Copyright © 2000 by the Genetics Society of America

Statistical Models for Estimating the Genetic Basis of Repeated Measures and Other Function-Valued Traits

Florence Jaffrézic* and Scott D. Pletcher†

**Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, Scotland and* †*Max Planck Institute of Demographic Research, D-18057 Rostock, Germany*

Manuscript received February 18, 2000
Accepted for publication June 26, 2000

ABSTRACT

The genetic analysis of characters that are best considered as functions of some independent and continuous variable, such as age, can be a complicated matter, and a simple and efficient procedure is desirable. Three methods are common in the literature: random regression, orthogonal polynomial approximation, and character process models. The goals of this article are (i) to clarify the relationships between these methods; (ii) to develop a general extension of the character process model that relaxes correlation stationarity, its most stringent assumption; and (iii) to compare and contrast the techniques and evaluate their performance across a range of actual and simulated data. We find that the character process model, as described in 1999 by Pletcher and Geyer, is the most successful method of analysis for the range of data examined in this study. It provides a reasonable description of a wide range of different covariance structures, and it results in the best models for actual data. Our analysis suggests genetic variance for Drosophila mortality declines with age, while genetic variance is constant at all ages for reproductive output. For growth in beef cattle, however, genetic variance increases linearly from birth, and genetic correlations are high across all observed ages.

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 advancements have made the genetic analysis of such *function-valued* traits readily accessible. Three methods have been advanced in the literature. First, random regression models have been widely used for the analysis of longitudinal data in the traditional statistical literature (Diggle *et al.* 1994) and recently have been applied in the animal breeding context (Jamrozik *et al.* 1997b). Second, the use of orthogonal polynomials to approximate covariance matrices was initially suggested by Kirkpatrick and Heckman (1989) and is closely related to the random regression models (Meyer and Hill 1997; Meyer 1998). Third, the character process model was recently proposed by Pletcher and Geyer (1999) and is based on theories of stochastic processes. 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, and they are freely available.

The aim of this article is to compare and contrast random regression, orthogonal polynomials, and character process models and evaluate their performance. We focus first on examining the underlying assumptions of the three methods, while emphasizing fundamental similarities and differences when appropriate. Next, we explore a variety of simulated data sets and describe the types of covariance structures (genetic, environmental, and otherwise) accommodated by each method. Last, 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.

THE GENETIC ANALYSIS OF FUNCTION-VALUED TRAITS

Detailed descriptions of the extension of classical quantitative genetics to the analysis of function-valued traits is given in 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

Corresponding author: Scott D. Pletcher, Department of Biology, Wolfson House, 4 Stephenson Way, University College, London NW1 2HE, England. E-mail: s.pletcher@ucl.ac.uk

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 in this article 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.

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

$$X(t) = \mu(t) + g(t) + e(t) + \varepsilon, \tag{1}$$

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 original development of the character process (CP) model did not include a residual variance term (Pletcher and Geyer 1999). Recently, however, we have found that data sets that exhibit a great deal of measurement error support a residual variance.

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, $r(s, t) = \mathrm{Cov}\{X(s), X(t)\}$. By the independence of $g(t)$ and $e(t)$, the phenotypic covariance function of $X(t)$ is given by $P(s, t)$ as

$$P(s, t) = G(s, t) + E(s, t), \tag{2}$$

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 three different methods suggested for estimating the desired covariance functions: orthogonal polynomials (Kirkpatrick and Heckman 1989), random regression (Meyer 1998), and the character process model (Pletcher and Geyer 1999). All three 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: Random regression (RR) models employ parametric forms for the unobserved functions in (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 modeled 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

$$\begin{aligned} G(s, t) &= \mathrm{Cov}(g(s), g(t)) \\ &= \mathrm{Cov}(a_1 + a_2 s, a_1 + a_2 t) \\ &= \mathrm{Var}(a_1) + (s + t)\mathrm{Cov}(a_1, a_2) + st\,\mathrm{Var}(a_2). \end{aligned} \tag{3}$$

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: 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 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

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

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.). 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).

We suggest an extension of the character process model for nonstationary correlations using a method proposed by Nunez-Anton (1998) and Nunez-Anton and Zimmerman (2000) in what they term structured antedependence models. The idea is to implement a nonlinear transformation upon the time axis, $f(t)$, such that correlation stationarity holds on the transformed scale—on the original scale the correlation is nonstationary. 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

$$f\lambda(t) = \begin{cases} (t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log t & \text{if } \lambda = 0, \end{cases} \quad (5)$$

where λ is a parameter to be estimated. Considering an absolute exponential correlation function, $\rho(s, t) = \theta^{|f(s) - f(t)|}$, the correlations on the subdiagonals are monotone increasing if $\lambda < 1$ or monotone decreasing if $\lambda > 1$. If $\lambda = 1$ the nonstationary 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 nonstationarity in the data. Additional flexibility in the nonstationary 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 subdiagonal of the covariance structure.

Orthogonal polynomials: Kirkpatrick and Heckman (1989) originally presented the use of orthogonal polynomials (OPs) as a nonparametric 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

$$G(s, t) = \sum_{i=0}^{m} \sum_{j=0}^{m} \phi_i(s)\phi_j(t)k_{ij}, \quad (6)$$

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 ith Legendre polynomial (Kirkpatrick *et al.* 1990). The environmental covariance function is modeled similarly. Meyer and Hill (1997) present a method for estimating covariance functions such as (6) 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). The analyst begins by considering the behavior 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 2), and the shapes of the individual deviations are a consequence of this structure. In some cases it may be possible to expose a duality between the two, as Meyer (1998) has done for certain RR and OP models. 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. At present, however, analytical difficulties preclude more general results for the character process models.

EXAMPLES AND ANALYSES

Estimation procedures: All models were estimated using REML. In all cases a nonparametric 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 nonnested models. BIC is

$$\text{log-likelihood} - \tfrac{1}{2} \times \text{number of parameters in the model} \times \log n^*,$$

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.

To determine the best-fitting model under each technique, a large number of models were fit to each data set. For the character process method, >100 different models (*i.e.*, different combinations of polynomial variance functions and stationary and nonstationary 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 behavior 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 on the order of 5–10 model fits for each method.

Estimates of the covariance structure based on random regression and orthogonal polynomial methods were obtained using the software package ASREML (Gilmour *et al.* 1997), while estimates of the character process model (and certain orthogonal polynomial models) were obtained using computer software developed by one of the authors (S. Pletcher; 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.

Simulated data: Many data sets were simulated according to various covariance structures from CP, RR, and OP models. 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 (Lynch and Walsh 1998). 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 magnitudes of the variance and covariances were different among the simulations, we set the residual variance for all simulations to ~10% of the total variance at age 0.

The first data set was simulated according to a stationary CP covariance structure, the purpose of which was to assess the behavior of 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 4] and "normal" correlation ($\rho(t_i, t_j) = \exp(-0.8(t_i - t_j)^2)$) (Figure 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 nonstationary correlation, we simulated data with genetic variance function identical to that in the stationary CP data, but with an arbitrary nonstationary correlation structure (Figure 1B). The environmental covariance was assumed identical to that in the stationary CP data. This data set is the nonstationary 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 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 6). 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 nonstationary (Figure 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.

Empirical data: *Drosophila reproduction and mortality:* 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 to insure 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 six records per cow. Analyses were 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).

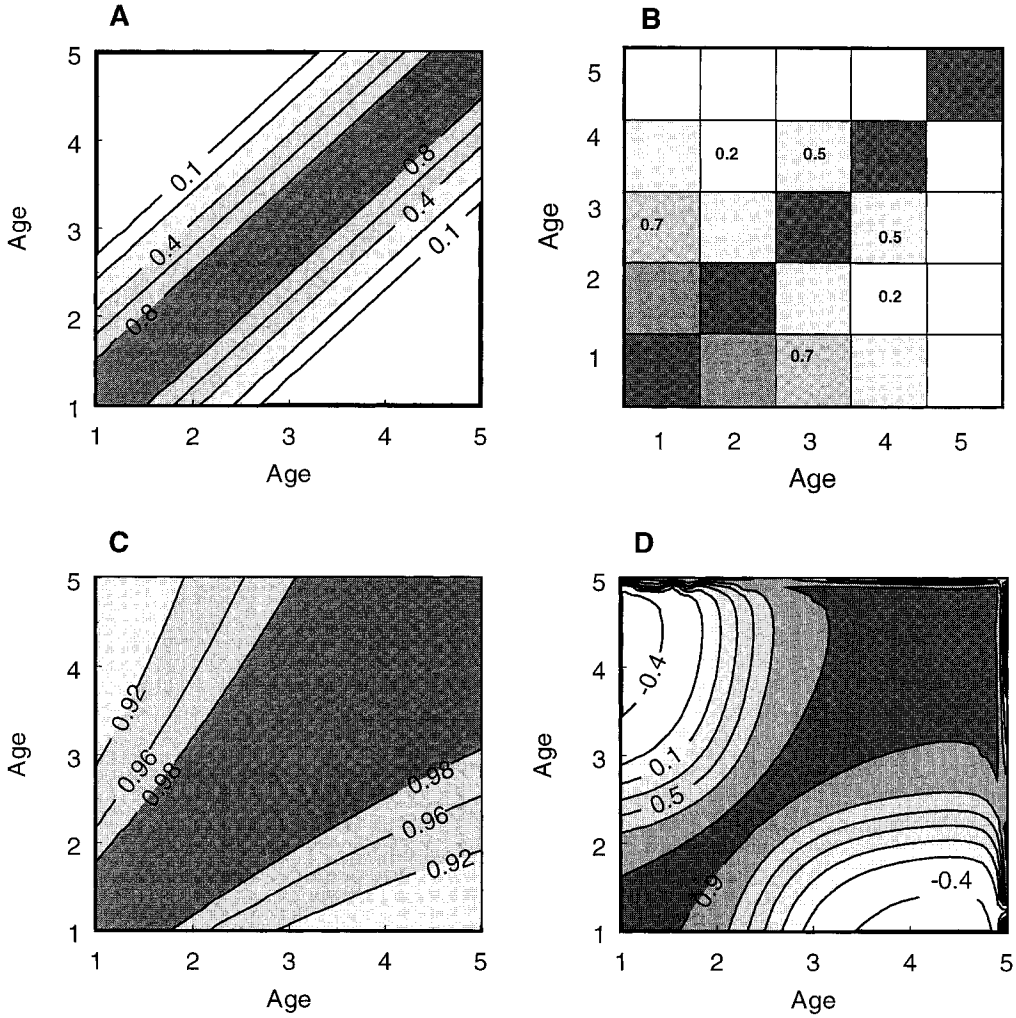

FIGURE 1.—Contour plots of the simulated genetic covariance structures for (A) data generated according to a stationary character process (CP) model, (B) data simulated according to a CP model with arbitrary and nonstationary correlation (this is a discrete value matrix rather than a continuous function), (C) data generated under a random regression (RR) model with linear deviations, and (D) data simulated assuming an orthogonal polynomial (OP) model of degree 2.

RESULTS

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 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 modeled, 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 behavior. 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).

The aim of the second simulated data set was to investigate the behavior of these models in the case of a rather simple nonstationary 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 nonstationary 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 nonstationary correlation pattern (Table 1). Our nonstationary extension did not significantly improve the goodness of fit (BIC = −4454 and −4456 for stationary and nonstationary models, respectively; $P = 0.052$ for a likelihood-ratio test of $\lambda = 1.0$). However, the goodness of fit of the fitted nonstationary 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

TABLE 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
	RR	0.96	0.36	0.93	0.87	−7414
	OP	0.98	0.61	0.98	0.98	−6605
Nonstationary CP	CP	0.91	0.03	0.99	1.0	−4454
	RR	0.95	0.10	0.94	0.81	−7397
	OP	0.84	0.70	0.98	0.97	−6628
Random regression	CP[a]	1.0	0.93	0.96	0.93	−3817
	RR	1.0	0.94	0.99	1.0	−3803
	OP	1.0	0.94	0.99	1.0	−3803
Orthogonal polynomial	CP[a]	0.86	0.10	0.69	0.94	−14334
	RR	0.30	0.15	0.94	0.90	−14371
	OP	0.99	0.83	0.99	1.0	−14272

Concordance values (see APPENDIX) for covariance functions estimated by three different methods on four representative covariance structures. The methods are CP, the character process model; RR, the random regression model; and OP, the orthogonal polynomial model. VarG represents the fit to age-specific genetic variances; CorrG refers to the fit to genetic correlations between ages; VarE represents the fit to environmental variances; and CorrE shows the fit to environmental correlations between ages. See text for details of the simulated covariance structures and details of the best-fitting models for each approach.

[a] The best-fitting correlation function was a nonstationary CP model.

retrospect, the nonstationarity 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 nonstationarity (see fly reproduction data). The genetic variance function and environmental covariance structure were identical to that for the stationary CP data and were well fit by all the methods (Table 1).

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 nonstationary 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 nonstationary 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 nonstationary CP models, respectively).

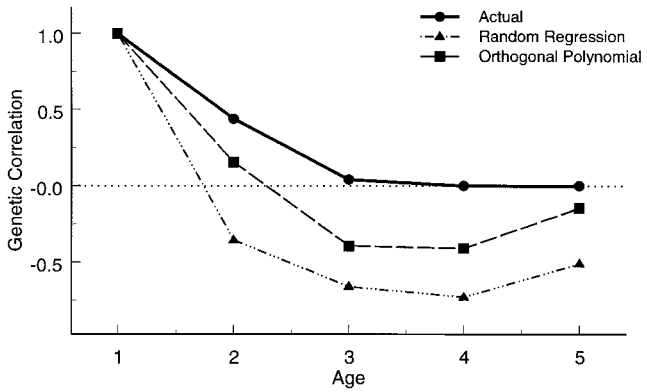

FIGURE 2.—Genetic correlations between age 1 and other for the simulated stationary character process data and fitted genetic correlations obtained from the random regression model with linear deviations and orthogonal polynomial of degree 3.

Although we have essentially no idea what a typical age-dependent genetic covariance function might look like, 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 1D). Such a structure 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 structures were described by a quadratic variance and nonstationary 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

TABLE 2

Results of covariance function estimation on empirical data

	Method	Genetic	Environmental	NPCov	Log-likelihood	BIC
			Fly mortality ($N = 955$)			
11 fixed effects	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[a]	Quad-Cauchy[a]	8	494.1	427.5
	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 genetic and environmental covariance functions for three different methods using empirical data on fruit fly mortality and reproduction and growth in beef cattle. Also presented is the log-likelihood of the models at their maximum and the BIC model selection criterion. NPCov represents the number of estimated parameters in the covariance structure for each model. The number of fixed effects reflects the number of different ages at which observations were obtained, and N is the total number of observations. Quad, quadratic; Const, constant; Exp, exponential; Lin, linear.

[a] The best-fitting correlation function was a nonstationary CP model.

1). The environmental covariance is better behaved and much less of a problem. As seen with the random regression simulations, the strong positive correlations across all ages are well fit by all the methods.

Empirical: *Drosophila reproduction and mortality:* For age-specific mortality and reproduction in Drosophila, the character process model provided a significantly better fit, according to the BIC criterion, than either the orthogonal polynomial or random regression methods (Table 2). In fact, the CP models achieved higher likelihoods despite containing significantly less parameters than the OP or RR models. For age-specific mortality, the best-fitting model for the genetic covariance was a quadratic variance with a Cauchy correlation function ($\rho_G(t_i, t_j) = 1/(1 + \theta(t_i - t_j)^2)$). For fly reproduction the best character process model was a constant variance at all ages coupled with a nonstationary correlation function described by the absolute exponential, $\rho_G(t_i, t_j) = \theta^{|f(t_i)-f(t_j)|}$ (see text following Equation 5). Parameter estimates and their standard errors for the CP model are presented in Table 3, and the fitted genetic covariance structures are presented in Figure 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 nonstationary parameter in the genetic correlation provides evidence for an increase in the correlation between two equidistant ages with increasing age.

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). 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 nonstationarity in the data. Parameter estimates and their standard errors for the CP model are presented in Table 3, and the fitted genetic covariance structure is shown in Figure 3C.

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

TABLE 3

Character process model estimates of genetic and environmental covariance functions for empirical data

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[a] (186.3)	0.0001[a] (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 estimates (and standard errors) for the best-fitting character process models for empirical data on fruit fly mortality and reproduction and growth in beef cattle. θ_0, θ_1, and θ_2 represent parameters of the variance function such that a quadratic variance is represented as $v^2(t) = \theta_0 + \theta_1 t + \theta_2 t^2$. In cases where the best-fitting model was constant or linear, the appropriate θ_i are omitted. θ_C and λ are parameters of the correlation function. A residual term is not always added in the model.

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

process theory (*i.e.*, the character process model) have provided important alternatives. Other types of random regression models (*e.g.*, nonlinear 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 the CP models provide the best description of the underlying covariance structure. It is clear from the simulation results that the CP model is the only method that adequately captures a correlation that declines rapidly to zero as character values become further separated in time. Both random regression models and orthogonal polynomials have noticeable problems approximating such a structure (Table 1, stationary CP data; Figure 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 behavior 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 work quite well when the correlation structure remains high over time (see Table 1, environmental correlation in CP simulated data).

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 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 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 (4 parameters) than random regression (10 parameters) or orthogonal polynomial (17 parameters) models (Table 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 regres-

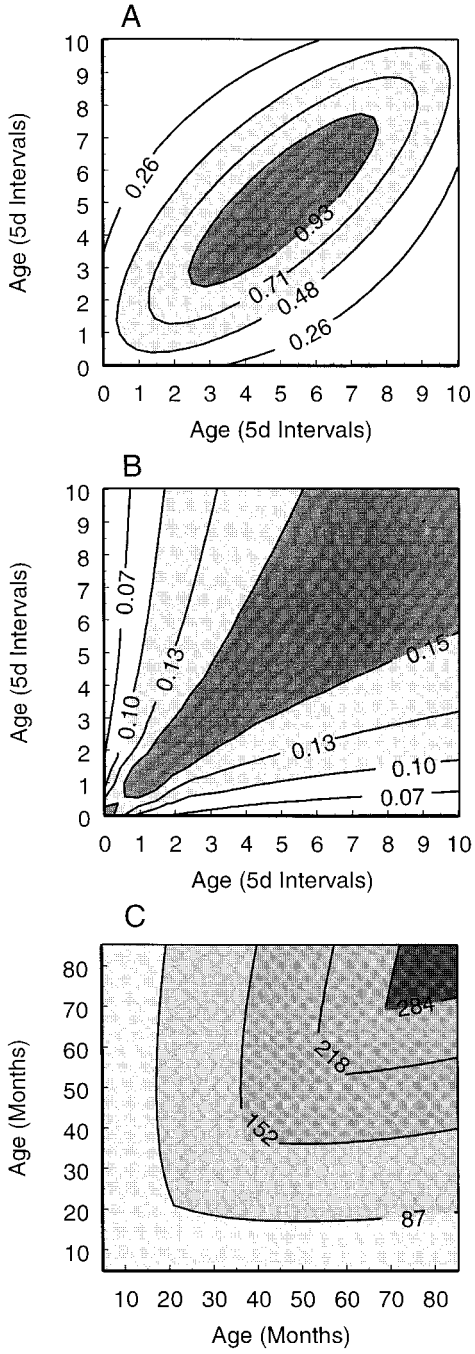

FIGURE 3.—Contour plots of genetic covariance functions fitted by the character process model. (A) Age-specific mortality in the fruit fly, *Drosophila melanogaster*; (B) age-specific reproduction in *D. melanogaster*, (C) age-specific growth in beef cattle.

sion 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. 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).

Last, the genetic analysis of two or more function-valued traits is an important goal. Generalization of regression models to multitrait 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.* 1997a). 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.

W. Hill, N. Barton, and two anonymous reviewers provided valuable comments on the manuscript. Thanks to J. Curtsinger and A. Khazaeli for generously providing published and unpublished data. F.J. thanks the INRA for support during this project.

LITERATURE CITED

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

DIGGLE, P. J., K. Y. LIANG and S. L. ZEGER, 1994 *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

GILMOUR, A. R., R. THOMPSON, B. R. CULLIS and S. J. WELHAM, 1997 *ASREML Manual.* New South Wales Department of Agriculture, Orange, 2800, Australia.

JAMROZIK, J., L. SCHAEFFER, Z. LIU and G. JANSEN, 1997a Multiple trait random regression test day model for production traits. Proceedings of 1997 Interbull Meeting, Vol. 16, pp. 43–47.

JAMROZIK, J., L. R. SCHAEFFER and J.-C. M. DEKKERS, 1997b Genetic evaluation of dairy cattle using test day yields and random regression model. J. Dairy Sci. **80:** 1217–1226.

KIRKPATRICK, M., and N. HECKMAN, 1989 A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. J. Math. Biol. **27:** 429–450.

KIRKPATRICK, M., and D. LOFSVOLD, 1992 Measuring selection and constraint in the evolution of growth. Evolution **46:** 954–971.

KIRKPATRICK, M., D. LOFSVOLD and M. BULMER, 1990 Analysis of the inheritance, selection and evolution of growth trajectories. Genetics **124:** 979–993.

KIRKPATRICK, M., W. G. HILL and R. THOMPSON, 1994 Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. Genet. Res. **64:** 57–69.

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

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

MEYER, K, 1998 Estimating covariance functions for longitudinal

data using a random regression model. Genet. Sel. Evol. **30:** 221–240.

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

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

Nunez-Anton, V., and D. L. Zimmerman, 2000 Modeling non-stationary logitudinal data. Biometrics **56** (in press).

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

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

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

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

Communicating editor: C. Haley

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

$$r_c = 1 - \frac{\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}(y_{ij} - \hat{y}_{ij})^2}{\sum_{i,j}(y_{ij} - \bar{y})^2 + \sum_{i,j}(\hat{y}_{ij} - \hat{y})^2 + T(T-1)(\bar{y} - \hat{y})^2/2}, \quad \text{(A1)}$$

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 means 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

$$r_c = 1 - \frac{\sum_{i=1}^{T}(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2 + \sum_i(\hat{y}_i - \hat{y})^2 + T(\bar{y} - \hat{y})^2}, \quad \text{(A2)}$$

where the y's 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} *vs.* $\hat{y}_{ij}$ 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.