Genet. Res., Camb. (2003), **82**, *pp.* 55–65. *With 3 figures.*
DOI: 10.1017/S0016672303006281 Printed in the United Kingdom

Structured antedependence models for genetic analysis of repeated measures on multiple quantitative traits

FLORENCE JAFFRÉZIC[1,2]*, ROBIN THOMPSON[3,4] AND WILLIAM G. HILL[2]
[1] *INRA Quantitative and Applied Genetics, 78352 Jouy-en-Josas Cedex, France*
[2] *Institute of Cell Animal and Population Biology, University of Edinburgh, West Mains Rd, Edinburgh, EH9 3JT, UK*
[3] *Rothamsted Research, Harpenden, Herts, AL5 2JQ, UK*
[4] *Roslin Institute (Edinburgh), Roslin, Midlothian, EH25 9PS, UK*

(*Received 1 August 2002 and in revised form 20 December 2002 and 6 March 2003*)

Summary

Simultaneous analysis of correlated traits that change with time is an important issue in genetic analyses. Several methodologies have already been proposed for the genetic analysis of longitudinal data on single traits, in particular random regression and character process models. Although the latter proved, in most cases, to compare favourably to alternative approaches for analysis of single function-valued traits, they do not allow a straightforward extension to the multivariate case. In this paper, another methodology (structured antedependence models) is proposed, and methods are derived for the genetic analysis of two or more correlated function-valued traits. Multivariate analyses are presented of fertility and mortality in *Drosophila* and of milk, fat and protein yields in dairy cattle. These models offer a substantial flexibility for the correlation structure, even in the case of complex non-stationary patterns, and perform better than multivariate random regression models, with fewer parameters.

1. Introduction

Genetic analysis of traits that change with time (or any other independent and continuous variable) is attracting increasing attention. Many studies have already been done for the analysis of longitudinal data on single traits including, for example, growth curves in beef cattle (Meyer, 2001), age-specific fitness components such as survival or reproductive output (Pletcher *et al*., 1998), and lactation curves in dairy cattle (Jaffrézic *et al*., 2002; Meuwissen & Pool, 2001).

However, the simultaneous analysis of two or more correlated function-valued traits can also be important in practice. It would be useful, for example, to study the changes in genetic correlation over time of fitness components such as survival or reproductive output in *Drosophila*, to have a more-precise genetic evaluation of milk, fat and protein yields, or to study the genetic correlation between milk production and somatic cell counts in dairy cattle.

Random regression models (Diggle *et al*., 1994) are the most commonly used at present for longitudinal data analysis. Previous comparisons of models in the univariate case, however, showed some drawbacks of this approach. Their focus is on choosing the most appropriate parametric function to fit individual deviations, and covariance functions are obtained directly from this regression model. Correlation functions are represented by quite complex polynomial functions that might not always be the most appropriate. Furthermore, because polynomials do not have asymptotes, they can hardly deal with correlation structures that decrease asymptotically to zero. Moreover, polynomials can have odd behaviours at the edges and high-order polynomials can show a 'wiggly' pattern, which is usually undesirable.

Extension of the random regression models to the multivariate case is straightforward and has already been used in several studies (Veerkamp & Thompson, 1998). It does, however, require a large number of parameters to describe the covariance structure. Moreover, similar difficulties to those observed in the univariate case are to be expected, and 'cross-covariance' functions obtained may not necessarily be most appropriate.

Character process (CP) models, proposed by Pletcher & Geyer (1999), have proved to compare

* Corresponding author. e-mail: jaffrezic@dga2.jouy.inra.fr

favourably to random regression models (Jaffrézic & Pletcher, 2000). They were in general better able to fit the covariance structure with fewer parameters, especially for asymptotic correlation patterns. This was due mainly to their ability to model variance and correlation separately. Extension of the CP models to the multivariate case is not straightforward, however, because parametric forms for 'cross-correlation' functions have still to be found.

Another method has recently been proposed by Nunez-Anton & Zimmerman (2000), namely structured antedependence (SAD) models. The idea of this approach is to model an observation at time t via a regression over the preceding observations. The number of parameters is considerably reduced in the SAD approach compared with the traditional antedependence models (Gabriel, 1962), thanks to a parametric modelling of the antedependence coefficients and innovation variances. They seem to offer similar advantages to CP models to fit the covariance structures adequately with few parameters.

Until now, very few comparisons have been performed for multivariate genetic longitudinal data analyses, because random regression was the only well-known approach available in this case. The aim of this paper is to propose an extension of SAD models to the genetic analysis of repeated measures and to the multivariate case. Properties of these models are studied, focusing especially on the shapes of the variance and correlation functions that can be obtained, and comparisons with CP models and random regression are presented. Data on fertility and mortality in *Drosophila* and on milk, fat and protein yields in dairy cattle are studied.

2. Model

(i) *Single-trait analysis*

The idea of antedependence models, as originally proposed by Gabriel (1962), is that an observation at time j can be explained by the previous ones. An antedependence structure of order r is defined by the fact that the jth observation ($j>r$) given the r preceding ones is independent of all other preceding observations (Gabriel, 1962). This concept will be generalized here to genetic analyses.

Although the process analysed is often continuous over time (such as growth), measurements are available only for a set of discrete times. Specification of the antedependence models relies on this discrete time scale. For simplicity, it will be assumed in the model presentation that measurements are equally spaced, but this assumption can be relaxed.

Assuming that the measurement times are on a discrete scale ($j=1, \ldots, J$), let Y_j be the phenotypic observation at time j. As in classical quantitative genetics, this can be decomposed as

$$Y_j = \mu_j + g_j + e_j, \tag{1}$$

where μ_j is a non random function that includes fixed effects and the mean curve in the population, g_j is the genetic effect and e_j the permanent environmental effect (including the residual).

Both the genetic and permanent environmental parts can be modelled with an antedependence structure, although both models will not necessarily be of the same order. Focusing on the genetic part, if a second-order SAD model is assumed, it can be written as

$$g_1 = \epsilon_1 \tag{2}$$

$$g_2 = \phi_1 g_1 + \epsilon_2 \tag{3}$$

$$g_j = \phi_1 g_{j-1} + \phi_2 g_{j-2} + \epsilon_j \tag{4}$$

for $j>2$. Here, ϕ_1 and ϕ_2 are antedependence parameters. For a SAD model of order r, r antedependence coefficients would be required ($\phi_1, \phi_2, \ldots, \phi_r$). The error terms ϵ_j are assumed to be normally distributed with mean zero and variance v_j^2, termed 'innovation variances', that can change with time j. In SAD models, Nunez-Anton & Zimmerman (2000) propose using a parametric function for innovation variances with, for example, a polynomial of time

$$\log v_j^2 = a + bj + cj^2. \tag{5}$$

In this case, only three parameters would be required to model the innovation variances v_j^2 regardless of the number of times of measurement J, whereas, in traditional antedependence models as originally proposed by Gabriel (1962), one parameter had to be estimated at each time and therefore J innovation variances v_j^2 would have to be estimated here.

It is also possible to deal with unequally spaced data by allowing the antedependence coefficients to depend on the lag between two measurements, for example as an exponential function as suggested by Nunez-Anton & Zimmerman (2000). In this case, if the measurement times are assumed to be ($t_1, t_2, \ldots, t_J$), a second order SAD model can be written as

$$g(t_1) = \epsilon(t_1) \tag{6}$$

$$g(t_2) = \phi_1(t_1, t_2) g(t_1) + \epsilon(t_2) \tag{7}$$

$$g(t_j) = \phi_1(t_j, t_{j-1}) g(t_{j-1}) + \phi_2(t_j, t_{j-2}) g(t_{j-2}) + \epsilon(t_j) \tag{8}$$

for $j=3, \ldots, J$. Any parametric function of time can be considered for the antedependence parameters, for example, an exponential function: $\phi_1(t_j, t_{j-1}) = \exp(-\theta_1(t_j - t_{j-1}))$ and $\phi_2(t_j, t_{j-2}) = \exp(-\theta_2(t_j - t_{j-2}))$.

The main difference from autoregressive models lies in the initial condition $g_1 = \epsilon_1$. Thanks to this

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

condition, antedependence parameters are unconstrained, whereas, for an autoregressive model with constant innovation variances and finite total variance, it is required that $|\phi|<1$.

SAD models require very few parameters for the covariance structure and increasing the order of antedependence involves only one extra parameter at each step. Higher-order antedependence models allow good flexibility in modelling the covariance structure, even for complex non-stationary correlations, which are not well accommodated by CP models, as will be shown in the example below.

Although SAD models do not, in general, allow simple analytic expressions for the covariance function, the covariance matrix can be calculated using the Cholesky decomposition of its inverse (Pourahmadi, 1999). Let $\boldsymbol{G}$ be the genetic covariance matrix, which has dimension $J\times J$, where J is the number of measurement times. It can be shown that

$$\boldsymbol{G}^{-1}=\boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}, \tag{9}$$

where $\boldsymbol{L}$ is a lower triangular matrix with 1s on the diagonal and the negatives of the antedependence coefficients ϕ_i ($i=1, \ldots, r$, for a SAD(r)) as below-diagonal entries, and $\boldsymbol{D}$ is a diagonal matrix with innovation variances v_j^2 ($j=1, \ldots, J$) as components. An interesting computational property is that the inverse, $\boldsymbol{G}^{-1}$, of these covariance matrices is sparse. Indeed, for a second-order antedependence model, for instance, only the diagonal and first two subdiagonals are non-zero. Antedependence and innovation variance parameters can be estimated by REML procedures.

For first-order SAD models, if innovation variances v^2 are assumed constant for all the error terms ϵ_j ($j=1, \ldots, J$), analytical forms for variance and correlation functions can be obtained. At time of measurement j, the genetic variance is given by

$$\mathrm{Var}(g_j)=\frac{1-\phi^{2j}}{1-\phi^2}\,v^2. \tag{10}$$

Therefore, even for constant innovation variances, the variance of the observed process can change with time, which is not the case for a simple first order autoregressive model where $\mathrm{Var}(g_j)=v^2/(1-\phi^2)$ (for $|\phi|<1$). For $j\geqslant k$, the covariance function of a SAD(1) is given by

$$\mathrm{Cov}(g_j, g_k)=\phi^{j-k}\frac{1-\phi^{2k}}{1-\phi^2}\,v^2 \tag{11}$$

and the correlation function is

$$\mathrm{Corr}(g_j, g_k)=\phi^{j-k}\frac{\sqrt{1-\phi^{2k}}}{\sqrt{1-\phi^{2j}}}. \tag{12}$$

Therefore, even for the simplest SAD model, the correlation function is non-stationary, i.e. $\mathrm{Corr}(g_j, g_k)$ does not depend only on the lag time $|j-k|$, in contrast to a first-order autoregressive model (AR(1)) in which the correlation function is given by $\mathrm{Corr}(g_j, g_k)=\phi^{|j-k|}$. When measurement times are equidistant, AR(1) is equivalent to a character process model with constant variance and exponential correlation.

For more complicated models, the relationship between the antedependence coefficients and innovation variances with the actual correlation and variance functions is far less straightforward than with CP models, in which they are modelled directly. However, in the SAD models, it is possible to increase the order of antedependence (increasing the number of antedependence parameters), which allows more flexibility than do CP models, especially for the modelling of non-stationary correlation functions.

(ii) *Bivariate analysis*

If two variables y_1 and y_2 are considered, it is possible to extend structured antedependence models to study the relationship between the two variables. As in the univariate case, measurements are decomposed into their genetic and permanent environmental components. Both parts are then modelled with an antedependence structure. For the genetic part, considering for instance a first order bivariate SAD, the model can be written as (for $j>1$)

$$g_{1j}=\phi_1 g_{1(j-1)}+\psi_1 g_{2(j-1)}+\epsilon_{1j} \tag{13}$$

$$g_{2j}=\phi_2 g_{2(j-1)}+\psi_2 g_{1(j-1)}+\epsilon_{2j} \tag{14}$$

with the initial condition $g_{11}=\epsilon_{11}$ and $g_{21}=\epsilon_{21}$. The error terms ϵ_{1j} and ϵ_{2j} are assumed to be bivariate normally distributed with mean zero and variances v_{1j}^2 and v_{2j}^2, respectively, and the correlation between ϵ_{1j} and ϵ_{2j} is assumed to change with time j. Several parametric functions of time could be considered. We propose to use here

$$r_j=\mathrm{Corr}(\epsilon_{1j}, \epsilon_{2j})=\exp(-\lambda_1 j)-\exp(-\lambda_2 j), \tag{15}$$

which is quite general and allows the correlation to be positive or negative depending on time j. It would also be possible to consider a constant correlation between the two error terms.

This bivariate specification allows to model various patterns of correlations between the two variables. For example, measurements within trait over time might not be highly correlated, but there might still be a strong correlation over time between the two traits. This will be achieved by low antedependence coefficients ϕ_1 and ϕ_2, and high values for ψ_1 and ψ_2, as well as for the correlation function r_j.

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

Assuming observations are ordered as: $\boldsymbol{y}=(y_{11}, y_{21}, \ldots, y_{1j}, y_{2j}, \ldots, y_{1J}, y_{2J})'$ the covariance matrices can be obtained as: $\boldsymbol{G}^{-1}=\boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}$ where $\boldsymbol{D}$ is a block-diagonal matrix with innovation variances (v_{1j}^2, v_{2j}^2) $(j=1, \ldots, J)$ along the diagonal and the covariance terms $\mathrm{Cov}(\epsilon_{1j}, \epsilon_{2j})=v_{1j}\ v_{2j}\ r_j$ on the subdiagonals. $\boldsymbol{L}$ can be written as

$$\begin{pmatrix} 1 & 0 & & & & & & \\ 0 & 1 & & & & & & \\ -\phi_1 & -\psi_1 & 1 & 0 & & & & \\ -\psi_2 & -\phi_2 & 0 & 1 & & & & \\ 0 & 0 & -\phi_1 & -\psi_1 & 1 & 0 & & \\ 0 & 0 & -\psi_2 & -\phi_2 & 0 & 1 & & \\ & & & & & & \ddots & \\ 0 & 0 & & \ldots & & -\phi_1 & -\psi_1 & 1 \ \ 0 \\ 0 & 0 & & \ldots & & -\psi_2 & -\phi_2 & 0 \ \ 1 \end{pmatrix}.$$

Matrix $\boldsymbol{G}$, obtained from $\boldsymbol{G}^{-1}=\boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}$ will always be positive semi-definite, as $\boldsymbol{D}$ is a positive semi-definite matrix and $\boldsymbol{L}$ is non-singular (Searle, 1982, Chapter 7).

A useful feature of the bivariate SAD models is that the cross-correlation functions are in general not symmetrical ($\mathrm{Corr}(g_{1j}, g_{2k})\neq\mathrm{Corr}(g_{2j}, g_{1k})$), because the correlation between, for example, body size at age 10 and weight at age 30 is not necessarily equal to the correlation between body size at age 30 and weight at age 10. Cross-correlation functions for bivariate SAD models do not have simple parametric expressions.

(iii) *Model selection*

A logical first step in model selection is to choose the appropriate SAD model for single function-valued trait analyses of variables y_{1j} and y_{2j}, by increasing the antedependence order (1, 2, …, r) until the additional antedependence coefficient ϕ_r is close to zero, meaning that no extra information on y_j is added by y_{j-r} (for $r<j$). Because models are nested, a likelihood ratio test can be used to check the significance of the additional antedependence parameters.

Once the model has been chosen for the two variables independently, the bivariate model can be selected. In general, antedependence orders for each trait will not be larger in the bivariate than in the univariate case, therefore the model selection procedure can start from the most complex model and reduce the order by a stepwise approach.

For example, assuming a first-order model for the genetic part in both univariate analyses

$$g_{1j}=\phi_1 g_{1(j-1)}+\epsilon_{1j} \tag{16}$$

$$g_{2j}=\phi_2 g_{2(j-1)}+\epsilon_{2j}, \tag{17}$$

it is then likely that a first order will also be sufficient for the bivariate model

$$g_{1j}=\phi_1 g_{1(j-1)}+\psi_1 g_{2(j-1)}+\epsilon_{1j} \tag{18}$$

$$g_{2j}=\phi_2 g_{2(j-1)}+\psi_2 g_{1(j-1)}+\epsilon_{2j}. \tag{19}$$

The cross-correlation between the two variables g_{1j} and g_{2j} at time j is modelled via the correlated error terms ϵ_{1j} and ϵ_{2j}, and so this model could even be reduced in some cases to

$$g_{1j}=\phi_1 g_{1(j-1)}+\epsilon_{1j} \tag{20}$$

$$g_{2j}=\phi_2 g_{2(j-1)}+\epsilon_{2j}, \tag{21}$$

with $\mathrm{Corr}(\epsilon_{1j}, \epsilon_{2j})=r_j$ as presented above. The likelihood criterion will be used here to choose the most appropriate model.

When higher-order antedependence models are chosen in the univariate case, adding antedependence terms into the bivariate analysis can reduce the initial order. For example, a third-order univariate model

$$g_{1j}=\phi_{11} g_{1(j-1)}+\phi_{12} g_{1(j-2)}+\phi_{13} g_{1(j-3)}+\epsilon_{1j} \tag{22}$$

can be reduced to a second-order model in a bivariate analysis

$$g_{1j}=\phi_{11} g_{1(j-1)}+\phi_{12} g_{1(j-2)}+\psi_1 g_{2(j-1)}+\epsilon_{1j}. \tag{23}$$

Indeed, adding the second trait $g_{2(j-1)}$ and a correlation between the error terms ϵ_{1j} and ϵ_{2j} can bring more information on the first trait than the third order $g_{1(j-3)}$.

The methodology section was based on modelling of the genetic part, but a similar approach will be used for the environmental part.

3. Examples

Two real examples were considered to illustrate these methodologies: first, a bivariate analysis of fertility and mortality rate in *Drosophila*; and second, bivariate analyses of milk and fat, and milk and protein yields for dairy cattle. Calculations were performed using ASREML (Gilmour *et al.*, 2000). In order to fit the SAD and the CP models, the OWN function, which allows the user to specify his own covariance matrix, was used. More details about the implementation of these models can be obtained from F. J.

(i) Drosophila *reproduction and mortality data*

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 genetic variation in longevity and reproduction (J. W. Curtsinger and A. A. Khazaeli, unpublished). Age-specific measures of mortality and average female reproductive output were collected simultaneously from two replicate

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

Table 1. *Likelihoods and parameter estimates for univariate phenotypic analyses of fertility and mortality rate in* Drosophila. *CP, character process model with quadratic variance and exponential correlation; RR2, quadratic random regression model; NPCov, number of parameters in the covariance structure*

		Fertility			Mortality		
			Parameters			Parameters	
Model	NPCov	Log L	ϕ_1	ϕ_2	Log L	ϕ_1	ϕ_2
SAD(1)	4	390·1	0·75		−256·6	0·73	
SAD(2)	5	390·6	0·73	0·03	−256·0	0·76	−0·04
CP	4	405·6			−259·4		
RR2	6	339·6			−342·0		

cohorts for each of 56 RI lines. Live/dead observations were made every day, and 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. This data set was previously analysed using univariate models by Jaffrézic & Pletcher (2000).

(ii) *Milk, fat and protein yields for dairy cattle*

These data comprised records on 9277 cows in first lactation from British herds, 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, fat and protein yields were available, with ten measurements per cow for each trait. Fixed effects considered were the age at calving, the proportion of North American Holstein genes and herd test month. To describe the mean, a non-parametric curve was used, fitting one mean at each test. Univariate genetic analyses for milk production using this data set were presented by Jaffrézic *et al*. (2002).

4. Results

(i) Drosophila *data*

(a) *Single trait phenotypic analyses*

Preliminary univariate analyses for fertility and mortality were performed in order to select the most appropriate SAD for both function-valued traits. Models of order r ($r=1, 2, \ldots, R$) were considered until the antedependence coefficient ϕ_R was close to zero. For all SAD models, a quadratic function was used to model innovation variances: $\text{Log } v_j^2 = a + bj + cj^2$ (where $j=1, \ldots, 11$ are the times of measurement). These models were compared with a character process with quadratic variance and exponential correlation (CP) as well as to a quadratic random regression model (RR2) (Jaffrézic & Pletcher, 2000).

Table 1 shows that, for both variables, a first-order SAD model (SAD(1)) would be appropriate, because there was no significant improvement in fit with a second-order model. SAD(1) fitted much better than a quadratic random regression model, despite having fewer parameters, and almost as well as the CP model.

(b) *Bivariate phenotypic analysis*

As explained in the model selection section, as first-order antedependence models fit well in the univariate case, the order of the bivariate model need not be larger. Estimates obtained for a first order bivariate SAD model were (for $j>1$)

$$\text{Mort}_P(j) = 0{\cdot}70\ \text{Mort}_P(j-1) - 0{\cdot}18\ \text{Fert}_P(j-1) + \epsilon_{P1j} \quad (24)$$

$$\text{Fert}_P(j) = 0{\cdot}68\ \text{Fert}_P(j-1) - 0{\cdot}07\ \text{Mort}_P(j-1) + \epsilon_{P2j}. \quad (25)$$

The correlation between the error terms at time j is given by

$$\text{Corr}(\epsilon_{P1j}, \epsilon_{P2j}) = \exp(-0{\cdot}37j) - \exp(-0{\cdot}18j). \quad (26)$$

Quadratic functions of time were used to model the logarithm of their variances

$$\text{Var}(\epsilon_{P1j}) = \exp(-0{\cdot}07 - 0{\cdot}03j - 0{\cdot}008j^2) \text{ and}$$
$$\text{Var}(\epsilon_{P2j}) = \exp(-0{\cdot}91 - 0{\cdot}31j + 0{\cdot}02j^2).$$

The likelihood value with this model was 183·8, with 12 parameters for the covariance structure. The fit obtained with this SAD model was therefore much better than with a bivariate quadratic random regression model, which involved 21 parameters and had a likelihood value of 67·7. SAD models seem to allow, in this case, a better flexibility to model the correlation structure between the two traits.

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

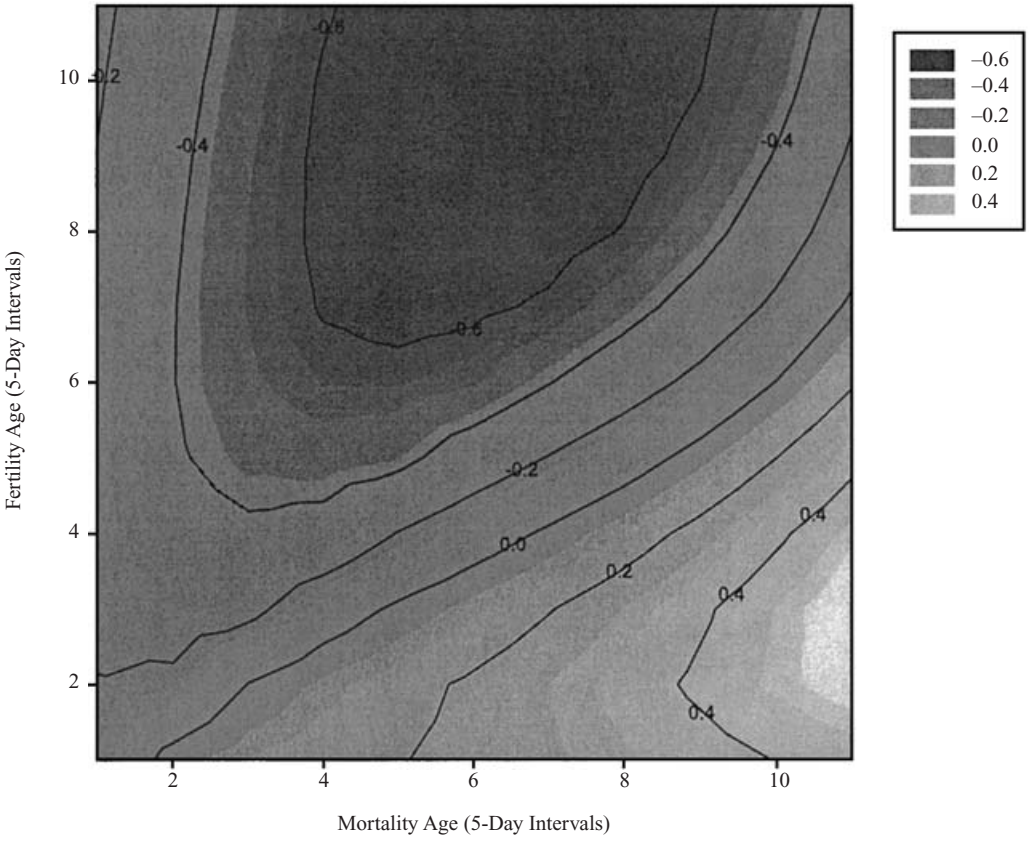

Fig. 1. Contour plot of the genetic cross-correlation between mortality and fertility in *Drosophila* for the chosen bivariate SAD model (Eqns 27–30).

Compared with univariate analyses for both variables, likelihood was considerably improved in the bivariate model, taking into account the dependence between fertility and mortality. The combined likelihood in univariate analyses was 133·5 (sum of the two univariate likelihood values given in Table 1 for the SAD(1) models), whereas it was 183·8 for the bivariate analysis, with only four extra parameters.

(c) *Bivariate genetic analysis*

The same bivariate SAD model was assumed for both genetic and environmental parts. Comparison with a quadratic random regression again showed a much higher likelihood for the antedependence model (Log L = 322·8) than for the RR model (Log L = 134·7), despite having many fewer parameters for the covariance structure (24 for the SAD model compared with 42 for the RR). Antedependence parameter estimates for the SAD models were (for $j>1$), for the genetic part

$$\mathrm{Mort_G}(j)=0{\cdot}87\ \mathrm{Mort_G}(j-1) + 0{\cdot}25\ \mathrm{Fert_G}(j-1)+\epsilon_{\mathrm{G1}j} \quad (27)$$

$$\mathrm{Fert_G}(j)=0{\cdot}80\ \mathrm{Fert_G}(j-1)-0{\cdot}08\ \mathrm{Mort_G}(j-1)+\epsilon_{\mathrm{G2}j} \quad (28)$$

with $\mathrm{Corr}(\epsilon_{\mathrm{G1}j}, \epsilon_{\mathrm{G2}j})=\exp(-0{\cdot}17j)-\exp(-0{\cdot}0001j)$ and

$$\mathrm{Var}(\epsilon_{\mathrm{G1}j})=\exp(-2{\cdot}05+0{\cdot}96j-0{\cdot}15j^2),$$

$$\mathrm{Var}(\epsilon_{\mathrm{G2}j})=\exp(-1{\cdot}0-0{\cdot}60j+0{\cdot}03j^2).$$

For the environmental part

$$\mathrm{Mort_E}(j)=0{\cdot}39\ \mathrm{Mort_E}(j-1)-0{\cdot}45\ \mathrm{Fert_E}(j-1)+\epsilon_{\mathrm{E1}j} \quad (29)$$

$$\mathrm{Fert_E}(j)=0{\cdot}54\ \mathrm{Fert_E}(j-1)-0{\cdot}02\ \mathrm{Mort_E}(j-1)+\epsilon_{\mathrm{E2}j} \quad (30)$$

with $\mathrm{Corr}(\epsilon_{\mathrm{E1}j}, \epsilon_{\mathrm{E2}j})=\exp(-0{\cdot}28\,j)-\exp(-0{\cdot}18\,j)$, and

$$\mathrm{Var}(\epsilon_{\mathrm{E1}j})=\exp(-0{\cdot}77-0{\cdot}08j+0{\cdot}002j^2),$$

$$\mathrm{Var}(\epsilon_{\mathrm{E2}j})=\exp(-2{\cdot}18-0{\cdot}08j+0{\cdot}009j^2).$$

The genetic correlation between fertility and mortality was found to be negative at most ages, as shown in Fig. 1, except between fecundity at early ages and mortality at late ages. The genetic correlation was strongly negative between fecundity at late ages and mortality. This correlation pattern is quite complex and would be difficult to model with a simple parametric function.

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

Table 2. *Likelihoods for univariate phenotypic analysis of 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. A constant term was added to all the likelihood values to make them easier to read and compare*

		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

(ii) *Dairy cattle data*

(a) *Single-trait phenotypic analyses*

The same steps were followed to choose the most appropriate antedependence model for milk, fat and protein yields. Different orders of SAD models were compared with character processes 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 ten in order to have variances of about the same order as for milk. Likelihoods for these different models are given in Table 2. For the analysis of these monthly records, times of measurement were supposed to be $j=1, \ldots, 10$.

Parameter estimates for the fourth-order SAD model (SAD(4)) were (for $j>4$)

$$\begin{aligned}\text{Milk}_P(j) = {} & 0{\cdot}50\ \text{Milk}_P(j-1) + 0{\cdot}22\ \text{Milk}_P(j-2) \\ & + 0{\cdot}10\ \text{Milk}_P(j-3) + 0{\cdot}05\ \text{Milk}_P(j-4) + \epsilon_{P1j}\end{aligned}$$

$$\begin{aligned}\text{Fat}_P(j) = {} & 0{\cdot}37\ \text{Fat}_P(j-1) + 0{\cdot}21\ \text{Fat}_P(j-2) \\ & + 0{\cdot}14\ \text{Fat}_P(j-3) + 0{\cdot}11\ \text{Fat}_P(j-4) + \epsilon_{P2j}\end{aligned}$$

$$\begin{aligned}\text{Prot}_P(j) = {} & 0{\cdot}51\ \text{Prot}_P(j-1) + 0{\cdot}22\ \text{Prot}_P(j-2) \\ & + 0{\cdot}10\ \text{Prot}_P(j-3) + 0{\cdot}06\ \text{Prot}_P(j-4) + \epsilon_{P3j}.\end{aligned}$$

Univariate analysis for milk, fat and protein yields showed that SAD models of order 1 were about equivalent to CP models (Table 2). Increasing orders of antedependence allowed more flexibility, especially to fit the highly non-stationary correlation patterns, and had a higher likelihood than CP models. SAD models of order 3 or 4 performed better than RR3 models, and also better than an RR4 in the milk yield analysis, while requiring far fewer parameters: seven parameters for a SAD(4) model, 15 parameters for an RR4 model. This difference in the number of parameters would be even larger in a bivariate analysis: 55 parameters would be required for a bivariate quartic regression but only 24 in a bivariate SAD(4) model.

Genetic univariate analysis for milk production using SAD models was also performed. It was found that the genetic part was quite simple to model and a first-order antedependence was sufficient, whereas a third-order antedependence was needed for the environmental part, which had a much more complex correlation structure. This model had a higher likelihood than an RR4, with many fewer parameters (11 for the SAD model, 31 for the RR).

(b) *Genetic bivariate analysis for milk and fat yields*

As in previous univariate studies (Jaffrézic *et al.*, 2003), it was found that the antedependence order required for the genetic part was lower than for the environmental part. The chosen model was, for the genetic part (for $j>1$)

$$\text{Fat}_G(j) = 0{\cdot}90\ \text{Fat}_G(j-1) + 0{\cdot}03\ \text{Milk}_G(j-1) + \epsilon_{G1j} \tag{31}$$

$$\text{Milk}_G(j) = 1{\cdot}02\ \text{Milk}_G(j-1) - 0{\cdot}24\ \text{Fat}_G(j-1) + \epsilon_{G2j} \tag{32}$$

with $\text{Corr}(\epsilon_{G1j}, \epsilon_{G2j}) = \exp(-0{\cdot}0001j) - \exp(-0{\cdot}38j)$, and

$$\text{Var}(\epsilon_{G1j}) = \exp(-1{\cdot}20 - 1{\cdot}10j + 0{\cdot}09j^2),$$

$$\text{Var}(\epsilon_{G2j}) = \exp(\pm 1{\cdot}62 - 1{\cdot}09j + 0{\cdot}09j^2).$$

For the environmental part

$$\begin{aligned}\text{Fat}_E(j) = {} & 0{\cdot}46\ \text{Fat}_E(j-1) + 0{\cdot}26\ \text{Fat}_E(j-2) \\ & + 0{\cdot}025\ \text{Milk}_E(j-1) + \epsilon_{E1j}\end{aligned} \tag{33}$$

$$\begin{aligned}\text{Milk}_E(j) = {} & 0{\cdot}69\ \text{Milk}_E(j-1) + 0{\cdot}23\ \text{Milk}_E(j-2) \\ & - 0{\cdot}32\ \text{Fat}_E(j-1) + \epsilon_{E2j},\end{aligned} \tag{34}$$

with $\text{Corr}(\epsilon_{E1j}, \epsilon_{E2j}) = \exp(-0{\cdot}007j) - \exp(-0{\cdot}57j)$, and

$$\text{Var}(\epsilon_{E1j}) = \exp(1{\cdot}36 - 0{\cdot}44j + 0{\cdot}03j^2),$$

$$\text{Var}(\epsilon_{E2j}) = \exp(3{\cdot}15 - 0{\cdot}47j + 0{\cdot}04j^2).$$

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

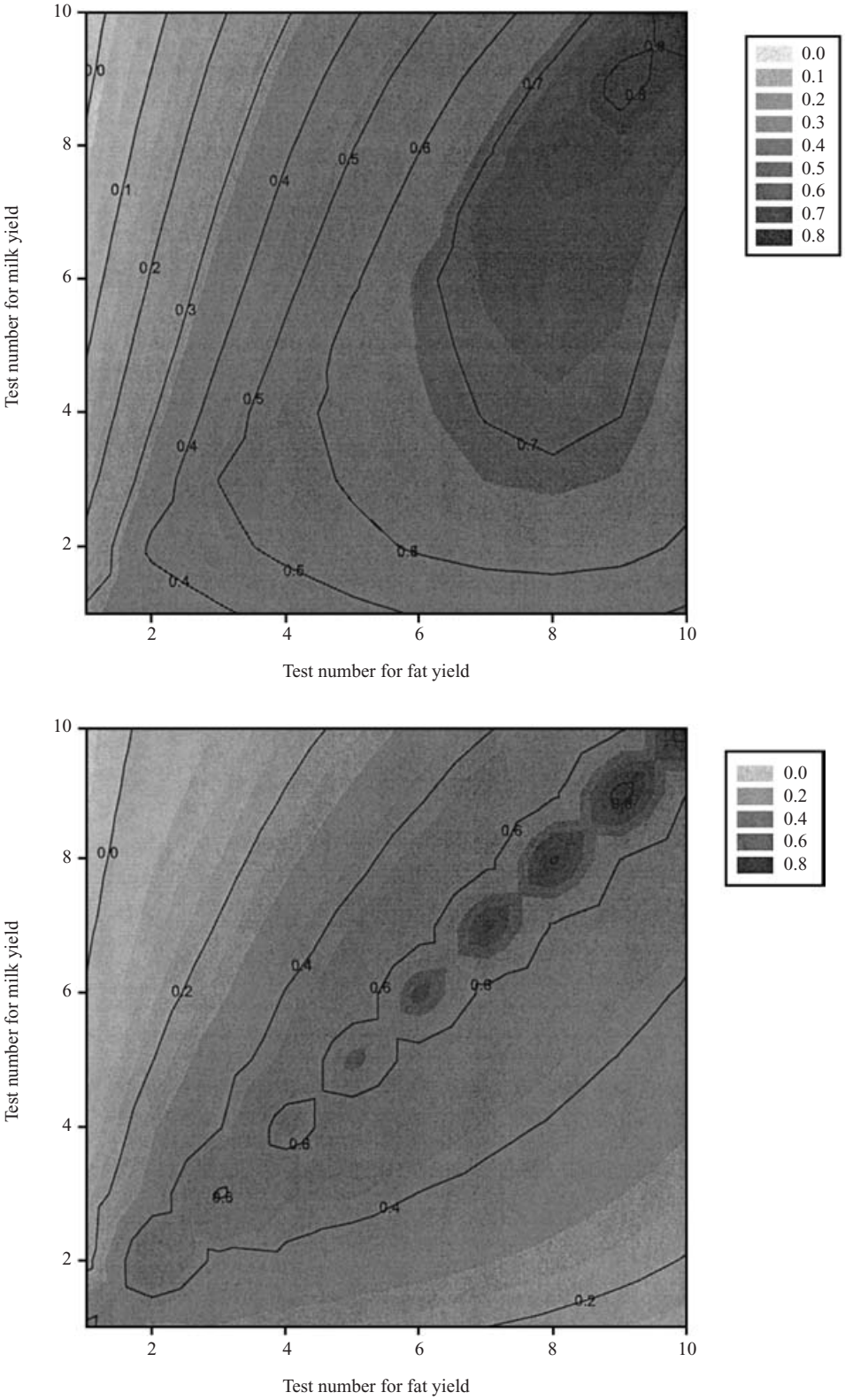

Fig. 2. Genetic (A) and environmental (B) cross-correlation between milk and fat yields with the chosen SAD model (Eqns 31–34).

For this model, the likelihood was 3664 with 26 parameters to model the covariance structure, and was higher than for a bivariate RR2 model with many more parameters (Log L = 1198 with 42 parameters).

Figure 2A, B gives the contour plots of the estimated genetic and environmental cross-correlation functions obtained with this model. As expected, the genetic correlation between milk and fat yields was quite high

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

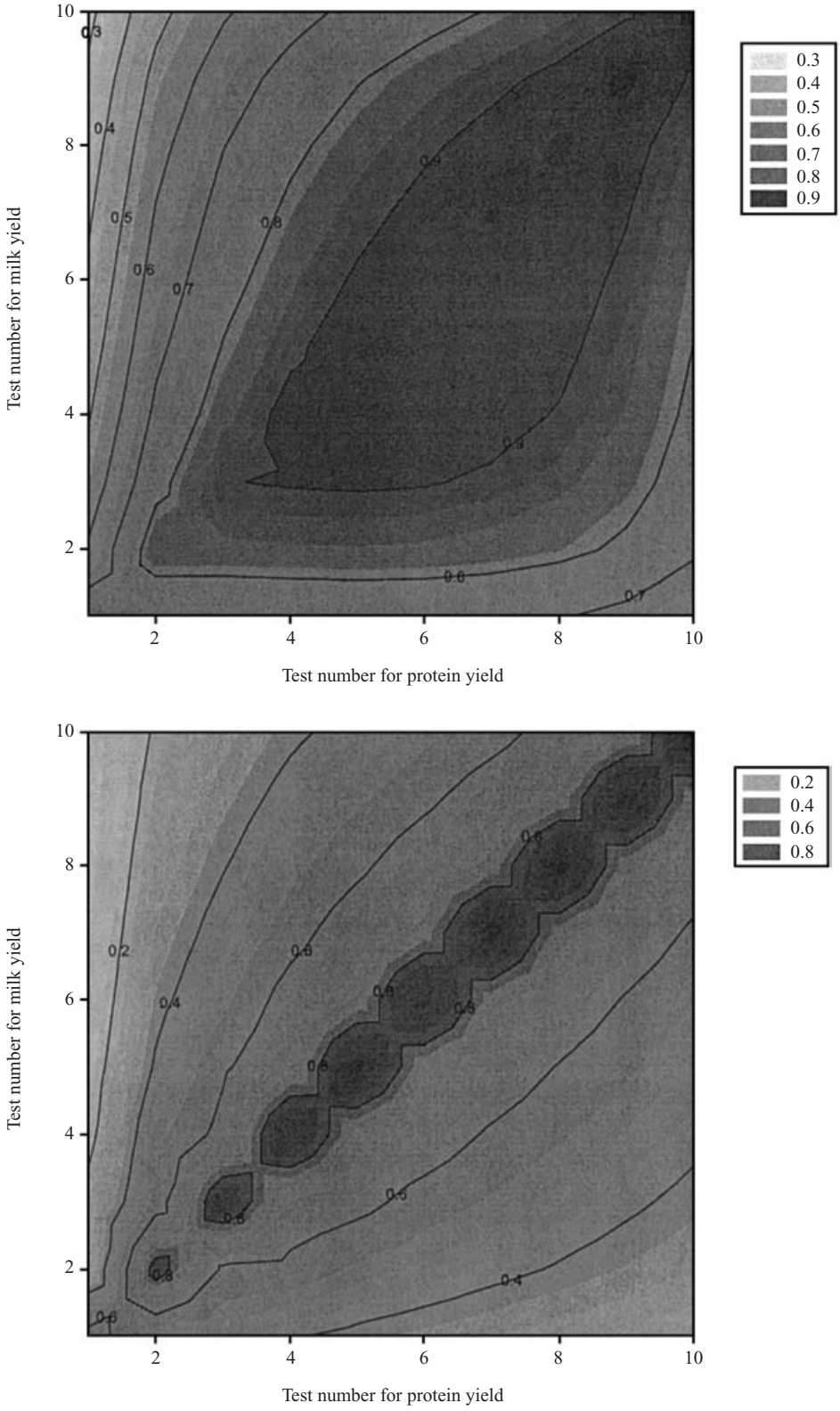

Fig. 3. Genetic (A) and environmental (B) cross-correlation between milk and protein yields with the chosen SAD model (Eqns 35–38).

throughout lactation. It was highest (between 0·7 and 0·8) for late stages of lactation, and lowest between the first test for fat yield and all the milk measurements. The environmental cross-correlation was a little lower, although it remained positive for the whole lactation period. The highest values were found along the diagonal (between milk and fat yields at the same test), and were the lowest for early lactation stages.

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

(c) *Genetic bivariate analysis for milk and protein yields*

The chosen model was, for the genetic part,

$$\text{Prot}_G(j) = 0{\cdot}78\ \text{Prot}_G(j-1) + 0{\cdot}06\ \text{Milk}_G(j-1) + \epsilon_{G1j} \quad (35)$$

$$\text{Milk}_G(j) = 1{\cdot}14\ \text{Milk}_G(j-1) - 0{\cdot}64\ \text{Prot}_G(j-1) + \epsilon_{G2j} \quad (36)$$

with $\text{Corr}(\epsilon_{G1j}, \epsilon_{G2j}) = \exp(-0{\cdot}0006j) - \exp(-1{\cdot}46j)$, and

$$\text{Var}(\epsilon_{G1j}) = \exp(-1{\cdot}31 - 1{\cdot}10j + 0{\cdot}10j^2),$$

$$\text{Var}(\epsilon_{G2j}) = \exp(1{\cdot}96 - 1{\cdot}29j + 0{\cdot}11j^2).$$

For the environmental part, it was

$$\text{Prot}_E(j) = 0{\cdot}19\ \text{Prot}_E(j-1) + 0{\cdot}24\ \text{Prot}_E(j-2) + 0{\cdot}11\ \text{Milk}_E(j-1) + \epsilon_{E1j} \quad (37)$$

$$\text{Milk}_E(j) = 0{\cdot}87\ \text{Milk}_E(j-1) + 0{\cdot}24\ \text{Milk}_E(j-2) - 1{\cdot}01\ \text{Prot}_E(j-1) + \epsilon_{E2j} \quad (38)$$

with $\text{Corr}(\epsilon_{E1j}, \epsilon_{E2j}) = \exp(-0{\cdot}006j) - \exp(-1{\cdot}10j)$, and

$$\text{Var}(\epsilon_{E1j}) = \exp(0{\cdot}24 - 0{\cdot}15j + 0{\cdot}008j^2),$$

$$\text{Var}(\epsilon_{E2j}) = \exp(2{\cdot}95 - 0{\cdot}32j + 0{\cdot}02j^2).$$

For this model, the likelihood was 4525, with 26 parameters to model the covariance structure, and was higher than for a bivariate RR2 model with many more parameters (Log L = 1281 with 42 parameters).

Figure 3A, B gives the contour plot of the estimated genetic and environmental correlations between milk and protein yields obtained with this model. The genetic cross-correlation was even higher than between milk and fat yields with values between 0·8 and 0·9 for most lactation stages. As previously, the lowest correlation values were found between protein yield at the first test and milk measurements. The environmental correlation had a similar pattern as that between milk and fat yields, with the highest values along the diagonal and a correlation decreasing as tests became further apart.

5. Discussion

SAD models have recently been proposed in the statistical literature (Nunez-Anton & Zimmerman, 2000) and seem to be a valuable alternative to other methodologies for the genetic analysis of longitudinal data. In particular, they offer a high degree of flexibility to model the covariance structure with very few parameters, and can even deal with complex non-stationary patterns that were not well accommodated with CP models, as observed in the milk production study.

This paper presents an extension of the SAD models to the multivariate case that proved, in most cases, to perform better than random regression with far fewer parameters. The multivariate extension of RR models requires a very large number of parameters; for example, a bivariate genetic analysis fitting only a quadratic model for both genetic and environmental parts requires 42 parameters, and if of cubic order for both parts, the number of parameters jumps to 72. By contrast, increasing the order of a SAD model adds only eight parameters at each step.

Therefore, SAD models seem to be very promising for the analysis of genetic repeated measures and other function-valued traits, in the univariate as well as multivariate cases. Further research is still needed, however, to study all the possible structures that can be fitted with these models. As shown in this paper, analytical formulae for variance and correlation functions can be worked out for a simple first-order SAD model when assuming constant innovation variances. In the general case, however, only recursive formulae for the covariances can be written. It is therefore much more difficult to obtain the relationships between antedependence parameters, innovation variances and the actual variance and correlation functions of the process, and to be able to study their properties. A complex simulation study would probably be required for that purpose, considering all the different possible models. Similar difficulties are encountered for the eigenfunctions of the process. Their relationship with matrices $\boldsymbol{L}$ and $\boldsymbol{D}$ is not at all straightforward and it is therefore extremely difficult to study their properties and the underlying assumptions, because no analytical expression is available.

Antedependence models are often suggested to analyse cumulative effects. It might therefore be useful in subsequent research to undertake analyses of cumulative mortality and milk yields to investigate this and to compare with alternative methods of analysing cumulative data. Estimates obtained here for the cross-correlation functions seem to be quite reasonable. However, additional bivariate genetic studies will have to be performed in order to validate these results.

In this study, antedependence coefficients were assumed constant over time. It would, however, be possible to relax this assumption as suggested by Nunez-Anton & Zimmerman (2000), and this would be particularly appropriate for unequally spaced data. In fact, the antedependence coefficient will not be the same – for example, when one observation and that preceding it are separated by one or two days.

https://doi.org/10.1017/S0016672303006281 Published online by Cambridge University Press

Parametric functions of time can be used for the antedependence coefficients, such as an exponential function of the lag between two measurements. Additional flexibility of the SAD models can also be obtained by incorporating heterogeneous innovation variances, and, because they were assumed here to change with time, it is also possible to include other factors of heterogeneity (such as herd, for the dairy cattle data). This extension is straightforward, because a structural model (Foulley & Quaas, 1995) is already used to model the time dependence of the innovation variances.

We are most grateful to I. White and S. D. Pletcher for helpful comments and ideas, to J. Curtsinger and A. Khazaeli for generously providing published and unpublished data, and to two anonymous referees for very useful suggestions. The dairy cattle data set was provided by the Milk Development Council (UK).

References

Diggle, P. J., Liang, K. Y. & Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford: Oxford University Press.

Foulley, J. L. & Quaas, R. L. (1995). Heterogeneous variances in gaussian linear mixed models. *Genetics, Selection, Evolution* **27**, 211–228.

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

Gilmour, A. R., Thompson, R., Cullis, B. R. & Welham, S. J. (2000). *ASREML Manual*. Orange, Australia: New South Wales Department of Agriculture.

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

Jaffrézic, F., White, I. M. S., Thompson, R. & Visscher, P. M. (2002). Contrasting models for lactation curve analysis. *Journal of Dairy Science* **84**, 968–975.

Jaffrézic, F., White, I. M. S. & Thompson, R. (2003). Use of the score test as a goodness-of-fit measure of the covariance structure in genetic analysis of longitudinal data. *Genetics, Selection, Evolution* **35**, 185–198.

Meuwissen, T. H. E. & Pool, M. H. (2001). Autoregressive versus random regression test-day models for prediction of milk yields. *Interbull Bulletin* **27**, 172–178.

Meyer, K. (2001). Estimating genetic covariance functions assuming a parametric correlation structure for environmental effects. *Genetics, Selection, Evolution* **33**, 557–585.

Nunez-Anton, V. & Zimmerman, D. L. (2000). Modelling non-stationary longitudinal data. *Biometrics* **56**, 699–705.

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

Pletcher, S. D., Houle, D. & 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.

Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York: John Wiley and Sons.

Veerkamp, R. F. & Thompson, R. (1998). Multi-trait covariance functions to model genetic variation in the dynamic relation between feed intake, live weight and milk yield during lactation. *49th Annual Meeting of the European Association of Animal Production*. Warsaw, Poland.