	3	4	5	6	7	8	9	10
1	1	0.82	0.75	0.71	0.69	0.67	0.66	0.65	0.63	0.61
2	0.66	1	0.92	0.87	0.85	0.83	0.81	0.79	0.77	0.75
3	0.62	0.79	1	0.95	0.92	0.90	0.88	0.86	0.84	0.82
4	0.60	0.75	0.85	1	0.97	0.95	0.93	0.91	0.89	0.86
5	0.56	0.71	0.80	0.88	1	0.98	0.96	0.94	0.91	0.89
6	0.52	0.67	0.76	0.83	0.89	1	0.98	0.96	0.94	0.91
7	0.49	0.63	0.71	0.78	0.83	0.89	1	0.98	0.96	0.93
8	0.45	0.58	0.66	0.72	0.77	0.82	0.87	1	0.98	0.95
9	0.41	0.52	0.60	0.65	0.70	0.75	0.79	0.85	1	0.97
10	0.36	0.46	0.52	0.57	0.61	0.65	0.69	0.74	0.81	1

In order to check the goodness-of-fit of the genetic part, an unstructured covariance was fitted for the environmental part and a first order antedependence model (SAD(1)) was used to model the genetic part. The likelihood difference compared to a completely unstructured model was equal to 17 whereas it was 281 for the chosen SAD model (SAD(1,3)). It can therefore be concluded that the genetic part was well fitted with the first order antedependence model, and that some improvement in the fit of the environmental part can still be achieved. It would for example be possible to consider a residual variance changing with time as proposed in Chapter 6.

5.4 Discussion

These analyses showed that structured antedependence models offer a high degree of flexibility in modelling covariance structure for genetic analysis of milk production in dairy cattle. The environmental correlation pattern is quite complex and highly non-stationary, but seems to be well modelled with a third order antedependence structure that performed better than a quartic random regression, with far fewer parameters. The genetic correlation pattern is simpler and can be captured with a first order SAD model.

There is however one limitation concerning the use of SAD models for national genetic evaluation: they do not provide simple individual genetic curves as random regression models, and at present one genetic value is estimated for each animal at each time of measurement. This may still be possible for monthly records, where at most 10 measures are available per animal over time, but will be a problem when the number of observations is larger. Chapter 2 showed that random regression models can deal well with a correlation that remains quite high over time. Therefore, a possible way to overcome this difficulty could be to consider a simple random regression model, either linear or quadratic, for the genetic part, and a structured antedependence model for the more complex environmental part. In that case, only two or three genetic parameters will be estimated for each animal, regardless the number of observations over time, while the number of parameters for the environmental covariance structure will be kept low with the SAD model.

It would also be possible to achieve more flexibility with structured antedependence models by allowing heterogeneous variances. In the previous analyses, variances were assumed to change as a polynomial function of time, but it would also be possible to incorporate other covariables as originally proposed by Foulley and Quaas (1995) in their structural models. This would improve the accuracy of estimation of the covariance structure and therefore the genetic value predictions.

Calculation of loss of efficiency of genetic response under the different models is presented in the following Appendix. Losses in the response to selection are found to be quite small. However, when applied to the dairy cattle selection scheme involving millions of animals, the slight differences can have a significant economic impact, as was already observed with the introduction of heterogeneous variances. The impact will be mainly on individual cow selection for which the ranking may be changed, and on early bull testings when not much information is available on daughters.

Appendix: Response to selection

Using the selection index theory as presented by Cameron (1997), let us consider the 10 tests as 10 correlated traits $Y_1, Y_2, ..., Y_{10}$. Each trait is assumed to have the same economic weight $a_i = 1$ for all $i = 1, ..., 10$. For individual selection on cows, the selection index is given by: $I = \sum_i b_i Y_i$. Let $\boldsymbol{P}$ be the phenotypic covariance matrix between the 10 traits and $\boldsymbol{G}$ the genetic covariance matrix. The selection criterion coefficients $\boldsymbol{b} = \{b_i\}_{i=1,...,10}$ can be calculated by:

$$\boldsymbol{b} = \boldsymbol{P}^{-1}\boldsymbol{G}\boldsymbol{a} \tag{5.5}$$

When genetic and phenotypic parameters are known, the response to selection is given by:

$$R = i\frac{\boldsymbol{b}'\boldsymbol{G}\boldsymbol{a}}{\sqrt{\boldsymbol{b}'\boldsymbol{P}\boldsymbol{b}}} \tag{5.6}$$

where i is the standardised selection differential in the selection criterion I. In general, however, genetic and phenotypic parameters have to be estimated, and coefficients of the selection criterion are given by:

$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{G}}\boldsymbol{a} \tag{5.7}$$

The actual response to selection given the estimated selection criterion coefficients is:

$$R^* = i\frac{\hat{\boldsymbol{b}}'\boldsymbol{G}\boldsymbol{a}}{\sqrt{\hat{\boldsymbol{b}}'\boldsymbol{P}\hat{\boldsymbol{b}}}} \tag{5.8}$$

If the estimated genetic and phenotypic parameters are not equal to the population parameters, then the response will be less than the maximum response. The loss in efficiency due to the difference between the estimated and population parameters (Sales and Hill, 1976) is:

$$d = 1 - \frac{R^*}{R} \tag{5.9}$$

Considering estimates provided by the unstructured model as true parameters, the loss in efficiency was calculated under each model for individual cow selection, and values are given in the table below. Although the loss of efficiency was small under each model, it was the smallest for structured antedependence models, that also have the highest likelihood.

Table 5.4: Loss of efficiency of the genetic response for individual cow selection under structured antedependence model (SAD(1,3): Model 1 in Table 5.2), quartic-quartic random regression (OP44: Model 13), and character process model (CP: Model 9).

Model	Loss of efficiency	LogL
SAD(1,3)	0.021	3845
OP44	0.023	3623
CP	0.060	3259

Assuming a balanced sire design, let n be the number of daughters per sire. The selection indices for sires are now based on averages of daughters observations: $I_s = \sum_i b_i \bar{Y}_i$ where $\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$. The phenotypic covariance matrix $\boldsymbol{P_s}$ is given by:

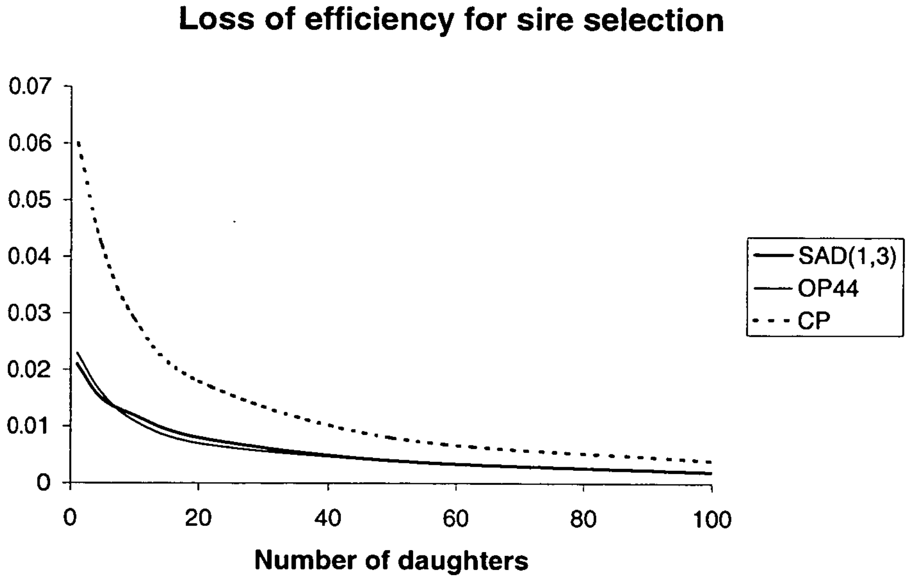
$$\boldsymbol{P_s} = \frac{1}{n}\boldsymbol{P} + (1 - \frac{1}{n})\boldsymbol{G} \tag{5.10}$$

where $\boldsymbol{G}$ is the between-sire covariance matrix. Replacing this value in formulae given above, it is possible to calculate the loss of efficiency for sire selection depending on the number of daughters per sire. Values are presented in the graph and table below.

Table 5.5: Loss of efficiency of the genetic response for sire selection under each model, depending on the number of daughters.

	Loss of efficiency						
Model	n=1	n=5	n=10	n=20	n=50	n=100	LogL
SAD(1,3)	0.021	0.015	0.012	0.008	0.004	0.002	3845
OP44	0.023	0.016	0.011	0.007	0.004	0.002	3623
CP	0.060	0.042	0.029	0.018	0.008	0.004	3259

Figure 5.2: Loss of efficiency for sire selection depending on the number of daughters.



Chapter 6

Residual variance modelling

6.1 Introduction

In the longitudinal data framework, part of the heterogeneity of variances across time in the population can be modelled via random regression (Jamrozik and Schaeffer, 1997 ; Verbeke and Molenberghs, 1997) or covariance functions (Kirkpatrick et al., 1994 ; Meyer and Hill, 1997). Nevertheless, heterogeneity usually remains in the residual variances. More specifically, in the case of the analysis of test day records for milk production in dairy cattle, different studies (Brotherstone et al., 1999 ; White et al., 1999) have shown that the residual variance changes over time. To cope with this heterogeneity, authors divide the lactation length in different intervals, assuming homogeneity within intervals and heterogeneity between them (Jamrozik and Schaeffer, 1997 ; Rekaya et al., 1999). However, this method can lead to a large number of variance parameters to be estimated and, moreover, requires the definition of arbitrary subclasses within which the variance is assumed constant, whereas the change of the residual variance is continuous over time.

Recently, Rekaya et al. (1998) proposed a changepoint technique to account for the heterogeneity of residual variances along lactation. This approach offers a way to model continuously the changes of the residual variance over time, but assumptions need to be made about the number of changepoints and the relationship between the residual variance and the number of days in milk. Moreover, the number of parameters that have to be estimated may still be quite large and the estimation (using for instance Bayesian techniques) time consuming.

The aim of this chapter is to propose another way to account for this heterogeneity, and to model the changes of the residual variance along lactation as a continuous function of time. In this purpose, a structural model, as proposed by Foulley and Quaas (1995) is assumed on the residual variances, and the covariates of this model are parametric functions of time. This procedure offers two main advantages: the number of parameters to be estimated for the residual variances is reduced compared to a purely heterogeneous model, and the changes in the residual variance are considered to be continuous over time, so there is no need to define arbitrary classes of heterogeneity.

The estimates of the parameters for this model on the variances were obtained using an EM-REML type algorithm. The equivalence between this system of equations and the GLM estimating equations was shown by Lee and Nelder (1999). This methodology is illustrated by an analysis of a real data set of monthly records for milk production in dairy cattle.

6.2 Materials and methods

6.2.1 Model

Consider a population with I individuals, with individual i having n_i observations. The time and the number of measurements may be different for each individual. For the sake of simplicity, a simple mixed model (Laird and Ware, 1982) for the analysis of longitudinal data was assumed:

$$y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{u}_i + e_{ij} \tag{6.1}$$

where y_{ij} is the j^{th} measurement on individual i at time t_{ij} ($i = 1, ..., I$ and $j = 1, ..., n_i$). $\boldsymbol{\beta}$ are the fixed effects associated to the incidence matrix $\mathbf{X}$ (of row $\mathbf{x}_{ij}^{'}$), and $\mathbf{u}_i$ is the vector of random effects for individual i, with incidence matrix $\mathbf{Z}$ (of row $\mathbf{z}_{ij}^{'}$). It is assumed that $\mathbf{u} = (\mathbf{u}_1^{'}, ..., \mathbf{u}_I^{'})' \sim \mathcal{N}(0, \mathbf{G})$, and that the residuals e_{ij} are independent and such that:

$$e_{ij} \sim \mathcal{N}(0, \sigma^2_{e_{ij}}) \tag{6.2}$$

In order to model the heterogeneity of the residual variances over time, a structural model (Foulley and Quaas, 1995 ; Foulley et al., 1998) was assumed:

$$\ln\sigma^2_{e_{ij}} = \mathbf{p}_{ij}^{'}\boldsymbol{\delta} \tag{6.3}$$

For instance, if a quadratic function of time is appropriate for the data studied, then

$$\ln\sigma^2_{e_{ij}} = a + bt_{ij} + ct^2_{ij} \tag{6.4}$$

and $\mathbf{p}_{ij}^{'} = (1 \; t_{ij} \; t^2_{ij})$. The model can easily be extended to higher order polynomials or other parametric functions of time. A step-wise procedure could be used to choose the covariates in the structural model, as discussed by Foulley and Quaas (1995).

Using an EM-REML procedure (Dempster et al., 1977) and Lee and Nelder's (1999) result (as detailed in the appendix), estimation of all the parameters in this model can be obtained by iterating between the following procedures which can be achieved with existing software (SAS, Genstat, AS-REML, etc.):

1. Mixed model equations (MME) are constructed assuming a fixed residual variance to obtain estimates of the factors in the model and residuals $\hat{e}_{ij}$.

2. A regression model is applied to the log of the squared residuals (GLM equations described in the appendix), to obtain an estimate of $\boldsymbol{\delta}$ in equation (6.3).

3. Mixed model equations are constructed again, but using the regression function to determine the appropriate residual variance for each time t_{ij}, $\sigma^2_{e_{ij}}$, the inverse of which is used as the weighting of the MME.

4. Back to step 2, until convergence is reached.

Other algorithms such as those proposed by Foulley et al. (1990), Verbyla (1993) and Schnyder et al. (1999) could also be used for estimating the parameters in the structural model, and may differ in convergence rate, ability to remain in the parameter space and computing time.

6.2.2 Application

The preceding theory was applied to the data set used by White et al. (1999). Lactation curves were fitted to test day records of milk production for 2885 progeny of 30 Holstein-Friesian sires in 503 herds. The lactation stage of animals entering the first test varied between 4 and 40 days, with successive tests at approximate 30-day intervals (10 tests for each cow). The fixed effects considered were the age at calving, the percentage of Holstein genes, and herd-test-month. White et al. (1999) considered a sire model, and modelled the mean curve of the population as well as the genetic and environmental effects non-parametrically using smoothing splines.

Here, the exponential curve of Wilmink (1987) was fitted as a fixed regression model for the general mean curve of the population:

$$g(t) = b_0 + b_1 t + b_2 \exp(-Dt) \qquad (6.5)$$

where t stands for days in milk (DIM). The parameter D was assumed to be known and equal to 0.068, chosen based on previous studies (Brotherstone et al., 1999 ; White et al., 1999). A sire model was considered, and quadratic random regressions were assumed to model both the genetic and environmental effects :

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + a_{k0} + a_{k1}t_{ij} + a_{k2}t^2_{ij} + b_{i0} + b_{i1}t_{ij} + b_{i2}t^2_{ij} + e_{ij} \tag{6.6}$$

where y_{ij} was the milk production of cow i taken at time t_{ij}, $\boldsymbol{x}'_{ij}\boldsymbol{\beta}$ were the fixed effects described above, $a_{k0} + a_{k1}t_{ij} + a_{k2}t^2_{ij}$ was the quadratic random regression for the genetic effect (sire k), $b_{i0} + b_{i1}t_{ij} + b_{i2}t^2_{ij}$ was the quadratic random regression for the environmental effect (for cow i within sire k). Parameters $\mathbf{a}_k = (a_{k0}, a_{k1}, a_{k2})'$ and $\mathbf{b}_i = (b_{i0}, b_{i1}, b_{i2})'$ were assumed to follow multivariate normal distributions, and e_{ij} was the residual term ($e_{ij} \sim \mathcal{N}(0, \sigma^2_{e_{ij}})$).

Two different models for the residual variances were considered:

Model 1: Ten classes were assumed for the residual variances, i.e. one for each measurement as considered by White et al. (1999).

Model 2: A structural model was assumed on the residual variances, as equation (6.4) a quadratic polynomial of time being considered. The estimates of the parameters for the latter model were obtained by iterating between AS-REML for the mixed model equations and SAS for the GLM equations, but this procedure could also easily be incorporated in a REML package.

6.3 Results

Fixed effect solutions were very similar for both models, and were also similar to those obtained by White et al. (1999), who fitted a 10-knot spline on the same data set. In fact, the breed difference (Holstein-Friesian) was estimated in the first model at 1.56 kg (SE=0.44) and in the second model at 1.51 kg (SE=0.44), and the effect of age at calving as 0.18 kg/mo (SE=0.02) in the two models. As seen in Table 6.1, the estimates of genetic parameters were also very similar in the two models and very close to the results of White et al. (1999).

Table 6.1: Mean DIM, variance estimates (kg^2), and heritabilities by test (G=Genetic, E=Environmental, R=Residual).

		Model 1				Model 2			
Test	DIM	G	E	R	h^2	G	E	R	h^2
1	18	3.08	9.17	4.93	0.21	3.11	9.20	5.16	0.21
2	48	3.02	7.90	4.10	0.24	3.05	7.91	4.17	0.24
3	78	3.11	7.42	3.74	0.26	3.13	7.42	3.49	0.27
4	109	3.25	7.31	3.23	0.29	3.26	7.31	3.01	0.29
5	139	3.36	7.33	2.36	0.32	3.37	7.32	2.72	0.31
6	169	3.41	7.33	2.69	0.31	3.42	7.32	2.54	0.32
7	199	3.43	7.33	2.49	0.32	3.44	7.32	2.46	0.32
8	229	3.45	7.49	2.43	0.32	3.47	7.49	2.47	0.32
9	259	3.56	8.09	2.07	0.32	3.60	8.11	2.57	0.31
10	290	3.90	9.63	3.56	0.28	3.96	9.69	2.78	0.29

Figure 6.1 shows that the quadratic function was a good representation for the changes of the residual variance. Table 6.2 gives the estimates of the parameters of the structural model (Model 2) with their standard errors. All were significantly different from 0, and the quadratic function for the residual variances was:

$$\ln\sigma^2_{e_{ij}} = 0.97 - 0.073\ t_{ij} + 0.018\ t_{ij}^2 \tag{6.7}$$

(with t_{ij}=(DIM-150)/30). Although this quadratic function seemed to be quite appropriate, the likelihood was higher for the first than the second model (difference of 32 for the Log-likelihood). Nevertheless, as there were fewer parameters for the residual variance to be estimated in the second (three) than in the first model (ten), a criterion such as Schwarz Bayesian Criterion (1978) is more appropriate. This criterion penalizes the likelihood with respect to the number of parameters and is defined by :

$$\text{Loglikelihood} - \frac{1}{2} \times \text{number of parameters in the model} \times \text{Log } n^*$$

where $n^* = n - p$ when using REML with n the number of observations in the data set and p the number of fixed effects. It showed a slightly better fit for the second than the first model (difference of 4).

Figure 6.1: Changes of the residual variance over time for the two models. Model 1 : 10 different classes of heterogeneity (1 for each test). Model 2 : Structural model on the residual variance ($\ln\sigma^2_{e_{ij}} = 0.97 - 0.073\ t_{ij} + 0.018\ t^2_{ij}$, where t_{ij}=(DIM-150)/30).

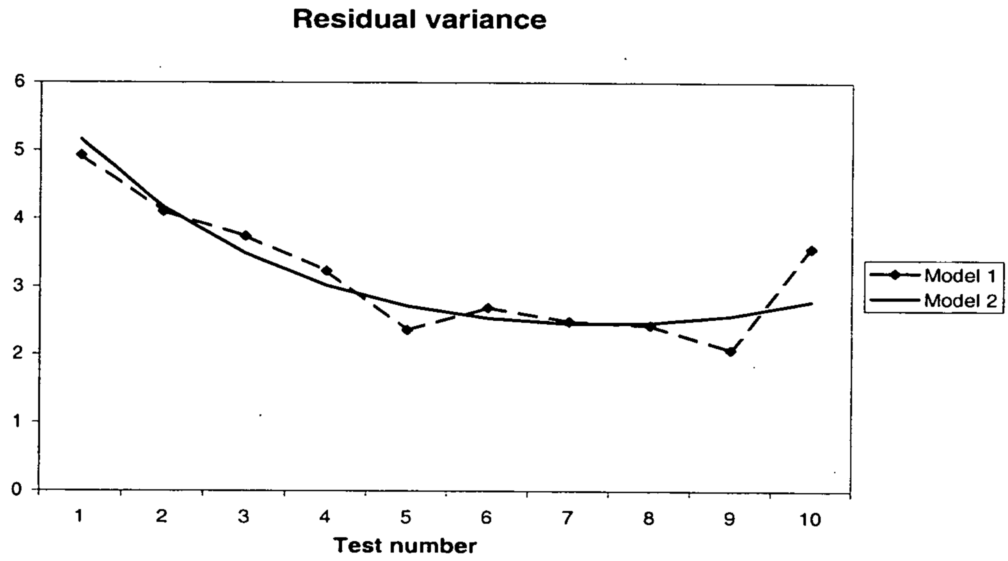


Table 6.2: Estimates and SE of the parameters of the structural model on the residual variances (Model 2) ($P < 0.001$ for each parameter).

Parameters	Estimate	SE
a	0.970	0.008
b	-0.073	0.002
c	0.018	0.001

6.4 Discussion

The improvement in fit of the structural model on the residual variance compared to the heterogeneous model (assuming ten different classes of heterogeneity) was not great. However this method would prove to be much better in the case of high heterogeneity within classes. For instance, with model 1 modified so that the lactation was divided into five intervals rather than ten, the likelihood was greater for the second than the first model (difference of 9 for the Log-likelihood), even though model 2 still had fewer parameters.

Nevertheless, the polynomial functions may not be the most appropriate, especially because of their lack of flexibility to model the variances at the beginning and at the end of the lactation. Other more flexible parametric functions could be considered using the same methodology.

This method offers two important advantages: there are fewer parameters to be estimated than in the classical heterogeneous model, and the variance is a continuous function of time, with no arbitrary classes. This approach could also be a useful alternative for other longitudinal studies that arise in animal breeding, for instance growth curve analyses.

Other factors of heterogeneity could be taken into account in the structural model on the residual variances (Foulley and Quaas, 1995), for instance the age at calving, month of calving, region, year, and even the herd-test-month (perhaps as a random effect). This aspect of the heterogeneity of variances, which applies to the residual variances as well as the genetic and permanent environmental variances, has to be investigated more thoroughly.

Appendix

The REML estimates of the parameters in the structural model for the residual variance were obtained using an EM algorithm (Dempster et al., 1977).

Letting $\mathrm{c} = (\mathbf{y}', \boldsymbol{\theta}')'$ be the complete set of data, and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{u}')'$ the vector of the missing values. The likelihood function of the complete data is:

$$\mathrm{p}(\mathrm{c}|\boldsymbol{\delta}, \mathbf{G}) = \mathrm{p}(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\delta})\mathrm{p}(\boldsymbol{\beta}, \mathbf{u}|\mathbf{G}) \tag{6.8}$$

Therefore, the log-likelihood is:

$$-2\ln\mathrm{p}(\mathrm{c}|\boldsymbol{\delta}, \mathbf{G}) = -2\mathrm{L}(\boldsymbol{\delta}, \mathbf{G}; \mathrm{c}) = -2\mathrm{L}(\boldsymbol{\delta}; \mathbf{e}) - 2\mathrm{L}(\mathbf{G}; \mathbf{u}) \tag{6.9}$$

and the estimation of $\boldsymbol{\delta}$ can then be separated from that of $\mathbf{G}$, considering the log-likelihood:

$$-2\mathrm{L}(\boldsymbol{\delta}; \mathbf{e}) = \text{const.} + \sum_{\mathrm{i}=1}^{\mathrm{I}}\sum_{\mathrm{j}=1}^{\mathrm{n_i}}[\ln\sigma^2_{\mathrm{e_{ij}}} + \frac{1}{\sigma^2_{\mathrm{e_{ij}}}}\mathrm{e}^2_{\mathrm{ij}}] \tag{6.10}$$

The E-step is defined as usual, i.e. at iteration (r) one calculates the conditional expectation of $\mathrm{L}(\boldsymbol{\delta}; \mathbf{e})$ given the data $\mathbf{y}$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^{(\mathrm{r})}$.

$$\begin{aligned} \mathrm{Q}(\boldsymbol{\delta}|\boldsymbol{\delta}^{(\mathrm{r})}) &= \mathrm{E}(-2\mathrm{L}(\boldsymbol{\delta}; \mathbf{e})|\mathbf{y}, \boldsymbol{\delta}^{(\mathrm{r})}) && (6.11) \\ &= \text{const.} + \sum_{\mathrm{i}=1}^{\mathrm{I}}\sum_{\mathrm{j}=1}^{\mathrm{n_i}}[\ln\sigma^2_{\mathrm{e_{ij}}} + \frac{1}{\sigma^2_{\mathrm{e_{ij}}}}\mathrm{E_c}(\mathrm{e}^2_{\mathrm{ij}})] && (6.12) \end{aligned}$$

where $\mathrm{E_c}(\mathrm{e}^2_{\mathrm{ij}})$ stands for the conditional expectation $\mathrm{E}(\mathrm{e}^2_{\mathrm{ij}}|\mathbf{y}, \boldsymbol{\delta}^{(\mathrm{r})})$, and

$$\begin{aligned} \mathrm{E}(\mathrm{e}^2_{\mathrm{ij}}|\mathbf{y}, \boldsymbol{\delta}^{(\mathrm{r})}) &= (\mathrm{E}(\mathrm{e_{ij}}|\mathbf{y}, \boldsymbol{\delta}^{(\mathrm{r})}))^2 + \mathrm{trace}(\mathrm{Var}(\mathrm{e_{ij}}|\mathbf{y}, \boldsymbol{\delta}^{(\mathrm{r})})) && (6.13) \\ &= \hat{\mathrm{e}}^2_{\mathrm{ij}} + \mathrm{Var}(\mathrm{e_{ij}}|\mathbf{y}, \boldsymbol{\delta}^{(\mathrm{r})}) && (6.14) \end{aligned}$$

The M-step consists of calculating the next value $\boldsymbol{\delta}^{(\mathrm{r}+1)}$ by minimizing the function $\mathrm{Q}(\boldsymbol{\delta}|\boldsymbol{\delta}^{(\mathrm{r})})$ with respect to $\boldsymbol{\delta}$,

$$\frac{\partial Q}{\partial \boldsymbol{\delta}} = \frac{\partial Q}{\partial \sigma^2_{e_{ij}}} \frac{\partial \sigma^2_{e_{ij}}}{\partial \ln \sigma^2_{e_{ij}}} \frac{\partial \ln \sigma^2_{e_{ij}}}{\partial \boldsymbol{\delta}} \tag{6.15}$$

Then

$$\frac{\partial Q}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} [1 - \frac{1}{\sigma^2_{e_{ij}}} E_c(e^2_{ij})] \mathbf{p}_{ij} \tag{6.16}$$

$$\frac{\partial Q}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (1 - w_{ij})(1 - \frac{1}{\sigma^2_{e_{ij}}} d^*_{ij}) \mathbf{p}_{ij} \tag{6.17}$$

where $w_{ij} = \frac{1}{\sigma^2_{e_{ij}}} \mathrm{Var}(e_{ij}|\mathbf{y}, \boldsymbol{\delta}^{(r)})$ and $d^*_{ij} = \hat{e}^2_{ij}/(1 - w_{ij})$.

Lee and Nelder (1999) showed that this system of equations is equivalent to the estimating equations for a GLM (McCullagh and Nelder, 1989) with response d^*_{ij} (where d^*_{ij} is the square of the residuals divided by the weight $(1 - w_{ij})$), mean $\sigma^2_{e_{ij}}$, error gamma, log-link $(\ln(\sigma^2_{e_{ij}}))$, linear predictor: $\xi_{ij} = \mathbf{p}'_{ij}\boldsymbol{\delta}$, and prior weight $(1 - w_{ij})$.

The values of $\hat{e}^2_{ij}$ and $\mathrm{Var}(e_{ij}|\mathbf{y}, \boldsymbol{\delta}^{(r)})$ can be calculated from the solutions of the MME (mixed model equations) as follows:

$$\hat{e}_{ij} = y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} - \mathbf{z}'_{ij}\hat{\mathbf{u}} \tag{6.18}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are the BLUP solutions.

Letting $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{u}')'$ and $\mathbf{b}_{ij} = (\mathbf{x}'_{ij}, \mathbf{z}'_{ij})$, then $e_{ij} = y_{ij} - \mathbf{b}_{ij}\boldsymbol{\theta}$ and therefore

$$\mathrm{Var}(e_{ij}|\mathbf{y}, \boldsymbol{\delta}^{(r)}) = \mathbf{b}_{ij}\mathrm{Var}(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\delta}^{(r)})\mathbf{b}'_{ij} \tag{6.19}$$

where $\mathrm{Var}(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\delta}^{(r)})$ corresponds to the inverse of the coefficient matrix in the MME.

Chapter 7

How to deal with incomplete and short lactations

7.1 Introduction

Classical longitudinal data models, such as random regression, are often said to be able to deal with missing values and unbalanced data sets. They are indeed able to deal with data that are *missing at random*, as defined by Little and Rubin (1987). Laird (1988) in fact showed that in this case inferences about parameters in the model are not influenced by the missing process that can then be ignored. This study concerns parameters in the population such as fixed effects or variance parameters. Individual predictions will however be affected by a missing process related to the observed character as illustrated by the examples below. In the case of milk production for dairy cattle, longitudinal models are able to extrapolate incomplete lactations and to predict a production at 305 days, for example. This prediction however does not take into account the fact that the cow may be dried off before day 305. Moreover, at present no special treatment is considered for cows with short lactations. As for incomplete lactations, the model predicts a production at day 305, that can even be negative with some random regression models, although production of the cow is known and equal to zero. An easy way to deal with short lactations would be to add zeros in the data set when the cow has been dried. However, for models such as random regression that try to fit a parametric individual curve, genetic value estimations would be greatly penalized at all times due to these zeros and the risk of having negative predictions for milk production at late stage of the lactation will be even greater. Moreover, variances at the end of the lactation would be underestimated. We therefore propose to keep analysing data as usual and to perform an a posteriori correction of phenotypic and genetic predictions using parameters estimated in the model combined with the probability for the cow to be dried off at each time. This procedure will be adapted to deal with both incomplete and short lactations and can be used with any longitudinal model. A simulation study illustrates the use of this methodology.

7.2 Model

There are two different ways by which missing data can arise in genetic evaluation for milk production based on test day records. The first one is when genetic analysis is decided at a given date and cows can be included in the analysis even if their lactation is not complete yet and they are still producing milk but the following records are still unknown (incomplete lactations). Another reason for missing data is when the cow is dried off (short lactations). Although data are not really missing as the production values for the cow are known and equal to 0, they appear as missing values in the data file.

Many papers have been published in the statistical literature about how to deal with missing data in longitudinal studies (Wu and Carroll, 1988; Diggle and Kenward, 1994). We are concerned here with the drop-out process, i.e. when the subject gets out of the study (either because the cow has been dried off or because measurements stopped at a given date). Little and Rubin (1987) propose to consider three kind of missingness:

(a) *Completely random dropout*: the dropout and measurement processes are independent, which is the case for incomplete lactations.

(b) *Random dropout*: the dropout process depends on the *observed* measurements, i.e. those preceding dropout. This is typically the case for short lactations when the cow is dried off by the breeder when its production drops below a given threshold.

(c) *Informative dropout*: the dropout process depends on the *unobserved* measurements, i.e. those that would have been observed if the subject had not dropped out.

Laird (1988) showed that a *completely random*, but also a *random* dropout can be ignored and will not affect inferences about parameters of the measurement process. Therefore, fixed effects and covariance parameters provided by the longitudinal model for lactation curve analysis will be valid. This conclusion about ignorability of the dropout process is based only on parameters in the population. Individual predictions, however, will have to be corrected for the dropout process, even for a randomly missing data. This will be illustrated later on in the case of short and incomplete lactations for evaluation of individual cumulative milk production.

Let y_{it}^* be the production for animal i at time t if the cow was assumed to be never dried off. Classical longitudinal models that do not take into account the dropout process work on variable y_{it}^* instead of working with the actually observed data y_{it}. For simplicity, it is assumed that the cow is dried off by the breeder when its production drops below a given threshold s:

$$y_{it} = \begin{cases} y^*_{it} & \text{if} \quad y^*_{i(t-1)} \geq s \\ 0 & \text{if} \quad y^*_{i(t-1)} < s \end{cases}$$

This model is very similar to the Tobit model used in econometrics. The cumulative production at time T is the sum of all the observed daily productions of cow i up to time T and is given by:

$$S_{iT} = \sum_{t=1}^{T} y_{it} \tag{7.1}$$

Classical longitudinal models approximate this sum by $\sum_{t=1}^{T} y^*_{it}$, which corresponds to the cumulative milk production if the cow was never dried off. This will therefore overestimate the cumulative milk production for cows with short lactations.

For short as well as incomplete lactations, observations for cow i are missing from a given time T_0. The idea would therefore be to use expectation instead of actual milk production in order to evaluate the cumulative value. It is also clear that short and incomplete lactations should not be treated the same way in order to evaluate the cumulative milk production. The model presented above, assuming that the cow is dried off when its production goes below a given threshold will allow to treat them differently, as cows with short lactations will have a lower production level (probably below threshold s) than cows with incomplete lactations, except for the ones that will be dried off early. So, this model seems to be flexible, and will correct predictions for both short and incomplete lactations.

The longitudinal model can be written as:

$$y^*_{ikt} = g(t) + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}_{it}\boldsymbol{a}_k + \boldsymbol{q}_{it}\boldsymbol{b}_i + e_{ikt} \tag{7.2}$$

where $g(t)$ is the mean curve in the population, that can be either parametric or non-parametric, $\boldsymbol{x}'_{it}\boldsymbol{\beta}$ are fixed effects, that can be time-dependent, $\boldsymbol{z}_{it}\boldsymbol{a}_k$ is the genetic value for sire k at time t and $\boldsymbol{q}_{it}\boldsymbol{b}_i$ is the permanent environmental effect for animal i at time t. It is assumed that $\boldsymbol{a}_k \sim \mathcal{N}(0, \boldsymbol{G})$, $\boldsymbol{b}_i \sim \mathcal{N}(0, \boldsymbol{P})$ and

$\boldsymbol{e}_i \sim \mathcal{N}(0, \boldsymbol{R})$. This model is very general and can be any longitudinal model such as random regression or character process models.

The corrected cumulative milk production for animal k is given by:

$$E(S_{iT}|\boldsymbol{a}_k) = \sum_{t=1}^{T} E(y_{it}|\boldsymbol{a}_k) \tag{7.3}$$

In a sire model, $\boldsymbol{a}_k$ is the genetic value for sire k of animal i. With the model given in equation (7.2), it follows that:

$$E(y^*_{it}|\boldsymbol{a}_k) = g(t) + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}_{it}\boldsymbol{a}_k = \mu_{it} \tag{7.4}$$

$$Var(y^*_{it}|\boldsymbol{a}_k) = \boldsymbol{q}_{it}\boldsymbol{P}\boldsymbol{q}'_{it} + \sigma^2_{eit} = \sigma^2_{it} \tag{7.5}$$

$$Cov(y^*_{it}, y^*_{it-1}|\boldsymbol{a}_k) = \boldsymbol{q}_{it}\boldsymbol{P}\boldsymbol{q}'_{it-1} = \sigma_{it(t-1)} \tag{7.6}$$

Using calculations presented in the appendix, it can be shown that:

$$E(y_{it}|\boldsymbol{a}_k) = \mu_{it}\ \Phi(\frac{\mu_{i(t-1)} - s}{\sigma_{i(t-1)}}) + \frac{\sigma_{it(t-1)}}{\sigma_{i(t-1)}}\ \phi(\frac{\mu_{i(t-1)} - s}{\sigma_{i(t-1)}}) \tag{7.7}$$

where $\Phi(.)$ and $\phi(.)$ are the cumulative probability and density functions of a $\mathcal{N}(0, 1)$, respectively. This expectation corresponds to a corrected prediction of the production of animal i at time t, taking into account the drying off process. It is easy and fast to compute from parameter estimates obtained in the longitudinal model (fixed effects, genetic values and variance parameters). As explained above, the dropout process is either *completely random* (for incomplete lactations), or *random* (for short lactations) and inference for parameters in the model remains valid. Estimates can therefore be used in the corrected individual predictions.

This procedure can also be used to correct the prediction of genetic values at time t. This correction can in fact simply be obtained by:

$$E(y_{it}|\boldsymbol{a}_k) - E(y_{it}) \tag{7.8}$$

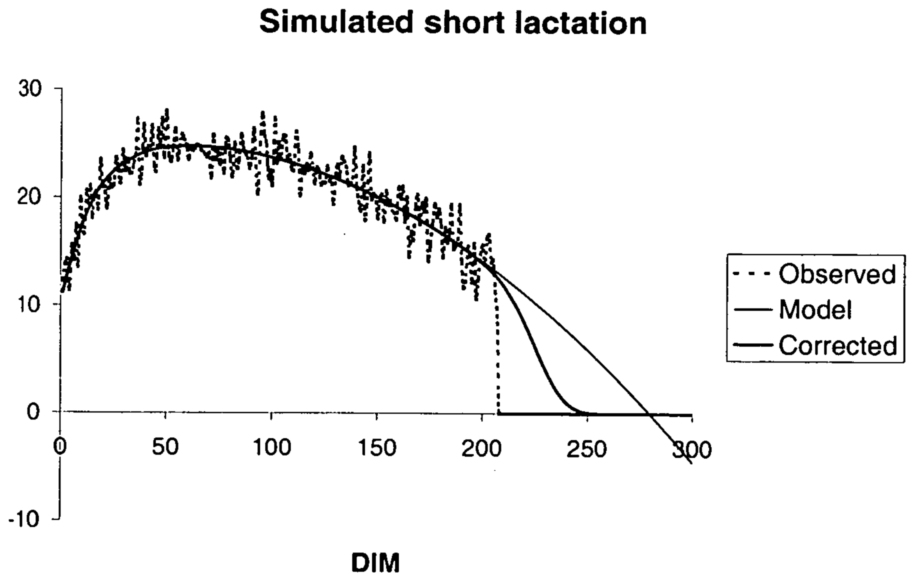
that corresponds to the sire deviation k from the mean curve.

7.3 Simulation study

7.3.1 Phenotypic analysis

A data set for daily records of milk production was simulated, considering 1000 cows with 300 measures per cow. Data were simulated according to a quadratic random regression model, using a Wilmink curve for the population mean. Parameters of the model were based on analyses of real data sets for milk production. The daily production threshold s for a cow to be dried off was set to 10 kg, and the minimum length of lactation curve was assumed to be 150 days.

Figure 7.1: Individual milk production: simulated curve, prediction obtained with the random regression model, corrected prediction taking into account the drying off process.



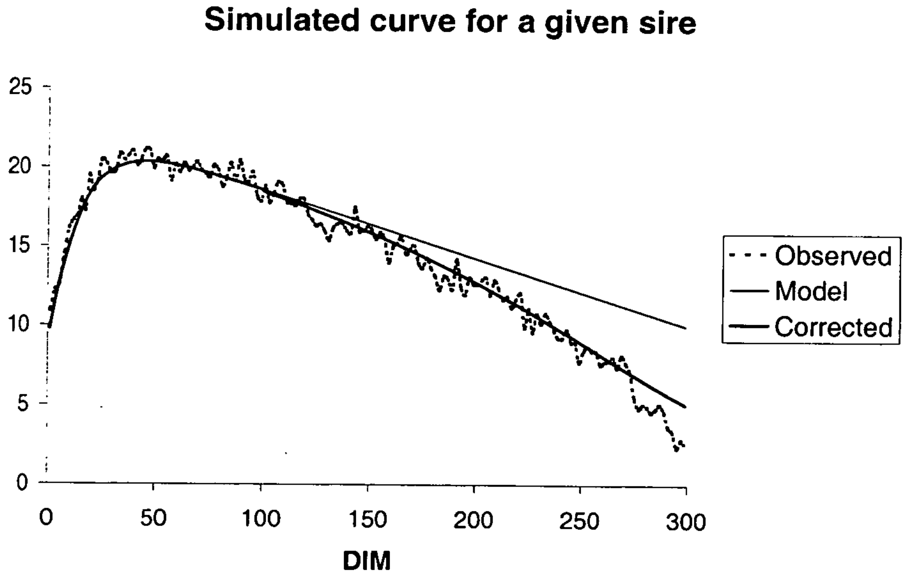
Expectation of total milk production at day 300 was calculated from the model and using the proposed corrected procedure. Figure 7.1 shows an individual curve for a simulated short lactation. Observed (simulated) production as well as curves estimated with the model and the proposed procedure are presented. It can be seen that the proposed methodology significantly improves the accuracy of the predicted production values compared to a classical random regression model. This is especially the case at the end of the lactation where expected production value is equal to 0 with the corrected approach, whereas it was predicted at -4.5 kg (with a negative expected production !). The methodology also provides a more accurate estimation of the cumulative production. The observed (simulated) cumulative production in this case was 4329 kg. It was overestimated by 454 kg by the random regression model, and only by 222 kg with the corrected procedure.

7.3.2 Genetic analysis

The problem of dealing with short lactations will also have an important impact on genetic evaluation, especially for genetic values at the end of the lactation. For example, for a sire having most of his daughters dried off before day 300, the genetic value predicted with classical longitudinal models will be overestimated if the drying off process is not taken into account.

In order to illustrate this problem, a data set was simulated as previously considering 100 sires, 10 daughters per sire and 300 observations per cow. Figure 7.2 shows the average milk production for the daughters of one of the simulated sires. Predicted curve obtained with the model as well as corrected curve are presented. This graph nicely illustrates the fact that the model predicts a production curve assuming that the cows are never dried off. The proposed correction seems to be able to model much more accurately the actual production, and will therefore provide a much more accurate genetic value for the given sire, especially in the late stage of lactation. It can also be seen that genetic values in the first part of the lactation will be the same with both procedures. This would not be the case if zeros were added in the data file, as genetic values would then be underestimated all over the lactation. Figure 7.2 also shows that the longitudinal model, when the drying off process is ignored, overestimates genetic values at the end of the lactation. The proposed corrected procedure therefore appears to be a good compromise between the two approaches as it keeps genetic values predicted by the longitudinal model in the first part of the lactation and penalizes them in the second part for shorter lactations.

Figure 7.2: Average milk production for the daughters of a given sire: simulated curve, prediction obtained with the random regression model, corrected predictions taking into account the drying off process.



7.4 Discussion

This chapter presents a simple procedure to correct predictions of milk production for lactation curve analysis taking into account the drying off process. It will be especially useful for incomplete lactations in order to have better estimates of individual cumulative production at 305 days. The proposed procedure is very simple and flexible as it can be adapted to any kind of longitudinal models: random regression, character processes, etc. Data used for genetic evaluation do not have to be modified, and calculations for the proposed correction are very simple and fast to compute.

The simulation study showed that the corrected procedure greatly improves the estimation of the cumulative milk production compared to a simple random regression that does not take into account the drying off process. It also avoids predicting negative production values, which is one drawback of classical longitudinal models. The procedure will also have an important impact for genetic value predictions and should help correcting the overestimations obtained with traditional models, especially at the end of the lactation.

More complicated models for the drying off probability could be used. For example, different values for the threshold could be considered for each herd and other covariables could be taken into account.

Appendix: Calculations

Let two variables y_1^* and y_2^* be defined as: $y_1 = y_1^*$ (y_1 is always observed), and

$$y_2 = \{ \begin{array}{ll} y_2^* & \text{if } y_1 > s \\ 0 & \text{otherwise} \end{array}$$

y_1^* and y_2^* are assumed to be correlated and normally distributed such as:

$$\begin{pmatrix} y_1^* \\ y_2^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right)$$

The aim is to calculate $E(y_2)$.

$$E(y_2) = E(y_2|y_1 > s)P(y_1 > s) + E(y_2|y_1 \leq s)P(y_1 \leq s) \tag{7.9}$$

When $(y_1 \leq s)$, $y_2 = 0$, therefore $E(y_2|y_1 \leq s) = 0$.

As

$$f(y_2|y_1 > s) = \frac{1}{P(y_1 > s)} \int_s^{+\infty} f(y_1, y_2)\, dy_1 \tag{7.10}$$

It follows that:

$$\begin{aligned}
E(y_2) &= \int_{-\infty}^{+\infty} y_2\, f(y_2|y_1 > s)\, dy_2 \times P(y_1 > s) \\
&= \frac{1}{P(y_1 > s)} \int_{-\infty}^{+\infty} y_2 [\int_s^{+\infty} f(y_1, y_2)\, dy_1]\, dy_2 \times P(y_1 > s) \\
&= \int_s^{+\infty} [\int_{-\infty}^{+\infty} y_2\, f(y_2|y_1)\, dy_2]\, f(y_1)\, dy_1 \\
&= \int_s^{+\infty} E(y_2|y_1)\, f(y_1)\, dy_1 \\
&= \int_s^{+\infty} [\mu_2 + \frac{\sigma_{21}}{\sigma_1^2}(y_1 - \mu_1)]\, f(y_1)\, dy_1 \\
&= (\mu_2 - \frac{\sigma_{21}}{\sigma_1^2}\mu_1)\, P(y_1 > s) + \frac{\sigma_{21}}{\sigma_1^2} \int_s^{+\infty} y_1\, f(y_1)\, dy_1
\end{aligned}$$

It can be shown that

$$\int_{s}^{+\infty} y_1 \; f(y_1) \; dy_1 = \mu_1 \; \Phi(\frac{\mu_1 - s}{\sigma_1}) + \sigma_1 \; \phi(\frac{\mu_1 - s}{\sigma_1}) \tag{7.11}$$

where $\Phi(.)$ and $\phi(.)$ are the cumulative probability and density functions of a $\mathcal{N}(0, 1)$, respectively.

This leads to:

$$E(y_2) = \mu_2 \; \Phi(\frac{\mu_1 - s}{\sigma_1}) + \frac{\sigma_{21}}{\sigma_1} \; \phi(\frac{\mu_1 - s}{\sigma_1}) \tag{7.12}$$

Chapter 8

General discussion

8.1 Thesis Overview

Modelling of time or age dependent traits is obtaining increasing attention. Random regression models are the most well-known and the most commonly used. In animal breeding, they have already been implemented for the genetic evaluation of dairy cattle for milk production. However, very few studies have been performed to compare them to other approaches. One of the aims of this work was therefore to search the statistical literature for other ways of analyzing longitudinal data. Two other approaches were considered, namely character process models (CP) and structured antedependence models (SAD). They rely on very different concepts from random regression and are not well known yet. However, the comparative study performed here proved that they offer very interesting characteristics and that they can model very different kinds of covariance structures, with few parameters. In most of the examples considered, they performed better than classical random regression models. More attention should therefore be given to these models for the genetic analysis of longitudinal data.

As originally proposed by Pletcher and Geyer (2000), character process models assumed stationarity of the correlation function. Ways of relaxing this quite stringent assumption were suggested in Chapter 2.

As the range of all possible models to be tested is very wide in practice, a non-parametric approach was proposed in Chapter 3 for preliminary exploratory analysis of the covariance structure. It is especially suitable when the number of measurements per subject over time is large, for example for daily record analysis for milk production. In this case, it is not possible to estimate completely unstructured covariance matrices with standard software. The proposed methodology, based on the variogram, is easy to implement and computing time required is small, even for large data sets.

In previous longitudinal studies, the residual variance was often found to be changing over time. A possible way to model it as a continuous function of time was proposed in Chapter 6 to avoid the problem of considering different discrete classes.

The practical aim of this research project was genetic analysis of dairy cattle for milk production. At present, the genetic evaluation is based on prediction for individual cumulative milk production at 305 days. Advantages of using longitudinal models as presented in this thesis are now well known. In particular, such models can easily provide individual cumulative predictions even for incomplete lactations without needing other extrapolation procedures. However, predictions obtained rely on the assumption that cows are never dried off, which is obviously not the case and can lead to overestimation of cumulative milk production. A way to correct these predictions by taking into account the drying off process was suggested in Chapter 7. The methodology is simple and easy to implement. Ways of improving it are possible, but the main objective of this study was to point out and illustrate potential problems in genetic evaluation.

All the analyses presented here were based on the normality assumption of the observations. Implementation of the character process methodology for non-normally distributed traits was considered and appended to the thesis. The estimation procedure is much more complex in this case as no analytical form for the likelihood is available, and methods based on Markov Chain Monte Carlo were used. An application to survival analysis was presented but other traits such as threshold characters could be analysed the same way.

8.2 Model comparison

Through extensive investigation of a variety of simulated covariance structures and empirical data it was found (see Chapter 2) that character process models provide under most conditions a better description of the underlying covariance structure than random regression models. This was especially clear for a correlation declining rapidly to zero as observations became further separated in time. Polynomials do not have asymptotes and the estimated correlation obtained with random regression models went negative instead of decreasing asymptotically to zero. This can be a serious drawback as such an asymptotic correlation pattern is often to be expected in practice.

A further advantage of character process models is their ability to model variance and correlation separately, whereas for random regression the entire covariance is implicitly determined by the shapes of the regression polynomials, and covariance surfaces described by orthogonal polynomials have a fixed relationship between variance and correlation. This was also a major factor contributing to the ability of CP models to give reasonable estimates of the covariance structure with a much smaller number of parameters than random regression models.

There are still however a few limitations of the character process models. The first concerns the stationary assumption of correlation functions. One possible way to relax this assumption was proposed in Chapter 2 based on a time-scale transformation. It seems to be a promising direction and offers reasonable flexibility in practice with only one extra parameter. However, it cannot deal with all kinds of non-stationary pattern as it assumes that correlations on the subdiagonals are monotone. Other more general parametric non-stationary correlation functions should therefore be investigated. The second limitation of process models is their extension to multi-trait analyses. Generalization of random regression models to the multivariate case is straightforward and has already been used (Jamrozik et al., 1997). Bivariate character process models might be implemented by defining a parametric cross-covariance function between the two traits but appropriate forms for this function are yet to be discovered.

Structured antedependence models (SAD), proposed by Zimmerman and Nunez-Anton (1997), seem to offer similar advantages to character processes to capture adequately the covariance structure with few parameters and to deal with asymptotic correlation patterns. They also overcome limitations of CP models concerning the non-stationary specification and the extension to multi-trait analyses, as shown in Chapter 4. They proved in particular to be able to deal with the non-stationary correlation pattern observed in the data analysis for milk production of dairy cattle, which was not well dealt with by the CP models. For the phenotypic analysis of milk production, SAD models performed even better than a quartic random regression with many fewer parameters (see Chapter 5). A first order structured antedependence model (SAD(1)) is in general about equivalent to CP models. Increasing orders of antedependence allow a high degree of flexibility to capture very different kinds of correlation patterns (see Chapter 4). Moreover, multivariate extension of random regression models requires a very large number of parameters. For example, a bivariate analysis with only a quadratic model for both genetic and environmental parts requires 45 parameters. With cubic order polynomials for both parts, the number of parameters jumps to 75! In contrast, increasing the order of structured antedependence model adds only two parameters at each step. As shown in Chapter 4, SAD models offer flexibility also for the cross-covariance structure and perform in general better for a bivariate analysis than random regression models, with fewer parameters. However, in the antedependence formulation presented in Chapter 4, it is required that times of measurement are on a discrete scale and equally spaced. This can be an important limitation in practice for the analysis of function-valued traits. A way to relax this assumption is proposed in the next section.

8.3 Continuous antedependence models

8.3.1 Model

As presented in Chapter 4, a third order antedependence model can be written as:

$$y_t = \theta_1 \, y_{t-1} + \theta_2 \, y_{t-2} + \theta_3 \, y_{t-3} + e_t \tag{8.1}$$

where y_t is an observation at time t and e_t is a random error assumed normally distributed with mean zero and variance σ_t^2 that can change with time as a polynomial function. As mentioned above, this antedependence parametrization however requires times of measurement to be on a discrete scale and equally spaced. This can be an important drawback when traits are expressed continuously over time and can be recorded at different time points for each individual. The proposed continuous extension of the SAD models is based on the following idea: instead of considering s correlation parameters $(\phi_1, ..., \phi_k, ..., \phi_s)$ for each lag time for an antedependence model of order s, a continuous parametric function $\phi(k)$ of the lag time k is fitted to the discrete points $(\phi_1, ..., \phi_s)$. For example, if a first order SAD model is considered, there is only one non zero parameter ϕ for lag time 1, therefore, the parametric function $\phi(k)$ should decrease very rapidly to have $\phi(k) \approx 0$ for $k > 1$. For example, an exponential function of order 4 could be appropriate: $\phi(k) = \exp(-\theta\ k^4)$. For an SAD(3), the function should try to fit points (ϕ_1, ϕ_2, ϕ_3), and the function should therefore decrease less rapidly to zero. An exponential function may fit well in this case: $\phi(k) = \exp(-\theta\ k)$. A very large range of parametric functions can be considered for this continuous extension depending on the dependence of the trait of interest on the previous observations.

8.3.2 Simulation study

In order to check the behaviour of the proposed continuous antedependence models, a data set was simulated with 100 sires, 20 progeny per sire and 10 measures per progeny. Data were simulated for a phenotypic analysis with the covariance matrix estimated from a multivariate analysis (unstructured covariance) of the cow data for milk production considered in Chapters 4 and 5. The correlation pattern was quite complex and non-stationary. Structured antedependence models up to order 3 were considered and compared to random regression models up to the cubic order and to character process models with an exponential correlation, either stationary or non-stationary. Variance functions were assumed quadratic for both CP and SAD models. Estimates for the correlation parameters in the third order antedependence model were:

$$\text{Trait}(t) = 0.51\ \text{Trait}(t-1) + 0.22\ \text{Trait}(t-2) + 0.11\ \text{Trait}(t-3) + \epsilon_t \qquad (8.2)$$

The best continuous antedependence model had an exponential *correlation function*:

$$\theta(k) = \exp(-0.77\ k) \qquad (8.3)$$

where k represents the lag time. In Figure 8.1 correlation parameters obtained in the SAD(3) model are plotted against the lag time and compared to the estimated exponential curve of the continuous antedependence model (CAD).

Table 8.1 shows that the continuous extension has a likelihood even a little higher than SAD(3), although it has only one parameter for the correlation function instead of three. Both SAD(3) and CAD performed better than random regression models up to the cubic order, although they require fewer parameters. Antedependence models in this case also proved to offer better flexibility than character process models to capture the correlation structure. In fact, the simulated non-stationary correlation pattern was quite complex, and was problematic for the non-stationary extension of character process models proposed in Chapter 2, as correlations on the subdiagonals were not monotone increasing or decreasing. It was however very well captured with the continuous antedependence model that has the same number of parameters as the CP model, as shown in Table 8.2, which gives the simulated correlation matrix as well as that estimated with the continuous antedependence model.

Figure 8.1: Correlation parameters for SAD(3) and exponential correlation function for CAD model.

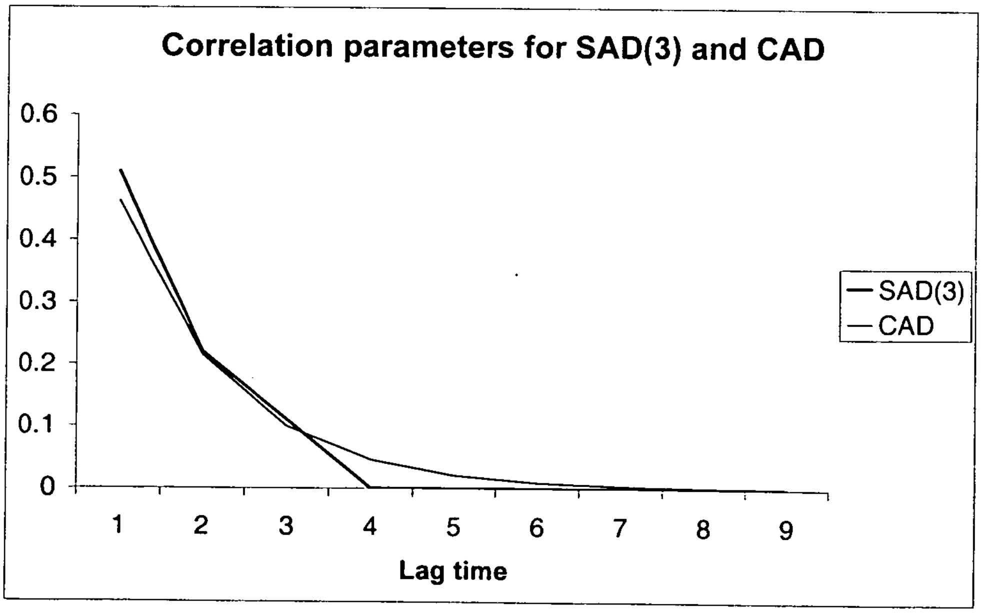


Table 8.1: Log-likelihood for the different models for the simulated non-stationary correlation structure (CAD = continuous antedependence model, CP and CPNS = stationary and non-stationary character process, RR2 and RR3 = quadratic and cubic random regression, NPCov = number of parameters in the covariance structure).

Model	NPCov	Log L
CAD	4	1613
SAD(3)	6	1566
SAD(2)	5	1511
SAD(1)	4	807
CP	4	764
CPNS	5	802
RR2	6	1447
RR3	10	1566

Table 8.2: Simulated non-stationary correlation structure (below diagonal) and estimated correlation with the continuous antedependence model (above diagonal).

1	0.49	0.47	0.45	0.43	0.40	0.37	0.34	0.31	0.27
0.60	1	0.60	0.57	0.54	0.51	0.47	0.43	0.39	0.34
0.53	0.69	1	0.67	0.63	0.59	0.55	0.51	0.46	0.40
0.49	0.64	0.71	1	0.71	0.67	0.62	0.57	0.51	0.45
0.47	0.62	0.69	0.73	1	0.73	0.68	0.63	0.56	0.50
0.44	0.58	0.65	0.71	0.77	1	0.74	0.68	0.61	0.54
0.42	0.56	0.62	0.68	0.74	0.78	1	0.73	0.66	0.58
0.40	0.52	0.59	0.64	0.70	0.74	0.78	1	0.72	0.63
0.36	0.48	0.55	0.59	0.64	0.67	0.71	0.76	1	0.69
0.29	0.38	0.44	0.48	0.52	0.55	0.58	0.62	0.70	1

8.4 Practical considerations

Analyses in this thesis have been performed on relatively small data sets compared to the extremely large size of data for national genetic evaluation. Estimation procedure used for the covariance parameters for all models was an average information algorithm as implemented in the ASREML program (Gilmour et al., 2000). A new function has been added to the program during the course of this work in order to allow the user to define his own covariance matrix. This was extremely useful and allowed all the models presented here to be fitted, which would not have been possible with standard software.

For large data set applications, structured antedependence models have a very interesting property: the inverse of the covariance matrix is sparse, and the Cholesky decomposition presented in Chapter 4 makes it easy to calculate.

At present, in some countries (France in particular), national evaluation incorporates information about heterogeneous variances, using for example structural models as presented by Foulley and Quaas (1995). It should be easy to keep this information in models such as SAD or CAD. In fact, the variance is assumed to change as a function of time, and it would be straightforward to include other covariates. It should therefore be necessary only to add an estimation procedure for correlation parameters to existing programs. Incorporation of heterogeneous variances is far less straightforward for random regression models.

Increasing power of computing facilities should now allow 10 tests for each animal to be dealt with in the national evaluation. However, the question about which measure has to be considered as breeding value still has to be solved. Is the predicted total milk production at 305 days the most appropriate measure? Should it be corrected for the drying off process? Should not information about length of the lactation be taken into account? Breeders would in fact probably prefer a cow that would give a certain amount of milk in 280 days rather than in 305 days.

8.5 Further directions of research

The proposed continuous antedependence models seem to present all the advantages observed for character process models over random regression. They can also deal with complex non-stationary correlation patterns as shown in the simulation study presented above while requiring very few parameters for the covariance structure. Like SAD models, they allow a straightforward extension to multi-trait analyses. There are still however some remaining difficulties. Firstly, analytical forms for variance and correlation functions for antedependence models are in general quite complex, as shown in the following appendix for a continuous antedependence model with an exponential correlation function. It is therefore difficult to have a clear idea of all the possible range of covariance structures that can be dealt with by these models. If an analytical study is not possible to answer this question, a simulation study should be performed, considering a wide range of covariance structures. With SAD or CAD models, it is also difficult to see clearly how the autoregressive coefficients and innovation variances are related to the estimated variance and correlation functions. Their parameter interpretations are not straightforward, and their biological meaning will have to be further investigated. In contrast to CP models, variance and correlation modelling are not completely separated. It would therefore be interesting to see how much flexibility can still be achieved for the two components and which are the combinations that cannot be fitted.

Secondly, with antedependence models, as with character processes, no simple analytical form for individual genetic curves has been found yet, and at present a genetic value has to be predicted at each time for each animal. More research is needed in order to make clearer the relationship between covariance matrices as modelled with CP or SAD and individual curves as considered in random regression models.

Results of Chapter 5 also suggest that the genetic part can be well fitted with simple models as the genetic correlation remains quite high over time. It could therefore be possible to consider a simple random regression model (linear or quadratic) for the genetic part while keeping a structured or continuous antedependence model for the environmental part. In that case, only two or three genetic parameters have to be estimated for each animal regardless of the number of observations over time, while the number of parameters for the complex environmental covariance structure would be kept low thanks to the antedependence structure.

Calculations have been presented in Chapter 5 concerning the loss of efficiency in response to selection under each model. However, more analyses should be performed to study the actual genetic improvement reached when using the different approaches, as well as changes that can be obtained in the lactation shape: improvement in persistency, or higher productivity and shorter lactations, etc.

As mentioned above, generalization of structured antedependence models to take into account heterogeneous variances as proposed by Foulley and Quaas (1995) is quite straightforward. Further avenues to improve their flexibility could also be obtained by allowing correlation parameters to depend on characteristics of the population as proposed by Pourahmadi (1999), or to change as a function of time as proposed by Nunez-Anton and Zimmerman (2000). These models offer a very promising direction for the genetic analysis of longitudinal data and will undoubtedly attract increasing attention in the near future.

Appendix: Analytical form for covariance functions with CAD models

Assuming an exponential function for the correlation in the continuous antedependence model: $\theta(t_j, t_k) = \theta^{t_j - t_k}$, observations y_{t_j} can be written as:

$$\begin{aligned}
y_{t_1} &= e_{t_1} \\
y_{t_2} &= \theta^{t_2-t_1}\ y_{t_1} + e_{t_2} \\
y_{t_3} &= \theta^{t_3-t_2}\ y_{t_2} + \theta^{t_3-t_1}\ y_{t_1} + e_{t_3} \\
y_{t_4} &= \theta^{t_4-t_3}\ y_{t_3} + \theta^{t_4-t_2}\ y_{t_2} + \theta^{t_4-t_1}\ y_{t_1} + e_{t_4} \\
& \vdots \\
y_{t_J} &= \theta^{t_J-t_{J-1}}\ y_{t_{J-1}} + \theta^{t_J-t_{J-2}}\ y_{t_{J-2}} + \ldots + \theta^{t_J-t_1}\ y_{t_1} + e_{t_J}
\end{aligned}$$

It is possible to express y_{t_j} for $j = (2, ..., J)$ as a function of the e_{t_j}'s:

$$y_{t_j} = 2^{j-2}\theta^{t_j-t_1}\ e_{t_1} + 2^{j-3}\theta^{t_j-t_2}\ e_{t_2} + 2^{j-4}\theta^{t_j-t_3}\ e_{t_3} + \ldots + \theta^{t_j-t_{j-1}}\ e_{t_{j-1}} + e_{t_j} \quad (8.4)$$

As $\text{Cov}(e_{t_j}, e_{t_k}) = 0$ for all $j \neq k$, it follows that covariance function for the observations is:

$$\text{Cov}(y_{t_j}, y_{t_k}) = 2^{j+k-4}\theta^{t_j+t_k-2t_1}\ v_{t_1} + 2^{j+k-6}\theta^{t_j+t_k-2t_2}\ v_{t_2} + \ldots + 2^{j-k-1}\theta^{t_j-t_k}\ v_{t_k} \quad (8.5)$$

for $k < j$ and $j > 2$.

$$\text{Var}(y_{t_j}) = 2^{2(j-2)}\theta^{2(t_j-t_1)}\ v_{t_1} + \ldots + \theta^{t_j-t_{j-1}}\ v_{t_{j-1}} + v_{t_j} \quad (8.6)$$

where $v_{t_j} = \text{Var}(e_{t_j})$.

Appendix A

Extension to non-normally distributed traits

Generalized character process models: Estimating the genetic basis of traits that cannot be observed and that change with age or environmental conditions

Scott D. Pletcher[†], Florence Jaffrézic[*]

[†] Department of Biology, Galton Laboratory, University College,
London, NW1 2HE.

[*] Institute of Animal, Cell and Population Biology, Edinburgh University,
Edinburgh, Scotland.

In press for *Biometrics*

Abstract

The genetic analysis of characters that change as a function of some independent and continuous variable has received increasing attention in the biological and statistical literature. Previous work in this area has focussed on the analysis of normally distributed characters that are directly observed. We propose a framework for the development and specification of models for a quantitative genetic analysis of function-valued characters that are not directly observed, such as genetic variation in age-specific mortality rates or complex threshold characters. We employ a hybrid Markov Chain Monte Carlo algorithm involving a Monte Carlo EM algorithm coupled with a Markov Chain approximation to the likelihood, which is quite robust and provides accurate estimates of the parameters in our models. The methods are investigated using simulated data and applied to a large data set measuring mortality rates in the fruit fly, *Drosophila melanogaster.*

A.1 Introduction

Function-valued quantitative genetics (Pletcher and Geyer, 1999) or the genetics of infinite-dimensional characters (Kirkpatrick and Heckman, 1989) is concerned with estimating the genetic contribution to observed variation in characters that change as a function of age or some other continuous variable. Taking advantage of observations from related individuals, observed variation in the function-valued character is decomposed into genetic and non-genetic contributions by estimating continuous, bivariate covariance functions (Kirkpatrick and Heckman, 1989). These models have been shown to be effective when applied to a variety of characters from age-dependent patterns of reproductive output in fruit flies to growth and lactation curves in cattle (Jaffrezic and Pletcher, 2000 ; Pletcher and Geyer, 1999 ; Kirkpatrick et al., 1994).

In this chapter, we present theory and implementation of the genetic analysis of survival and other threshold characters thought to be influenced by a continuously distributed underlying trait, commonly termed "frailty" or "liability," which is unobserved and changes as a function of some continuous variable, such as age. An important example is inference concerning age-specific mortality rates, which are genetically influenced but unobserved (Shaw et al., 1999). Other applications include estimating the genetic component of variation in the appearance of an environmentally induced phenotype across different environmental conditions (Roff and Bradford, 2000) or in the expression of an ordered categorical character across age and space (Wright, 1934). As a foundation for our development of a generalized function-valued quantitative genetics, we have chosen the character process model (Pletcher and Geyer, 1999). It has several desirable properties; most important for us is its improved efficiency—this model fits many observed covariance structures better and with fewer parameters than other popular models such as random regression and other repeated measures type analyses (see Chapter 2 ; Jaffrezic and Pletcher, 2000).

A.2 Generalized Process Models

We are interested in inferring the genetic basis of some character Y, which is not observed, given a series of measurements on an observed trait, which is denoted by X. We assume that some reasonable model for the relationship between X and Y is available and that all genetic and shared environmental effects are modeled with respect to the Y value. This is in keeping with the standard interpretation of threshold characters (Wright, 1934) and of correlated frailty (Yashin et al., 1999).

When considering function-valued traits, it is assumed that the *trajectory* (over some continuous variable) of the character is random and influenced by one or more unobservable factors. For the additive model, we assume the unobserved character can be decomposed as

$$y(t) = \mu(t) + g(t) + e(t) + \epsilon, \tag{A.1}$$

where t is some continuous measure, $g(t)$ and $e(t)$ are Gaussian random functions, which are independent of one another and have an expected value of zero at each age (Kirkpatrick and Heckman, 1989 ; Pletcher and Geyer, 1999). These represent genetic and environmental deviations at each value of t. The mean function is $\mu(t)$, and ϵ is the residual variation.

In practice, a finite number of observations (each associated with a particular value of the continuous variable) are made on a number of individuals i of varying relatedness. Thus, let $y_i(t_j)$, etc. denote the effects for individual i at point t_j and y be a vector containing all data on all individuals in the order $y_1(t_1)$, $y_1(t_2)$, ..., $y_2(t_1)$, ..., then the distribution of y, $f_{\theta_1}(y)$, is multivariate normal with density

$$f_{\theta_1}(y) = C_1|V^{-1}|^{1/2}\exp\{-1/2 \ \ (y-\mu)^T V^{-1}(y-\mu)\}. \tag{A.2}$$

where

$$V = A \otimes Z + I \otimes R \tag{A.3}$$

where $\otimes$ denotes the Kroneker product, A is a matrix containing coefficients of relatedness and I is the identity matrix. The remaining matrices, Z and R, are discrete representations of the covariance functions for the genetic and environmental processes given in (A.1). If $G(s,t) = Cov\{g(s), g(t)\}$ and $E(s,t) = Cov\{e(s), e(t)\}$ then $Z[i,j] = G(t_i, t_j)$, and $R[i,j] = E(t_i, t_j)$ (Pletcher and Geyer, 1999). The vector μ describes the mean function non-parametrically by specifying a unique parameter for each value of t in the data set.

Parametric forms for the covariance functions are based on the character process model where, taking $G(s,t)$ as an example, the functions are written as

$$G(s,t) = v_G(s)v_G(t)\rho_G(|s-t|) \tag{A.4}$$

where $v_G(t)^2$ describes how the genetic variance changes with age and $\rho_G(|s-t|)$ describes the genetic correlation between two ages. There are no restrictions on the form of $v_G(\cdot)$, and it is often modeled using simple polynomials (linear, quadratic, etc.). If the correlation between two ages is a function only of the time distance ($|s-t|$) between them (correlation stationarity) then numerous choices for $\rho(\cdot)$ are available, all of which satisfy several theoretical requirements (for a list see Pletcher and Geyer, 1999). Strict correlation stationarity can be relaxed by implementing a non-linear transformation upon the time axis, $f(t)$ (Nunez-Anton, 1998 ; Jaffrezic and Pletcher, 2000). The correlation function is then defined as $\rho(s,t) = \rho(|f(s) - f(t)|)$, and the functions suggested by Pletcher and Geyer (1999) remain valid.

The elements of the observed vector x are conditionally independent given y, and

$$f_{\theta_2}(x|y) = \prod_{i=1}^{N} f_{\theta_2}(x_i|y_1,\ldots,y_N) = \prod_{i=1}^{N} f_{\theta_2}(x_i|y_i). \tag{A.5}$$

The likelihood associated with the observed data is

$$f_\theta(x) = \int \prod_{i=1}^{N} f_{\theta_2}(x_i|y_i) f_{\theta_1}(y)dy \tag{A.6}$$

where θ_2 is a vector of parameters describing the relationship between X and Y, and θ_1 contains parameters describing the distribution of Y, which includes parameters of the variance functions, mean function, and potential fixed effects (Meyer and Hill, 1997 ; Pletcher and Geyer, 1999).

A.3 Likelihood maximization

The likehood was maximized using a hybrid algorithm composed of Markov Chain Monte Carlo EM (MCEM) (McCulloch, 1997) and Markov Chain Monte Carlo integration/maximization (MCMLE) (Shaw et al., 1999 ; Geyer, 1995). The computational cost of the MCEM algorithm is much lower than that of the MCMLE.

However, parameter estimates obtained from MCEM show a good deal of variation (McCulloch, 1997 ; S. Pletcher, unpublished results) and confidence intervals are not easily obtained. The MCMLE provides accurate parameter estimates and confidence intervals, but it is computationally expensive and requires a reference point in the parameter space of $\theta = \{\theta_1, \theta_2\}$ that is close to the MLE (Shaw et al., 1999). We found the following three step procedure combines the strengths of both methods. First, the MCEM is used to determine the reference point, call it θ_0, for the MCMLE. Second, a single chain of random deviations from $f_{\theta_0}(y|x)$ are obtained using a Metropolis algorithm (Shaw et al., 1999). These deviates are used to approximate the likelihood function (A.6) through a Monte Carlo evaluation of the integral (Geyer, 1995). Third, the approximation is maximized, and estimates and standard errors of the parameters are obtained. Details of the computational algorithms and relevant computer code are available from the first author (or see http://www.ucl.ac.uk/biology/goldstein/scott_index.html).

A.4 Example

For the following examples, the character we are interested in, $y(t)$, is the age-specific mortality function for a specific cohort of genetically identical individuals. The observed character $x(t)$ is the number of individuals dying in that cohort at age t. Shaw and colleagues assumed parametric forms for the unobserved mortality curves using Gompertz and Logistic functions (Shaw et al., 1999), which is analogous to a random regression on the age-dependent trajectories (Jaffrezic and Pletcher, 2000). Because the character process models have been shown to perform better than random regression models for observed function-valued characters (see Chapter 2 ; Jaffrezic and Pletcher, 2000), we extend the Shaw model to the generalized character process theory.

Measurements are taken at a finite number of ages, and therefore we observe a "census vector" $\{x_{ij}\}$, which contains the number of individuals alive in cohort

i at census number j. Similarly, we assume each cohort has a log-mortality rate y_{ij} at census number j, and t_{ij} is the age at which census number j was taken. We estimate a separate mean mortality rate for each t_{ij} (call these parameters μ_j) as well as genetic and environmental covariance functions.

Assuming a piece-wise exponential hazard function, the probability of an individual alive at the start of census $j-1$ surviving the interval $[t_{i(j-1)}, t_{ij})$ is

$$p_{t_{ij}} = \exp\{-H(\Delta t_{ij})e^{y_{ij}}\} \tag{A.7}$$

where $\Delta t_{ij} = t_{ij} - t_{i(j-1)}$ and

$$H(\Delta t) = \frac{e^{\mu_{j-1}}}{\mu_j - \mu_{j-1}}(e^{(\mu_j - \mu_{j-1})\Delta t} - 1)\Delta t. \tag{A.8}$$

Equation A.8 is valid for uneven census intervals and intervals over which mortality rates change substantially.

The number of deaths in the interval is binomially distributed with frequency $p_{t_{ij}}$ and number of trials $x_{i(j-1)}$. Writing $x_i = \{x_{i1}, x_{i2}, \ldots\}$ and y_i similarly, the conditional probability of observing a specific census vector for a specific cohort is

$$f(x_i|y_i) = C_1 \prod_{j=1}^{J} p_{t_{ij}}^{x_{ij}}(1 - p_{t_{ij}})^{x_{i(j-1)} - x_{ij}}. \tag{A.9}$$

where J is the number of census times (Shaw et al., 1999). This distribution is substituted into A.5 and combined with A.2 to yield the likelihood A.6 for use in the Metropolis algorithm and in likelihood maximization.

A.4.1 Simulated Data

Simulated ages at death were generated for 600 distinct cohorts (20 replicate cohorts from each of 30 genetically distinct lines) of 500 individuals each. The data were simulated using a covariance function with a constant variance (i.e., $v_G^2(t) = 0.2$ in equation A.4) and standard normal correlation function (i.e., $\rho(s,t) = e^{-\theta_C(s-t)^2}$) for both the genetic and environmental parts ($\theta_C = 0.1$ and

$\theta_C = 0.4$ for the genetic and environmental correlations, respectively). Similar results were obtained using other covariance functions and experimental designs.

The small number of lines in the simulated data leaves open the possibility that the realized genetic variance and covariance among lines may deviate significantly from the target values. To compensate for this, we estimated covariance functions for the realized y-values themselves (i.e., the unobserved age-dependent mortality rates), which are saved during the course of the simulation, and we used these estimates as metrics for determining the accuracy of the covariance functions estimated from the x-values (the observed ages at death).

The MCEM routines provided an excellent θ_0 for the MCMLE routines. The sample paths for the four covariance parameters (two for both the genetic and environmental covariance functions) show a rapid convergence to the neighborhood of the simulated value, with the genetic variance converging less quickly than the others (data not presented). θ_0 for the MCMLE was obtained by averaging the values from the last 200 (of a total of 500) iterations.

The MCMLE routines were then used to obtain estimates and confidence intervals for the genetic and environmental covariance functions and for the mean mortality trajectory. The approximation of the likelihood was based on a MCMC sample size of 1000 random deviates sampled from the chain every 1000 steps. The genetic covariance function obtained from this analysis is in complete agreement with that obtained when standard methods are used on the unobserved y-values themselves (Table A1.1). The environmental covariance functions, which are estimated accurately with smaller sample sizes, are essentially identical. As expected, the asymptotic standard errors on the parameter estimates are much larger (up to five times larger) when obtained from the observed data (Table A1.1). It may be that increasing the MCMC sample size would reduce this difference. It is more likely, however, that there is simply more uncertainty in the estimates.

Table A1.1: Estimated genetic and environmental covariance functions for simulated data. Covariance functions are composed of a constant variance across ages ($v(t)^2 = \theta$) and a normal correlation function $\rho(s,t) = e^{-\theta_C(s-t)^2}$. Asymptotic standard errors of the estimates are in parentheses. y-values indicate results from a function-valued analysis directly on the unobserved frailty. x-values represent the results of the Markov Chain Monte Carlo models on the observed ages at death. Parameter estimates from the two methods are nearly identical, but standard errors are higher for the MCMC analysis.

	V_G		V_E	
Data	θ	θ_C	θ	θ_C
y-values (Unobserved)	0.15 (0.03)	0.095 (0.041)	0.20 (0.006)	0.402 (0.006)
x-values (Observed)	0.17 (0.16)	0.083 (0.140)	0.20 (0.030)	0.403 (0.014)

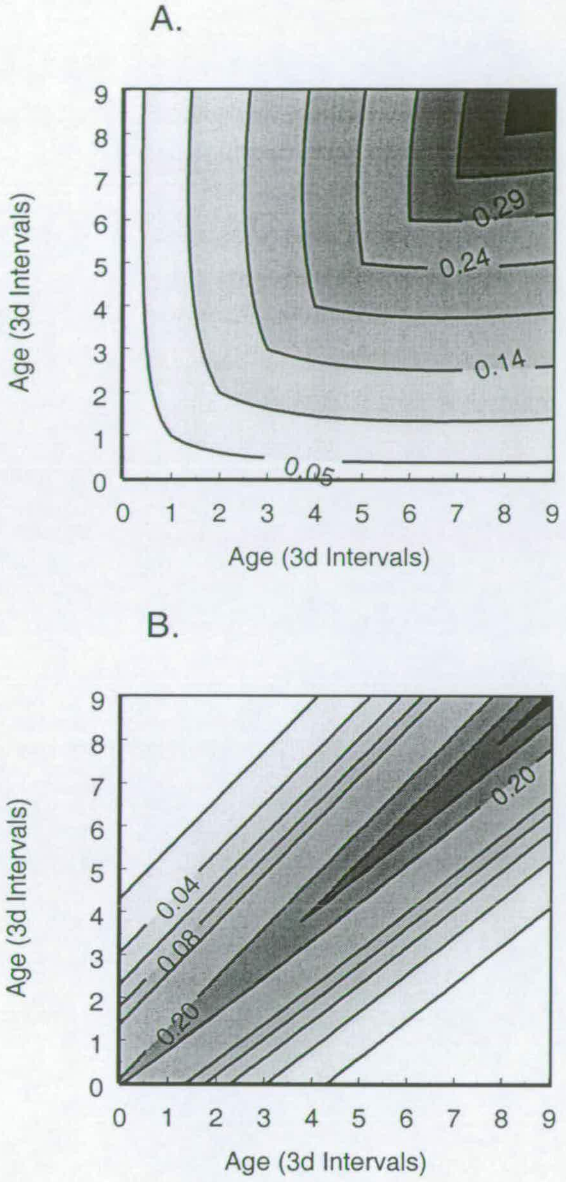
A.4.2 Mortality in Drosophila melanogaster

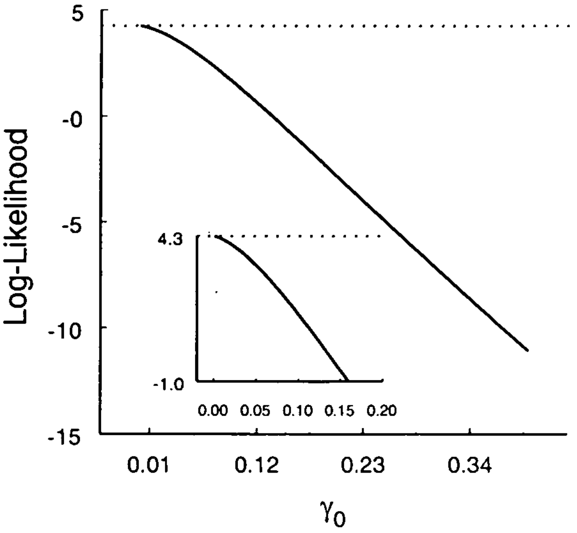
The *Drosophila* data are taken from a large mortality experiment composed of 29 genetically distinct lines of flies. The lines were created using an experimental mutagenesis technique whereby single mutational events were initiated in a genetically homogeneous background (S. D. Pletcher, unpublished data). Experimental populations differed among themselves genetically via one mutational event. Genetic variation in mortality rates as a function of age and genetic covariation in mortality between ages provide important insights into the age-specific properties of these mutations (Pletcher et al., 1998). Ages at death were recorded for four replicate cohorts (each of approximately 300 males) from each line, and pooled into 3 day intervals for analysis.

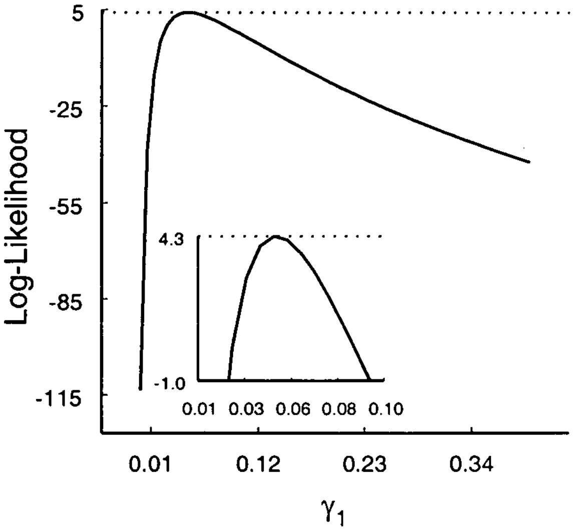
Exploratory analyses, including an examination of the phenotypic covariance structure and an estimate of the genetic variogram cloud (see Chapter 3), suggested the use of genetic and environmental covariance functions that are composed of a linear variance function (i.e., $v^2(t) = \gamma_0 + \gamma_1 t$ in equation A.4) and a normal correlation function (i.e., $\rho(s,t) = e^{-\gamma_C(s-t)^2}$). The θ_0 value for the MCMLE procedures was obtained by averaging 500 consecutive iterations of the MCEM algorithm after it was determined to have converged to a stable region for each parameter. The MCMLE routines were then executed with a MCMC sample size of 2000, and the chain was sampled every 1000 steps.

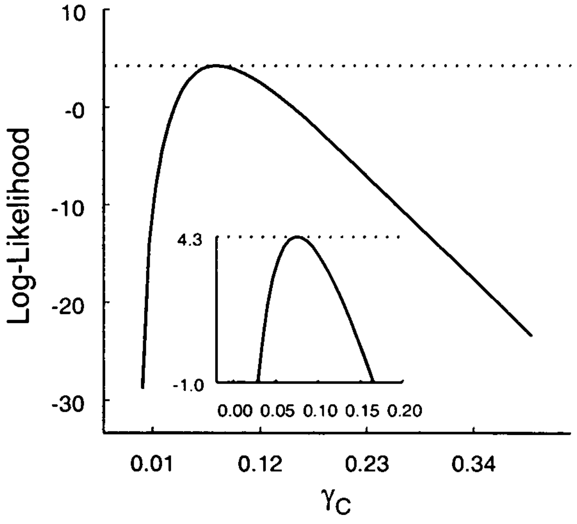
Figure A1.1: Contour plots of the genetic and environmental covariance functions estimated from a large mortality experiment using the fruit fly, *Drosophila melanogaster.* Functions represent age-dependent covariance in log-mortality rates. Both genetic and environmental covariance functions are described by a linear variance function $v_G^2(t) = \gamma_0 + \gamma_1 t$ and normal correlation function $\rho(s,t) = e^{-\gamma_C(s-t)^2}$. A: Estimated genetic covariance function. $\hat{\gamma_0} < 0.0001$, $\hat{\gamma_1} = 0.047$, $\hat{\gamma_C} = 0.075$. γ_0 was estimated at its lower boundary (≈ 0). B: Estimated environmental covariance function. $\hat{\gamma_0} = 0.21$, $\hat{\gamma_1} = 0.024$, $\hat{\gamma_C} = 0.59$.

Figure A1.2: Likelihood profiles for the parameters of the genetic covariance function estimated from a large mortality experiment in *Drosophila.* The estimated genetic covariance function is described by a linear variance function $v_G^2(t) = \gamma_0 + \gamma_1 t$ and normal correlation function $\rho(s,t) = e^{-\gamma_C(s-t)^2}$. Estimated values are: $\hat{\gamma_0} < 0.0001$, $\hat{\gamma_1} = 0.047$, and $\hat{\gamma_C} = 0.075$. γ_0 was estimated at its lower boundary (≈ 0). Insets focus on a narrow range of parameter values and provide guidance for the construction of 99% confidence intervals on the estimates.


A.
Age (3d Intervals)
0
1
2
3
4
5
6
7
8
9
0.05
0.14
0.24
0.29
Age (3d Intervals)
0
1
2
3
4
5
6
7
8
9
B.
Age (3d Intervals)
0
1
2
3
4
5
6
7
8
9
0.04
0.08
0.20
0.20
Age (3d Intervals)
0
1
2
3
4
5
6
7
8
9


5
-0
-5
-10
-15
Log-Likelihood
4.3
-1.0
0.00 0.05 0.10 0.15 0.20
0.01
0.12
0.23
0.34
γ_0


5
-25
-55
-85
-115
Log-Likelihood
4.3
-1.0
0.01 0.03 0.06 0.08 0.10
0.01
0.12
0.23
0.34
γ_1


10
-0
-10
-20
-30
Log-Likelihood
4.3
-1.0
0.00 0.05 0.10 0.15 0.20
0.01
0.12
0.23
0.34
γ_C

We found that both genetic and environmental variance for age-specific mortality increased with age. Early in life environmental variance was very high, $\gamma_0 = 0.21$ and $\gamma_0 < 0.0001$ for the environmental and genetic variances, respectively, but the rate of increase in genetic variance with age was faster (Figure A1.1). The correlation parameter was much higher in the environmental correlation function than it was in the genetic function (0.59 vs. 0.075) implying that environmental covariance decreases much more rapidly as ages become more and more separated in time. This suggests a rather high degree of pleiotropy (single genes affecting mortality at more than one age) and a relatively transient influence of the environment. The degree of uncertainly in the parameter estimates of the genetic function is illustrated by profile likelihoods (Figure A1.2).

A.5 Discussion

We present a flexible approach for examining the genetic basis of function-valued characters that are unobserved but that influence the expression of an observed character through some arbitrary, hypothesized form. The complexity of the models necessitated the use of stochastic methods for model specification, and we rely heavily on Markov Chain Monte Carlo methods, which can be troublesome and difficult to implement. To alleviate some of the difficulties we implemented a composite algorithm consisting of a Markov Chain EM algorithm (MCEM) followed by a Markov Chain approximation to the actual likelihood (MCMLE). This combination was found to work well for generalized linear mixed models (McCulloch, 1997), and many of the properties of convergence discussed in McCulloch (1997) apply to the models we develop here. The MCEM algorithm robustly provided excellent reference values (i.e., θ_0) from a wide range of starting points, which were then used in the MCMLE to estimate parameters and to obtain likelihood statistics and confidence intervals. Results obtained through the analysis of simulated data and of mortality rates in the fruit fly, *Drosophila melanogaster*, show that variation accumulated through heterogeneity of starting values and through the stochastic nature of the Markov chain algorithms is surprisingly small (essentially inconsequential) in comparison to the support of the parameter estimates provided by the data (data not shown).

Although the algorithms are successful in recovering the underlying genetic structure in simulated data sets and in capturing the variation in *Drosophila* mortality rates, some limitations are apparent. Despite the large number of individuals in our data sets, the asymptotic standard errors (and profile likelihood functions) of the estimates in our analyses are considerable. This suggests that large sample sizes may be required for inference regarding the genetic basis of unobserved characters. In addition, our choice of covariance model was based on exploratory algorithms that will not apply in all situations. The development of model selection criteria similar to those used for observed function-valued traits is an important issue and work is currently underway in this area.

Our examples have focussed exclusively on age-specific mortality rates. However, precisely the same theory applies to any non-normally distributed phenotype that is thought to be determined by an unobserved, normally distributed character (Wright, 1934). An example may be the expression of a threshold character, such as the occurrence of a disease, over space or time. The distribution of the observed trait given the unobserved liability $f_{x|y}$ is the only aspect of the theory and computer code that requires change. Furthermore, although we prefer the character process model for describing the covariance structure of the unobserved character, random regression or orthogonal polynomial models could be implemented with small modifications to the procedure described above.

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on automatic control* **19** : 716-723.

Brotherstone, S., White, I.M.S. and Meyer K. 1999. Genetic modelling of weekly milk yield using orthogonal polynomials and parametric curves. *Animal Science* **70**: 407-415.

Cameron, N.D. 1997. *Selection Indices and Prediction of Genetic Merit in Animal Breeding.* CAB International, London.

Davidian, M. and Giltinan, D.M. 1995. *Nonlinear Models for Repeated Measurement Data.* Chapman and Hall, London.

Dempster, A., Laird, N. and Rubin, R. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**:1-20.

Diggle, P.J. and Verbyla, A.P. 1998. Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* **54**: 401-415.

Diggle, P.J. and Kenward, M.G. 1994. Informative drop-out in longitudinal data analysis. *Applied Statistics* **43**: 49-93.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. 1994. *Analysis of Longitudinal Data.* Oxford Science Publications, Clarendon Press, Oxford.

Foulley, J.L., Quaas, R.L., and Thaon d'Arnoldi, C. 1998. A link function approach to heterogeneous variance components. *Genetic Selection Evolution* **30**:27-43.

Foulley, J.L. and Quaas, R.L. 1995. Heterogeneous variances in gaussian linear mixed models. *Genetic Selection Evolution* **27**: 211-228.

Foulley, J.L., Gianola D., San Cristobal, M. and Im, S. 1990. A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models. *Journal of Dairy Science* **73**:1612-1624.

Gabriel, K.R. 1962. Ante-dependence analysis of an ordered set of variables. *Annals of Mathematical Statistics* **33**: 201-212.

Geyer, C.J. 1995. Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice*, pages 241-258, Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., editors. Chapman and Hall, London.

Gilmour, A.R., Thompson, R., Cullis, B.R. and Welham, S.J. 2000. *ASREML Manual.* New South Wales Department of Agriculture, Orange, 2800, Australia.

Gilmour, A.R., Thompson, R., and Cullis, B.R. 1995. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics.* **51**:1440-1450.

Jaffrezic, F. and Pletcher, S.D. 2000. Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. *Genetics* **156**: 913-922.

Jaffrezic, F., White, I.M.S., Thompson, R. and Hill, W.G. 2000. A link function approach to model heterogeneity of residual variances over time in lactation curve analyses. *Journal of Dairy Science* **83**:1089-1093.

Jamrozik, J. and Schaeffer, L.R. 1997. Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. *Journal of Dairy Science* **80**:762-770.

Jamrozik, J., Schaeffer, L.R and Dekkers, J.C.M. 1997. Genetic evaluation of dairy cattle using test day yields and random regression model. *Journal of Dairy Science* **80**: 1217-1226.

Jamrozik, J., Schaeffer, L.R., Liu, Z. and Jansen, G. 1997. Multiple trait random regression test day model for production traits. *Proceedings of 1997 Interbull Meeting* **16**: 43-47.

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

Kirkpatrick, M. and Lofsvold, D. 1992. Measuring selection and constraint in the evolution of growth. *Evolution* **46**: 954-971.

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

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

Laird, N.M. 1988. Missing data in longitudinal studies. *Statistics in Medicine* **7**: 305-315.

Laird, N.M. and Ware, J.H. 1982. Random effects models for longitudinal data. *Biometrics* **38**:963-974.

Lee, Y. and Nelder, J.A. 1999. Extended REML using GLM technology: a new formulation. *Technical Report for the Department of Mathematics of the Imperial College*, London.

Lindstrom, M.J. and Bates, D.M. 1990. Non-linear mixed effects models for repeated measures data. *Biometrics* **46**: 673-687.

Little, R.J.A. and Rubin, D.B. 1987. *Statistical Analysis with Missing Data.* New York: Wiley.

Lynch, M. and Walsh, B. 1998. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

McCullagh, P. and Nelder, J.A. 1989. *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

McCulloch, C.E. 1997. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Society* **92**: 162-170.

Meyer, K. 1998. Estimating covariance functions for longitudinal data using a random regression model. *Genetic Selection Evolution* **30**: 221-240.

Meyer, K. and Hill, W.G. 1997. Estimation of genetic and phenotypic covariance functions for longitudinal or "repeated" records by Restricted Maximum Likelihood. *Livestock Production Science* **47**: 185-200.

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

Nunez-Anton, V. and Zimmerman, D.L. 2000. Modelling non-stationary longitudinal data. *Biometrics* **56**: 699-705.

Nunez-Anton, V. 1998. Longitudinal data analysis: non-stationary error structures and antedependent models. *Applied Stochastic Models Data Analysis* **13**: 279-287.

Pletcher, S.D. and Geyer, C.J. 1999. The genetic analysis of age-dependent traits: modelling a character process. *Genetics* **153**: 825-833.

Pletcher, S.D., Houle, D. and Curtsinger, J.W. 1998. Age-specific properties of spontaneous mutations affecting mortality in *Drosophila Melanogaster. Genetics* **148**: 287-303.

Pourahmadi, M. 1999. Joint mean-covariance models with applications to longitudinal data: unconstrained parametrisation. *Biometrika* **86**: 677-690.

Rekaya, R., Carabano, M.J. and Toro, M.A. 1999. Use of test day yields for the genetic evaluation of production traits in Holstein-Friesian cattle. *Livestock Production Science* **57**:203-217.

Rekaya, R., Carabano, M.J. and Toro, M.A. 1998. Assessment of heterogeneity of residual variances using changepoint techniques. *49th Annual Meeting of the European Association of Animal Production.* Warsaw, Poland.

Robertson, A. 1962. Weighting in the estimation of variance components in the unbalanced single classification. *Biometrics* **18**: 413-417.

Roff, D.A. and Bradford, M.J. 2000. A quantitative genetic analysis of phenotypic plasticity of diapause induction in the cricket allonemobius socius. *Heredity* **84**: 193-200.

Sales, J. and Hill, W.G. 1976. Effect of sampling errors on efficiency of selection indices: use of information from relatives for single trait improvement. *Animal Production* **22**: 1-17.

Schnyder, U., Hofer, A., Labroue, F. and Kunzi, N. 1999. Genetic parameters of a random regression model for daily feed intake of performance tested French Landrace and Large White growing pigs. *50th Annual Meeting of the European Association of Animal Production.* Zurich, Switzerland.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics*

6: 461-464.

Shaw, F., Promislow, D.E.L., Tatar, M., Hughes, K. and Geyer, C.J. 1999. Towards reconciling inferences concerning genetic variation in senescence. *Genetics* **152**: 553-566.

Stram, D.O. and Lee, J.W. 1994. Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**: 1171-1177.

Verbeke, G. and Molenberghs, G. 1998. *Linear Mixed Models in Practice.* Springer.

Verbyla, A.P. 1993. Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society* **55**:493-508.

Vonesh, E.F., Chinchilli, V.M. and Pu, K. 1996. Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics* **52**: 572-587.

White, I.M.S., Thompson, R. and Brotherstone, S. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. *Journal of Dairy Science* **82**: 632-638.

Wilmink, J.B.M. 1987. Adjustement of test day milk, fat and protein yield for age, season and stage of lactation. *Livestock Production Science* **16**: 335-348.

Wright, S. 1934. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics* **19**: 506-536.

Wu, M.C. and Carroll, R.J. 1988. Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics* **44**: 175-188.

Yashin, A.A., Iachine, I.A. and Harris, J.R. 1999. Half of variation in susceptibility to mortality is genetic: Findings from swedish twin survival data. *Behavior Genetics* **29**: 11-19.

Zimmerman, D.L. and Nunez-Anton, V. 1997. Structured antedependence models for longitudinal data. In *Modelling Longitudinal and Spatially Correlated Data. Methods, Applications, and Future Directions* (Gregoire, T. G., D. R. Brillinger, P. J. Diggle, E. Russel-Cohen, W. G. Warren, and R. Wolfinger, Eds.) Lecture Notes in Statistics No. 122, New York: Springer-Verlag.