Genet. Res., Camb. (2002), **80**, *pp.* 47–53. *With* 3 *figures.*
DOI: 10.1017/S0016672302005700 Printed in the United Kingdom

Non-parametric exploratory analysis of the covariance structure for genetic analysis of repeated measures and other function-value traits

FLORENCE JAFFRÉZIC[1,3]*, SCOTT D. PLETCHER[2] AND WILLIAM G. HILL[3]
[1] *INRA Quantitative and Applied Genetics, 78352 Jouy-en-Josas Cedex, France*
[2] *Department of Biology, Galton Laboratory, University College, London NW1 2HE, UK*
[3] *Institute of Cell Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK*

(*Received 30 May 2001 and in revised form 22 October 2001 and 5 March 2002*)

Summary

In the analysis of longitudinal data, before assuming a parametric model, an idea of the shape of the variance and correlation functions for both the genetic and environmental parts should be known. When a small number of observations is available for each subject at a fixed set of times, it is possible to estimate unstructured covariance matrices, but not when the number of observations over time is large and when individuals are not measured at all times. The non-parametric approach, based on the variogram, presented by Diggle & Verbyla (1998), is specially adapted for exploratory analysis of such data. This paper presents a generalization of their approach to genetic analyses. The methodology is applied to daily records for milk production in dairy cattle and data on age-specific fertility in *Drosophila*.

1. Introduction

Animal breeders and evolutionary geneticists are often faced with the problem of analysing traits that change as a function of age or some other independent and continuous variable. This is the case, for example, for lactation curve analysis in dairy cattle, growth curve analysis of laboratory and agricultural species, or the study of age-specific fitness components such as reproductive output. Many techniques have already been proposed to deal with this kind of data. The most commonly used at present are random regression models (Diggle *et al.*, 1994). Another approach, called 'character process models', has recently been proposed by Pletcher & Geyer (1999), and corresponds to a parametric modelling of the covariance structure. An overview of these techniques is presented by Jaffrézic & Pletcher (2000).

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

When a small number of measures with common times of measurement is available for each subject, an unstructured covariance matrix may be estimated with standard software. However, when the number of measurements per subject is large and when data are unbalanced, for example for daily records for milk production in dairy cattle, estimation may not be feasible. The aim of this paper is to describe a non-parametric procedure that deals with this kind of data, and makes no *a priori* assumption about the model for the covariance structure. This methodology is based on the 'variogram' (Diggle & Verbyla, 1998), which is easy to implement as it only requires calculation of simple functions of the observations, and provides a useful representation of the covariance structure to help choose an appropriate parametric model.

2. Variogram approach

The focus is on the analysis of repeated measures over time, but this approach can also be applied to traits that change as a function of another independent and

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

continuous variable. In order to present the variogram methodology, first consider the case of a phenotypic analysis, and then an extension to a genetic analysis.

(i) *Phenotypic analysis*

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

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

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

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

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

where $P(s, t)$ is the phenotypic covariance function. This description of the variogram does not assume stationarity, i.e. that the variogram is a function only of the difference $|s - t|$ between the two times s and t as in classical definitions.

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

$$v_{ijk} = \tfrac{1}{2} [(y_{ij} - \mu_i(t_j)) - (y_{ik} - \mu_i(t_k))]^2. \tag{3}$$

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

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

In most applications, $\mu_i(t_j)$ is unknown and will have to be replaced by an appropriate estimate. In practice, data (y_{ij}) are pre-corrected for fixed effects considering $(y_{ij} - \hat{\mu}_i(t_j))$ with $\hat{\mu}_i$ estimated by a simple regression model. Most simply, a non-parametric mean curve is fitted in the variogram using averages $\bar{y}_{\cdot j}$. Diggle *et al.* (1994, chapter 4) provide a discussion about fixed effects estimation; but as the variogram is to be used for exploratory purposes, the aim of this estimation procedure is to be simple and computationally fast rather than statistically efficient.

(ii) *Genetic analysis*

The observed phenotypic process $Y(t)$ is assumed to be a Gaussian process and can be decomposed as:

$$Y(t) = \mu(t) + g(t) + e(t) \tag{4}$$

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

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

The linear mixed model can be written as:

$$y_{sij} = \mu_j + u_{sj} + e_{sij} \tag{5}$$

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

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

(b) *Sample variograms.* Extending results for single times, two sample variograms can be defined:

$$\bar{v}_{1jk} = \frac{1}{2(S-1)} \sum_{s=1}^{S} [(\bar{y}_{s\cdot j} - \bar{y}_{\cdot\cdot j}) - (\bar{y}_{s\cdot k} - \bar{y}_{\cdot\cdot k})]^2 \tag{6}$$

https://doi.org/10.1017/S0016672302005700 Published online by Cambridge University Press

and

$$\bar{v}_{2jk} = \frac{1}{2S(n_s - 1)} \sum_{s=1}^{S} \sum_{i=1}^{n_s} [(y_{sij} - \bar{y}_{s\cdot j}) - (y_{sik} - \bar{y}_{s\cdot k})]^2. \quad (7)$$

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

$$\gamma_1(t_j, t_k) = \tfrac{1}{2} [(v_G(t_j) + v_G(t_k) - 2G_{jk}) + \frac{1}{n_s} (v_E(t_j) + v_E(t_k) - 2E_{jk})] \quad (8)$$

and the second provides estimates for:

$$\gamma_2(t_j, t_k) = \tfrac{1}{2}(v_E(t_j) + v_E(t_k) - 2E_{jk}) \quad (9)$$

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

Extension to the unbalanced case is given in Appendix A. Implementation of the variogram genetical analysis was easy and, consequently, calculations were fast. Fortran code is available from the first author.

3. Application

(i) *Daily records in dairy cattle*

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

Fig. 1 shows the estimates of genetic and environmental variances. In order to check these non-parametric estimates, as well as their ability to deal with unbalanced data and fixed effects estimation, REML estimates were calculated with the program ASREML (Gilmour *et al.*, 2000) assuming a sire model. Diagonal covariance matrices were considered and one variance was estimated at each time (230 values). REML1 represents estimates obtained while estimating fixed effects at the same time, and REML2 represents estimates obtained on the data set previously corrected for fixed effects. Variance estimates obtained here with the three methodologies were extremely close. A similar analysis was performed for the covariance estimates. Unstructured covariance matrices for both the genetic and environmental parts were obtained using the package REMLPK (Meyer, 1985). However, as it cannot provide estimates for unstructured covariance matrices of size 230 by 230, this analysis was performed for only a few given days. Covariance estimates obtained with the non-parametric approach and with REML were also very similar.

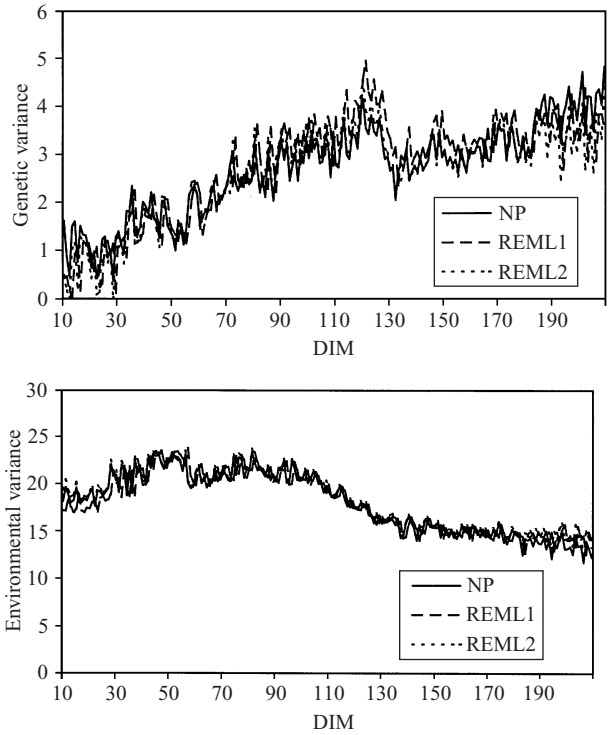

Fig. 1. Genetic and environmental variances for daily records for milk production in dairy cattle, given in kg². DIM, days in milk; NP, non-parametric estimates; REML1, REML with fixed effects estimated at the same time; REML2, REML on the data set pre-corrected for fixed effects.

As completely unstructured covariance matrices cannot be obtained with standard software for all the observed ages, this non-parametric methodology should prove to be extremely useful for studying the covariance and correlation structure for these daily records. Fig. 2 shows estimates of genetic and environmental correlations for days in milk 10, 80 and 210. As expected from previous analyses (White *et al.*, 1999), the genetic correlation was quite high for all pairs of ages (about 0·8), except for the early stage of lactation. For example, the correlation between day 210 and day 10, as well as that between day 80 and

https://doi.org/10.1017/S0016672302005700 Published online by Cambridge University Press

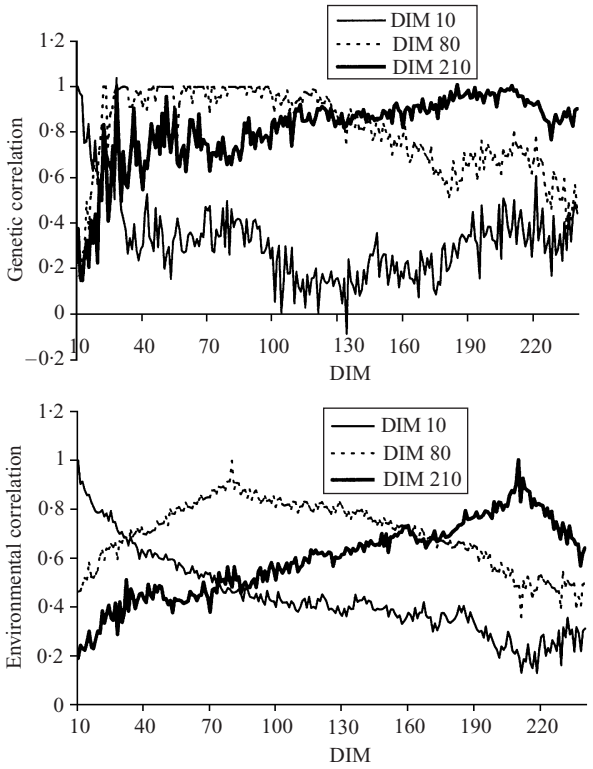

Fig. 2. Genetic and environmental correlations for daily records for milk production in dairy cattle. DIM 10, correlation between day in milk 10 and others; DIM 80, correlation between day in milk 80 and others; DIM 210, correlation between day in milk 210 and others.

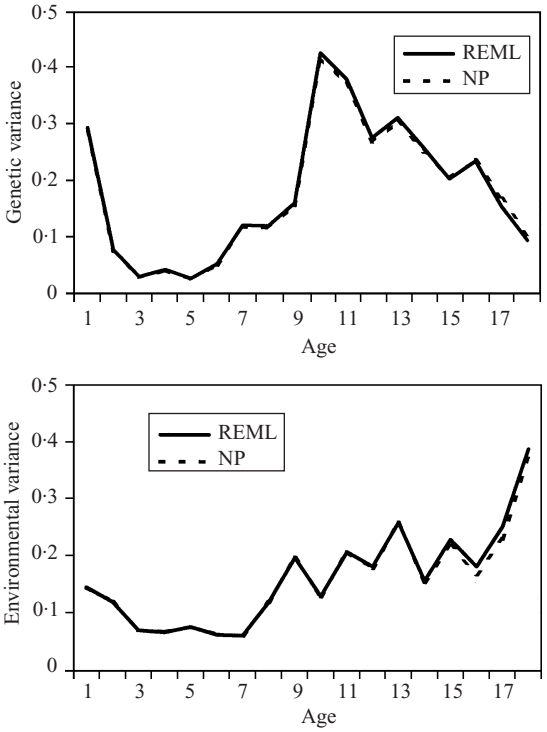

Fig. 3. Genetic and environmental variances for fertility data in *Drosophila*. Each age corresponds to a 2-day interval. NP, non-parametric estimation.

day 10, was about 0·2. For all stages of lactation, the environmental correlation was high for days in milk close in time (e.g. 0·8 between day 210 and day 190), and decreased steadily as days became further apart (e.g. 0·6 between day 210 and day 130, and 0·2 between day 210 and day 10). Raw estimates provided by the variogram were plotted on these graphs. In order to choose an appropriate parametric model it may be useful to use a smoothing procedure as proposed by Diggle & Verbyla (1998).

(ii) *Fertility data in* Drosophila

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

Fig. 3 shows estimates of genetic and environmental variances using both the non-parametric procedure presented above and a REML analysis performed with the software ASREML. For the latter, a sire model was assumed and diagonal covariance matrices were used to estimate one variance at each time. The procedures showed very similar results for both genetic and environmental parts. If a parametric model were to be chosen, a quadratic function would probably be appropriate for the environmental variance; but for the genetic variance, the choice of a parametric function may be more difficult. In fact, in a previous study (Jaffrézic & Pletcher, 2000), data were pooled into 5-day intervals, and the best parametric model for the genetic variance, using a likelihood based criterion, was a constant function estimated at 0·18. However, the variation observed here for the genetic variance with both the non-parametric and REML methodologies may be worthwhile to study. The genetic variance dropped quickly for early ages, then increased rapidly at about age 10, and decreased thereafter. The biological causes of these large changes may therefore be worth investigating.

Table 1 gives non-parametric estimates for the correlation matrices. Both the genetic and environmental correlations seem to be non-stationary, as was also found by Jaffrézic & Pletcher (2000).

https://doi.org/10.1017/S0016672302005700 Published online by Cambridge University Press

Table 1. *Non-parametric estimates for genetic* (*above diagonal*) *and environmental* (*below diagonal*) *correlation for fertility data in* Drosophila (*table gives correlation for every 4-day interval*)

	1	2	3	4	5	6	7	8	9
1	1	0·34	0·50	0·37	0·34	0·47	0·44	0·34	0·56
2	0·03	1	1·0	0·44	0·54	0·32	0·20	0·33	0·22
3	0·30	0·16	1	0·87	0·68	0·80	0·67	0·50	0·73
4	0·17	0·31	0·32	1	0·76	0·77	0·58	0·46	0·45
5	0·33	0·04	0·48	0·35	1	1·0	0·90	0·87	0·91
6	−0·04	0·22	0·20	0·17	0·07	1	0·92	1·0	0·93
7	0·16	0·09	0·35	−0·03	0·30	0·37	1	0·95	1·0
8	0·05	−0·20	0·14	0·08	0·22	0·12	0·37	1	1·0
9	−0·05	0·03	−0·13	0·03	0·00	0·24	−0·07	0·17	1

4. Discussion

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

The method presented here is easy to implement as it implies mainly sum and average calculations. The computing time required is small even for a large data set such as the daily records for milk production, especially because it is a non-iterative procedure. A large number of observations can be handled over time, and estimates for covariances and correlations between all ages are provided, which was not possible with the usual software. Moreover, it was found in a simulation study (results not shown) that the methodology can also deal with non-stationary covariance and correlation structures, and could therefore be useful to check the stationarity assumption for the parametric model.

The non-parametric procedure is equivalent to the REML in the balanced case and when no fixed effects are considered. It should, however, be used mainly for exploratory purposes as the estimates are not always statistically efficient. As pointed out by Diggle & Verbyla (1998), one of the main difficulties of this approach can be fixed effects estimation. Nevertheless, when only a few fixed effects are considered, as for the Langhill data, the non-parametric analysis on pre-corrected data performs well compared with REML, which estimates fixed effects at the same time. Another point that needs to be further investigated concerns extension to an animal model that would take into account the relationship matrix. This does not seem to be straightforward, and requires further study.

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

Appendix A. Unbalanced analysis

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

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

Let n_{dj} be the average number of daughters per sire with measures at time t_j. The sample variance $\bar{v}_{1j}$ will provide estimates for:

$$\gamma_1(t_j) = v_G(t_j) + \frac{1}{n_{dj}} v_E(t_j).$$

https://doi.org/10.1017/S0016672302005700 Published online by Cambridge University Press

A straightforward extension of this result to covariance estimates is:

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

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

This sample variogram will give estimates for:

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

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

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

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

Appendix B. Small example

A small example of simulated data is provided to allow the reader to check his or her own non-parametric program using the estimation procedure presented above. Data are given in Table B1. They were simulated according to a non-stationary character process structure with a linear variance, $\sigma_S^2(t) = \mathrm{Var}_G(t) = 0{\cdot}3 + 0{\cdot}4t$, and a non-stationary exponential correlation, $\rho_G(t, s) = \exp(-0{\cdot}8(|\frac{t^\lambda - s^\lambda}{\lambda}|))$, with $\lambda = 0{\cdot}5$ for the genetic part, and a quadratic variance, $\mathrm{Var}_E(t) = 0{\cdot}5 + 0{\cdot}6t + 0{\cdot}2t^2$, and a non-stationary exponential correlation, $\rho_E(t, s) = \exp(-0{\cdot}1(|\frac{t^\lambda - s^\lambda}{\lambda}|))$, with $\lambda = 0{\cdot}5$ for the environmental part.

A balanced sire design was considered with 10 sires, 3 daughters per sire and 4 measures per daughter. In order to check the proposed unbalanced extension, 1 daughter was deleted for 5 of the sires. Covariance parameter estimates obtained in the balanced and unbalanced cases with the non-parametric procedure are given in Table B2. As no fixed effects were considered, these estimates are equivalent to the REML in the balanced case. No additional smoothing was performed and results given in Table B2 correspond to the raw data obtained with the sample variogram.

Table B1. *Simulated data set analysed in the small example* (*animals in bold were deleted in the unbalanced case*)

Sire	Animal	Y1	Y2	Y3	Y4	Sire	Animal	Y1	Y2	Y3	Y4
1	1	1·01	0·96	1·69	0·75	6	16	−1·41	−2·63	−0·68	−1·62
	2	**0·66**	**0·94**	**1·69**	**1·63**		17	−0·83	0·05	1·29	0·88
	3	0·79	1·39	0·74	0·38		18	−0·76	−1·09	−0·04	−0·78
2	4	0·48	2·46	4·62	2·82	7	19	−1·98	−0·19	−0·27	−1·94
	5	0·51	1·55	3·37	1·36		20	−0·65	0·76	−0·15	−1·46
	6	−0·74	−0·74	−0·38	−2·95		**21**	**−0·96**	**0·64**	**0·19**	**−0·33**
3	7	2·24	1·86	1·65	3·10	8	22	2·03	−2·48	−2·12	−1·55
	8	**0·87**	**0·70**	**−0·16**	**0·41**		23	−0·93	−4·35	−5·17	−4·69
	9	1·97	2·49	3·76	5·34		24	1·66	−1·87	−3·25	−2·22
4	10	1·20	0·44	−0·54	−0·22	9	**25**	**1·63**	**0·80**	**0·68**	**1·28**
	11	−0·53	−2·50	−4·43	−5·59		26	1·46	0·51	2·06	5·35
	12	1·67	1·95	2·56	2·43		27	1·21	0·91	2·66	3·47
5	13	−0·79	−1·25	−0·88	0·95	10	28	0·33	0·44	2·49	4·01
	14	−2·13	−1·85	−1·21	−0·22		29	0·40	−0·39	0·91	2·54
	15	**0·20**	**1·17**	**3·15**	**4·27**		30	−0·07	−0·38	1·50	3·99

https://doi.org/10.1017/S0016672302005700 Published online by Cambridge University Press

Table B2. *Non-parametric covariance parameter estimates for the small simulated example* (*data given in Table B1*) *in the balanced and unbalanced cases*

		Balanced		Unbalanced	
Times		GEN	ENV	GEN	ENV
1	1	0·8243	0·7098	0·9375	0·7441
2	1	0·0949	0·8741	0·2588	0·8745
2	2	1·2975	1·5301	1·6197	1·6062
3	1	−0·2278	1·1735	0·0212	1·1188
3	2	1·3056	2·0979	1·8420	2·1076
3	3	1·9377	3·4194	2·8359	3·2696
4	1	0·2524	1·3638	0·5266	1·3519
4	2	0·9537	2·3791	1·3648	2·4523
4	3	1·7043	3·8626	2·6130	3·6564
4	4	3·0951	4·7524	4·4081	4·4444

We are most grateful to Professor Robin Thompson, Dr Ian White, Dr Peter Visscher and Dr Vincent Ducrocq for helpful comments and ideas. This work was supported by the Department of Animal Genetics of the INRA (National Institute of Agronomical Research), Jouy-en-Josas, France.

References

Diggle, P. J. & Verbyla, A. P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* **54**, 401–415.

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

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

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.

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

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

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

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