Genet. Sel. Evol. 32 (2000) 383–394



Original article

# Assessment of heterogeneity of residual variances using changepoint techniques

Romdhane REKAYA[a,*], Maria J. CARABAÑO[b], Miguel A. TORO[b]

[a] Department of Dairy Science, University of Wisconsin, Madison 53706, USA
[b] Area de Mejora Genética Animal, CIT-INIA, Carretera de la Coruna Km 7, 28040 Madrid, Spain

(Received 25 January 2000; accepted 25 April 2000)

**Abstract** – Several studies using test-day models show clear heterogeneity of residual variance along lactation. A changepoint technique to account for this heterogeneity is proposed. The data set included 100 744 test-day records of 10 869 Holstein-Friesian cows from northern Spain. A three-stage hierarchical model using the Wood lactation function was employed Two unknown changepoints at times $T_1$ and $T_2$, $(0 < T_1 < T_2 < t_{max})$, with continuity of residual variance at these points, were assumed Also, a nonlinear relationship between residual variance and the number of days of milking $t$ was postulated. The residual variance at a time $t$ ($\sigma^2_{et}$) in the lactation phase $i$ was modeled as: $\sigma^2_{et} = t^{\lambda_i}\sigma^2_{ei}$ for $(i = 1, 2, 3)$, where $\lambda_i$ is a phase-specific parameter. A Bayesian analysis using Gibbs sampling and the Metropolis-Hastings algorithm for marginalization was implemented. After a burn-in of 20 000 iterations, 40 000 samples were drawn to estimate posterior features. The posterior modes of $T_1$, $T_2$, $\lambda_1$, $\lambda_2$, $\lambda_3$, $\sigma^2_{e1}$, $\sigma^2_{e2}$, $\sigma^2_{e3}$ were 53.2 and 248.2 days; 0.575, −0.406, 0.797 and 0.702, 34.63 and 0.0455 kg$^2$, respectively. The residual variance predicted using these point estimates were 2.64, 6.88, 3.59 and 4.35 kg$^2$ at days of milking 10, 53, 248 and 305, respectively. This technique requires less restrictive assumptions and the model has fewer parameters than other methods proposed to account for the heterogeneity of residual variance during lactation.

**changepoint / heterogeneity / residual variance**

**Résumé – Évaluation de l'hétérogénéité de la variance résiduelle durant la lactation en utilisant la technique de changement de points.** La technique de changement de points a été utilisée pour étudier l'hétérogénéité de la variance résiduelle durant la lactation en considérant 100 744 observations de production laitière le jour du contrôle issues de 10 898 vaches dans le nord de l'Espagne. Un modèle Bayésien à trois étapes utilisant la function de Wood a été mis en place. Deux points de changement aux temps inconnus $T_1$ et $T_2$, $(0 < T_1 < T_2 < t_{max})$ ont été adoptés. Nous avons également supposé la continuité de la variance résiduelle aux points de changement. Une relation non linéaire entre la variance résiduelle et la durée de la lactation a été postulée. La variance résiduelle à un moment $t$ ($\sigma^2_{et}$) durant la phase $i$ de la lactation est donnée par $\sigma^2_{et} = t^{\lambda_i}\sigma^2_{ei}$ pour $(i = 1, 2, 3)$. L'estimation

* Correspondence and reprints
E-mail: rekaya@calshp.cals.wisc.edu

de Gibbs et l'algorithme de Metropolis-Hastings ont été utilisés pour l'échantillonage des distributions conditionelles *a posteriori* des paramètres du modèle. Après une période d'échauffement de 20 000 échantillons, 40 000 itérations supplémentaires ont été réalisées. Les modes *a posteriori* de $T_1$, $T_2$, $\lambda_1$, $\lambda_2$, $\lambda_3$, $\sigma^2_{e1}$, $\sigma^2_{e2}$ et $\sigma^2_{e3}$ étaient de 53,21 et 248,16 jours, 0,575, − 0,406, 0,797, 0,702, 34,63 et 0,0455, respectivement. La variance résiduelle estimée utilisant ces estimateurs ponctuels était de 2,64, 6,88, 3,59 et 4,35 $kg^2$ aux jours 10, 53, 248 et 305 de la lactation, respectivement. La technique de changement de points est d'une part moins restrictive et d'autre part permet de réduire le nombre de paramètres à estimer par rapport à d'autres méthodes utilisées pour étudier l'hétérogénéité de la variance résiduelle durant la lactation.

**changement de points / hétérogénéité / variance résiduelle**

## 1. INTRODUCTION

Heterogeneous residual variance in the course of lactation has been observed when studying test-day records [4,5,7,8]. Authors have argued that this heterogeneity might be associated with factors such as the stage of pregnancy, calving conditions, and the length of the dry period. Incorporation of such explanatory effects in genetic evaluation models may be difficult, mainly due to lack of information. The impact of heterogeneity of the residual variance on evaluation goes through the weight given to information in each part of the lactation. If homogeneous variance is assumed, information from parts of the lactation having lower residual variance would, implicitly, receive lower weight. On the contrary, test-days from periods of lactation with higher residual variance would have a higher impact on the estimation.

It is not clear how to account for this heterogeneity in genetic evaluation. Jamrozik *et al.* [4], Jamrozik and Schaeffer [3] and Rekaya *et al.* [8] divided the lactation length into 10 intervals, and assumed homogeneity of variance within intervals, and heterogeneity between them. Drawbacks of this approach are that intervals are decided in an arbitrary way, and that a large number of residual variance components needs to be estimated, in many cases with low precision, especially at the beginning and the end of lactation.

An alternative is to employ a changepoint identification technique [10]. Typically, in time series (longitudinal data), such as with economic data or milk production in the course of lactation, changes in the generation process of the data can take place as a result of changes in the assumed model or in the parameters of the model that describes the process. The technique of changepoint identification allows to make inferences about the time at which changes occur, and about possible changes of parameter values or of the assumed model. In a previous study with test day milk yields collected in the Spanish Friesian population, residual variances estimated for 33 consecutive intervals along lactation followed a pattern that might indicate that residual variance changes in successive phases. Figure 1 suggests three consecutive phases: ascending, descending and finally an ascending phase again.

In this paper, we extend the changepoint identification technique described by Stephens [10] to the situation of multiple changes affecting the dispersion parameters. The effects of considering this heterogeneity of residual variance on estimates of variance components of parameters of the lactation curve and

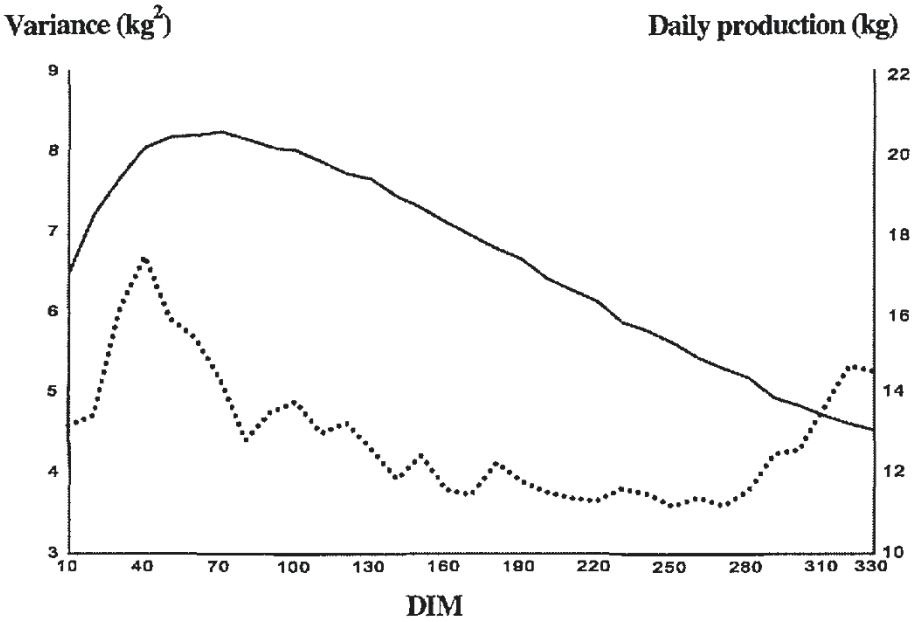


**Figure 1.** Changes in residual variance (····) and daily production (—) observed in a Spanish Friesian population.

on the genetic evaluation for these parameters are assessed as well. A Bayesian implementation using Gibbs Sampling and the Metropolis-Hastings algorithm was used for this purpose.

## 2. MATERIALS AND METHODS

### 2.1. Data

Test-day records from first-lactation Holstein-Friesian cows in four regions in northern Spain obtained from 1982 through 1994 were used in this study. Only data from complete lactations were considered. Requirements for a cow to be included in the analysis were: first test between 4 and 34 days post parturition, time interval between successive test-day less than 40 days (except for holiday periods), test-day records with milk production between 5 kg and 55 kg and total yield in 305 days over 2 000 kg. The final data set had 100 744 test-day milk yield records from 10 890 cows. A summary description of the data set is presented in Table I. The pedigree file included 42 882 animals.

### 2.2. Methods

A Bayesian analysis using a nonlinear model (the Wood incomplete gamma function [11]) to describe the shape of the lactation curve, accounting for heterogeneity of residual variance, was considered. Implementation of the same model but assuming homogeneity of the residual variance, has been described by Rekaya [6]. Hence, the description will focus mainly on new methodological aspects resulting from consideration of heterogeneity of residual variance and from the changepoint technique.

**Table I.** Number of cows, mean milk yield and standard deviation by test-day.

| Test | No cows | Mean yield (kg) | SD |
|---|---|---|---|
| 1 | 10 869 | 22.41 | 4.90 |
| 2 | 10 869 | 22.79 | 5.26 |
| 3 | 10 869 | 21.74 | 5.40 |
| 4 | 10 869 | 20.71 | 5.37 |
| 5 | 10 869 | 19.77 | 5.24 |
| 6 | 10 869 | 18.78 | 5.25 |
| 7 | 10 363 | 17.92 | 5.14 |
| 8 | 9 622 | 16.87 | 5.06 |
| 9 | 7 650 | 15.94 | 4.98 |
| 10 | 4 381 | 15.36 | 4.89 |
| 11 | 2 161 | 14 83 | 4.97 |
| 12 | 1 028 | 14.55 | 5.04 |
| 13 | 319 | 15.38 | 5.30 |
| 14 | 6 | 18 77 | 5.16 |

### 2.3. Analysis

As suggested by the pattern described in Figure 1, a hierarchical nonlinear model with two unknown changepoints ($T_1$ and $T_2$) for residual variance along the lactation was fitted. The relationship between residual variance and days in milk was assumed to be positive in the first part of lactation (0–$T_1$), negative in the second interval ($T_1$–$T_2$) and positive in the last part of lactation ($T_2$–end of lactation). The following models were assumed for the residual variance in each of the three phases of lactation:

$$\begin{aligned}
\sigma^2_{et} &= t^{\lambda_1}\sigma^2_{e1} \qquad t = 1, \ldots, T_1 \\
\sigma^2_{et} &= t^{\lambda_2}\sigma^2_{e2} \qquad t = T_1 + 1, T_2 \\
\sigma^2_{et} &= t^{\lambda_3}\sigma^2_{e3} \qquad t = T_2 + 1, t_{\max}
\end{aligned}$$

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are parameters relating the residual variance in each interval to the scale parameters or base-line variance $\sigma^2_{e1}$, $\sigma^2_{e2}$ and $\sigma^2_{e3}$; $t$ are days in milk (DIM); $T_1$ and $T_2$ are the two unknown changepoints and $t_{\max}$ is the time of the last test-day in the data file.

The first stage of the Bayesian hierarchy describes the conditional distribution of the observations in each of the three intervals, given the parameters of the model. It was assumed that:

$$\mathbf{y}|\beta, \alpha, \mathbf{p}, \lambda, \mathbf{T}, \sigma^2_{e1}, \sigma^2_{e2}, \sigma^2_{e3} \sim \begin{cases} N\left(\mathbf{X}\beta + f(\alpha, t) + \mathbf{W}\mathbf{p}, t^{\lambda_1}\sigma^2_{e1}\mathbf{I}\right) & 1 \leq t \leq T_1 \\ N\left(\mathbf{X}\beta + f(\alpha, t) + \mathbf{W}\mathbf{p}, t^{\lambda_2}\sigma^2_{e2}\mathbf{I}\right) & T_1 < t \leq T_2 \\ N\left(\mathbf{X}\beta + f(\alpha, t) + \mathbf{W}\mathbf{p}, t^{\lambda_3}\sigma^2_{e3}\mathbf{I}\right) & T_2 < t \leq t_{\max} \end{cases} \tag{1}$$

where $\mathbf{y}$ is the vector of observations, $\beta$ is a vector of herd-test-day effects and $\mathbf{X}$ is an incidence matrix. Further, $f(\alpha, t) = \alpha_1 t^{\alpha_2} \exp(-\alpha_3 t)$ is the Wood function representing the shape of the lactation curve at the phenotypic level with $\alpha = [\alpha_1, \alpha_2, \alpha_3]$; $\mathbf{p}$ is a permanent environmental effect common to all test-days of a cow, and $\mathbf{W}$ is an incidence matrix.

Due to the requirement of continuity of the residual variance at the changepoint, the following equalities need to be satisfied:

$$\begin{aligned} T_1^{\lambda_1}\sigma_{e1}^2 &= T_1^{\lambda_2}\sigma_{e2}^2 \\ T_2^{\lambda_2}\sigma_{e2}^2 &= T_2^{\lambda_3}\sigma_{e3}^2. \end{aligned} \tag{2}$$

The continuity of the residual variance at the changepoints has the consequence that some parameters become a combination of the remaining ones. An adequate reparametrization can reduce the number of parameters in the model and, in some cases, the resulting conditional distributions are easier to handle. We opted for the re-parameterizing of $\sigma_{e2}^2$ and $\sigma_{e3}^2$ as a function of the remaining parameters used to model the residual variance ($\sigma_{e1}^2$, $T_1$, $T_2$, $\lambda_1$, $\lambda_2$ and $\lambda_3$). New reparameterization was used together with the restriction: $1 \leq T_1 \leq T_2 \leq t_{\max}$ to avoid infinite solutions for the changepoints.

Therefore,

$$\begin{aligned} \sigma_{e2}^2 &= T_1^{(\lambda_1-\lambda_2)}\sigma_{e1}^2 = K_1\sigma_{e1}^2 \\ \sigma_{e3}^2 &= T_2^{(\lambda_2-\lambda_3)}T_1^{(\lambda_1-\lambda_2)}\sigma_{e1}^2 = K_2\sigma_{e1}^2. \end{aligned} \tag{3}$$

After re-parameterization and taking into account that $1 \leq T_1 \leq T_2 \leq t_{\max}$ the likelihood function is proportional to:

$$\begin{aligned} \mathbf{y}|\beta, \mathbf{p}, \alpha, \lambda, \mathbf{T}, \sigma_{e1}^2 \propto & \prod_{i=1}^{n_1}\left(\frac{1}{t^{\lambda_1}}\right)^{0.5} \prod_{i=n_1+1}^{n_2}\left(\frac{1}{t^{\lambda_2}K_1}\right)^{0.5} \prod_{i=n_2+1}^{N}\left(\frac{1}{t^{\lambda_3}K_2}\right)^{0.5} \\ \times\sigma_{e1}^{-N}\exp\Bigg(-0.5\sigma_{e1}^2\Bigg[ & \sum_{j=1}^{q}\sum_{t\leq T_1}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_1}} \\ & +\sum_{j=1}^{q}\sum_{T_1<t\leq T_2}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_2}K_1} \\ & +\sum_{j=1}^{q}\sum_{t>T_2}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_3}K_2}\Bigg]\Bigg) \end{aligned} \tag{4}$$

where $q$ is the number of animals with data ($q = 10\,890$); $n_1$ and $n_2$ represent the number of test-day records for all animals with data realized at a time $t$ smaller than the first ($T_1$) and the second ($T_2$) changepoints; $N$ is the total number of observations in the data file; $K_1$ and $K_2$ are unknown as defined before; $HTD_{jt}$ is the herd-test day effect for cow $j$ at time $t$, $p_j$ is the permanent environmental effect peculiar to cow $j$, and $f_j(\alpha, t) = \alpha_{1j}t^{\alpha_{2j}}\exp(-\alpha_{3j}t)$ is the Wood function for the cow $j$ evaluated at time $t$.

At the second stage of the hierarchy, prior distributions were specified for first stage parameters. The priors were:

$$\beta \sim U\left[\beta_{\min}, \beta_{\max}\right] \tag{5}$$

$$\alpha|\mathbf{m}, \Sigma_0 \sim N\left(\mathbf{m}, \mathbf{I} \otimes \Sigma_0\right) \tag{6}$$

with $\mathbf{m} = \mathbf{Hb} + \mathbf{Zu}$

$$\mathbf{p}|\sigma_p^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_p^2) \tag{7}$$

$$\sigma_{e1}^2|\upsilon_e, s_e^2 \sim \chi^{-2}(\upsilon_e, \upsilon_e s_e^2) \tag{8}$$

$$T \sim U[1, t_{\max}], \text{ subject to } 1 \leq T_1 \leq T_2 \leq t_{\max}. \tag{9}$$

Where $\Sigma_0$ is a $3 \times 3$ matrix of residual (co)variances between parameters of the Wood function, $\mathbf{b}$ is the age-season of calving effect and $\mathbf{u}$ the additive genetic value for the lactation curve parameters and $\mathbf{H}$ and $\mathbf{Z}$ are the corresponding incidence matrices. Values adopted for the hyper-parameters were: $\beta_{\min} = -200$, $\beta_{\max} = 200$ and $t_{\max} = 345$.

In the third stage, prior distributions for $\mathbf{b}$, $\mathbf{u}$, $\Sigma_0$ were specified as:

$$\mathbf{b} \sim U[\mathbf{b}_{\min}, \mathbf{b}_{\max}] \tag{10}$$

$$u|\Sigma_g \sim N\left(0, A \otimes \Sigma_g\right) \tag{11}$$

$$\sigma_p^2|\upsilon_p, s_p^2 \sim \chi^{-2}(\upsilon_p, \upsilon_p s_p^2) \tag{12}$$

$$\Sigma_0|\upsilon_0, \mathbf{S}_0^2 \sim W^{-1}(\upsilon_0, \upsilon_0 \mathbf{S}_0^2) \tag{13}$$

$$\Sigma_g|\upsilon_g, \mathbf{S}_g^2 \sim W^{-1}(\upsilon_g, \upsilon_g \mathbf{S}_g^2). \tag{14}$$

A value of 4 was given to $\upsilon_i$ $(i = e, 0, p, g)$ in order to assign a low weight to the prior information. Values for the scaling factors ($s_e^2$, $s_p^2$, $\mathbf{S}_0^2$ and $\mathbf{S}_g^2$) were obtained from results in a previous study under a similar model but assuming homogeneity of residual variance [6].

### 2.4. Conditional distributions

The joint posterior density was obtained as the product of the likelihood function in (4) and the prior densities in (5–14). Conditional posterior distributions for $\beta$, $\mathbf{b}$, $\mathbf{u}$, $\Sigma_0$ and $\Sigma_g$ were normal for the position parameters $(\beta, \mathbf{p}, \mathbf{b}, \mathbf{u})$, and scaled inverted Wishart distribution for the dispersion matrices $\Sigma_0$ and $\Sigma_g$.

The conditional posterior distribution of $\sigma_{e1}^2$ is a scaled-inverted chi square where the scaling factor is a weighted sum of the residual terms in the three phases of lactation

$$p(\sigma_{e1}^2|\beta, \mathbf{b}, \mathbf{u}, \alpha, \lambda, \mathbf{T}, \mathbf{y}) \sim \chi^{-2}(\upsilon_e + N, \upsilon_e s_e^2 + \mathbf{e}'e)$$

where,

$$\mathbf{e}'e = \left[\sum_{j=1}^{q}\sum_{t\le T_1}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_1}} + \sum_{j=1}^{q}\sum_{T_1<t\le T_2}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_2}K_1} + \sum_{j=1}^{q}\sum_{t>T_2}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_3}K_2}\right].$$

The conditional distribution of the first parameter of the Wood function is normal, as in the case of homogeneity of residual variance:

$$p(\alpha_1|\beta,\mathbf{b},\mathbf{u},\alpha_2,\alpha_3,\lambda,\mathbf{T},\sigma^2_{e1},\Sigma_0,\Sigma_g\mathbf{y}) \sim N\left[\hat{\alpha}_1,(\Lambda' R_{ti}^{-1}\Lambda + r^{11}\mathbf{I})^{-1}\right]$$

where,

$$\hat{\alpha}_1 = \left(\Lambda' R_{ti}^{-1}\Lambda + r^{11}\mathbf{I}\right) \times \left[\Lambda' R_{ti}^{-1}(\mathbf{y}-\mathbf{X}\beta-\mathbf{W}\mathbf{p}) + r^{11}\mathbf{m}_1 - r^{12}(\alpha_2-\mathbf{m}_2) - r^{13}(\alpha_3-\mathbf{m}_3)\right]$$

where $\Lambda$ is a matrix of order $Nxq$ with elements $t^{\alpha_{2k}}\exp(-\alpha_{3k}t)$ ($\alpha_{2k}$ and $\alpha_{3k}$ evaluated at their current values for animal $k$ and corresponding DIM $t$) in column $k$, and zeroes in any other column; $r^{ij}$ is the $(i,j)$ element of the inverse of the residual matrix $\Sigma_0$, and $\mathbf{m}_i$ is the mean of $\alpha_i$ as defined in (6).

The conditional posterior distributions of the remaining parameters of the model ($\alpha_2$, $\alpha_3$, $T_1$, $T_2$, $\lambda_1$, $\lambda_2$ and $\lambda_3$) are not in closed forms, as a result of non-linearity. The conditional distribution of the second and third parameter of the Wood function, $\alpha_i$ $(i=2,3)$ is:

$$p\left(\alpha_i|\beta,\mathbf{b},\mathbf{u},\mathbf{p},\alpha,\lambda,\mathbf{T},\alpha_1,\alpha_{j\neq i},\sigma^2_{e1},\Sigma_0,\Sigma_g,\mathbf{y}\right) \propto \prod_{i=1}^{n_1}\left(\frac{1}{t^{\lambda_1}}\right)^{0.5}\prod_{i=n_1+1}^{n_2}\left(\frac{1}{t^{\lambda_2}K_1}\right)^{0.5}\prod_{i=n_2+1}^{N}\left(\frac{1}{t^{\lambda_3}K_2}\right)^{0.5}\sigma_{e1}^{-N} \times \exp\left(-0.5\sigma^2_{e1}\left[\sum_{j=1}^{q}\sum_{t\le T_1}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_1}} + \sum_{j=1}^{q}\sum_{T_1<t\le T_2}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_2}K_1} + \sum_{j=1}^{q}\sum_{t>T_2}\frac{\left(y_{jt}-HTD_{jt}-f_j(\alpha,t)-p_j\right)^2}{t^{\lambda_3}K_2}\right]\right) \times \exp\left(-0.5(\alpha-\mathbf{H}\mathbf{b}-\mathbf{Z}\mathbf{u})'\left(I\otimes\Sigma_0\right)^{-1}(\alpha-\mathbf{H}\mathbf{b}-\mathbf{Z}\mathbf{u})\right). \quad (15)$$

Before reparameterization, the conditional distribution of the changepoints $T_i$ $(i = 1, 2)$ depends on the sequence of data between times $T_{i-1}$ and $T_{i+1}$ (in our case, $T_0 = 1$ and $T_3 = t_{\max}$). After reparameterization, this holds just for the last changepoint $T_2$. This avoids absurd estimates for the changepoints. Thus:

$$p\left(T_1|\beta, \mathbf{b}, \mathbf{u}, \mathbf{p}, \alpha, T_2, \sigma^2_{e1}, \Sigma_0, \Sigma_g, \mathbf{y}\right) \propto p\left(y|\beta, \alpha, \mathbf{p}, \lambda, \sigma^2_{e1}\right) \tag{16}$$

where the right hand side of (16) is the likelihood function viewed as a function of $T_1$ only, and,

$$\begin{aligned} p\left(T_2|\beta, \mathbf{b}, \mathbf{u}, \mathbf{p}, \alpha, T_1, \sigma^2_{e1}, \Sigma_0, \Sigma_g, \mathbf{y}\right) \\ \propto \prod_{i=n_1+1}^{n_2} \left(\frac{1}{t^{\lambda_2} K_1}\right)^{0.5} \prod_{i=n_2+1}^{N} \left(\frac{1}{t^{\lambda_3} K_2}\right)^{0.5} \sigma_{e1}^{-N} \\ \times \exp\left(-0.5\sigma^2_{e1}\left[\sum_{j=1}^{q} \sum_{T_1 < t \leq T_2} \frac{\left(y_{jt} - HTD_{jt} - f_j(\alpha, t) - p_j\right)^2}{t^{\lambda_2} K_1}\right.\right. \\ \left.\left. + \sum_{j=1}^{q} \sum_{t > T_2} \frac{\left(y_{jt} - HTD_{jt} - f_j(\alpha, t) - p_j\right)^2}{t^{\lambda_3} K_2}\right]\right). \end{aligned} \tag{17}$$

The conditional posterior distributions of the first two exponents ($\lambda_1$ and $\lambda_2$) have the same form as those of $T_1$ and $T_2$, respectively, but are viewed as a function of the $\lambda$'s. The conditional distribution of $\lambda_3$ depends only on the test-day data collected after the second changepoint ($T_2$):

$$\begin{aligned} p\left(\lambda_3|\beta, \mathbf{b}, \mathbf{u}, \mathbf{p}, \alpha, \mathbf{T}, \lambda_1, \lambda_2, \sigma^2_{e1}, \Sigma_0, \Sigma_g, \mathbf{y}\right) \propto \prod_{i=n_2+1}^{N} \left(\frac{1}{t^{\lambda_3} K_2}\right)^{0.5} \sigma_{e1}^{-N} \\ \times \exp\left(-0.5\sigma^2_{e1}\left[\sum_{j=1}^{q} \sum_{t > T_2} \frac{\left(y_{jt} - HTD_{jt} - f_j(\alpha, t) - p_j\right)^2}{t^{\lambda_3} K_2}\right]\right). \end{aligned} \tag{18}$$

Sampling from (15–18) was *via* the adaptive rejection Metropolis algorithm [2] and the sampling-resampling algorithm [9]. Rekaya [6] presented a full description and implementation of both algorithms within a Gibbs sampling scheme. A single chain of 60 000 samples was run with a burn-in period of 20 000 samples. Analysis was based on 40 000 samples, drawn without thinning.

## 3. RESULTS

A summary of the marginal posterior distributions of the parameters defining the residual variance is presented in Table II. The signs of the exponents indicate, as it was expected, a positive relationships between the residual variance and DIM until the day $T_1 = 53.21$ ($\lambda_1 = 0.575$), a negative relationship ($\lambda_2 = -0.406$) from that day to day $T_2 = 248.16$, and a positive relationship ($\lambda_3 = 0.797$) in the last part of lactation.

**Table II.** Summary of the marginal distributions of the parameters defining the change of residual variance under the changepoint model.

| Parameter [(a)] | Mean | Mode | SD | HPD (95%) |
|---|---|---|---|---|
| $\sigma^2_{e1}$ | 0.707 | 0.702 | 0.11 | 0.54 , 0.83 |
| $T_1$ | 54.55 | 53.21 | 4.08 | 42 83 , 69 38 |
| $T_2$ | 246.8 | 248 16 | 11.57 | 224.18 , 262.47 |
| $\lambda_1$ | 0.569 | 0.575 | 0.049 | 0.464 , 0.668 |
| $\lambda_2$ | −0.413 | −0.406 | 0.064 | −0.512 , −0.332 |
| $\lambda_3$ | 0 789 | 0.797 | 0.082 | 0.646 , 0.913 |

[(a)] $\sigma^2_{e1}$, $\lambda_1$, $\lambda_2$, and $\lambda_3$ are the factor and exponents that provide the value of the residual variance at each time, and $T_1$ and $T_2$ are the changepoints.
HPD stands for High Posterior Density interval.

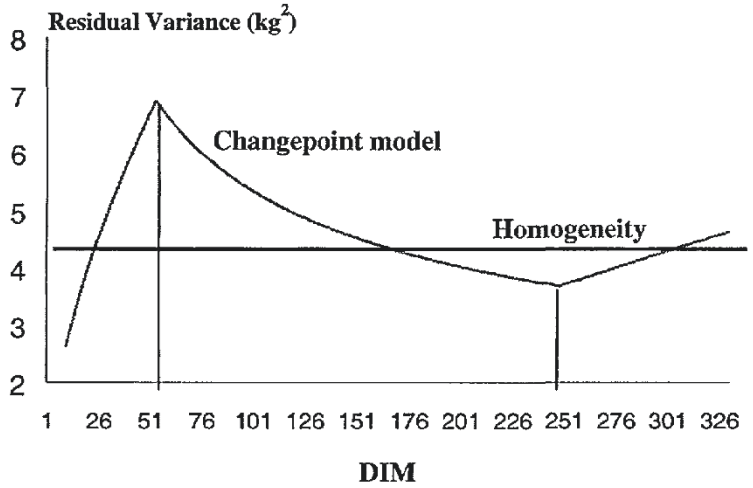


**Figure 2.** Residual variance along lactation as predicted by the changepoint model and under a homogeneous variance model.

The predicted behavior of the residual variance using the posterior mode of the parameters in Table II is presented in Figure 2. This is similar to the one observed in Figure 1 where the residual variance was estimated within 10-day intervals. A similar pattern was observed by Jamrozik *et al.* [4], assuming homogeneity within 10-day intervals and heterogeneity between intervals. However, the absolute values were lower in this case. A somehow different trend, with no initial increase of the residual variance, was found by Jamrozik and Schaeffer [3] and by Rekaya *et al.* [8], with a similar data set but using linear random regressions for genetic and permanent effects.

The highest and lowest predicted values of the residual variance were 6.89 $kg^2$ and 3.59 $kg^2$ at 53.21 and 248.16 days in milk, respectively. Both values are similar to those found when the residual variance was estimated within 10-day intervals. The value of 4.48 $kg^2$ found by Rekaya *et al.* [6] when homogeneity of residual variance was assumed, agrees well with results from this study, given

**Table III.** Summary of the marginal distributions of the permanent environmental variance and residual and genetic (co)variances for the parameters of the Wood function under the changepoint model.

| | Mean | Mode | SD | HPD (95%) |
|---|---|---|---|---|
| $\sigma_p^2$ | 4.98 | 4.96 | 0.30 | 4.45 , 5.37 |
| $r_{11}$ | 6 28 | 6.33 | 0.38 | 5.18 , 7.48 |
| $r_{22}$ | 9.13E − 4 | 9.19E − 4 | 7.87E − 5 | 7.56E − 4 , 1.12E − 3 |
| $r_{33}$ | 3.91E − 7 | 3.87E − 7 | 2.55E − 8 | 2.43E − 7 , 4.67E − 4 |
| $r_{12}$ | −4 27E − 2 | −4.29E − 2 | 4.30E − 3 | −6.17E − 2 , −3.76E − 2 |
| $r_{13}$ | 4.63E − 4 | 4.60E − 4 | 4 83E − 5 | 5.34E − 4 , 6.15E − 4 |
| $r_{23}$ | 8.13E − 6 | 8.11E − 6 | 7.50E − 7 | 6.92E − 6 , 1.07E − 5 |
| $g_{11}$ | 2.40 | 2.43 | 0.24 | 1.82 , 3.19 |
| $g_{22}$ | 4.84E − 4 | 4.82E − 4 | 5.12E − 5 | 3.77E − 4 , 5.97E − 4 |
| $g_{33}$ | 6.36E − 8 | 6.49E − 8 | 7.13E − 9 | 5.21E − 8 , 8.14E − 8 |
| $g_{12}$ | −2.97E − 2 | −3.09E − 2 | 5.09E − 3 | −4.87E − 2 , −2.08E − 2 |
| $g_{13}$ | −1.48E − 4 | −1.50E − 4 | 3 84E − 5 | −2.77E − 4 , −9.89E − 5 |
| $g_{23}$ | 2.67E − 6 | 2.61E − 6 | 2 65E − 6 | 1.35E − 6 , 3.74E − 6 |

$\sigma_p^2$ is the permanent environmental variance, $r_{ij}$ and $g_{ij}$ are the residual and genetic variance components associated to parameters of the Wood function, respectively. HPD stands for High Posterior Density interval.

that this value represents a weighted average of the residual variance along lactation.

Table III shows a summary of the posterior distributions of variance of permanent effect and the genetic and residual (co)variances associated with the parameters of the Wood function. The posterior mode of variance of permanent effect ($\sigma_p^2$) was 4.96 kg$^2$, clearly lower than the value of 6.6 kg$^2$ obtained using the same data set with a repeatability model and homogeneity of residual variance [8]. Point estimates of genetic and residual (co)variance for the Wood function parameters show similar tendency to those found using the same data and model but assuming homogeneity of residual variance [6]. Genetic correlation was negative between the first and second, and first and third parameters of the Wood function and of opposite sign between second and third parameters. Residual covariance was negative between the first and second parameter and positive for the remaining two covariances. However, the absolute values for these genetic and residual (co)variances indicate some differences when compared with those obtained assuming homogeneity of residual variance [6]. A reduction of genetic and residual variance of the first parameter of the Wood function and an increase for the other two parameters were observed. The major difference was noted for the second parameter of the Wood function. This was probably because the predicted residual variance in the period from the beginning of lactation to 160 days of milking, where this

**Table IV.** Pearson correlations between parameters of the Wood function ($Cp$) and between the genetic values associated with those parameters ($Cg$) assuming homogeneity and heterogeneity of residual variance.

| Parameter | $Cp$ | $Cg$ |
|---|---|---|
| $\alpha_1$ | 0.879 | 0.881 |
| $\alpha_2$ | 0.791 | 0.863 |
| $\alpha_3$ | 0.935 | 0.900 |

parameter has more relative weight to describe variation in milk production, was higher than the homogenous residual variance (Fig. 2).

Heritabilities of the parameters of the Wood function were similar to those obtained assuming homogeneity of residual variance. The major difference was noted for the third parameter as a consequence of the disproportional increase in its residual variance in this study.

Table IV presents Pearson correlations among the same parameters of the Wood function as well as the correlations between the genetic values associated with those parameters assuming homogeneity and heterogeneity of residual variance. Correlation coefficients indicated that taking into account the heterogeneity of residual variance caused significant changes both on the estimation of the Wood function parameters and on the breeding values associated with them which will affect the computation of production function of economic interest like persistency, peak yield, and total milk yield.

## 4. CONCLUSIONS

An additional complication of using individual test-day information is caused by heterogeneity of residual variance along lactation, this being lower during mid-lactation and higher at the two ends. Early lactation results are less clear, probably due to estimation problems. Some studies indicate an increase of residual variance, whereas others show a more complex pattern, as the one presented in Figure 1, with an increase from the beginning of lactation to days 40–50.

The changepoint technique assuming two changepoints and a simple model for the residual variance in each interval of lactation, was adequate for predicting the behavior of residual variance with a significant reduction in the number of parameters to be estimated, and avoiding subjectivity.

The main objective of this study was to illustrate the use of the changepoint models within a Bayesian framework to account for heterogeneous residual variances. The assumptions made with respect to the number of changepoint or about the relationship between the residual variance and days of milking may not be suitable in other situations. In fact, in a more complex residual variance behavior, a structural model allowing the inclusion of other sources of heterogeneity as suggested by Foulley *et al.* [1] together with the changepoint

technique can be more appropriate. Changepoint models introduce an interesting alternative in the analysis of other longitudinal data that occur in animal breeding.

## ACKNOWLEDGEMENTS

This work was supported by a grant of Programa Sectorial I+D of MAPA, Spain (SC96-046). Data were provided by CONAFE (Spanish Friesian Association). The first author was supported by AECI-ICMA. We are grateful to Dr. D. Gianola for his kind assistance in reading this manuscript.

## REFERENCES

[1] Foulley J.L., Gianola D., San Cristobal M., Im S., A method for assessing extent and sources of heterogeneity of residual variances in mixed linear model, J Dairy Sci 73 (1990) 1612–1624.

[2] Gilks W.R., Best N.G , Tan K.K C , Adaptive rejection Metropolis sampling within Gibbs sampling, Appl. Statist 44 (1995) 455–472

[3] Jamrozik J., Schaeffer L.R., Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins, J. Dairy Sci. 80 (1997) 762–770.

[4] Jamrozik J., Kistemaker G.J., Dekkers G.C.M., Schaeffer L.R., Comparison of possible covariates for use in a random regression model for analyses of test day yields, J. Dairy Sci. 80 (1997) 2550–2556.

[5] Ptak E., Schaeffer L.R., Use of test day yields for genetic evaluation of dairy sires and cows, Livest. Prod. Sci. 34 (1987) 23–34.

[6] Rekaya R., Análisis Bayesiano de datos de producción en los días del control para la selección de caracteres lecheros, Tesis doctoral, Universidad Politécnica de Madrid, Spain, 1997.

[7] Rekaya R., Carabaño M.J., Toro M A., Análisis bayesiano de la heterogeneidad de la varianza residual durante la lactación utilizando la técnica de los puntos de cambio, VII Jornadas sobre Producción Animal, Zaragoza, ITEA, Vol. Extra, 18(I) (1997) 400–402.

[8] Rekaya R., Carabaño M.J., Toro M.A., Use of test yields for the genetic evaluation of production traits in Holstein-Friesian cattle, Livest. Prod. Sci. 57 (1999) 203–217.

[9] Smith A F.M., Gelfand A.E., Bayesian statistics without tears: A sampling resampling perspective, J. Am Stat. Assoc. 46 (1992) 84–88.

[10] Stephens D.A., Bayesian Retrospective Multiple-changepoint Identification, Appl. Stat 43 (1994) 159–178.

[11] Wood P D.P., Algebraic model of the lactation curve in cattle, Nature 216 (1967) 164–166