DOI: 10.1534/genetics.103.019554

Multivariate Character Process Models for the Analysis of Two or More Correlated Function-Valued Traits

Florence Jaffrézic,*[,1] Robin Thompson[†,‡] and Scott D. Pletcher[§]

*INRA Quantitative and Applied Genetics, 78352 Jouy-en-Josas Cedex, France, [†]Rothamsted Research, Harpenden, Herts AL5 2JQ, United Kingdom, [‡]Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, United Kingdom and [§]Huffington Center on Aging and Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030

Manuscript received July 1, 2003
Accepted for publication May 17, 2004

ABSTRACT

Various methods, including random regression, structured antedependence models, and character process models, have been proposed for the genetic analysis of longitudinal data and other function-valued traits. For univariate problems, the character process models have been shown to perform well in comparison to alternative methods. The aim of this article is to present an extension of these models to the simultaneous analysis of two or more correlated function-valued traits. Analytical forms for stationary and nonstationary cross-covariance functions are studied. Comparisons with the other approaches are presented in a simulation study and in an example of a bivariate analysis of genetic covariance in age-specific fecundity and mortality in Drosophila. As in the univariate case, bivariate character process models with an exponential correlation were found to be quite close to first-order structured antedependence models. The simulation study showed that the choice of the most appropriate methodology is highly dependent on the covariance structure of the data. The bivariate character process approach proved to be able to deal with quite complex nonstationary and nonsymmetric cross-correlation structures and was found to be the most appropriate for the real data example of the fruit fly *Drosophila melanogaster*.

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 growing. Important examples of these so-called function-valued traits include growth curves (Meyer 2001), age-specific components of organismal fitness such as survival or reproductive output (Pletcher *et al.* 1998), lactation curves in dairy cattle (Meuwissen and Pool 2001; Jaffrézic *et al.* 2002), and gene expression profiles across age or environmental treatments (DeRisi *et al.* 1997; Pletcher *et al.* 2002).

Several techniques have been proposed for single-trait (univariate) analyses. These include random regression models, which are based on a parametric modeling of individual curves (Diggle *et al.* 1994), character process models, which focus on parametric modeling of the covariance structure (Pletcher and Geyer 1999), and structured antedependence models (SAD; Nunez-Anton and Zimmerman 2000; Jaffrézic *et al.* 2003), where an observation at time *t* is modeled via a regression over the preceding observations. The number of parameters is considerably reduced in the SAD approach compared to the traditional antedependence models (Gabriel 1962), thanks to a parametric modeling of the antedependence coefficients and innovation variances. A comparison among these methods revealed that, in many cases, character process models performed well in comparison to alternative methods, especially random regression, often providing a better fit to the covariance structure (genetic and nongenetic) with fewer parameters (Jaffrézic and Pletcher 2000).

A parsimonious method for the analysis of two or more correlated function-valued traits is needed. Although a multivariate extension of random regression models is straightforward, their sometimes poor performance in the univariate case argues for the development of alternative methods. Moreover, the nature of the parameterization results in a dramatic increase in the number of parameters required to describe complicated covariance structures, which is often problematic. The data sets that are generated in experimental sciences, such as genetics, and that are used to estimate different types of covariance structures (*e.g.*, genetic and nongenetic) are often too small to support the estimation of many parameters (Pletcher *et al.* 1998). This would also preclude the use of other models such as spline functions.

The aim of this article is to investigate an extension of the character process (CP) models (Pletcher and Geyer 1999) to the multivariate case. The advantages that apply to the CP models in the univariate setting, *i.e.*, a small number of parameters to model the covariance structure and a high degree of flexibility, are crucial for developing practical multivariate models. Several

[1] *Corresponding author:* INRA-SGQA, 78352 Jouy-en-Josas Cedex, France. E-mail: florence.jaffrezic@dga2.jouy.inra.fr

cross-correlation and cross-covariance functions are studied, and their behavior is compared to multivariate random regression and structured antedependence models in a simulation study and in an example for the genetic analysis of age-specific fecundity and mortality in the fruit fly, *Drosophila melanogaster*.

MATERIALS AND METHODS

Bivariate character process models: A detailed description of the quantitative genetic model for univariate function-valued traits is given by Jaffrézic and Pletcher (2000) and Pletcher and Geyer (1999). In the genetic analysis of two correlated function-valued traits, it is assumed that the observed phenotypic characters can be decomposed as

$$\boldsymbol{Y}(\boldsymbol{t}) = \boldsymbol{\mu}(\boldsymbol{t}) + \boldsymbol{g}(\boldsymbol{t}) + \boldsymbol{e}(\boldsymbol{t}), \tag{1}$$

where $\boldsymbol{Y}(\boldsymbol{t}) = (Y_1(t), Y_2(t))'$ represent the observed phenotypic trajectories for the two characters $Y_1(t)$ and $Y_2(t)$, t represents any continuous independent variable, which for clarity we assume is time, $\boldsymbol{\mu}(\boldsymbol{t}) = (\mu_1(t), \mu_2(t))'$ are nonrandom functions that correspond to the genotypic mean functions of $Y_1(t)$ and $Y_2(t)$, respectively, and $\boldsymbol{g}(\boldsymbol{t}) = (g_1(t), g_2(t))'$ represent the genetic deviations for the two characters. Both deviations are correlated over time and $\boldsymbol{g}(\boldsymbol{t})$ is a bivariate Gaussian process. Similarly, $\boldsymbol{e}(\boldsymbol{t}) = (e_1(t), e_2(t))'$ are the environmental deviations. Processes $\boldsymbol{g}(\boldsymbol{t})$ and $\boldsymbol{e}(\boldsymbol{t})$ are assumed independent of one another, with mean zero at each age and with covariance functions $\boldsymbol{G}(\boldsymbol{t}, \boldsymbol{s})$ and $\boldsymbol{E}(\boldsymbol{t}, \boldsymbol{s})$. Focus is on the modeling of these covariance functions.

In the univariate character process approach, there is only one function-valued trait, $Y(t)$, and its covariance functions (genetic and environmental) are modeled as

$$G(t, s) = v(t)v(s)\rho(t, s), \tag{2}$$

where $v^2(t)$ represents the variance function and is usually a parametric function of the continuous variable such as a polynomial and $\rho(t, s)$ is the correlation function. Assuming stationarity in the correlations, Pletcher and Geyer (1999) proposed parametric forms for the correlation function including an exponential ($\rho(t, s) = \exp(-\theta|t - s|)$), a Gaussian ($\rho(t, s) = \exp(-\theta(t - s)^2)$), and a Cauchy ($\rho(t, s) = 1/(1 + \theta(t - s)^2)$) function. Jaffrézic and Pletcher (2000) suggested a nonstationary extension of the models based on a nonlinear transformation of the timescale, $f(t)$ (Nunez-Anton and Zimmerman 2000). Correlation stationarity is assumed to hold on the transformed scale $\rho(t, s) = \rho(|f(t) - f(s)|)$.

Models for bivariate Gaussian processes have been investigated previously (Sy *et al.* 1997) as, for example, the bivariate Ornstein-Uhlenbeck process. It corresponds to a continuous-time extension of a first-order autoregressive process [AR(1)], which is also equivalent to a CP model with an exponential correlation and a constant variance. We adapt these ideas to extend the character process methodology.

Let the continuous variable of interest be time and the object of analysis be the genetic covariance function. In the bivariate case, let $\boldsymbol{g}(t) = (g_1(t), g_2(t))'$ be the genetic character process, where $g_1(t)$ is associated with trait 1 and $g_2(t)$ with trait 2. The bivariate covariance function of the process can be written as

$$\mathrm{Cov}(\boldsymbol{g}(t), \boldsymbol{g}(s)') = \boldsymbol{V}(t)^{1/2}\boldsymbol{\Omega}(t - s)(\boldsymbol{V}(s)^{1/2})' \tag{3}$$

(for $0 \leq s \leq t$), where

$$\mathrm{Cov}(\boldsymbol{g}(t), \boldsymbol{g}(s)') = \begin{pmatrix} \mathrm{Cov}(g_1(t), g_1(s)) & \mathrm{Cov}(g_1(t), g_2(s)) \\ \mathrm{Cov}(g_2(t), g_1(s)) & \mathrm{Cov}(g_2(t), g_2(s)) \end{pmatrix}. \tag{4}$$

As the covariance function has to be symmetric, it is required that

$$\mathrm{Cov}(\boldsymbol{g}(s), \boldsymbol{g}(t)') = \mathrm{Cov}(\boldsymbol{g}(t), \boldsymbol{g}(s)'). \tag{5}$$

Definition of matrix $\boldsymbol{\Omega}(t - s)$: In the bivariate case, matrix $\boldsymbol{\Omega}(t - s)$ is of dimension 2×2. The requirements on this matrix are that it is positive definite, equal to the identity matrix when $t = s$, and should verify the symmetry property $\boldsymbol{\Omega}(t - s) = \boldsymbol{\Omega}(s - t)$. It corresponds to a bivariate extension of the correlation functions proposed for univariate character process models by Pletcher and Geyer (1999). All the functions proposed in their article can be extended. Among them, however, the most commonly used are the exponential, the Gaussian, and the Cauchy correlations. These functions are defined as follows:

Exponential: $\boldsymbol{\Omega}(t - s) = \mathbf{exp}(-\boldsymbol{\Theta}(|t - s|))$.
Gaussian: $\boldsymbol{\Omega}(t - s) = \mathbf{exp}(-\boldsymbol{\Theta}(t - s)^2)$.
Cauchy: $\boldsymbol{\Omega}(t - s) = (\boldsymbol{I} + \boldsymbol{\Theta}(t - s)^2)^{-1}$.

In the bivariate case, $\boldsymbol{I}$ is the 2×2 identity matrix and $\boldsymbol{\Theta}$ is a 2×2 matrix, not necessarily symmetric, with positive eigenvalues. The matrix exponentiation corresponds to a series expansion and can be calculated using an eigenvalue decomposition as shown in appendix a.

The bivariate exponential function is also used in the statistical literature for the Ornstein-Uhlenbeck process (Sy *et al.* 1997).

Further extension to this framework includes a relaxation of stationarity of the correlation function. The nonstationary extension of the CP models proposed by Jaffrézic and Pletcher (2000) is implemented by replacing time lags $(t - s)$ by a transformation $(f(t) - f(s))$. Considering a Box-Cox transformation, as suggested by Nunez-Anton and Zimmerman (2000), and an exponential CP model, the correlation function can be written as

$$\boldsymbol{\Omega}(t, s) = \mathbf{exp}(-\boldsymbol{\Theta}((t^{\ell} - s^{\ell})/\ell)) \tag{6}$$

for $\ell \neq 0$ and

$$\boldsymbol{\Omega}(t, s) = \mathbf{exp}(-\boldsymbol{\Theta}(\mathrm{Log}(t) - \mathrm{Log}(s))) \tag{7}$$

when $\ell = 0$.

Definition of matrix $\boldsymbol{V}(t)$: In the bivariate case, matrix $\boldsymbol{V}(t)$ is also of dimension 2×2. The requirements for this matrix are that it is symmetric and positive definite. It in fact corresponds to the covariance of the process at a given time t, as matrix $\boldsymbol{\Omega}(t - s)$ is the identity matrix when $t = s$:

$$\boldsymbol{V}(t) = \mathbf{Var}(\boldsymbol{g}(t)) = \begin{pmatrix} \mathrm{Var}(g_1(t)) & \mathrm{Cov}(g_1(t), g_2(t)) \\ \mathrm{Cov}(g_1(t), g_2(t)) & \mathrm{Var}(g_2(t)) \end{pmatrix}. \tag{8}$$

We present here two possible ways of modeling matrix $\boldsymbol{V}(t)$.

It is possible to use a polynomial of time to model function $\boldsymbol{V}(t)$. That would correspond to a direct bivariate extension of the variance function of the character process model (Pletcher and Geyer 1999).

When considering, for example, a quadratic function of time, the bivariate variance function can be written as

$$\mathbf{ln}(\boldsymbol{V}(t)) = \boldsymbol{A} + \boldsymbol{B}t + \boldsymbol{C}t^2, \tag{9}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are 2×2 symmetric matrices. The ln( ) of the variance again corresponds to a series expansion and can be calculated as the exponential in the $\boldsymbol{\Omega}$ matrix by using an eigenvalue decomposition as explained in appendix a.

The covariance matrix $\boldsymbol{V}(t)$ can also be decomposed in terms of variance and correlation functions such as

TABLE 1

Likelihood values for the simulated data sets based on unstructured covariance matrices

Model	NPCov	Example 1	Example 2	Example 3
US	55	2746.7	2401.8	3801.9
CP Quad-Exp	13	551.0	799.4	588.1
CP Quad-ExpNS	14	566.3	1478.9	703.0
SAD(1)	12	262.2	1008.4	545.0
SAD(2)	14	430.8	1380.2	864.4
RR1	13	980.0	200.7	472.6

US, unstructured covariance matrix; CP Quad-ExpNS, quadratic polynomial used to model $\boldsymbol{V}(t)$, exponential function for $\boldsymbol{\Omega}(t - s)$ with the nonstationary extension (Equation 6); RR1, linear random regression model with three additional parameters for the residual structure; NPCov, number of parameters in the covariance structure.

$$\boldsymbol{V}(t) = \begin{pmatrix} v_1^2(t) & v_1(t)v_2(t)\rho_{12}(t) \\ v_1(t)v_2(t)\rho_{12}(t) & v_2^2(t) \end{pmatrix}. \quad (10)$$

Variance functions can be modeled as for univariate character process models with polynomial functions of time. For a quadratic function, for instance, $v_1^2(t) = \mathrm{Var}(g_1(t)) = \exp(a_1 + b_1 t + c_1 t^2)$ and $v_2^2(t) = \mathrm{Var}(g_2(t)) = \exp(a_2 + b_2 t + c_2 t^2)$.

Function $\rho_{12}(t)$ represents the cross-correlation between the two traits at a given time t. A possible parametric modeling for this cross-correlation function is

$$\mathrm{Corr}(g_1(t), g_2(t)) = \rho_{12}(t) = \exp(-\lambda_1 t) - \exp(-\lambda_2 t) \quad (11)$$

for $\lambda_1, \lambda_2 > 0$. For practical purposes, it is interesting to note that this correlation function is equal to 0 at $t = 0$, increases to a maximum at $t = [\ln(\lambda_2/\lambda_1)]/(\lambda_2 - \lambda_1)$, and then decreases to 0 at infinity.

A likelihood-ratio test can be used to examine specific hypotheses about the parameters. For example, testing if the cross-correlation between the two processes at all times t is equal to zero is equivalent to testing if $\lambda_1 = \lambda_2$. The cross-correlation function $\rho_{12}(t)$ can also be assumed constant: $\rho_{12}(t) = r$, which would imply that the cross-correlations are equal for all t.

Estimation procedure: Parameters of these bivariate character process models can be estimated with REML procedures, using, for example, the OWN function of ASREML (Gilmour *et al.* 2002) as presented in appendix a. The nonstationary parameter ℓ (Equation 6) is estimated at the same time as the other covariance parameters with standard REML procedures. The properties of the proposed bivariate covariance function are studied in appendix b.

EXAMPLE

Simulation study: A simulation study was performed to understand better the analogies between the different methodologies: the bivariate CP model proposed here, the bivariate structured antedependence models presented in Jaffrézic *et al.* (2003), and the random regression models. In a first set of simulations, data were generated according to a bivariate CP model, with an exponential "correlation" function ($\mathbf{exp}(-\boldsymbol{\Theta}(t - s))$) and a $\boldsymbol{V}(t)$ structure defined as $\mathbf{ln}\ \boldsymbol{V}(t) = \boldsymbol{A} + \boldsymbol{B}t + \boldsymbol{C}t^2$. Different assumptions on parameters of $\boldsymbol{\Theta}$, $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ were investigated, setting some elements to zero or giving various values to these parameters. It was found that the first-order bivariate structured antedependence model [SAD(1)] was well able to capture the covariance structures simulated under all these different assumptions (results not shown). The similarity between these two approaches had already been pointed out in the univariate case for SAD(1) models and CP with an exponential correlation function (Jaffrézic *et al.* 2003). On the other hand, random regression models dealt poorly with all the different covariance structures considered here, even when a cubic polynomial was used (involving 36 parameters for the covariance structure).

Simulations with unstructured covariance models: To understand better the abilities and limitations of the different models, several patterns of covariance structures were investigated. To avoid favoring any of the methodologies, data were simulated with unstructured covariance matrices. A total of 2000 animals were considered with five observations for each trait. As focus was on the cross-correlation modeling, quite simple structures for the variances and correlations of both variables were chosen. Three examples are presented here.

In the first case, the data were generated using a cross-correlation that was stationary, symmetric, with quite high values. With regard to the likelihood value (see Table 1), a simple bivariate linear random regression model was found to be the most appropriate, followed by the bivariate CP models and then the SAD models (all models had about the same number of parameters: from 12 to 14). Estimated cross-correlations obtained with the unstructured model and the bivariate linear random regresssion model are presented in Figure 1.

In the second example, the cross-correlation was more complex. Although the correlations between the traits were still quite high, they were nonstationary and nonsymmetric. The bivariate quadratic random regression model did not converge and, on the other hand, the linear bivariate model was not able to deal adequately with this cross-correlation pattern. It was found for the character process model that the nonstationary extension, using only one extra parameter (parameter ℓ in Equation 6), considerably improved the fit as shown in

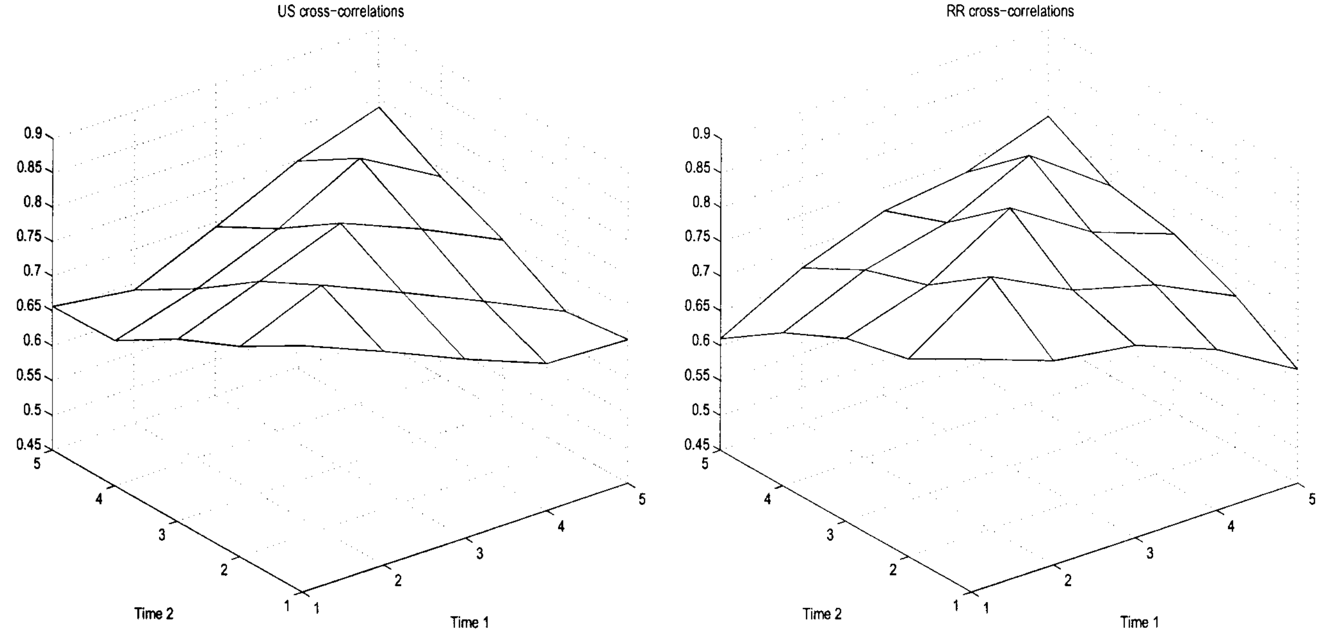


Figure 1.—Estimated cross-correlations for example 1 of the simulation study for the unstructured model (US) and a bivariate linear random regression model (RR).

Table 1. The likelihood value was then higher than that for the second-order SAD model with the same number of parameters. Figure 2 gives the estimated cross-correlations obtained with the unstructured model and with the chosen bivariate CP model.

In the third example, the data were also generated with nonsymmetric and nonstationary cross-correlations, with lower values than those for the first two examples. The diagonal cross-correlations were lower for early ages and then increasing and decreasing for late ages. The likelihood value was higher for SAD(2) than for all the other models. It can be seen, however, in Figure 3, that this model was not able to adequately fit the diagonal cross-correlation terms. On the other hand, although the likelihood value was a little lower than that with the second-order SAD model, the character process model was better able to capture the diagonal cross-correlation pattern. These figures do show, how-

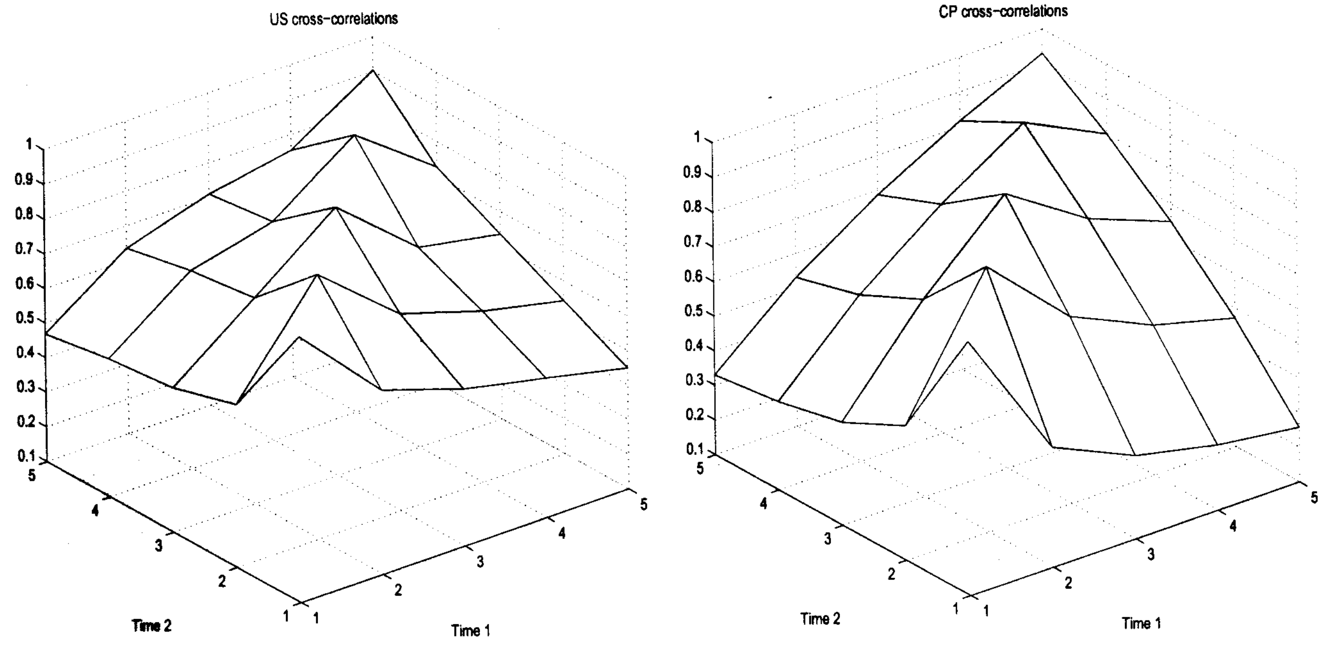
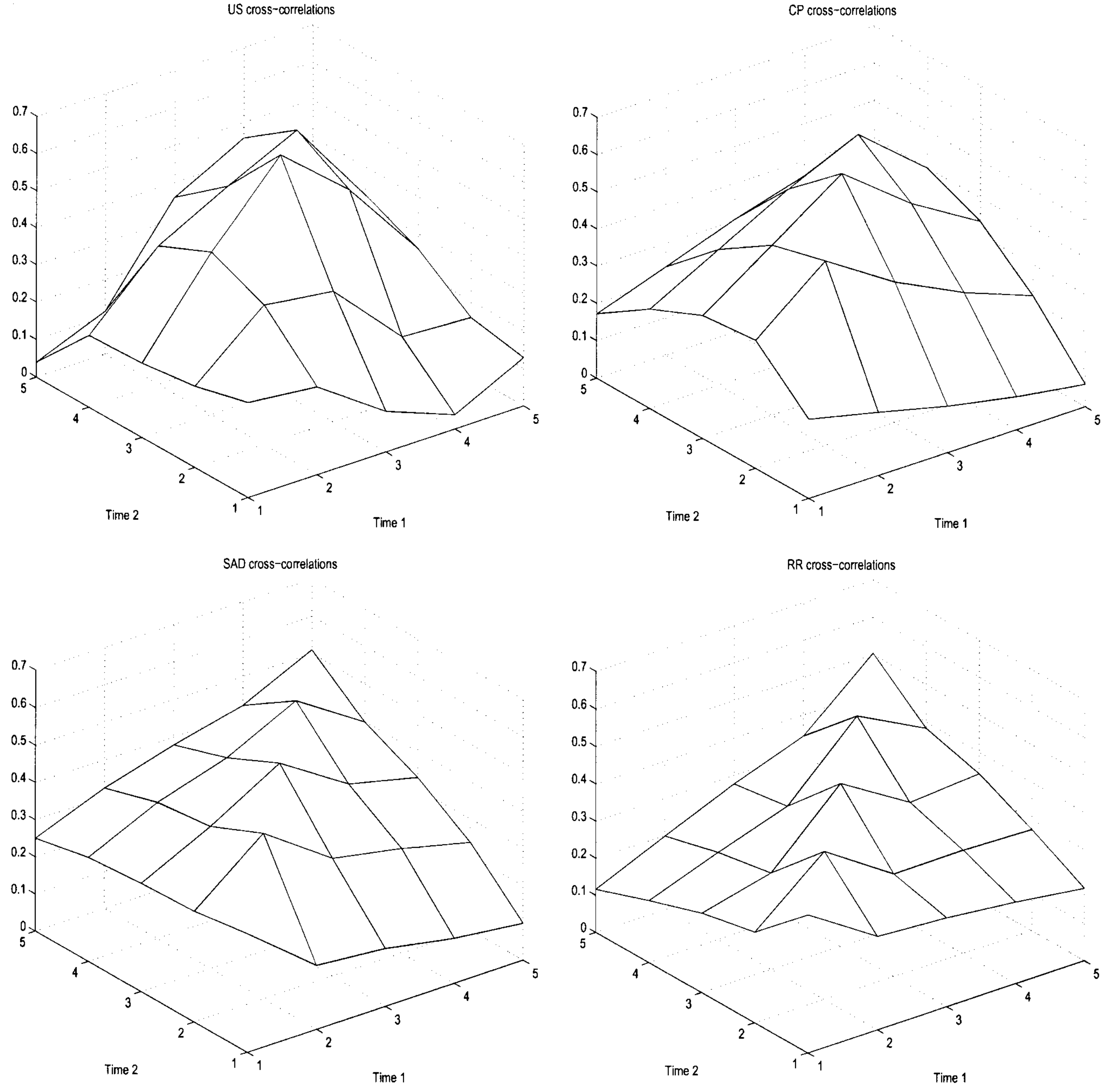


Figure 2.—Estimated cross-correlations for example 2 of the simulation study for the unstructured model (US) and the chosen bivariate CP model: quadratic polynomial used to model $V(t)$, exponential function for $\mathbf{\Omega}(t-s)$ with the nonstationary extension (Equation 6).


US cross-correlations

CP cross-correlations

SAD cross-correlations

RR cross-correlations

FIGURE 3.—Estimated cross-correlations for example 3 of the simulation study, with data simulated with an unstructured covariance matrix. [US, unstructured covariance matrix; CP, quadratic polynomial used to model $V(t)$, exponential function for $\mathbf{\Omega}(t - s)$ with the nonstationary extension (Equation 6); SAD, second-order bivariate structured antedependence model; RR, linear random regression model.]

ever, that even for the chosen models, there is still scope for improving the fit, although this might be difficult while keeping the number of parameters reasonably low.

Empirical data—joint analysis of fecundity and mortality in Drosophila: Age-specific measurements of reproduction and mortality rates were obtained from 56 different recombinant inbred (RI) lines of *D. 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. Deaths were observed 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.

Parameter estimates for the different methodologies were obtained with ASREML using the OWN function (GILMOUR *et al.* 2002). Models were compared using the

TABLE 2

Likelihood values and BIC criterion (Schwarz 1978) for univariate and bivariate genetic analyses of fecundity and mortality in Drosophila

	Genetic		Environmental				
	Corr.	Var.	Corr.	Var.	NPCov	Log L	BIC
Univariate							
Mortality	Cauchy	Quad.	Cauchy	Lin.			
Fecundity	Exp. NS	Const.	Cauchy NS	Quad.	15	329.0	186.7
Mortality	Cauchy NS	Quad.	Cauchy NS	Quad.			
Fecundity	Cauchy NS	Quad.	Cauchy NS	Quad.	20	337.2	175.6
Bivariate	Cauchy NS	Quad-Const.	Cauchy NS	Lin-Quad.	23	377.9	204.8
	Cauchy NS	Quad.	Cauchy NS	Quad.	28	380.6	188.2
	Exp. NS	Quad.	Cauchy NS	Quad.	28	370.2	177.8
	Cauchy	Quad.	Cauchy	Quad.	26	352.9	168.2
	Exp. NS	Quad.	Exp. NS	Quad.	28	354.6	162.2

In both cases the logarithms of the variances were modeled, such as $\ln v^2(t) = a + bt + ct^2$ and $\mathbf{ln}(\boldsymbol{V}(t)) = \boldsymbol{A} + \boldsymbol{B}t + \boldsymbol{C}t^2$ with $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ 2×2 symmetric matrices. Corr., correlation; Var., variance; Quad., quadratic; Lin., linear; Exp., exponential; Const, constant.

BIC criterion (Schwarz 1978; Jaffrézic and Pletcher 2000): BIC $= \ln L - 0.5 n_c \ln(N - p)$, where $\ln L$ is the REML likelihood value, n_c is the number of covariance parameters in the model, p is the number of fixed effects, and N is the total number of observations. Standard likelihood-ratio tests could be used for nested models. Specific cases include testing if certain parameters in matrices $\boldsymbol{V}(t)$ or $\boldsymbol{\Theta}$ are equal to zero. A nonparametric mean function was used for both traits (*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).

The best models chosen in the univariate analyses are given in the first part of Table 2. For the genetic part, a Cauchy correlation with quadratic variance was chosen for mortality and a nonstationary exponential correlation with a constant variance was chosen for fecundity. Many different correlation and variance functions were investigated for the bivariate analysis and the best ones regarding the likelihood value and BIC criterion are given in Table 2. In the bivariate model, the correlation function has to be the same for the two variables and was chosen here to be a nonstationary Cauchy correlation (with parameter ℓ of the nonstationary extension as in Equation 6). For the variance function, more flexibility can be achieved in the choice of the function by setting some parameters of matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ to zero. In the bivariate model, the chosen function was, as in the univariate case, quadratic for mortality and constant for fecundity. Estimates obtained for the variance and correlation functions for fecundity and mortality were very similar with the univariate and bivariate models (although their analytical forms were different, as shown in the methodology section). The main improvement of the bivariate model lies in its ability to model the cross-covariance structure. The likelihood value of the bivariate model (Log $L = 377.9$) was indeed much higher than that for the two univariate analyses (Log $L = 329.0$). Therefore, taking into account the correlation function between the two variables fits the actual process much better. Estimates obtained for the chosen bivariate model are given in Table 3 and the first graph of Figure 4 gives the genetic cross-correlation estimates. They were found to be negative at all ages, nonstationary and nonsymmetric. Fecundity and mortality were more strongly negatively correlated at a similar age (diagonal terms), and the correlation intensity decreased when ages became farther apart.

As they allow a simple and straightforward extension to the multivariate case, random regression models (RRM) are most often used for multivariate analyses of longitudinal data. They may not always, however, be the most appropriate methodology. In this example, for instance, the likelihood value was much higher for the character process approach (Log $L = 377.9$) than for a bivariate quadratic random regression model (Log $L = 134.7$), despite having far more parameters (42 for the RRM compared to 23 for the CP model). Moreover, increasing the order of the polynomials dramatically increases the number of parameters (for instance, from quadratic to cubic: 42 to 72 parameters).

Although the difference was not as important as for random regression models, the likelihood value was also higher, in this example, for the bivariate CP model than for a bivariate structured antedependence model (Jaffrézic *et al.* 2003; Log $L = 322.8$, 24 parameters).

TABLE 3

Parameter estimates (and standard errors) for the bivariate genetic analysis of fecundity and mortality in Drosophila with the best-fitting bivariate character process model, for the BIC criterion, given in Table 2

Parameters	Genetic	Environmental	Parameters	Genetic	Environmental
θ_1	0.49(0.22)	5.82(2.45)	b_1	14.91(2.14)	−0.04(0.19)
θ_2	1.20(0.61)	18.17(8.57)	b_2	0.0	0.04(0.70)
γ_1	−0.71(0.44)	−1.16(0.30)	b_3	0.0	−2.22(0.44)
γ_2	0.18(0.31)	2.32(0.83)	c_1	−16.64(2.19)	0.0
a_1	−2.71(0.46)	−0.92(0.13)	c_2	0.0	0.75(0.56)
a_2	−1.99(0.14)	−2.40(0.18)	c_3	0.0	1.46(0.36)
a_3	−0.59(0.07)	0.41(0.12)	ℓ	0.37(0.14)	0.43(0.11)

The variance functions are defined by $\mathbf{ln}(\boldsymbol{V}(t)) = \boldsymbol{A} + \boldsymbol{B}t + \boldsymbol{C}t^2$, where t = age/10 and

$$\boldsymbol{A} = \begin{pmatrix} a_1 & a_3 \\ a_3 & a_2 \end{pmatrix}$$

and similarly for matrices $\boldsymbol{B}$ and $\boldsymbol{C}$. Parameters θ_1, θ_2, γ_1, and γ_2 define matrix $\boldsymbol{\Theta}$ as specified in APPENDIX A for the Cauchy correlation function, and ℓ is the nonstationary parameter (Equation 6).

The estimated genetic cross-correlations obtained with the three methodologies are presented in Figure 4. Their patterns were found to be very different, even between the bivariate CP and SAD models, although there was only a small difference in their likelihood values. As the true genetic cross-correlations are not known, it is difficult, however, to know which pattern is the closest to reality and how much discrepancy still remains compared to the actual values.

To address these issues, a phenotypic analysis was performed on these data, which allows us to obtain estimates for an unstructured covariance matrix (22 × 22). This was not possible in the genetic study due to the very large number of parameters to be estimated. Estimated phenotypic cross-correlations obtained with the different models are presented in Figure 5 and the unstructured estimates were considered as the reference model. Once again, the four estimated patterns were found to be very different. As in the genetic analysis, the likelihood value was the highest for the character process model (= 197.1 with a nonstationary Cauchy correlation function and quadratic $V(t)$ function, with 14 parameters, BIC = 58.6), compared to a bivariate SAD(1) model (Log L = 183.8, with 12 parameters, BIC = 53.0), a bivariate SAD(2) model (Log L = 185.9, with 14 parameters, BIC = 47.4), and a quadratic bivariate random regression model (Log L = 67.7, 21 parameters, BIC = −97.7). The highest likelihood value, obtained here with the bivariate CP model, is still, however, quite far away from that of the unstructured model (Log L = 535.6). But as the number of parameters in the unstructured model is very large (= 253), its BIC value is extremely low (= −522.4), and the best model with regard to the BIC criterion here, therefore, is the bivariate CP model.

To have a measure of the discrepancy between the estimated phenotypic cross-correlations and the unstructured estimates, the Vonesh concordance coefficient (VONESH *et al.* 1996) was used, as presented by JAFFRÉZIC and PLETCHER (2000), considering the unstructured estimates as the correct values.

The concordance coefficients were 0.77 for the CP model, 0.52 for the SAD model, and 0.73 for the RR model (a perfect fit being at 1.0). As shown with the likelihood value, the bivariate character process model fit best the phenotypic cross-correlation structure. On the other hand, the goodness of fit was found higher for the bivariate random regression model than for the structured antedependence model (0.73 compared to 0.52), although the likelihood value was much higher for the SAD model (Log L = 183.8) than for the RR model (Log L = 67.7). The SAD models were therefore in this case better able to model the covariance structure for each trait separately, as in univariate analysis, whereas the random regression models were better able to fit the cross-correlation structure. The choice of the model should therefore not be made regarding the likelihood value only, but also depends on the priorities of the study. In any case, in this particular study, the character process model was more appropriate than the other two methodologies.

Figure 5 shows, however, that the obtained cross-correlation patterns were still all quite different from the unstructured phenotypic estimates and that there is still, therefore, scope for improvement.

DISCUSSION

The character process model, originally proposed by PLETCHER and GEYER (1999) to analyze function-valued traits, is based on a parametric modeling of the variance and correlation functions of a stochastic process. It mod-

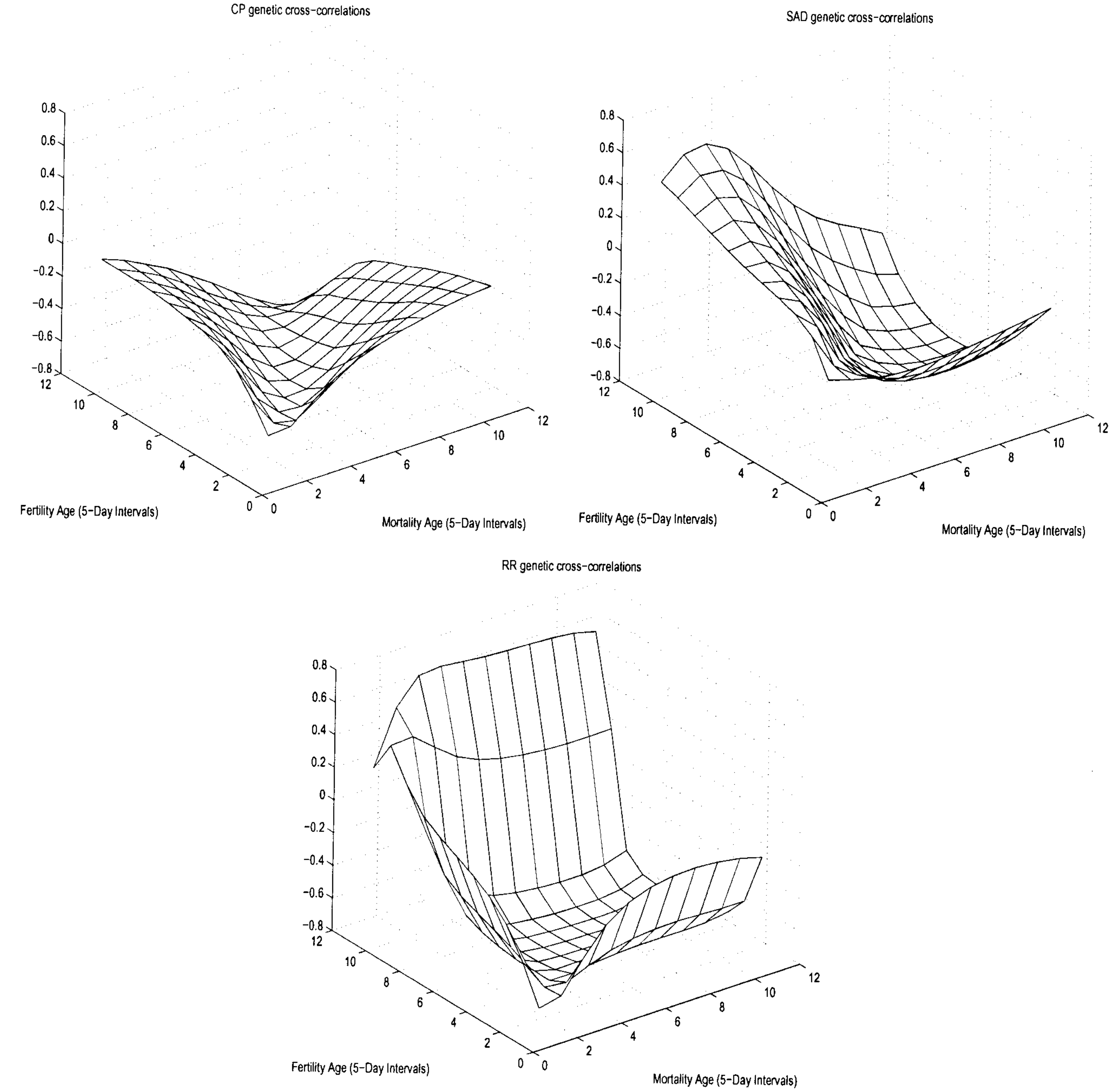


Figure 4.—Estimated genetic cross-correlations between fecundity and mortality obtained with the chosen CP model, a bivariate SAD(1) model, and a quadratic random regression model.

els the covariance structure with a small number of interpretable parameters. A special case of these models has been independently proposed in the statistical literature, namely the Ornstein-Uhlenbeck process (Taylor *et al.* 1994). It is equivalent to a character process model with an exponential correlation function and constant variances and represents a continuous time extension of a first-order autoregressive model.

We proposed an extension of the univariate character process model to the multivariate case. Our goal was to develop a method of analysis for two or more correlated function-valued traits that would retain all the desirable properties of the univariate character process approach and simultaneously allow a parametric modeling of the cross-covariance structure. The proposed extension was based on an idea presented by Sy *et al.* (1997) for the Ornstein-Uhlenbeck process and was generalized to other kinds of correlation functions, including those that are nonstationary.

Models were presented here in the bivariate case, but extension to the analysis of more than two correlated function-valued traits is straightforward and accomplished by increasing the dimensions of matrices V and Ω in accord with the number of traits analyzed.

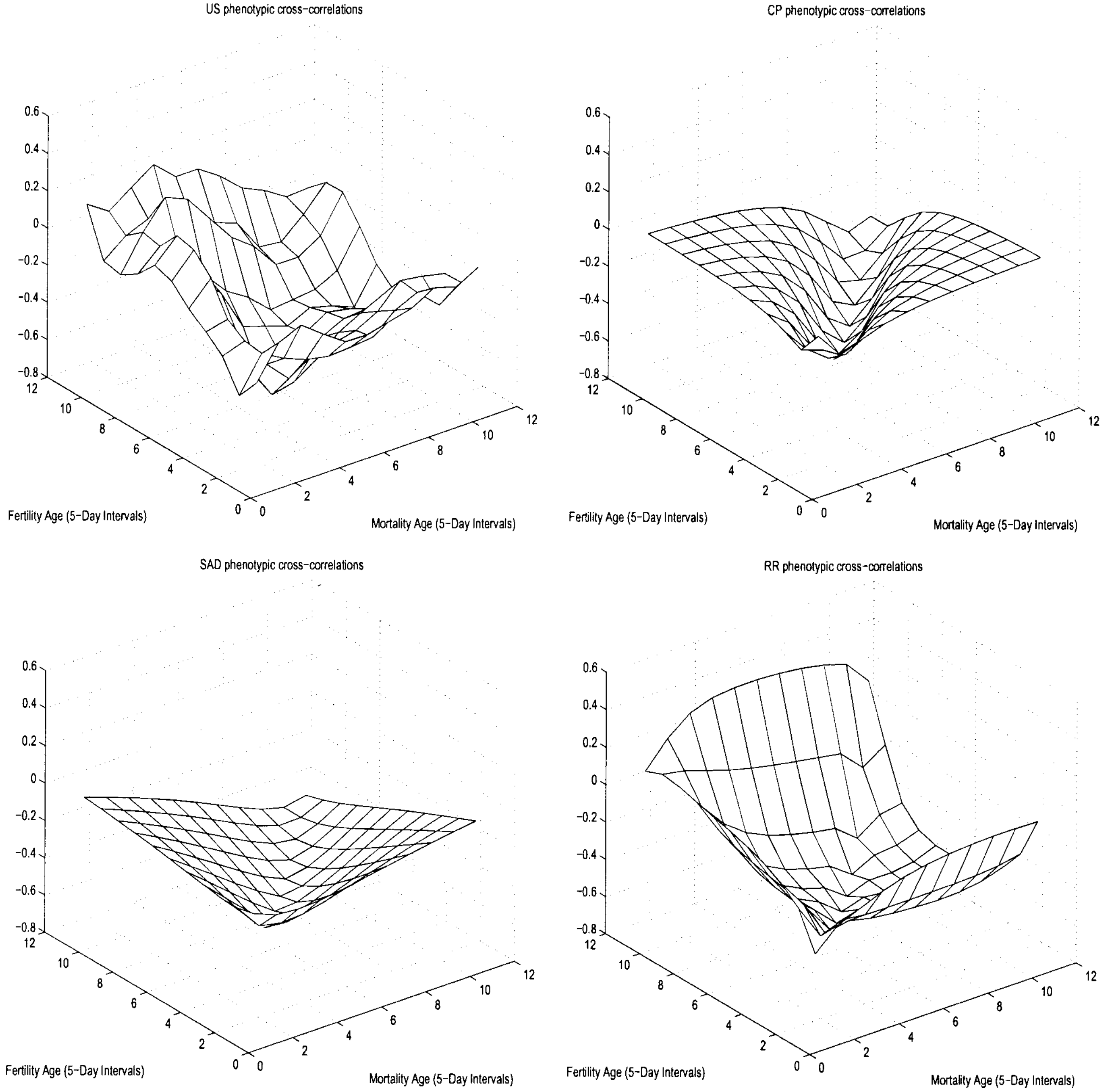


FIGURE 5.—Estimated phenotypic cross-correlations between fecundity and mortality obtained with the unstructured model (US); a character process model CP Quad-CauchyNS: quadratic polynomial used to model $V(t)$, Cauchy function for $\mathbf{\Omega}(t - s)$ with the nonstationary extension; a bivariate SAD(1) model; and a quadratic random regression model.

The first part of the simulation study highlighted the similarities between the bivariate CP models with an exponential correlation and bivariate first-order SAD models (JAFFRÉZIC *et al.* 2003), as in the univariate case. Further differences between the two approaches appear when higher orders of antedependence are considered or when other parametric correlation functions are used in the CP models.

It was found in the second part of the simulation study that the choice of the most appropriate methodology is highly dependent on the covariance structure of the data and that the three models (random regression, structured antedependent, or character process) can be worthwhile depending on the particular biological phenomenon studied. When the cross-covariance structure is symmetric and stationary with quite high correlations, the most appropriate model to use might be a simple random regression model. When the cross-correlation structure becomes more complex it should be either structured antedependence or character process models, especially because the number of parameters required in a more complex random regression model dramatically increases. For the Drosophila analysis, the bivariate character process model proved to be the most appropriate.

The multivariate extension of the character process models represents a flexible and powerful technique for the genetic analysis of two or more function-valued traits. Although the observed measurements are available only on a discrete timescale, this approach can model the fact that the underlying process is continuous and therefore can deal with highly unbalanced data. As variance parameters are assumed to change with time, other environmental factors of heterogeneity could be included in the variance modeling, as suggested by Foulley and Quaas (1995). Further research might extend these multivariate models to include the genetic analysis of nonnormally distributed traits, as studied by Pletcher and Jaffrézic (2002) in the univariate case.

We are most grateful to Jean-Louis Foulley, William G. Hill, Nancy Heckman, Jay Beder, and two anonymous referees for very interesting comments and ideas. Thanks go to J. Curtsinger and A. Khazaeli for generously providing published and unpublished data.

LITERATURE CITED

DeRisi, J. L., V. R. Iyer and P. O. Brown, 1997 Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278:** 680–686.

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

Foulley, J. L., and R. L. Quaas, 1995 Heterogeneous variances in Gaussian linear mixed models. Genet. Sel. Evol. **27:** 211–228.

Gabriel, K. R., 1962 Ante-dependence analysis of an ordered set of variables. Ann. Math. Stat. **33:** 201–212.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham and R. Thompson, 2002 *ASREML User Guide Release 1.0.* VSN International, Hemel Hempstead, UK.

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

Jaffrézic, F., I. M. S. White, R. Thompson and P. M. Visscher, 2002 Contrasting models for lactation curve analysis. J. Dairy Sci. **84:** 968–975.

Jaffrézic, F., R. Thompson and W. G. Hill, 2003 Structured antedependence models for genetic analysis of multivariate repeated measures in quantitative traits. Genet. Res. **82:** 55–65.

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

Meyer, K., 2001 Estimating genetic covariance functions assuming a parametric correlation structure for environmental effects. Genet. Sel. Evol. **33:** 557–585.

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

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

Pletcher, S. D., and F. Jaffrézic, 2002 Generalized character process models: estimating the genetic basis of traits that cannot be observed and that change with age or environmental conditions. Biometrics **58:** 157–162.

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

Pletcher, S. D., S. J. Macdonald, R. Marguerie, U. Certa, S. C. Stearns *et al.*, 2002 Genome-wide transcript profiles in aging and calorically restricted *Drosophila melanogaster.* Curr. Biol. **12** (9): 712–723.

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Stat. **6:** 461–464.

Sy, J. P., J. M. G. Taylor and W. G. Cumberland, 1997 A stochastic model for the analysis of bivariate longitudinal AIDS data. Biometrics **53:** 542–555.

Taylor, J. M. G, W. G. Cumberland and J. P. Sy, 1994 A stochastic model for analysis of longitudinal AIDS data. J. Am. Stat. Assoc. **89:** 727–736.

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

Communicating editor: M. K. Uyenoyama

APPENDIX A: IMPLEMENTATION

As suggested by Sy *et al.* (1997), to calculate the matrix exponentiation used in the correlation functions, diagonalization of matrix $\boldsymbol{\Theta}$ is used,

$$\boldsymbol{\Theta} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}, \tag{A1}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of the distinct eigenvalues θ_1 and θ_2 of $\boldsymbol{\Theta}$, and $\boldsymbol{\Gamma}$ is a 2×2 matrix whose columns are the right eigenvectors. The matrix exponential is then written and evaluated as

$$\boldsymbol{e}^{-\boldsymbol{\Theta}(t-s)} = \boldsymbol{\Gamma}\boldsymbol{e}^{-\boldsymbol{\Lambda}(t-s)}\boldsymbol{\Gamma}^{-1}. \tag{A2}$$

For the exponential correlation,

$$\mathbf{exp}(-\boldsymbol{\Theta}(t-s)) = \begin{pmatrix} 1 & \gamma_2 \\ \gamma_1 & 1 \end{pmatrix} \begin{pmatrix} e^{-\theta_1(t-s)} & 0 \\ 0 & e^{-\theta_2(t-s)} \end{pmatrix} \begin{pmatrix} 1 & \gamma_2 \\ \gamma_1 & 1 \end{pmatrix}^{-1}, \tag{A3}$$

where parameters γ_1 and γ_2 are the elements of matrix $\boldsymbol{\Gamma}$ (Sy *et al.* 1997). The Gaussian is similar, with $(t-s)$ being replaced by $(t-s)^2$. For the Cauchy correlation, taking advantage of the fact that $\boldsymbol{\Theta}^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^{-1}$, it follows that

$$\begin{aligned}(\boldsymbol{I} + \boldsymbol{\Theta}(t-s)^2)^{-1} &= \begin{pmatrix} 1 & \gamma_2 \\ \gamma_1 & 1 \end{pmatrix} \\ &\quad \times \begin{pmatrix} 1/(1+\theta_1(t-s)^2) & 0 \\ 0 & 1/(1+\theta_2(t-s)^2) \end{pmatrix} \\ &\quad \times \begin{pmatrix} 1 & \gamma_2 \\ \gamma_1 & 1 \end{pmatrix}^{-1}.\end{aligned}$$

For the variance functions $\boldsymbol{V}(t)$, an eigenvalue decomposition, $\mathbf{ln}\ \boldsymbol{V}(t) = \boldsymbol{P}(t)\boldsymbol{\Delta}(t)\boldsymbol{P}'(t)$, can also be used. It follows that $\boldsymbol{V}(t)^{1/2} = \boldsymbol{P}(t)\mathbf{exp}(\tfrac{1}{2}\boldsymbol{\Delta}(t))\boldsymbol{P}(t)'$.

Parameter estimations were obtained using the OWN function of ASREML (Gilmour *et al.* 2002), which requires us to provide the first derivatives of the covariance matrix with respect to each parameter. The nonstationary parameter ℓ of Equation 6 is obtained at the same time as the other parameters of the covariance matrix.

APPENDIX B: PROPERTIES OF THE DEFINED BIVARIATE CHARACTER PROCESS COVARIANCE FUNCTION

When J times of measurement are available for two variables Y_1 and Y_2, and for each individual i, observations are ordered as $\boldsymbol{y}_i = (y_{i11}, y_{i21}, \ldots, y_{i1J}, y_{i2J})'$. The whole genetic covariance matrix $\boldsymbol{G}$ of dimension $(2J \times 2J)$ can be written as $\boldsymbol{G} = \boldsymbol{V\Omega V'}$. By construction (Equation 5), matrix $\boldsymbol{G}$ will be symmetric. Matrix $\boldsymbol{V}$ is block diagonal: $\boldsymbol{V} = (\boldsymbol{V}_j)_{j=1,J}$, where $\boldsymbol{V}_j$ are 2×2 matrices defined by $\boldsymbol{V}_j = (\boldsymbol{V}(t_j))^{1/2}$, where $\mathbf{ln}\ \boldsymbol{V}(t_j) = \boldsymbol{A} + \boldsymbol{B}t_j + \boldsymbol{C}t_j^2$, or is specified as in Equation 10. In both cases, matrices $\boldsymbol{V}_j$, for $j = 1, \ldots, J$, are positive definite. Matrix $\boldsymbol{\Omega}$ is a $2J \times 2J$ symmetric matrix defined, for $(i, j = 1, \ldots, J)$, by $\boldsymbol{\Omega}(2(i-1) + 1{:}2i, 2(j-1) + 1{:}2j) = \boldsymbol{\Omega}_{ij}$, where $\boldsymbol{\Omega}_{ij} = (\mathbf{exp}(-\boldsymbol{\Theta}(t_i - t_j)))_{1 \leq j \leq i}$ and $\boldsymbol{\Omega}_{ji} = \boldsymbol{\Omega}'_{ij}$, if an exponential function is considered. In this case, matrix $\boldsymbol{\Omega}$ is defined as for the bivariate Ornstein-Uhlenbeck process (Sy *et al.* 1997) and therefore satisfies the positive definiteness property. When considering other functions as proposed in the univariate case by Pletcher and Geyer (1999), such as Gaussian or Cauchy, the property is maintained. Therefore, the proposed function for the bivariate CP model satisfies the theoretical requirements of a covariance function as it is symmetric and positive definite.