Preface

In recent years cattle herd sizes have been growing rapidly. Because each farmer is getting an increasing number of cows to look after there is an equivalent growing need for more automated management of the herds. In this context use of statistics can be very helpful in a number of ways, e.g. in detecting diseases and heat in cows. This PhD thesis is a contribution to the ongoing research in analysing on-line data with the focus on using state space models for automated management.

The thesis consists of a review together with three independently written papers and is submitted to the Faculty of Science, University of Aarhus. The review provides an introduction to state space models and a presentation of the main results of the accompanying papers. The core of the thesis is the three enclosed papers. Two of the papers present approaches for analysing online data with emphasis on applications. The third paper covers new results on theoretical aspects of state space models.

In the three years of my PhD study I found the whole process to be interesting, challenging and occasionally also frustrating. Luckily, numerous people have supported and helped me in my work throughout this period. First of all I would like to thank my supervisors, Professor Jens Ledet Jensen and Senior scientist Søren Højsgaard, for excellent supervision, many fruitful discussions and for encouraging me when all things seemed to be a bit too hard. Next, I want to thank Senior scientist Nicolas Friggens for giving me the necessary insight into the biology involved in my work and for many fruitful discussions on the applications of my work to biology. I also wish to thank Professor Paul Fearnhead, Lancaster University for hospitality and inspiration in my work during my stay abroad. Finally, I would like to thank my family and friends for being there for me and encouraging me.

Århus, June 2nd, 2008 Jørgen Vinsløv Hansen

Summary

In the last few decades management of cattle herds has become a much more automated process because of the increasing availability of robotic devices such as milking machines. These machines partly take over the work that the farmer earlier had to do himself. In addition, the robot in many cases provide built-in machinery for measuring a number of biological entities online. For example, in this thesis one of the main focuses is on measurements of progesterone concentration in cow milk, which can be obtained from milking machines. Often data of this type holds information that can be very useful for a farmer in the management of the cows. Therefore, with the increasing use of these technologies, there is also an increasing demand for developing statistical methods to extract as much information as possible from the data. Statistical tools for analysing time series has been developed through the last 50 years. State space models is a large class of models that can be applied to time series and much of the work behind this thesis is based on state space models. A thorough overview on state space space can be found in e.g. West and Harrison (1989), Durbin and Koopman (2001) and Brockwell and Davis (2002).

Two large datasets containing online records from milking machines have been analysed as a part of the study behind this thesis. The first data set contain daily milk yields and is analysed using a parametric model to quantify the effects of breed and parity on the milk yield. Also the acceleration in the yield in the beginning of lactation is studied in order to provide an indicator of physiological stress. The second data set contain progesterone concentrations in cow milk. The concentration of progesterone in the milk is closely connected to the reproductive status of a cow because the cow produces progesterone with a varying intensity through its reproductive cycle. Therefore, progesterone measurements can be useful to detect when a cow is in oestrus and thus receptive of insemination. A state space model is developed for the purpose of analysing the progesterone data. Though developed with the aim of analysing a specific dataset, this state space model can be used to model any time series that has a cyclic behavior where the

mean of the observations in each cycle is continuous and piecewise linear. For some of the cows in the dataset the time points of successful inseminations is known. Using these, the ability of the model to predict when cows go into oestrus can be evaluated.

Asymptotic results for parameter estimates from state space models have been studied for some years. Under a set of regularity conditions the maximum likelihood estimate has been found to be asymptotic normal. An alternative method to estimate parameters is to use estimating equations. In connection with the progesterone data we use estimating equations to estimate parameters. The asymptotic behavior of these estimates is studied and for a class of estimating equations asymptotic normality is verified.

The thesis consist of a review and three independently written papers. One of the papers has already been published and the two other papers have been submitted. The co-authors of the papers are my supervisors Jens L. Jensen from the Department of Mathematical Sciences, University of Aarhus and Søren Højsgaard and Nicolas Friggens from the Faculty of Agricultural Sciences, University of Aarhus.

Contents

Accompanying Papers

A Hansen, J.V., Friggens, N.C. and Højsgaard, S.(2006).
The influence of breed and parity on milk yield, and milk yield acceleration curves
Livestock Science, 104:1-2, 53-62.

B Hansen, J.V., Jensen, J.L., Friggens, N.C. and Højsgaard, S.(2008).
A state space model exhibiting a cyclic structure with an application to progesterone concentration in cow milk
To be submitted (primo June 2008) to J. Roy. Stat. Soc. Series C.
To appear as Thiele Research Report, Department of Mathematical Sciences, University of Aarhus.

C Hansen, J.V. and Jensen, J.L. (2008).
Asymptotics for estimating equations in hidden Markov models
Submitted to Statistica Sinica.
To appear as Thiele Research Report, Department of Mathematical Sciences, University of Aarhus.

Chapter 1

Introduction

Technology that collects online data consisting of different physical and biological entities has developed quickly in the last decades. Therefore in many cases vast amounts of data can now be collected very cheaply. Because these online data often contain valuable information there is a growing interest in the development of statistical tools to extract the information from data. State space models is a class of models that can be applied to time series and here we focus on this class of models. We use state space models to analyse certain datasets and study as well more theoretical aspects of this class of models.

This review is organised as follows. In Chapter 2 state space models and corresponding concept of filtering are presented together with some methods for parameter estimation. Especially the technique of estimation using estimating equation is important in this thesis as it is both used for estimation and the asymptotic behavior of this kind of estimates are investigated.

In Chapter 3, the analysis of a dataset consisting of daily milk yields is presented. The aim was to quantify effects of breed and parity on lactation curves and data was analysed using a parametric model suggested in Emmans and Fisher (1986).

In Chapter 4, a state space model developed to handle time series with a cyclic behavior is presented. Further it is assumed that the mean of the observations in each cycle is continuous and piecewise linear. For such a model a corresponding filter has been developed and estimation of parameters will be discussed.

Finally in Chapter 5, we present the results of a study where the asymptotic normality of a class of estimates, obtained using estimating equations, is established.

Chapter 2

State Space Models

State space models is a flexible class of models for analysing data that is collected as online records. This class of models generalises the classical autoregressive moving average (ARMA) models introduced by Box and Jenkins (1970) for analysing time series. Often when a model for a time series can be given both a state space and e.g. an ARMA representation for computational reasons the ARMA representation is preferred because there exist better developed software for analysis of this smaller class of models. However, there can also be good reasons for choosing the state space approach. For example, depending on the formulation of a state space model the parameters can often be given a better interpretation in biological or physical terms as opposed to an ARMA model where the parameters may be hard to interpret. In engineering the system equation of a state space model will often correspond to differential equation in which case it is natural to use the state space approach. Comprehensive treatments of the state space approach to time series analysis can be found in e.g. West and Harrison (1989), Durbin and Koopman (2001) and Brockwell and Davis (2002).

A state space model consists of two processes. An unobserved *latent process* and an *observation process*. Often the observer is interested in assessing the latent process. The latent process is assumed to be a Markov process allowing for serial correlation among the observed variables. To model the data it is assumed that the observations are independent in the conditional distribution given the latent process.

Here the most fundamental concepts concerning state space models are introduced through the example of the well known Gaussian state space model. This model is presented in a Bayesian setting using the terminology of West and Harrison (1989). Firstly, a formal definition of the Gaussian model is given. Then the Kalman filter and smoother for assessing the latent process is presented. Finally, a short presentation of a few techniques for

estimation of unknown parameters is given.

2.1 Definition

This section provides the definition of a Gaussian state space model. State space models can be applied to multivariate responses $\{Y_t\}$. Let d denote the dimension of the process. The process $\{Y_t\}_{t=1}^n$ is called the *observation process*. We consider here only processes with a discrete time parameter t, so that the index t is an enumeration of the observations. We also introduce the p-dimensional *latent process* $\{\theta_t\}_{t=1}^n$. The following three descriptions now define the Gaussian state space model

$$Y_t = F_t^\top \theta_t + v_t, \qquad v_t \sim N_d(0, V_t) \tag{2.1}$$

$$\theta_t = G_t \theta_{t-1} + w_t, \qquad w_t \sim N_p(0, W_t) \tag{2.2}$$

$$\theta_0 | D_0 \sim N_p(m_0, C_0), \tag{2.3}$$

for $t = 1, \ldots, n$, where $F^\top$ denotes the transpose of the matrix F. The equation (2.1) is traditionally called the *observation equation* and (2.2) is called the *system equation*. In this definition the following notation has been introduced.

$F_t^\top$: The $d \times p$ design matrix at time t.

G_t : The $p \times p$ *evolution transfer matrix* at time t.

v_t : The d-dimensional *observational error* vector at time t with variance matrix V_t.

w_t : The p-dimensional *evolution error* vector at time t with variance matrix W_t.

D_0 : The information available on the state vector at time zero. The information is stated as a prior distribution with mean m_0 and variance C_0.

The observational errors $\{v_t\}_{t=1}^n$ are assumed to be mutually independent as are the evolution errors $\{w_t\}_{t=1}^n$. Additionally, these two processes are assumed to be independent.

The matrices $F_t^\top, G_t, V_t$ and W_t, $t = 1, \ldots, n$ are usually assumed to be known. However, it is possible to let the matrices depend on a parameter ψ, which can be estimated using the maximum likelihood method, see Section 2.3

2.2 Assessment of the state vector

When an experimenter is using a state space model the primary interest usually is to assess the state vector at different time points. As more data becomes available the information about the states increases. To formalise this we introduce the information set D_t which is the information available about the state process at time t. The experimenter may have some knowledge about the process prior to collecting data. This information is denoted D_0 and effectively consists of the mean and variance of the prior distribution (2.3) for the initial state θ_0. We assume that the only additional information to D_0 that is gained after time $t = 0$ is the observed values of the observation process. That is, given D_0 the information set is recursively defined as $D_t = D_{t-1} \cup \{y_t\}$.

2.2.1 Kalman Filtering

During the process of collecting data, at time $t \leq n$, we are often interested in assessing the current state θ_t. That is, data $y^t = \{y_i\}_{i=1}^t$, has been collected and we want to determine the conditional distribution $\theta_t|D_t$. As new data y_{t+1} arrives one aims at a simple updating to find $\theta_{t+1}|D_{t+1}$.

The famous *Kalman filter* solves this problem through a set of updating equations, that was first derived by Thiele (1880) (see Lauritzen (2002)). However, the significance of this work was not understood until Kalman (1960) was published. Therefore, the filter is named after Kalman.

The start of the filter is simply the assumption (2.3). Writing $\theta_{t-1}|D_{t-1} \sim N_p(m_{t-1}, C_{t-1})$ we find from the system equation (2.2) that $\theta_t|D_{t-1} \sim N_p(a_t, R_t)$ with

$$\begin{aligned} a_t &= G_t m_{t-1} \\ R_t &= G_t C_{t-1} G_t^\top + W_t. \end{aligned}$$

Next, using the observation equation (2.1) to obtain $y_t|D_{t-1} \sim N_d(f_t, Q_t)$ where

$$\begin{aligned} f_t &= F_t^\top a_t \\ Q_t &= F_t^\top R_t F_t + V_t. \end{aligned}$$

Using standard normal distribution theory we find that $\theta_t|D_t \sim N_p(m_t, C_t)$ with

$$\begin{aligned} m_t &= a_t + A_t(y_t - f_t) \\ C_t &= R_t - A_t Q_t A_t^\top, \end{aligned}$$

where $A_t = R_t F_t Q_t^{-1}$.

2.2.2 Kalman Smoothing

Having observed all data $\{y_t\}_{t=1}^n$ up until time n the experimenter may also be interested in assessing the states at previous time points. This objective is called *smoothing*. For Gaussian state space models the conditional distributions $\theta_t|D_n$ can be obtained using the *Kalman smoother*. The Kalman smoother also consists of a set of recursive equations. Contrary to the Kalman filter these recursions are carried out backwards in time. The smoother takes the values m_t, C_t, and R_t of the Kalman filter as input and begins with $\theta_n|D_n = N_p(\tilde{m}_n, \tilde{C}_n)$. The Kalman filter immediately yields $\tilde{m}_n = m_n$ and $\tilde{C}_n = C_n$. The remaining smoothed states $\theta_t|D_n = N_p(\tilde{m}_t, \tilde{C}_t)$, $t = 1, \ldots, n-1$ are given by the recursions:

$$\tilde{m}_t = m_t + B_t(\tilde{m}_{t+1} - G_{t+1}m_t)$$
$$\tilde{C}_t = C_t + B_t(\tilde{C}_{t+1} - R_{t+1})B_t^\top$$

where $B_t = C_t G_{t+1}^\top R_{t+1}^{-1}$. A sketch of a proof that these recursions yield the desired result goes as follows: From the construction of the process we have that $(\theta_t|\theta_{t+1}, D_n)$ equals $(\theta_t|\theta_{t+1}, D_t)$. The latter can be calculated from $\theta_t|D_t$, the system equation and $\theta_{t+1}|D_t$.

2.3 Estimation of parameters

Until now we have assumed that the matrices F_t, G_t, V_t and W_t defining the state space model are known. It is possible to let these matrices depend on an unknown parameter ψ. In Sections 2.3.1, 2.3.2 and 2.3.3 three methods for estimating the unknown parameter are described.

2.3.1 Likelihood estimation

Here we consider maximum likelihood estimation of the parameter. Assuming that the initial prior is known we can determine the joint distribution of the observation process as a function of ψ. Thus we get the log likelihood function

$$\begin{aligned} l(\psi|y) &= \log p(y|\psi) \\ &= \sum_{t=1}^{n} \log p(y_t|y_1, \ldots, y_{t-1}, \psi) \\ &= -\frac{nd}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}(\log|Q_t| + (y_t - f_t)^\top Q_t^{-1}(y_t - f_t)), \end{aligned}$$

where f_t and Q_t are obtained by running the Kalman filter with a fixed value of ψ. The likelihood function can be maximised numerically yielding the maximum likelihood estimate $\hat{\psi}$.

2.3.2 EM algorithm

The EM algorithm (Dempster et al. 1977) is an algorithm designed to find maximum likelihood estimates in situations in which there are missing data. This scenery is seen in connection with state space models if we think of the unobserved state variables to be missing data. Here we illustrate how this algorithm work with an example where the variances V_t and W_t are unknown but assumed to be constant over time. Therefore we let $V_t = V$ and $W_t = W$ for all t. Also let $Y = \{Y_t\}_{t=1}^n$ and $\theta = \{\theta_t\}_{t=1}^n$. Then the likelihood density $p(Y, \theta|V, W)$ is Gaussian. To start the algorithm initial estimates V^0 and W^0 are chosen. Then in each iteration of the algorithm we go through the following two steps

- *E-step* (Expectation)
  Calculate the conditional expectation

$$E\left[\log p(Y, \theta|\ V, W)|Y, V^{(m)}, W^{(m)}\right] \tag{2.4}$$

  as a function of V and W.

- *M-step* (Maximisation) Maximise 2.4 wrt. V and W to obtain new estimates $V^{(m+1)}$ and $W^{(m+1)}$.

The main feature of this algorithm is that the likelihood increases at each iteration. That is

$$L(V^{(m+1)}, W^{(m+1)}|Y) \geq L(V^{(m)}, W^{(m)}|Y),$$

so that under mild assumptions the algorithm will converge to a local maximum of the likelihood. Essentially, each iteration of the updates the parameter estimates by solving

$$\frac{\partial}{\partial(V, W)} E\left(\log(L(V, W|Y, \theta))\right) = 0$$

In the example with constant variances the E-step can be simplified using the the conditional independence of the observations given the state.

Since we think of m_0 and C_0 as known we find that

$$\begin{aligned}\log p(Y,\theta|V,W) &= \text{constant} \\ &= \sum_{t=1}^{n} \log p(Y_t|\theta,V) \qquad (2.5)\\ &= \sum_{t=1}^{n} \log p(\theta_t|\theta_{t-1},W). \qquad (2.6)\end{aligned}$$

Hence V and W can be estimated separately from (2.5) and (2.6). The updated estimates can be found to be

$$\hat{V} = \frac{1}{n}\sum_{t=1}^{n} F_t^\top \tilde{C}_t F_t + \frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t^\top \tilde{m}_t)(Y_t - F_t^\top \tilde{m}_t)^\top$$

and

$$\hat{W} = \frac{1}{n}\sum_{t=1}^{n} L_t + \frac{1}{n}\sum_{t=1}^{n}(\tilde{m}_t - G_t\tilde{m}_{t-1})(\tilde{m}_t - G_t\tilde{m}_{t-1})^\top$$

where

$$L_t = \tilde{C}_t + G_t\tilde{C}_{t-1}G_t^\top - \tilde{C}_t B_{t-1}^\top G_t^\top - G_t B_{t-1}\tilde{C}_t^\top.$$

A detailed derivation of these results can be found in Klein (2003).

2.3.3 Estimating equations

Here we consider an alternative method for estimation of the unknown parameter ψ. Let $\nu(\psi,\bar{\theta},y)$ be a function of the parameter ψ, a triple $\bar{\theta}$ of consecutive states and an observed variable y. Let $\nu_i(\psi) = \nu(\psi,\bar{\theta}_i,y_i)$ where $\bar{\theta}_i = (\theta_{i-1},\theta_i,\theta_{i+1})$. We think of $\sum_{i=1}^n \nu_i(\psi) = 0$ as an estimating equation had both the variables $\{\theta_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ been observed. Having observed only the process $\{y_i\}_{i=1}^n$ we use the estimating equation

$$\sum_{i=1}^{n} E_\psi[\nu_i(\psi)|(1,n)] = 0, \qquad (2.7)$$

where $E_\psi[\cdot|(1,n)]$ is the conditional mean given $y_1,\ldots,y_n$. To solve (2.7) one often uses an EM-type algorithm. That is, $\sum_{i=1}^n E_\psi[\nu_i(\tilde{\psi})|(1,n)] = 0$ is solved with respect to $\tilde{\psi}$, and this defines a new value improving on the old value ψ. In the actual case where ν_i is chosen to be the terms of score function $\frac{dl}{d\psi}(\psi,\bar{\theta}_i,y_i)$ this algorithm corresponds to the EM algorithm. This EEE-algorithm (EEE is short for Expectation-Estimating-Equation)

has been considered in Heyde and Morton (1996), Rosen et al. (2000) and Elashoff and Ryan (2004). In Hansen et al. (2008) we use this algorithm to estimate the parameters of a state space model and in Hansen and Jensen (2008) asymptotic results of this kind of estimates are studied. In Section 4.5 below a specific example of estimation using estimating equations is presented.

Chapter 3

The influence of breed and parity on milk yield, and milk acceleration curves

This paper is based on an experimental study which had two purposes. Firstly, we wanted to determine the effects of breed and parity[1] on milk yield from cows. Secondly, milk yield acceleration[2] was examined for an introductory investigation of the possibility of using acceleration as an indicator of physiological stress and therefore also health problems. Here the statistical analysis involved in the study is briefly presented.

3.1 The Emmans and Fisher model

We analysed data consisting of daily recordings of milk yield for a large group of cows of three different breeds and three different parities. The daily milk yield at t days past calving is denoted $\mu(t)$. For each individual cow-lactation the model of Emmans and Fisher (1986)

$$\mu(t) = \exp\left(\tilde{a} - te^{\tilde{c}} - \exp(G_0 - te^{\tilde{b}})\right) \tag{3.1}$$

was fitted to data. The parameters $\tilde{a}, \tilde{b}, \tilde{c}$ and G_0 of the model was estimated using the ordinary least square procedure. In Figure 3.1 some typical examples of the development in daily milk yield through a lactation period are shown together with the corresponding fits of the model (3.1).

[1]Parity denotes the number of calves a cow has given birth to. That is, a cow of first parity has just given birth to its first calf.

[2]Denoting the daily yield at time t by $\mu(t)$, the yield acceleration is defined to be the derivative $\mu'(t)$.

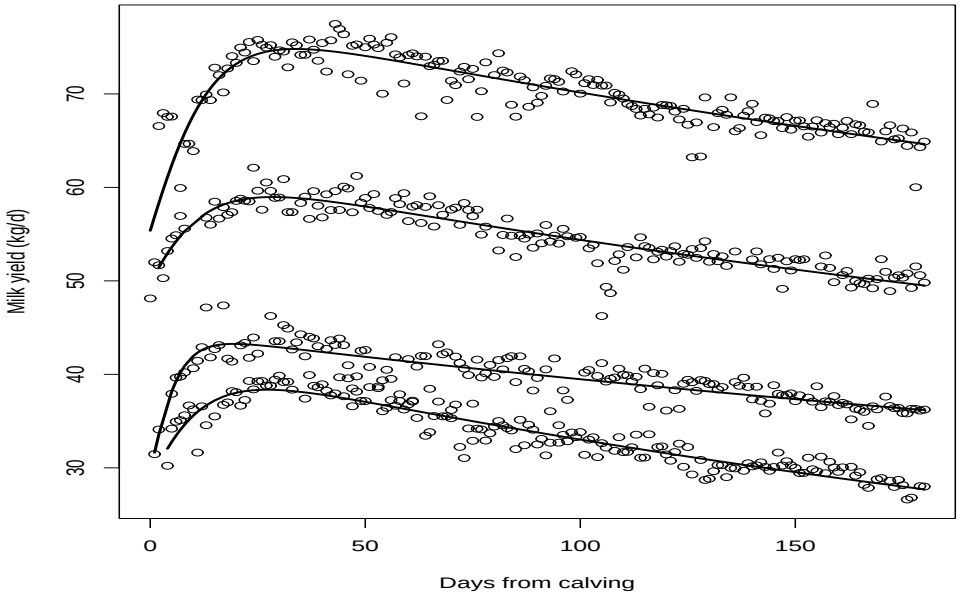

Figure 3.1: Plots of milk yield (kg/d) (d=day) records for four cows of second parity. The curves are vertically displaced by 0, 10, 20 and 30 kg/d to avoid overlap. For each cow the corresponding fit of the Emmans and Fisher model is shown.

3.2 Analysis of parameters

The investigation of the effect of breed and parity on the four parameters $\tilde{a}, \tilde{b}, \tilde{c}$ and G_0 of the model were carried out using a mixed linear model. Breed and parity were included as fixed effects and cow id was included as a random effect. E.g. with b denoting breed, p parity and c cow the parameter $\tilde{a}$ was analysed as

$$\tilde{a}_{bpc} = \alpha_b + \beta_p + \gamma_{bp} + u_{bc} + \epsilon_{bpc}$$

with $u_{bc} \sim N(0, \omega^2)$ and $\epsilon_{bpc} \sim N(0, \sigma^2)$.

The acceleration in milk yield at time t is defined to be the derivative $\mu'(t)$. Based on the estimates of the four parameters in model (3.1) for each cow-lactation we could calculate the maximal acceleration and the maximal daily milk yield as well as the two corresponding time points. These four entities were analysed for breed and parity effects in the same way as the original four parameters above.

3.3 Results

The effect of breed and parity found from the analysis is illustrated in Figure 3.2. The differences found between breeds and between parities are typical of the values reported in previous literature. Furthermore no significant interactions were found between the effects of breed and parity. With regard to the introductory investigation of milk yield acceleration indexing the degree of physiological stress experienced by cows, it was found that the acceleration was highest immediately after calving and that acceleration was highest for higher yielding cows. This means that the first demands for acceleration to be considered an indicator of physiological stress are fulfilled as literature says that the highest incidences of diseases occur immediately after calving and that higher yielding cows are more likely to have health problems. However, further study is required to show if milk yield acceleration is a useful indicator of stress.

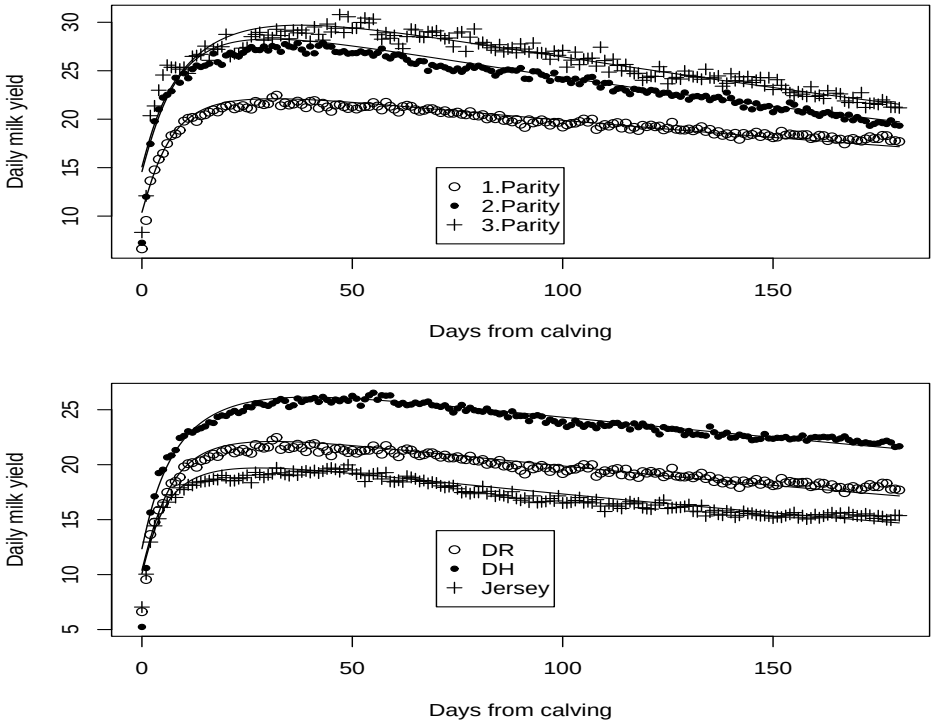

Figure 3.2: Plots of mean daily milk yield. In the upper plot the mean daily milk yield is plotted for cows of breed Danish Red for each parity. In the lower plot the mean daily milk yield is plotted for cows of first parity for Danish Red, Danish Holstein and Jersey. Along each group of points the curve obtained by taking mean of the corresponding curve fits is drawn.

Chapter 4

A state space model exhibiting a cyclic structure with an application to progesterone concentration in cow milk

The aim of this study was to evaluate the concentration of progesterone as an indicator for oestrus in cows. In its reproductive cycle the progesterone secreted by a cow shows a cyclic pattern. For biological reasons the concentration of progesterone in the milk decreases before the cow goes into oestrus (Peters and Ball 1995). This drop in progesterone is very rapid and ends about 70 hours before ovulation (Roelofs et al. (2006)). To accomplish the aim of the study a state space model, which supposedly should be able to capture the decrease in progesterone concentration, was developed. The ability of the model to predict when a cow is in oestrus was evaluated using a dataset containing measurements of progesterone concentration in the milk and knowledge of successful inseminations. In the following a short overview of the model, the corresponding filter and the estimation techniques used in the paper is given.

4.1 Cyclic state space model

The state space model that was developed incorporates the idea of m different stages each describing a linear development in the mean of the observations, such that the mean is a continuous function of time. In Figure 4.1 a possible development of the mean of the observations is shown. The model is considered in discrete time $t \in \mathbf{Z}$ and has five hidden variables that hold

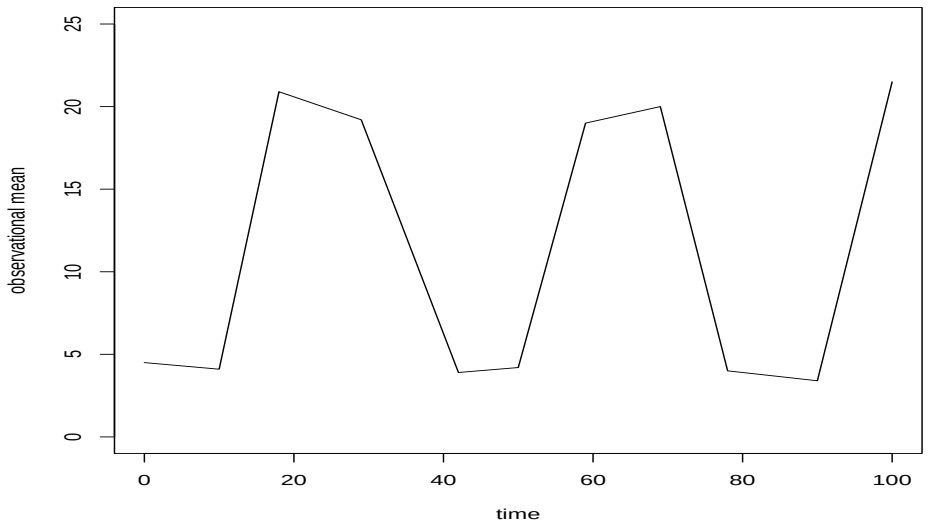

Figure 4.1: Example of possible development in the mean of the observations.

information about the end points of the current stage and the mean of the observations during the stage. The five variables are

R_t : the point in time prior to t with the most recent change of stage, $(R_t < t)$.

S_t : the stage entered at time R_t with value in $\{1, 2, \ldots, m\}$.

N_t : the point in time for the next change of stage after R_t, $(N_t \geq t)$.

a_t : the mean of an observation at time R_t.

b_t : the mean of an observation at time N_t.

4.1.1 Stochastics of the state variable

With this notation the development of the state variables can now be described. Given that there is a change to stage q at time t, the distribution of the waiting time to the next change depends on q only. This distribution is denoted W_q and the only restriction we put on W_q is that it has finite support. Formally

$$S_{t+1} = S_t + 1 (\text{mod } m) \tag{4.1}$$

and

$$(N_{t+1} - N_t | N_t = t, S_{t+1} = q) \sim W_q. \tag{4.2}$$

Thus (4.1) and (4.2) describes the dynamics of the three discrete hidden variables (R_t, S_t, N_t).

Next, the stochastic behavior of the continuous hidden variables a_t and b_t is described. These two variables hold information about the mean of the observations within a stage. For each time point t where there is a change of stage, let $x(t)$ denote the hidden mean of a possible observation y_t. If the stage of a time interval beginning at time t is q, then we assume

$$(x(t) | N_t = t, S_{t+1} = q) \sim N(\mu_q, \omega_q^2), \tag{4.3}$$

where μ_q and ω_q^2 are parameters. The hidden variables a_t and b_t are then defined to be

$$a_t = x(R_t) \text{ and } b_t = x(N_t). \tag{4.4}$$

That is, a_t is the mean at the beginning and b_t the mean at the end of the stage entered at time R_t. This means that if there is no change of stage at time t then $(a_{t+1}, b_{t+1}) = (a_t, b_t)$. On the other hand if there is a change of stage at time t, then $a_{t+1} = b_t$ and $b_{t+1} = x(N_{t+1})$.

4.1.2 Stochastics of observations

The mean at any time point $s \in \mathbf{R}$ is defined by linear interpolation using the mean at the end points of the stage. That is,

$$x(s) = a_t + \frac{b_t - a_t}{N_t - R_t}(s - R_t), \quad s \in \mathbf{R}, t = \lceil s \rceil,$$

where $\lceil s \rceil$ is the smallest integer greater than or equal to s. In this way the underlying mean $x(s)$ of the observation given the hidden variables is continuous and piecewise linear. Figure 4.1 shows a realisation of $x(\cdot)$.

Data $\{(y_i, s_i) | i = 1, \ldots, n\}$ consist of a set of observations y_i recoded at time points s_i, where n is the number of observations. To define the distribution of data, assume that we have an observation y_i at time $s_i \in \mathbf{R}$. Note that we do not restrict the observations to occur at time points that are multiples of the time unit. If the stage at time s_i is q, that is, $S_{\lceil s_i \rceil} = q$, we assume

$$\left(y_i | (R, S, N, a, b)_{\lceil s_i \rceil}\right) \sim N(x(s_i), \sigma_q^2),$$

where $\sigma_q^2, q = 1, 2, \ldots, m$ are parameters.

To summarize the parameters of the model are

μ_q : the mean of the hidden stochastic mean at a time point where the stage changes to q,

ω_q^2 : the variance of the hidden stochastic mean at a time point where the stage changes to q,

σ_q^2 : residual variance of observations within stage q,

for $q = 1, 2, \ldots, m$. Note that the waiting time distributions $W_q, q = 1, 2, \ldots, m$ may depend on an unknown parameter θ.

4.2 Approximate filter

An approximate filter has been developed for the model defined in Section 4.1. Using the notation $y^s = \{y_i | s_i \leq s\}$ and $y_r^s = \{y_i | r < s_i \leq s\}$ our goal is to determine the filter densities which we write as

$$p(R_t = j, S_t = q, N_t = l, a_t, b_t | y^t) = p_t(j, q, l, a_t, b_t) \text{ for all t } \in \mathbf{N}. \quad (4.5)$$

We use the approximation

$$p_t(j, q, l, a_t, b_t) = p_t(j, q, l)\phi(a_t, b_t;\ \mu_t(j, q, l), \Sigma_t(j, q, l)), \quad (4.6)$$

where $p_t(\cdot, \cdot, \cdot)$ on the right hand side of (4.6) is the marginal density of (R_t, S_t, N_t) and $\phi(\cdot, \cdot;\ \mu, \Sigma)$ is the normal density with mean μ and variance Σ. Therefore the filter densities (4.5) are specified by $p_t(j, q, l)$, $\mu_t(j, q, l)$ and $\Sigma_t(j, q, l)$. If the filter distribution at time t is of the form (4.6), the distribution

$$p(a_{t+1}, b_{t+1} |\ R_{t+1}, S_{t+1}, N_{t+1}, y^{t+1}) \quad (4.7)$$

is again Gaussian if $R_{t+1} \neq t$. That is if there is no change of stage at time t. But if the stage changes at time t this is not the case. The distribution of (4.7) is then a mixture of Gaussian distribution. In this case we make an approximation and therefore the filter is only an approximate filter. We do not state the updating equations of the filter here as they are given in Hansen et al. (2008).

4.3 Parameter estimation

Maximum likelihood can be used to estimate the residual variances $\sigma_q^2, q = 1, \ldots, m$. This estimation procedure is described in Section 4.4. The waiting time distributions can be estimated using an EEE-algorithm as described in Section 4.5. We do not suggest any general procedures for estimation of the parameters μ_q and ω_q^2, $q = 1, \ldots, m$.

4.4 Estimation of the residual variance

The residual variance parameters σ_q^2, $q = 1, \ldots, m$ can be estimated using maximum likelihood. From the derivation of the filter in Hansen et al. (2008) it is seen that

$$\frac{p(y^t)}{p(y^{t+1})} = \frac{1}{p(y_t^{t+1}|y^t)}.$$

can be calculated approximately using entities found from the filter updates. Therefore, using the approximate filter, we can calculate an approximation to the likelihood function

$$L(\sigma_1^2, \ldots, \sigma_m^2) = p(y^n) = \prod_t p(y_t^{t+1}|y^t),$$

which can be maximized using numerical techniques to find estimates $\hat{\sigma}_q^2$ of the residual variances.

4.5 Estimation of the distribution of waiting times

Given a model for the waiting time distributions $W_q, q = 1, \ldots, m$, we can estimate the parameter θ of this model using an EEE algorithm as discussed in Section 2.3.3. For a parameter θ of the waiting time distribution we use an estimating function of the form $\sum_1^n \psi_i$, where $\psi_i = \psi(z_i, z_{i-1}; \theta)$ for a function ψ, and where $z_i = (R_i, S_i, N_i)$. The E (expectation) step is to calculate

$$E\Big(\sum_1^n \psi_i | y^n\Big),$$

where n is the number of observations. As an example consider the case where the waiting time probabilities are modeled with no other restriction than $\sum_{l=1}^M W_q(l) = 1$ for all q. We can then use the estimating functions

$$\begin{aligned}&\psi(z_t, z_{t-1}; q, l)\\ &\quad = 1(R_t = t-1, S_t = q, N_t = t-1+l) - W_q(l)1(R_t = t-1, S_t = q),\end{aligned}$$

where $1(\cdot)$ is the indicator function. In the EE (estimating equation) step the new value of $W_q(l)$ becomes

$$W_q(l) = \frac{E\left(\sum_1^n 1(R_t = t-1, S_t = q, N_t = t-1+l)|y_1^n\right)}{E\left(\sum_1^n 1(R_t = t-1, S_t = q)|y_1^n\right)},$$

where the nominator and denominator have been found in the E step.

Chapter 5

Asymptotics for estimating equations in hidden Markov models

State space models and the Kalman filter was introduced more than fifty years ago and since then the asymptotic behavior of parameter estimates has been studied. Several contributions on this subject have been made. Among these are the pioneering paper of Baum and Petrie (1966) where asymptotic normality of the maximum likelihood estimator is established for the case when the state spaces for both the hidden and the observed variables are finite. This result is generalised to a general state space for the observed variable in Bickel et al. (1998). In Jensen and Petersen (1999) a further generalisation to a non-discrete state space for the hidden variable is given. Here the contributions of Hansen and Jensen (2008) to this line of results is shortly described in Section 5.1. In Section 5.2 an example to which these new results can be applied is presented.

5.1 New asymptotic results

In Hansen and Jensen (2008) we generalise the existing results on asymptotic normality of the maximum likelihood estimator in state state models in two ways. Firstly, we introduce the possibility of non-stationarity of the hidden process through a covariate process $\{z_i\}$ influencing the hidden Markov process. Stationarity has been an important assumption for most of the previous results on asymptotics in state space models. Secondly, we consider not only the maximum likelihood estimate, but instead a class of estimates obtained using estimating equations as in Section 2.3.3 above.

Formally, we consider an observed process $y_1, \ldots, y_n$ controlled by an unobserved Markov process $\{x_i\}$. Conditionally on the x-process the y_is are independent. Both the observed y_i and the unobserved x_i may be influenced by a covariate z_i, making the process inhomogeneous. Therefore, we write the transition probabilities as $p_\theta(x_i|x_{i-1}; z_i)$ and $p_\theta(y_i|x_i; z_i)$, where θ is the unknown parameter for which we want to establish asymptotic normality. For estimation of θ we consider an estimating function $\psi(\theta, \bar{x}, y, z)$ where $\bar{x}$ is a triple of consecutive states. Letting $\psi_i(\theta) = \psi(\theta, \bar{x}_i, y_i, z_i)$, where $\bar{x}_i = (x_{i-1}, x_i, x_{i+1})$, we think of $\sum_{i=1}^n \psi_i(\theta) = 0$ as an estimating equation in the same way as in Section 2.3.3.

Under a set of conditions on the transitions probabilities and the estimating function we establish asymptotic normality of the EEE estimate of θ. Essentially, the proof is split into two halves where first mixing results of the chain are studied before convergence results for the 'observed information':

$$J_n(\theta) = -\frac{\partial}{\partial\theta} E_\theta[\sum_{i=1}^n \psi_i(\theta)|(1, n)]$$

are established. Here $E_\theta(\cdot|(1, n))$ is the conditional mean given $y_1, \ldots, y_n$. With θ_0 denoting the true value of the parameter, in its final form the asymptotic result for the parameter estimate is stated as

Corollary 1. *Assume that the conditions imposed on the transition probabilities and the estimating function hold. Assume that the covariates $\{z_i\}$ are such that the variance of $S_n = \sum_{i=1}^n E_{\theta_0}(\psi_i(\theta_0)|(1, n))/\sqrt{n}$ converges to a positive definite limit $V(\theta_0)$, and also $\frac{1}{n}J_n(\theta_0)$ converges to a positive definite limit $I(\theta_0)$. Then there exists a sequence $\hat{\theta}_n$ solving the estimating equation such that $\hat{\theta} \to \theta_0$ in probability and $\sqrt{n}(\hat{\theta} - \theta_0)$ has a limiting normal distribution with mean zero and variance $I(\theta_0)^{-1}V(\theta_0)I(\theta_0)^{-1}$.*

5.2 Example

In the example of Section 3 in Hansen and Jensen (2008) asymptotic normality of an estimate obtained using estimating equations is established. The model considered corresponds to the cyclic model of Hansen et al. (2008) used to analyse the progesterone data as described in Chapter 4 above. There are however two differences. Firstly, in Hansen and Jensen (2008) we assume that the parameter controlling the distribution of waiting times between points with a change of stage is cow specific. This is not the case in Hansen et al. (2008) where it is assumed that the distribution is the same

for all cows. However, this difference is not important for the asymptotic result. The other difference in the two models is that we assume that the hidden variables corresponding to a_t and b_t in (4.4) belong to compact sets. This is not the case in the model (4.3) for the progesterone data, but the assumption is needed to secure asymptotic normality. Had we chosen a distribution with compact support instead of the normal distribution in (4.3) the assumptions of Hansen and Jensen (2008) would have been met.

Chapter 6

Conclusions

In this thesis, we have studied two datasets consisting of online records. The first dataset consisting of daily milk yields from cows was analysed using a parametric model with the aim of quantifying the effects of breed and parity on lactation curves and evaluating the acceleration in milk yield as an indicator of physiological stress. The second dataset consisting of progesterone measurements was analysed using a state space model with the aim of evaluating progesterone as an indicator of oestrus. Finally, asymptotic behavior of estimates obtained using estimating equations were studied.

Bibliography

Baum, L. and T. Petrie (1966). Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist. 37*, 1554–1563.

Bickel, P., Y. Ritov, and T.Rydén (1998). Asymptotic normality of the maximum likelihood estimator for general hidden markov models. *Ann. Statist. 26*, 1614–1635.

Box, G. and G. Jenkins (1970). *Time series analysis, forecasting and control*. Holden Day, San Francico.

Brockwell, P. and R. Davis (2002). *Introduction to Time Series and Forecasting*. Springer-Verlag, New York.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B 39(1)*, 1–38.

Durbin, J. and S. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Elashoff, M. and L. Ryan (2004). An EM algorithm for estimating equations. *J. Comput. Graph. Statist. 13*, 48–65.

Emmans, G. and C. Fisher (1986). Problems in nutrinional theory. *Butterworths, London*, 9–39.

Hansen, J. and J. Jensen (2008). Asymptotics for estimating equations in hidden markov models. *Submitted to Statistica Sinica*.

Hansen, J., J. Jensen, N. Friggens, and S. Højsgaard (2008). A state space model exhibiting a cyclic structure with an application to progesterone concentration in cow milk. *Submitted to J. Roy. Stat. Soc. Series C.*.

Heyde, C. and R. Morton (1996). Quasi-likelihood and generalizing the EM-algorithm. *J. Roy. Statist. Soc. B 58*, 317–327.

Jensen, J. and N. Petersen (1999). Asymptotic normality of the maximum-likelihood estimator in state space models. *Ann. Statist. 27*, 514–535.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*, 25–45.

Klein, B. (2003). *State Space Models for Exponential Family Data*. Ph. D. thesis, Department of Statistics, University of Southern Denmark.

Lauritzen, S. L. (2002). *Thiele: Pioneer in Statistics*. Oxford University Press.

Peters, A. R. and P. J. H. Ball (1995). *Reproduction in cattle*. Blackwell Science, Oxford, UK.

Roelofs, J., F. V. Eerdenburg, W. Hazeleger, N. Soede, and B. Kemp (2006). Relationship between progesterone concentrations in milk and blood and time of ovulation in dairy cattle. *Animal Reproduction Science 91*, 337–343.

Rosen, O., W. Jiang, and M. Tanner (2000). Mixtures of marginal models. *Biometrika 87*, 391–404.

Thiele, T. (1880). Sur la compensation de quelques erreurs quasi-systématiques par la méthodes de moindre carrés. *Copenhagen: Reitzel 43*.

West, M. and J. Harrison (1989). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

Paper

A

Hansen, J.V., Friggens, N.C.,
and Højsgaard, S.(2006)
The influence of breed and parity on milk yield, and milk yield acceleration curves.

Livestock Science, 104:1-2, 53-62.

The influence of breed and parity on milk yield, and milk yield acceleration curves

J.V. Hansen[a,*], N.C. Friggens[b] and S. Højsgaard[a]

[a]*Department of Genetics and Biotechnology, Danish Institute of Agricultural Sciences, Research Centre Foulum, P.O. Box 50, DK-8830 Tjele, Denmark*

[b]*Department of Animal Health, Welfare and Nutrition, Danish Institute of Agricultural Sciences, Research Centre Foulum, P.O. Box 50, DK-8830 Tjele, Denmark*

Abstract

This study had two aims. Firstly, we wanted to quantify the effects of breed and parity on lactation curves. A parametric model for describing milk yield for cows (Friggens et al. (1999)) was used. The data contained 155051 daily records of milk yield from 318 cows of three different breeds; Danish Red, Danish Holstein and Jersey. There were 276, 230, and 98 lactation curves for parities 1, 2 and 3 respectively. For every cow lactation, the parameters of the model were estimated using a least squares procedure for non-linear models. The resulting parameters were analysed in a mixed linear model. Significant effects of parity were observed on the same two parameters as in Friggens et al. (1999). Breed was also found to have a significant effect on some of the parameters. However, there was no significant interaction between breed and parity. The second aim of the study was to evaluate the properties of acceleration in milk yield in the context of providing an indicator for physiological stress and subsequent health problems. Milk yield acceleration was highest around calving and also reflected trends for higher stress/risk for higher yielding cows.

Key words: Milk yield, Lactation, Parity, Breed, Cows.

*Corresponding author: Tel.: +45-8999-1340; fax: +45-8999-1300;
E-mail address: JorgenV.Hansen@agrsci.dk(J.V.Hansen)

1 Introduction

Being able to predict the potential milk production of a cow through lactation period is an important pre-requisite for designing feed rations that will allow this potential to be expressed and feed efficiency maximized. In this context, lactation curve models that can predict potential milk yield using limited information are of relevance. Under commercial conditions the available information is frequently limited to factors such as breed, parity and composite estimates of previous yield. Both breed and parity effects have been shown to exist on lactation curves (e.g. Wood (1980), Collins-Lusweti (1991), Friggens et al. (1999) and Rekaya et al. (2001)) and can now easily be included as fixed factors in test-day models and other linear models that incorporate time trends (Van der Werf et al. (1998) and Macciotta et al. (2005)). However, in such models it is not usually possible to relate the breed or parity effects to the underlying biological processes (see Vetharaniam et al. (2003)). This makes it difficult to build these effects into prediction models that allow the consequences of different potential yields to be evaluated.

An alternative approach is to estimate breed and parity effects in biologically derived lactation curve models (e.g. Dijkstra et al. (1997), Friggens et al. (1999) and Pollott (2000)). These are usually non-linear. In this context, Friggens et al. (1999) provided estimates of parity effects on the different phases of lactation. However, this study was based on data from only one herd and did not estimate how parity effects were affected by breed. The first aim of the present study was to quantify parity effects on lactation curves in different breeds.

It has been suggested that the degree of physiological stress experienced by cows in early lactation can be indexed by the acceleration in milk yield, which is the daily rate of change in yield, during this period (Ingvartsen et al. (2003)). Formally, denoting the daily milk yield at time t by $\mu(t)$ the acceleration in milk yield at that same time point is the derivative $\mu'(t)$. The characteristics that acceleration in milk yield must have if it is to be an indicator of stress, and thus of susceptibility to metabolic diseases, can be deduced from litterature pertaining to production factors affecting disease incidences in early lactation. The highest incidences of diseases occur immediately after calving, substantially before peak yield (Ingvartsen et al. (1999) and Ingvartsen et al. (2003)). Higher yielding cows have also been found to be more likely to have health problems (Pryce et al. (1999) and Hansen (2000)). Thus, for milk yield acceleration to provide a suitable indicator of physiological stress it should reflect these findings; being highest immediately post calving and, at a given time point in lactation, higher for

higher producing animals. To our knowledge there is no published information concerning the properties of milk yield acceleration curves in real datasets. Thus the second aim of this study was to characterise the properties of milk yield acceleration curves.

2 Materials and methods

2.1 Data material

The data set used consisted of daily recordings of milk yield for 604 cow lactations. The cows were of three breeds; Danish red (106 cows), Danish Holstein (129 cows), and Jersey (83 cows). The total number of lactations in first, second and third parities were 276, 230 and 98 respectively. Breed and parity are the factors studied here to determine the effects on the parameters and quantities derived from these parameters. The data were collected between January 1996 and October 2001 at the Danish Cattle Breeders Organisation research farm, Ammitsbøl Skovgård, as a part of a long-term ongoing genetic evaluation programme. The design and methods for the production aspects of the experiment has been described in detail in Nielsen et al. (2003). The experiment was focused on genetic evaluation and therefore environmental conditions including feeding conditions were kept as constant as possible. Cows received one standard total mixed ration, fed ad libitum, throughout lactation containing either 12.88 MJ/kg dry matter or 13.55 MJ/kg dry matter. In Figure 1 on the following page some typical examples of the development in milk yield through a lactation period are shown. Observations where *days from calving* was greater than 180 were discarded to exclude the depressing effects of pregnancy on milk yields at the end of lactation. Those lactation periods with no more than 140 daily milk yield recordings (in those first 180 days) and those with no measurement made before the fifth day after calving were also excluded. These exclusions reduced the number of lactation periods from 604 to 409.

2.2 Lactation curve coefficients

Several models with different functional forms have been proposed in the litterature to model yield data (Rook et al. (1993), Olori et al. (1999) and Val-Arreola et al. (2004)). The Wood's model (Wood (1967)) has been the most widely used. The models used for describing yield can be divided into two groups. Those that aim to describe the underlying biology, and the empirical models that are solely based on the actual data available.

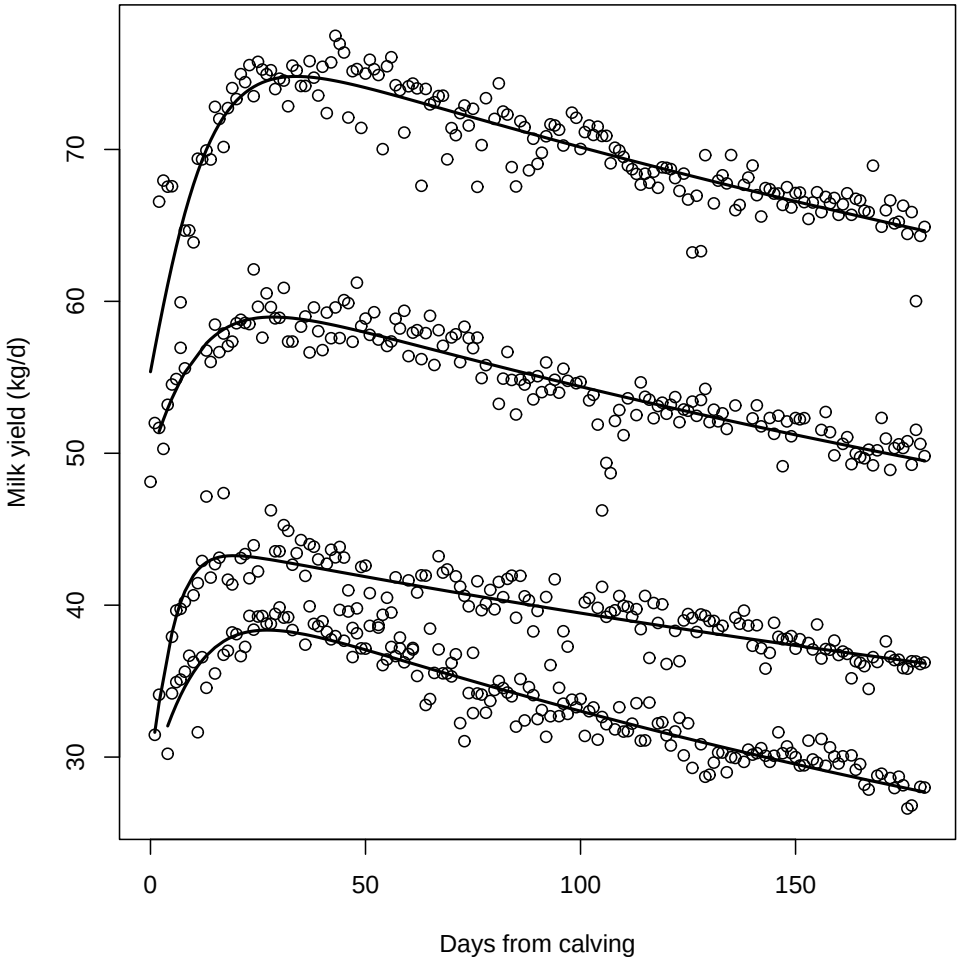

Figure 1: Plots of milk yield (kg/d) records for four cows of second parity. The curves are vertically displaced by 10 kg/d to avoid overlap. For each cow the corresponding fit of the Emmans and Fisher model is shown.

The empirical models include those of (Wood (1967), Grossman and Koops (1988) and Grossman et al. (1999)), whereas the models of Emmans and Fisher (Emmans and Fisher (1986)), Dijkstra (Dijkstra et al. (1997)), Pollott (Pollott (2000) and Pollott (2004)) and (Grossman and Koops (2003)) are amenable to biological interpretation. In this paper, the model of Emmans and Fisher (hereafter referred to as the EF model), which is mathematically equivalent to the model of Dijkstra (see Friggens et al. (1999)), has been used. This model and parametrization was chosen for two reasons. Firstly we chose the model, because it fitted data slightly better than the Wood's model (see Section 3.1) and allows for non-zero yield at time $t = 0$. Secondly, we used the parametrization of the EF model to allow direct comparison with the results of this study with that of Friggens et al. (1999). Also in this model the different parameters determine the development in yield in different regions of time through lactation which is a useful property making interpretation of the parameters possible.

To evaluate the effects of breed and parity on milk yield curves, a two step procedure was used. Firstly, for each cow lactation we estimated the parameters of the EF model. Secondly, these parameter estimates were analysed as observations divided into nine groups by breed and parity as described in Section 2.3. As an alternative to this procedure a non-linear mixed effect model could have been used to model data. However, large models like that will often fit data poorly. A consequence of using the simpler two step procedure compared to a non-linear mixed effect model is that the variance of the parameters a, b, c and G_0 will be overestimated. This will make it harder to find significant effects on the parameters. Therefore conclusions about significant effects will tend to be more robust in this case.

The EF model states that the milk yield of a cow t days after calving is given by the expression

$$\mu(t) = a \cdot \exp\left(-\exp\left(G_0 - bt\right)\right) \cdot \exp\left(-ct\right). \tag{1}$$

Here a, b, c and G_0 are the four parameters specific for every cow lactation that are of interest. In the following, a slightly different parametrization of this model was used: Instead of a, b and c, the logarithms of these three parameters were used, while G_0 was left unchanged. So, let $\tilde{a} = \log(a)$ and similarly let $\tilde{b}$ and $\tilde{c}$ be the logarithms of b and c. Then (1) becomes

$$\begin{aligned}\mu(t) &= \exp\left(\tilde{a}\right) \cdot \exp\left(-\exp\left(G_0 - te^{\tilde{b}}\right)\right) \cdot \exp\left(-te^{\tilde{c}}\right) \\ &= \exp\left(\tilde{a} - te^{\tilde{c}} - \exp\left(G_0 - te^{\tilde{b}}\right)\right). \end{aligned} \tag{2}$$

The reason for these log-transforms of the parameters is that the transformed estimates are much closer to being normally distributed than the

original ones. This is of importance in the subsequent analysis of the parameter estimates. Parametrization (2) was used for all analyses in this paper.

The milk yield acceleration, μ', as a function of time is obtained by differentiation of (2):

$$\begin{aligned}\mu'(t) &= \exp\left(\tilde{a} - te^{\tilde{c}} - \exp\left(G_0 - te^{\tilde{b}}\right)\right) \cdot \left(e^{\tilde{b}} \exp\left(G_0 - te^{\tilde{b}}\right) - e^{\tilde{c}}\right) \\ &= \mu(t) \cdot \left(e^{\tilde{b}} \exp\left(G_0 - te^{\tilde{b}}\right) - e^{\tilde{c}}\right). \end{aligned} \tag{3}$$

The time point for peak yield is found by equating (3) to zero, which gives

$$t_{PY} = \frac{G_0 + \tilde{b} - \tilde{c}}{e^{\tilde{b}}}. \tag{4}$$

Similarly the time point for the maximum milk yield acceleration can be found by differentiating (3) and equating to zero, which yields

$$t_{MA} = \frac{G_0 - \log(\frac{2e^{\tilde{c}} + e^{\tilde{b}} + \sqrt{e^{\tilde{b}}(4e^{\tilde{c}} + e^{\tilde{b}})}}{2e^{\tilde{b}}})}{e^{\tilde{b}}}. \tag{5}$$

For each cow lactation the parameters $\tilde{a}, \tilde{b}, \tilde{c}$ and G_0 were estimated using the ordinary least squares method. Having estimated the four parameters for one cow lactation, the time points for peak yield, t_{PY}, and maximum acceleration, t_{MA}, can be calculated as well as the peak yield, PY, and the maximal acceleration, MA, in milk yield. The time points, t_{PY} and t_{MA}, can be found by (4) and (5). The peak yield and maximum acceleration can then be evaluated as $PY = \mu(t_{PY})$ and $MA = \mu'(t_{MA})$ respectively.

2.3 Effects on the parameters

All statistical analysis were carried out using the statistical programming language **R** (R Development Core Team (2004)) and the MIXED procedure in SAS (SAS Institute Inc (2001)). We investigated the effects of breed and parity on the parameters and functions of these (i.e. t_{PY}, t_{MA}, PY and MA). For each parameter, these effects were modelled using a mixed linear model. Parity and breed were included as fixed effects of the parameters and cow was included as a random effect, so that with b indexing breed, p parity and c cow within breed, the parameter $\tilde{a}$ was analysed as

$$\tilde{a}_{bpc} = \alpha_b + \beta_p + \gamma_{bp} + u_{bc} + \epsilon_{bpc} \tag{6}$$

with $u_{bc} \sim N(0, \omega^2)$ and $\epsilon_{bpc} \sim N(0, \sigma^2)$. The values of α_b and β_p determine the fixed effects of breed and parity respectively, with γ_{bp} being the interaction between the two factors. The parameter ω^2 is the covariance of estimates of $\tilde{a}$ from different lactations on the same cow and σ^2 is the residual variance for each estimate. This means that $\omega^2/(\omega^2 + \sigma^2)$ is the correlation between estimates from different lactations coming from one cow. When estimating the parameters $\tilde{a}, \tilde{b}, \tilde{c}$ and G_0 using the `nls`-function of **R** (R Development Core Team (2004)) one also gets standard errors of these estimates. These standard errors allow the possibility of weighting the observations in the model (6) so that estimates with a high standard error are given a correspondingly low weight.

3 Results and Discussion

3.1 Curve fitting

For each cow lactation the parameters of the EF model were estimated. Out of the 409 cow lactations, the `nls`-function did not converge for 37 of them. Missing data in the phase of acceleration and around peak yield caused these convergence problems. Some sets of estimates were regarded as outliers: Sets where one or more of the four parameter estimates deviated from the mean by ± 2.5 standard deviations were excluded. This removed 26 cow lactations from further study so that 346 sets of estimates remained. In most of these outlier lactations the cows had severe diseases that caused milk yield to deviate substantially from the normal, potential lactation curve shape. In Table 1 on the next page the mean of the parameter estimates are given, grouped according to breed and parity. The average R^2 of these curve fits was 0.648 (first quartile=0.54, median=0.67, third quartile=0.77) and the average residual standard deviation of the curve fit was 1.856 kg/d (1.31, 1.67, 2.24). To compare with Wood's model which converged for all these 346 cow lactations we found an average R^2 of 0.613 (0.50, 0.63, 0.74) and an average residual standard deviation of 1.941 kg/d (1.39, 1.75, 2.33) indicating that the EF model fitted the data slightly better than the Wood's model.

3.2 Parity and breed effects

The effect of breed and parity on milk yield is illustrated in Figure 2 on page 9. The plots show that both factors have an effect on the yield. The differences between breeds and between parities are typical of the values

Table 1: The mean of the estimates of each parameter are given together with the standard deviation of the estimates (in the parenthesis). The means are calculated groupwise for every breed and also for every parity. In the last row means and standard deviations of all estimates are given.

Group	Parameter			
	$\tilde{a}$	$\tilde{b}$	$\tilde{c}$	G_0
Danish Red	3.34(0.24)	−2.00(0.67)	−6.20(0.61)	−0.18(0.33)
Danish Holstein	3.48(0.25)	−2.02(0.78)	−6.37(0.59)	−0.19(0.38)
Jersey	3.17(0.20)	−1.69(0.73)	−6.12(0.60)	−0.22(0.40)
Parity 1	3.22(0.21)	−1.88(0.71)	−6.45(0.63)	−0.19(0.35)
Parity 2	3.43(0.24)	−1.90(0.71)	−6.13(0.55)	−0.20(0.39)
Parity 3	3.49(0.27)	−2.04(0.89)	−5.94(0.46)	−0.20(0.37)
Overall	3.34(0.26)	−1.91(0.74)	−6.24(0.60)	−0.20(0.37)

reported in the literature, see e.g. Nielsen et al. (2003). Figure 2 suggests that the time points for peaks of the mean yield curves do not match the mean of the corresponding time points for peak yields though. This is caused by the empirical fact that the milk yield acceleration a short while before peak yield is larger in numerical value than the negative acceleration shortly after peak yield (see Figure 3 on page 10). Therefore the peaks of the curves in Figure 2 on the facing page are shifted to the right of the mean of time points for peak yield. Now we will apply the models described in Section 2.3 to the 346 sets of estimates in order to find out in which way breed and parity affects the parameters of the EF model. Twelve tests for no interaction between breed and parity were carried out. Two for each of the four parameters of the EF model (with and without using the weights as described in Section 2.3), and one test for each of the four functions of the parameters. The lowest of the twelve p-values of these tests was 0.0099. This means, no clear interactions between breed and parity was found. Also, the significant effects of parity and breed on each of the parameters were the same independent of the use of weights. Because there is no similar way to easily obtain standard errors for the functions MA, PY, t_{MA} and t_{PY} of the four parameters, the following results are from the analysis using model (6) without the weights. So for $\tilde{a}$ we use (6) with γ_{bp} eliminated:

$$\tilde{a}_{bpc} = \alpha_b + \beta_p + u_{bc} + \epsilon_{bpc} \tag{7}$$

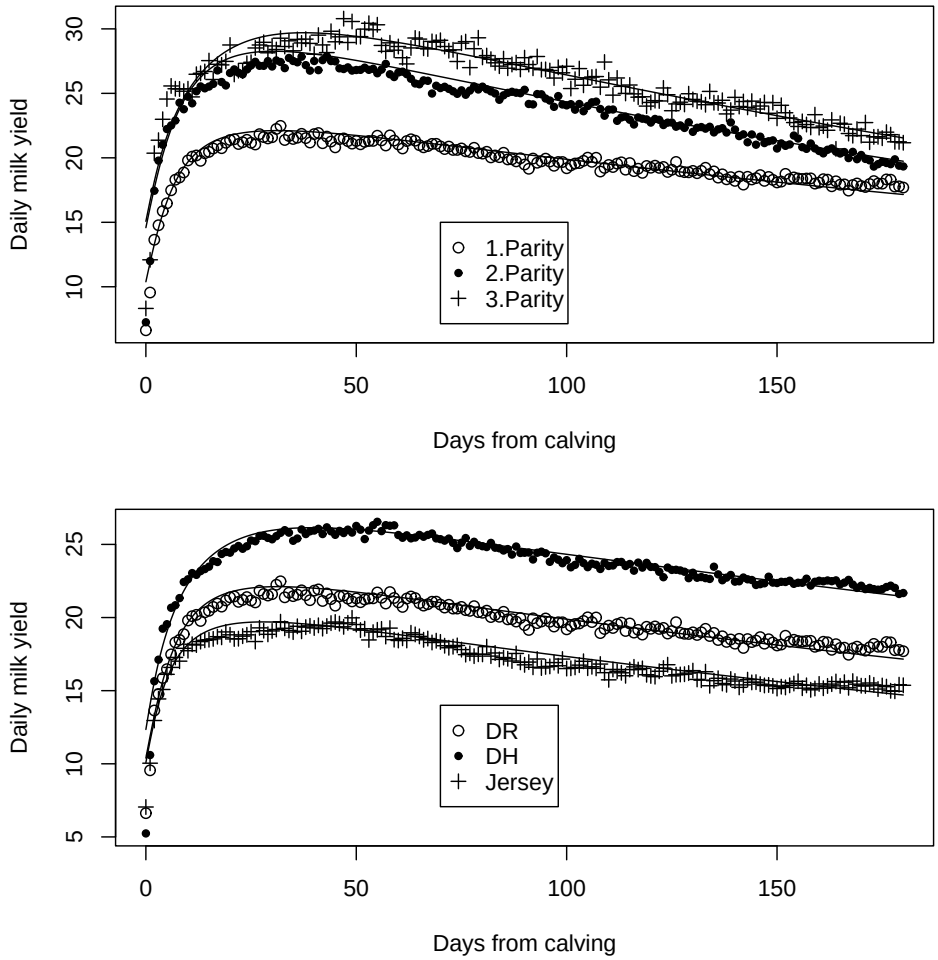

Figure 2: Plots of mean daily milk yield. In the upper plot the mean daily milk yield is plotted for cows of breed Danish Red for each parity. In the lower plot the mean daily milk yield is plotted for cows of first parity for Danish Red, Danish Holstein and Jersey. Along each group of points the curve obtained by taking mean of the corresponding curve fits is drawn.

where, as before, b denotes breed, p parity and c cow within breed. Similarly the same model (7) is used for the other parameters and for the functions of these.

The significance of the effects of breed and parity on the lactation curve parameters are given in Table 2 on page 11. From the table it can also be seen that the variance component due to cow, ω, is only significantly different from zero in four of eight cases. This means, that for these four

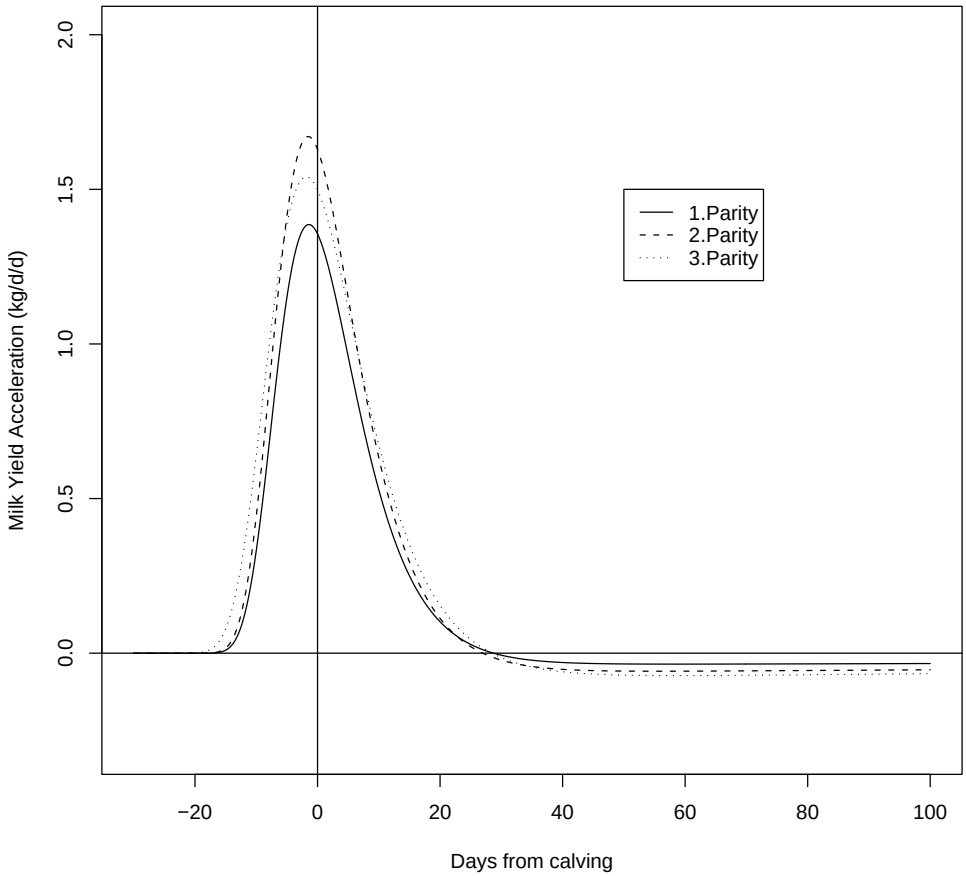

Figure 3: The milk yield acceleration function, $\mu'(t)$, plotted for all three parities. For each parity the corresponding means of the four parameters are used to obtain a curve.

parameters; $\tilde{a}$, maximum acceleration (MA), peak yield (PY) and time to peak (t_{PY}), a significant proportion of the individual cow deviation from the breed-parity mean was repeatable across parities.

In all cases, the effects of breed and parity were evaluated in the model including the cow variance component. It was found that parity had a highly significant effect on the two parameters $\tilde{a}$ and $\tilde{c}$ of the EF model but no significant effect on $\tilde{b}$ and $\tilde{G}_0$. This agrees with the previous findings of Friggens et al. (1999). From Table 2 on the facing page it can also be seen that breed had a highly significant effect on $\tilde{a}$ and $\tilde{b}$. There was also a minor effect on $\tilde{c}$ and no significant effect of breed on G_0. All the p-values given in Table 2 on the next page are the results of testing three parameters (breed DR, DH and Jersey or parity 1,2,3) to be equal. There is no exact

Table 2: The significance of effects of breed and parity on the lactation curve coefficients and on milk yield acceleration. In the last column the p-value for the likelihood ratio test of the hypothesis that the between lactation variance within cow, ω^2, is zero, is given.

Parameter[a]	Breed	Parity	P($\omega^2 = 0$)
$\tilde{a}$	<0.0001	<0.0001	<0.0001
$\tilde{b}$	0.0007	0.19	0.51
$\tilde{c}$	0.021	<0.0001	0.050
G_0	0.77	0.99	0.27
$MA(kg/d^2)$	0.096	0.062	0.016
$PY(kg/d)$	<0.0001	<0.0001	<0.0001
$t_{MA}(d)$	0.67	0.009	0.54
$t_{PY}(d)$	<0.0001	0.63	0.022

[a] $\tilde{a}, \tilde{b}, \tilde{c}$ and G_0 are the curve coefficients of model (2). MA, PY, t_{MA} and t_{PY} are the maximum acceleration, the peak yield and the time points to these events based on the lactation curve coefficients.

correspondance between a factor being significant in such a test and all three parameters being pairwise different. Therefore we also considered the tests for the contrasts of parameters in the model (7). The results of testing contrasts equal to zero revealed first of all, and not surprisingly, that second and third parity cows were much more similar to each other than to cows of first parity. It was found that the contrast between these parities 2 and 3 was barely significant for the parameter $\tilde{c}$ ($p = 0.0359$) even though parity had a highly significant overall effect on this parameter.

Considering contrasts in a similar fashion for the breed effects, the conclusions in most cases, where significant breed effects were observed (Table 2), were that the parameter of Danish Red was in between the parameters of the other two breeds, being slightly more similar to that of Danish Holstein than that of Jersey. This was particularly seen for the parameter $\tilde{b}$. Here, no significant differences was seen between Danish Red and Danish Holstein ($p = 0.80$) while the other two contrasts were significantly different from zero. In one case a different pattern was observed. This was for the parameter $\tilde{c}$ which in Table 2 is seen to be slightly affected by breed. Here we found that Danish Red was more similar to Jersey than to Danish Holstien. The three tests for the contrasts being zero yielded ($p = 0.65$, DR and Jersey), ($p = 0.011$, DH and Jersey) and ($p = 0.032$, DR and DH). It is interesting to speculate on these breed differences, the Jersey

and the Danish Holstein are both specialized dairy breeds whilst the Danish Red is a dual purpose breed. The Danish Holstein and Danish Red are of similar mature body size, considerably larger than the Jersey. Given that size-scaling for lactational output has been reported across species (Linzell (1972) and Taylor and Murray (1987)) it would be tempting to suggest that the size differences between these breeds explain the breed differences in lactation curve coefficients. However, given the differences in selection history of these breeds for different purposes it is difficult to draw clear conclusions.

In summary, the effect of parity on the lactation curve coefficients were the the same as those previously reported by Friggens et al. (1999). Further, as there were no significant interactions between breed and parity, for all breeds one common adjustment for parity effects will suffice when predicting parity adjusted milk yields within breed. There were however, significant breed effects on 3 of the 4 curve coefficients. The breed specific estimates found in this study provide a means to allow comparison of potential milk yield curves across breeds.

3.3 Milk yield acceleration

For the four biologically interpretable functions of the parameter estimates; maximum yield acceleration (MA), peak yield (PY), and the corresponding time points, t_{MA} and t_{PY}, estimates and confidence intervals for the fixed effects parameters are given in Table 3 on the facing page.

Breed was found to have significant effects on both the peak yield and the time point, t_{PY}, for peak yield but not on the maximum acceleration and the corresponding time point (Table 2 on the previous page). Finally, Table 2 on the preceding page shows that parity significantly affects peak yield and also has slight effects on maximum acceleration and t_{MA}.

When pairwise contrasts between breeds and parities were tested for, the same general picture emerged as was found for the curve coefficients: The significant breed effects were between (DR+DH) and Jersey and the main parity differences were between 1 and (2+3). There was one exception to this picture which was seen for the maximum acceleration. Here we found DR and DH to be significantly different ($p = 0.034$) but Jersey not to be different from either DR ($p = 0.15$) or DH ($p = 0.56$).

In Figure 3 on page 10, the acceleration curve is plotted for each parity. Each curve is derived from the averages of the four parameters $\tilde{a}, \tilde{b}, \tilde{c}$ and G_0 within each parity using (3). It should be noted that the levels of the peaks of the curves in the Figure do not match the estimates in Table 3 on the facing page. It is seen from the expressions (3) and (5) that the

Table 3: Estimates and 95% confidence intervals for fixed effects parameters for the functions MA, PY, t_{MA} and t_{PY}. For every function significant effects are included in the model. Though parity was not a significant effect on maximum acceleration ($p = 0.062$) it is included here. Because no interactions between breed and parity was observed estimates and confidence intervals of the mean of the peak yield are not given for all nine combinations of the two factors.

Function	Parameter mean	lower	estimate	upper
$MA(kg/d^2)$	β_1 [a]	1.47	1.72	1.97
	β_2	1.77	2.04	2.30
	β_3	1.80	2.20	2.61
$PY(kg/d)$	$\alpha_{DR} + \beta_1$ [b]	21.73	22.70	23.68
	$\alpha_{\mathrm{DR}} + \beta_2$	27.26	28.24	29.21
	$\alpha_{\mathrm{DR}} + \beta_3$	28.53	29.78	31.04
	$\alpha_{\mathrm{DH}} + \beta_1$	26.08	26.99	27.91
	$\alpha_{\mathrm{Jer}} + \beta_1$	17.68	18.74	19.80
$t_{MA}(d)$	β_1 [c]	−3.69	−2.76	−1.82
	β_2	−4.27	−3.29	−2.31
	β_3	−6.85	−5.32	−3.78
$t_{PY}(d)$	α_{DR} [d]	27.78	30.36	32.94
	α_{DH}	30.77	33.18	35.59
	α_{Jer}	21.44	24.17	26.90

[a] For $i = 1, 2$ and 3, β_i is the mean of the maximum acceleration (MA) for cows of parity i.

[b] For $i = 1, 2$ and 3, $j =$ DR,DH and Jer, $\alpha_j + \beta_i$ is the mean peak yield (PY) of cows from breed j and parity i.

[c] For $i = 1, 2$ and 3, β_i is mean of the time to maximum acceleration (t_{MA}) for cows of parity i.

[d] For $j =$ DR,DH and Jer, α_j is the mean of the time to peak yield (t_{PY}) for breed j.

maximum acceleration is not a linear function of the parameters of the EF model. As the parameters show a great deal of variation we can not expect that the curves in the Figure peak at the average peak levels. If, instead, the medians of the estimates of the maximum acceleration for each parity (1.41, 1.54 and 1.57) are considered it is seen that the Figure 3 on page 10 levels match these numbers very well.

From Figure 3 on page 10, it can be seen that in the post calving period the acceleration in milk yield is highest immediately post calving. This means that the first demand for acceleration to be a possible indicator of physiological stress and subsequent disease is fulfilled. The second criterion for acceleration to be an indicator is that it is higher in higher yielding cows at any given time-point in lactation (Ingvartsen et al. (2003)). If time to peak yield was strongly related to peak yield then this criterion would not be met. In Table 4 the correlation between the peak yield and the time to peak yield is given for each combination of breed and parity. There is no strong evidence for a relationship between peak yield and time to peak yield. Only one of these nine correlations is seen to be significantly different from zero. Thus, cows that reach peak yield early must have a correspondingly high acceleration in early lactation as all cows have only a small yield at calving. The Table also suggests that cows reaching their peak yield later than average must have a positive acceleration for a longer time. These findings fits with the hypothesis that at any time point in lactation higher yielding cows have a higher acceleration.

Table 4: For each combination of breed and parity the observed correlation between the the peak yield and the time for peak yield is given. The p-value of Pearson's test for no correlation is given in the parenthesis.

Parity	Breed		
	DR	DH	Jersey
1	0.101(0.494)	0.015(0.906)	0.467(0.002)
2	0.024(0.868)	0.021(0.885)	0.354(0.025)
3	0.260(0.369)	−0.082(0.738)	−0.039(0.859)

In Table 5 on the facing page, the correlation between the maximum milk yield acceleration and the peak yield is given for each combination of breed and parity. None of these correlations are significantly different from zero, so no clear dependence is seen between the maximum acceleration and the peak yield. This means that there is no clear difference in maximum acceleration between cows reaching a high level of yield at peak and cows

reaching a lower level of yield. Now one possibility is that all cows reach maximum acceleration at the same time i.e., they have identical lactation curves in early lactation. That would mean that cows reaching a high peak level are in the phase of high acceleration for a longer time than other cows. The other possibility is that the maximum acceleration is reached at different time points for different cows due to differences in the shape of the growth phase of the lactation curve in which case there need not be a relation between maximum acceleration and peak yield. If the first of these was the case we would have seen a clearer dependence between peak yield and time to peak yield in Table 4 on the preceding page. Since this was not the case we believe the second possibility is more likely to be true. This second case fits the assumption that at any time point in lactation higher yielding cows have a higher acceleration in yield. Thus, this first analysis of the properties of milk yield acceleration using real data indicated that acceleration has the appropriate properties to be considered as a possible indicator of physiological stress and subsequent health problems.

Table 5: For each combination of breed and parity the observed correlation between the maximum milk yield acceleration and the peak yield is given. The p-value of Pearson's test for no correlation is given in the parenthesis.

Parity	Breed		
	DR	DH	Jersey
1	−0.128(0.384)	0.031(0.812)	−0.197(0.216)
2	0.035(0.810)	0.124(0.403)	−0.155(0.340)
3	−0.026(0.930)	0.098(0.691)	0.313(0.146)

3.4 Further considerations

The present study has confirmed the findings of Friggens et al. (1999) that parity affects only two out of the few lactation curve coefficients, having no effect on the coefficients controlling the rate of "growth" of milk yield. Further, it was found that these effects were independent of breed although there were significant effects of breed on lactation curve coefficients. The magnitude of the breed differences reported here should be treated with caution as these estimates are based on relatively few animals per breed (approx. 80). However, because the three breeds were compared on the same farm under identical management conditions, the relative differences between breeds are well estimated. It should also be remembered that

the effects on lactation curve parameters, and indeed the lactation curve function used, relate to lactation curves that do not deviate significantly in shape from potential lactation curves.

With respect to the properties of milk yield acceleration, this study presented the first evaluation of the properties of acceleration curves. These properties were assessed in terms of whether or not they fit the criteria for providing a possible indicator of physiological stress and subsequent risk of health and reproductive problems. Milk yield acceleration was greatest around calving and no relationship was found between peak yield and time to peak yield. These findings are prerequisites for any further evaluation of milk yield acceleration as a risk indicator. Thus, the results presented in this study suggest that milk yield acceleration warrants further study. However, they do not, on their own, establish that acceleration is a useful indicator. Evidence to show this requires careful relation of milk yield acceleration to disease incidences in very substantial datasets, a task beyond the scope of this paper.

4 Conclusions

The first aim of this study was to quantify breed and parity effects on lactation curves. Here, the effects of parity on the coefficients of the lactation curve were found to be independent of the breed of cow in question. Further, the effects of parity matched the findings of Friggens et al. (1999) and the lactation curves were seen to be also affected by breed. The second aim was to evaluate the properties of acceleration in milk yield. This part revealed that acceleration in milk yield has the appropriate properties for consideration as an indicator of risk of health and reproduction problems.

Bibliography

Collins-Lusweti, E., 1991. Lactation curves of Holstein-Friesian and Jersey cows in Zimbabwe. S. Afr. J. Anim. Sci. 21, 11–15.

Dijkstra, J., France, J., Dhanoa, M., Maas, J., Hanigan, M., Rook, A., Beever, D., 1997. A model to describe growth patterns of the mammary gland during pregnancy and lactation. J. Dairy Sci. 80, 2340–2354.

Emmans, G., Fisher, C., 1986. Problems in nutritional theory. Butterworths, London Nutrient Requirements of Poultry and Nutritional Research, 9–39.

Friggens, N., Emmans, G., Veerkamp, R., 1999. On the use of simple ratios between lactation curve coefficients to decribe parity effects on milk production. Livest. Prod. Sci. 62, 1–13.

Grossman, M., Hartz, S., Koops, W., 1999. Persistancy of lactation yield: A novel approach. J. Dairy Sci. 82 (2), 2192–2197.

Grossman, M., Koops, W., 1988. Multiphasic analysis of lactation curves in dairy cattle. J. Dairy Sci. 71, 1598–1608.

Grossman, M., Koops, W., 2003. Modelling extended lactation curves of dairy cattle: A biological basis for the multiphasic approach. J. Dairy Sci. 86, 988–998.

Hansen, L., 2000. Consequences of selection for milk yield from a geneticist's viewpoint. J. Dairy Sci. 83, 1145–1150.

Ingvartsen, K., Dewhurst, R., Friggens, N., 2003. On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? a position paper. Livest. Prod. Sci. 83, 277–308.

Ingvartsen, K., Friggens, N., Faverdin, P., 1999. Food intake regulation in late pregnancy and early lactation. British Society of Animal Science 24, 37–54.

Linzell, J., 1972. Milk yield, energy loss in milk, and mammary gland weight in different species. Dairy Science Abstracts 34, 351–360.

Macciotta, N., Vicario, D., Capplo-Borlino, A., 2005. Detection of different shapes of lactation curve for milk yield in dairy cattle by empirical mathematical models. J. dairy Sci. 88, 1178–1191.

Nielsen, H., Friggens, N., Løvendahl, P., Jensen, J., Ingvartsen, K., 2003. Influence of breed, parity, and stage of lactation on lactational performance and relationship between body fatness and live weight. Livest. Prod. Sci. 79, 119–133.

Olori, V., Brotherstone, S., Hill, W., McGuirk, B., 1999. Fit of standard models of the lactation curve to weekly records of cows in a single herd. Livest. Prod. Sci. 58, 55–63.

Pollott, G., 2000. A biological approach to lactation curve analysis for milk yield. J. Dairy Sci. 83, 2448–2458.

Pollott, G., 2004. Deconstructing milk yield and composition during lactation using biologically based lactation models. J. Dairy Sci. 87, 2375–2387.

Pryce, J., Nielsen, B., Veerkamp, R., Simm, G., 1999. Genotype and feeding system effects and interactions for health and fertility traits in dairy cattle. Livest. Prod. Sci. 57, 193–201.

R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3.
URL `http://www.R-project.org`

Rekaya, R., Weigel, K. A., Gianola, D., 2001. Heirarchical nonlinear model for persistency of milk yield in the first three lactations of Holsteins. Livest. Prod. Sci. 68, 181–187.

Rook, A., France, J., Dhanoa, M., 1993. On the mathematical description of lactation curves. J. Agric. Sci. Camb. 121, 97–102.

SAS Institute Inc, 2001. SAS/STAT Software: Changes and Enhancements, Release 8.2. SAS Institute Inc, Cary, NC.

Taylor, S., Murray, J., 1987. Genetic aspects of mammalian growth and survival in relation to body size. Butler Memorial Lecture, Academic Press. University of Queensland , 1–58.

Val-Arreola, D., Kebreab, E., Dijkstra, J., France, J., 2004. Study of the lactation curve in dairy cattle on farms in central Mexico. J. Dairy Sci. 87, 3789–3799.

Van der Werf, J., Goddard, M., Meyer, K., 1998. The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. J. Dairy Sci. 81, 3300–3308.

Vetharaniam, I., Davis, S., Upsdell, M., Kolver, E., Pleasants, A., 2003. Modelling the effect of energy status on mammary gland growth and lactation. J. Dairy Sci. 86, 3148–3156.

Wood, P., 1967. Algebraic model of the lactation curve in cattle. Nature 216, 164–165.

Wood, P., 1980. Breed variations in the shape of the lactation curve of cattle and their implications for efficiency. Anim. Prod. 31, 131–141.

Paper

B

Hansen, J.V., Jensen, J.L., Friggens, N.C. and Højsgaard, S. (2008).

A state space model exhibiting a cyclic structure with an application to progesterone concentration in cow milk.

To be submitted (primo June 2008) to *J. Roy. Stat. Soc. Series C.*,
To appear as Thiele Research Report, Department of Mathematical Sciences, University of Aarhus.

A State Space Model Exhibiting a Cyclic Structure with an Application to Progesterone Concentration in Cow Milk

Jørgen V. Hansen[1]
University of Aarhus, Foulum and Århus, Denmark.
Jens L. Jensen
University of Aarhus, Århus, Denmark.
Nicolas C. Friggens
and Søren Højsgaard
University of Aarhus, Foulum, Denmark.

Abstract

Progesterone is a hormone linked to the reproductive status of dairy cows. Hence, with the increasing availability of on-line records of the concentration of progesterone in cow milk, there is a need for new tools to analyse such data. The aim is to find techniques for better determination of the time when cows are in oestrus to increase the rate of succesful inseminations. In this paper we propose a state space model for data with a continuous and cyclic trend in the mean. Furthermore a matching Kalman filter is developed. The model is tested on progesterone data from 112 cow-lactations with the purpose of evaluating the use of progesterone for detection of oestrus.

Keywords: cyclic model, dairy cow, Kalman filter, oestrus detection.

1 Introduction

Data from many biological processes exhibit a clear cyclic nature. A classical example is the yearly number of lynx in Canada (Elton and Nicholson, 1942) where the cyclic nature is caused by a predator-prey relationship. The example of main interest to us in this paper is one where the oestrus cycle in cows generates a cyclic behavior of the concentration of progesterone in cow milk. A short presentation of the biology behind this process is given in Section 2. In models for such data it is often natural to introduce hidden

[1]Jørgen V. Hansen, Department of Genetics and Biotechnology, Research Centre Foulum, Blichers Allé, Postbox 50, DK-8830 Tjele.

variables which have certain biological or physical interpretations. We study a state space model which generates a cyclic behavior with continuous and piecewise linear mean of the univariate observations. The continuity and cyclic nature implies that a period of increase in the mean must at some point be followed by a decrease in the mean. It can be useful to think of a model with four stages, where the four stages correspond to an increase in the mean, a high level of the mean, a decrease in the mean and finally a low level of the mean. However, this does not exhaust the variety of possible cyclic models. In this way the time axis will be divided into segments where all observations within a segment belong to the same stage. The hidden variables will include one variable holding the current stage and two variables holding the time points for the beginning and end of the current stage. The inference concerning time points at which paramters change is known as changepoint detection, a subject on which Page (1954) wrote one of the first papers. The state space model we consider can be seen as a modification of the model of Fearnhead and Liu (2007) and Fearnhead and Vasileiou (2008). The precise mathematical formulation of the model is given in Section 3. In Section 4 we present an approximate filter for the hidden variables of the state space model, leaving the detailed derivation of this filter to Appendix B. In Section 5 we discuss how to estimate the parameters of the model. In Section 6 we show how a model of the class presented in Section 3 can be used to describe the level of progesterone in cows. Also in Section 6 we describe an algorithm for finding the optimal time point for insemination and we study how well the algorithm works in practice. Finally, in Section 7 we discuss the results of this analysis in the perspective of creating better tools to assess the optimal time point for artificial insemination.

2 Biological background

The main motivation for this work is to detect oestrus in cows by modelling the progesterone concentration in cow milk. Oestrus is usually defined to be the period of low progesterone in the cyclic pattern of the hormone (Peters and Ball, 1995). When we apply the cyclic model to the progesterone data we therefore are specifically interested in determining when a cow is in the stage corresponding to low progesterone measurements. Only during oestrus can the cow be succesfully inseminated and then produce a calf. Cows in oestrus usually exhibit physiological signs of sexually receptive behavior. However, the traditional, visual, detection of oestrus signs (i.e. not using progesterone) is becoming more difficult as genetic selection of cows for high

milk yields has reduced the intensity of the oestrus (Dobson et al., 2008). Also, in modern dairy cow herds there is a growing need for automated management of the cows due to the large herd sizes. Therefore, detection of the time the cow is in oestrus is one such problem where a farmer could benefit from improved techniques.

The reproductive cycle of dairy cows is approximately 21 days (ranging from 18 to 26 days), and for maximal chance of success insemination should take place 12-24 hours before ovulation (Roelofs et al., 2006). Though several indicators exist for determining when cows are in oestrus (Fulkerson et al., 1983; Xu et al., 1998; Cavalieri et al., 2003) there is still a need for better prediction of the time when cows are most susceptable for insemination. Progesterone, which can now be measured automatically in the milk, is the accepted gold standard for assessing the reproductive status (Peters and Ball, 1995; Cavalieri et al., 2001; Roelofs et al., 2006).

The oestrus cycle in a cow is initiated by the creation of a follicle in the ovary. The follicle grows in size until ovulation where the follicle ruptures and releases the egg which is tranported down the oviduct toward the uterus. After the ovulation the remains of the follicle, called the corpus luteum, stay in the ovary. The cells of the corpus luteum begin to secrete progesterone approximately 4 days after ovulation. Progesterone is required for the maintance of pregnancy (Peters and Ball, 1995). The presence of a fertilized egg (embryo) blocks regression of the corpus luteum which then continue to secrete progesterone throughout pregnancy. If the egg is not fertilized the corpus luteum will regress and stop producing progesterone approximately 17 days after ovulation. The following drop in progesterone then causes a new follicle to grow and the cycle repeats itself.

Progesterone can be measured in the milk throughout lactation which is the period in time where a cow produces milk following a calving. In this paper we define the term *cow-lactation* to be the statistical term of the cross factor of cow and parity, where parity is the number of calves a cow has given birth to. That is, a cow of first parity has just had its first calf and so on. The concentration of progesterone in the milk is measured in ng/ml and varies in the range from 0 to 30 ng/ml. An example showing the development of the concentration of progesterone in milk from a cow is seen in Figure 1. In this example the cyclic behavior of the hormone is observed through several cycles.

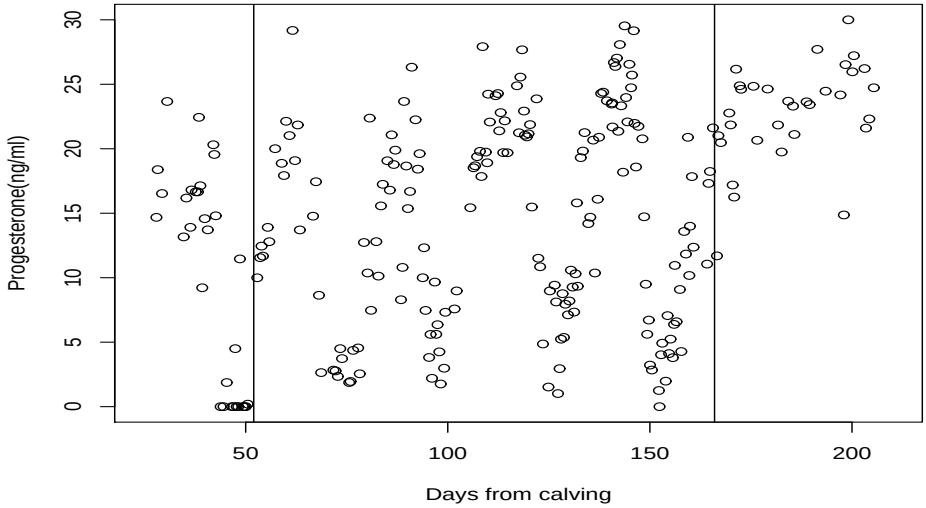

Figure 1: Progesterone measurements through a cow-lactation showing a cyclic behavior through several cycles. The model apply to data only when the cow is in its oestrus cycle. Therefore observations in the beginning and the end of lactation are excluded in the analysis. This is indicated by the vertical lines.

3 Cyclic model

In the following, the model we consider in this paper is defined. We use a state space model incorporating the idea of several different stages each describing a linear development in the mean of the observations, such that the mean as a function of time is continuous. The number of different stages is denoted by m. The m stages follow each other in the same order, so that one round of the m stages constitute a cycle of the process. More specifically, a time segment of stage $q = 1$ will be followed by a segment of stage $q = 2$, and so forth, until a segment of stage $q = m$ is again followed by stage $q = 1$. An illustration of a possible development in the mean of the observation is shown in Figure 2. In this illustration the number of stages is $m = 4$, which is the value we use for modeling the progesterone data in Section 6.

3.1 Hidden variables

The model has five hidden variables. The hidden process is considered in discrete time, $t \in \mathbf{Z}$. In applications these discrete times will constitute a scaling of real time. To each time point t these hidden variables contain in-

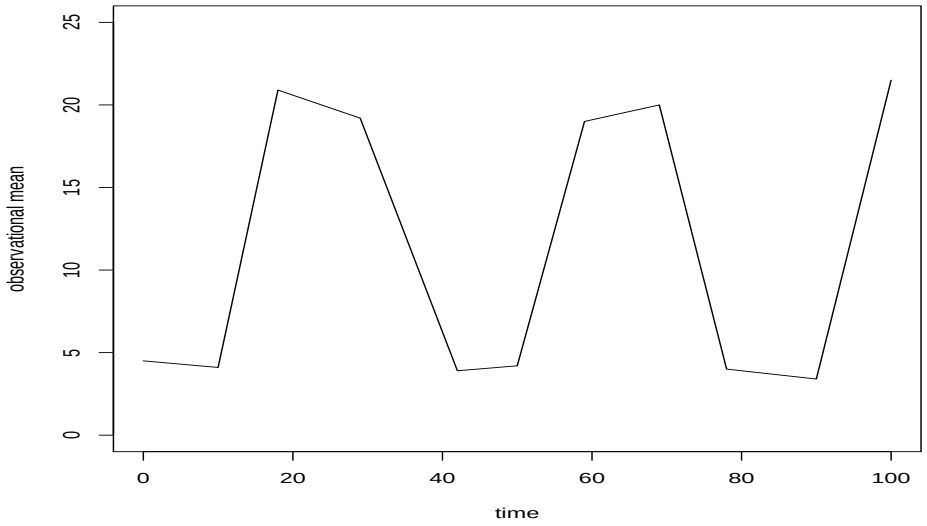

Figure 2: Example of possible development in the mean of the observations.

formation about the position of the change points separating the stages, and infomation on the mean level of the observations. The five hidden variables are:

R_t : the point in time prior to t with the most recent change of stage, $(R_t < t)$.

S_t : the stage entered at time R_t with value in $\{1, 2, \ldots, m\}$.

N_t : the point in time for the next change of stage after R_t, $(N_t \geq t)$.

a_t : the mean of an observation at time R_t.

b_t : the mean of an observation at time N_t.

3.2 Stochastics of the state $(R_t, S_t, N_t, a_t, b_t)$

The starting point of describing the stochastics governing the state variables is to find the positions of the changepoints. Given that there is a change to stage q at t, the distribution of the waiting time for the next change depends on q only. This distribution is denoted W_q, where q is the new stage. The

only restriction we put on W_q is that it has finite support. Formally, we write

$$(N_{t+1} - N_t | N_t = t, S_{t+1} = q) \sim W_q. \tag{1}$$

Throughout the paper we use the notation $W_q(r) = P(w = r)$ if $w \sim W_q$. Thus (1) describes the dynamics of the three discrete hidden variables (R_t, S_t, N_t). In terms of one step transition probabilities we have that

$$\begin{aligned} &P(R_{t+1} = j', S_{t+1} = q', N_{t+1} = l' | R_t = j, S_t = q, N_t = l) \\ &\qquad = \begin{cases} W_{q'}(l' - t) & \text{if } l = t \text{ and } j' = t, q' = q + 1 (\text{mod } m), \\ 1 & \text{if } l > t \text{ and } j' = j < t, q' = q, l' = l, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

The triple $(R_t - t, S_t, N_t - t)$ as defined above constitutes by itself a Markov chain. In Appendix A it is shown that the stationary distribution for $(R_t - t, S_t, N_t - t)$ is

$$\pi(j, q, l) = \frac{W_q(l - j + M)}{\sum_r \nu(r)},$$

where

$$M = \max\{r \mid \exists q : W_q(r) > 0\} \tag{3}$$

is the maximal possible waiting time between two consecutive changepoints and $\nu(q) = \sum_{i=1}^{\infty} i W_q(i)$ is the mean of the waiting time distribution W_q. The stationary distribution can be used as a prior when no information about the state variables is at hand at the time of the first recording.

Next, the stochastic behavior of the continuous hidden variables a_t and b_t is described. These two variables hold information about the mean of the observations within a stage. For each changepoint t, let $x(t)$ denote the hidden mean of a possible observation y_t. If the stage of a time interval beginning at time t is q, then we assume

$$(x(t) | N_t = t, S_{t+1} = q) \sim N(\mu_q, \omega_q^2), \tag{4}$$

where μ_q and ω_q^2 are parameters. The hidden variables a_t and b_t are then defined to be

$$a_t = x(R_t) \text{ and } b_t = x(N_t). \tag{5}$$

That is, a_t is the mean at the beginning and b_t the mean at the end of the stage entered at time R_t. This means that if there is no change point at time t then $(a_{t+1}, b_{t+1}) = (a_t, b_t)$. On the other hand if there is a change of stage at time t, then $a_{t+1} = b_t$ and $b_{t+1} = x(N_{t+1})$.

3.3 Stochastics of observations

The mean at any time point $s \in \mathbf{R}$ we define by linear interpolation using the mean at the end points of the stage. That is,

$$x(s) = a_t + \frac{b_t - a_t}{N_t - R_t}(s - R_t), \quad s \in \mathbf{R}, t = \lceil s \rceil,$$

where $\lceil s \rceil$ is the smallest integer greater than or equal to s. In this way the underlying mean $x(s)$ of the observation given the hidden variables is continuous and piecewise linear.

Data $\{(y_i, s_i) | i = 1, \ldots, n\}$ consist of a set of observations y_i recoded at time points s_i, where n is the number of observations. To define the distribution of data, assume that we have an observation y_i at time $s_i \in \mathbf{R}$. Note that we do not restrict the observations to occur at time points that are multiples of the time unit. If the stage at time s_i is q, that is, $S_{\lceil s_i \rceil} = q$, we assume

$$\left(y_i | (R, S, N, a, b)_{\lceil s_i \rceil}\right) \sim N(x(s_i), \sigma_q^2),$$

where $\sigma_q^2, q = 1, 2, \ldots, m$ are parameters.

To summarize the parameters of the model are

μ_q : the mean of the hidden stochastic mean at a time point where the stage changes to q,

ω_q^2 : the variance of the hidden stochastic mean at a time point where the stage changes to q,

σ_q^2 : residual variance of observations within stage q,

for $q = 1, 2, \ldots, m$. Please note that the waiting time distributions $W_q, q = 1, 2, \ldots, m$ may depend on an unknown parameter θ.

4 An approximate filter

An approximate filter for the state space model presented in Section 3 is described below. By using the notation $y^s = \{y_i \mid s_i \leq s\}$ and $y_r^s = \{y_i \mid r < s_i \leq s\}$ our goal is to determine the filter densities which we write as

$$p(R_t = j, S_t = q, N_t = l, a_t, b_t | y^t) = p_t(j, q, l, a_t, b_t) \text{ for all t } \in \mathbf{N}. \quad (6)$$

We use the approximation

$$p_t(j, q, l, a_t, b_t) = p_t(j, q, l)\phi(a_t, b_t;\ \mu_t(j, q, l), \Sigma_t(j, q, l)), \quad (7)$$

where $p_t(\cdot,\cdot,\cdot)$ on the right hand side of (7) is the marginal density of (R_t, S_t, N_t) and $\phi(\cdot,\cdot;\ \mu,\Sigma)$ is the normal density with mean μ and variance Σ. Therefore the filter densities (6) are specified by $p_t(j,q,l)$, $\mu_t(j,q,l)$ and $\Sigma_t(j,q,l)$. Here we only state the recursions for updating the filter. A detailed derivation of the filter is given in Appendix B.

Let $\sum_{(t)}$ (and $\prod_{(t)}$) denote the sum (and the product) over the set of observations $\{y_i\}$ in the time interval from t to $t+1$. When $j = R_{t+1} < t$ the filter recursions are given as follows:

$$p_{t+1}(j,q,l) = c_{t+1}(y^{t+1})p_t(j,q,l)\frac{\phi(0;\ \mu_t(j,q,l),\Sigma_t(j,q,l))}{\phi(0;\ \mu_{t+1}(j,q,l),\Sigma_{t+1}(j,q,l))}\prod_{(t)}\phi(y_i;\ 0,\sigma_q^2), \quad (8)$$

with

$$\Sigma_{t+1}(j,q,l)^{-1} = \Sigma_t(j,q,l)^{-1} + \frac{1}{\sigma_q^2(l-j)^2}\sum_{(t)}\begin{pmatrix} (l-s_i)^2 & (s_i-j)(l-s_i) \\ (s_i-j)(l-s_i) & (s_i-j)^2 \end{pmatrix} \quad (9)$$

and

$$\Sigma_{t+1}(j,q,l)^{-1}\mu_{t+1}(j,q,l) = \Sigma_t(j,q,l)^{-1}\mu_t(j,q,l) + \frac{1}{\sigma_q^2}\sum_{(t)}\begin{pmatrix} y_i(l-s_i)/(l-j) \\ y_i(s_i-j)/(l-j) \end{pmatrix}. \quad (10)$$

The constant of proportionality $c_{t+1}(y^{t+1})$ is found from (8) and (11) below together with the condition $\sum_{j,q,l} p_{t+1}(j,q,l) = 1$.

When $R_{t+1} = t$ the recursions are

$$p_{t+1}(t,q,l) = c_{t+1}(y^{t+1})W_q(l-t)\prod_{(t)}\phi(y_i;\ 0,\sigma_q^2)\sum_{j'<t}p_t(j',q',t) \times \frac{\phi(0;\ \mu_t(j',q',t)_2,\Sigma_t(j',q',t)_{22})\phi(0;\ \mu_{\tilde{q}},\omega_{\tilde{q}}^2)}{\phi(0;\ \bar{\mu}_t(j',q',t,l),\bar{\Sigma}_t(j',q',t,l))}, \quad (11)$$

$$\mu_{t+1}(t,q,l) = \sum_{j'<t}\alpha_t(j',q',t,l)\bar{\mu}_t(j',q',t,l), \quad (12)$$

and

$$\Sigma_{t+1}(t,q,l) = \sum_{j'<t}\alpha_t(j',q',t,l)[\bar{\Sigma}_t(j',q',t,l) + \bar{\mu}_t(j',q',t,l)\bar{\mu}_t(j',q',t,l)^T] - \mu_{t+1}(t,q,l)\mu_{t+1}(t,q,l)^T \quad (13)$$

where $\tilde{q} = q+1 (\text{mod } m)$ and where $\bar{\mu}_t(j', q', t, l)$, $\bar{\Sigma}_t(j', q', t, l)$ and $\alpha_t(j', q', t, l)$ are given in Appendix B.

5 Parameter estimation

Maximum likelihood can be used to estimate the residual variances $\sigma_q^2, q = 1, \ldots, m$. This estimation procedure is described in Section 5.1. The waiting time distributions can be estimated using an EEE-algorithm as described in Section 5.2. We do not suggest any general procedures for estimation of the parameters μ_q and ω_q^2, $q = 1, \ldots, m$, but in Section 6 we describe how to find crude estimates of these parameters to use for the modelling of the progesterone data.

5.1 Estimation of the residual variance

The residual variance parameters σ_q^2, $q = 1, \ldots, m$ are estimated using maximum likelihood. The constant of proportionality $c_{t+1}(y^{t+1})$ in (8) and (11) can, from the derivation of the filter in Appendix B, be seen to be

$$c_{t+1}(y^{t+1}) = \frac{p(y^t)}{p(y^{t+1})} = \frac{1}{p(y_t^{t+1}|y^t)}.$$

Therefore, using the approximate filter, we can calculate an approximation to the likelihood function

$$L(\sigma_1^2, \ldots, \sigma_m^2) = p(y^n) = \prod_t p(y_t^{t+1}|y^t),$$

which can be maximized using numerical techniques to find estimates $\hat{\sigma}_q^2$ of the residual variances.

5.2 Estimation of the distribution of waiting times

Given M as defined in (3) and a model for the waiting time distributions $W_q, q = 1, \ldots, m$, we can estimate the parameter θ of this model using an EEE algorithm which is proposed and discussed in e.g. Heyde and Morton (1996), Rosen et al. (2000) and Elashoff and Ryan (2004). Fundamentally, an EEE algorithm works similar to an EM algorithm (Dempster et al., 1977). The difference is that the M-step of maximizing the likelihood is replaced by a step where an estimating equation is solved. In the special case where the estimating equation is the likelihood equation the EEE algorithm is an EM algorithm. For a parameter θ of the waiting time distribution we

use an estimating function of the form $\sum_1^n \psi_i$, where $\psi_i = \psi(z_i, z_{i-1}; \theta)$ for a function ψ, and where $z_i = (R_i, S_i, N_i)$. The E (expectation) step is to calculate

$$E\Big(\sum_1^n \psi_i | y^n\Big),$$

where n is the number of observations. In Appendix C it is shown that $E\big(\sum_1^k \psi_i | y^k\big)$ can be calculated iteratively in k using the filter probablities of Section 4. As an example consider the case where the waiting time probabilities are modelled with no other restriction than $\sum_{l=1}^M W_q(l) = 1$ for all q. We can then use the estimating functions

$$\begin{aligned}&\psi(z_t, z_{t-1}; q, l)\\ &\quad = 1(R_t = t-1, S_t = q, N_t = t-1+l) - W_q(l)1(R_t = t-1, S_t = q),\end{aligned}$$

where $1(\cdot)$ is the indicator function. In the EE (estimating equation) step the new value of $W_q(l)$ becomes

$$W_q(l) = \frac{E\left(\sum_1^n 1(R_t = t-1, S_t = q, N_t = t-1+l) | y_1^n\right)}{E\left(\sum_1^n 1(R_t = t-1, S_t = q) | y_1^n\right)},$$

where the nominator and denominator have been found in the E step.

6 Application

6.1 Progesterone data

The objective of this part of the study is to test the ability of the model to predict the time of oestrus in cows from on-line progesterone concentration measurements. We consider a data set where progesterone measurements were made on milk samples taken from all milking cows in one research herd (Danish Cattle Research Centre) during the period 12 Sept. 2002 to 30 Sept. 2006. In the this dataset 123 cow-lactations included an oestrus that was identified as a confirmed oestrus, i.e. an oestrus at which insemination resulted in a confirmed pregnancy. A detailed description of the collection of data can be found in Friggens et al. (2008).

Parts of the dataset were collected before the cows had entered their oestrus cycle or after the cows had succesfully been inseminated as seen in Figure 1. Because only data collected when the cows are in their oestrus cycle are of interest, we excluded parts of the data at the beginning and at the end of each cow-lactation. For a small number of cows no data was

left by this reduction of data. This was primarily the case when the cow was succesfully inseminated at the first oestrus which is not preceeded by a high progesterone stage. This reduction left us with 112 cow-lactations.

6.2 Model for progesterone data

The concentration of progesterone in milk has a cyclic nature with an average cycle length of about 21 days (ranging from 18 to 26 days). There is from cycle to cycle a small variation in the cycle length within cows. Roughly the cyclic nature of the progesterone content can be described in the following way. In each cycle we see four different stages for the concentration of progesterone. Each with a different time length. The four stages we enumerate as follows: 1. Low level of progesterone, 2. Slow increase in progesterone, 3. High level of progesterone, 4. Rapid decrease in progesterone.

It seems reasonable to assume that the mean level at the beginning and at the end of a low stage or a high stage is roughly the same. According to the model specification in (4) we can formulate this as the restriction $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. Also we let $\omega_1^2 = \omega_2^2$ and $\omega_3^2 = \omega_4^2$. The parameters μ_q and ω_q^2, $q = 1, 2, 3, 4$ will be the same for all cows and we estimate them as described in Section 6.3.1.

We model the waiting time distributions $W_q, q = 1, 2, 3, 4$ as discretized gamma distributions truncated at M. The gamma distributions are parameterized with parameters α_q and β_q, such that the means and variances are α_q/β_q and α_q/β_q^2, respectively. The waiting time distributions are also asumed to be the same for all cows. To estimate the α_q's and β_q's we use the estimating functions

$$\psi_1(z_t, z_{t-1}) = 1(R_t = t-1, S_t = q)\log\beta_q - 1(R_t = t-1, S_t = q)\frac{\Gamma'(\alpha_q)}{\Gamma(\alpha_q)} + \sum_l \log(l)\cdot 1(R_t = t-1, S_t = q, N_t = t-1+l) \qquad (14)$$

and

$$\psi_2(z_t, z_{t-1}) = 1(R_t = t-1, S_t = q)\frac{\alpha_q}{\beta_q} - \sum_l l\cdot 1(R_t = t-1, S_t = q, N_t = t-1+l) \qquad (15)$$

chosen so as to resemble the likelihood equations for a sample from a gamma distribution.

Finally we take the residual variances to be the same for all four stages, but specific for each cow. The cow specific variances are estimated as described in Section 5.1.

6.3 Result of analysis

6.3.1 Estimation of parameters

We estimate the parameters μ_q and σ_q^2, $q = 1, \ldots, 4$ in the following way. For each cow-lactation we find the 85 percent quantile of all observations in that cow-lactation and regard this as an outcome of a $N(\mu_3, \omega_3^2)$-distributed variable. Similarly, we regard the 15 percent quantile as an outcome of a $N(\mu_1, \omega_1^2)$-distributed variable. In Figure 3, the 15 percent and the 85 percent quantile of all progesterone measurements for the cow-lactation of Figure 1 is indicated by horizontal lines. Calculating the mean and variance of these quantiles from all cow-lactations we obtain estimates for the μ_q's and the ω_q^2's. These estimates will then be used as inputs when we make inference using the filter. The estimates are given in the second and third column of Table 1.

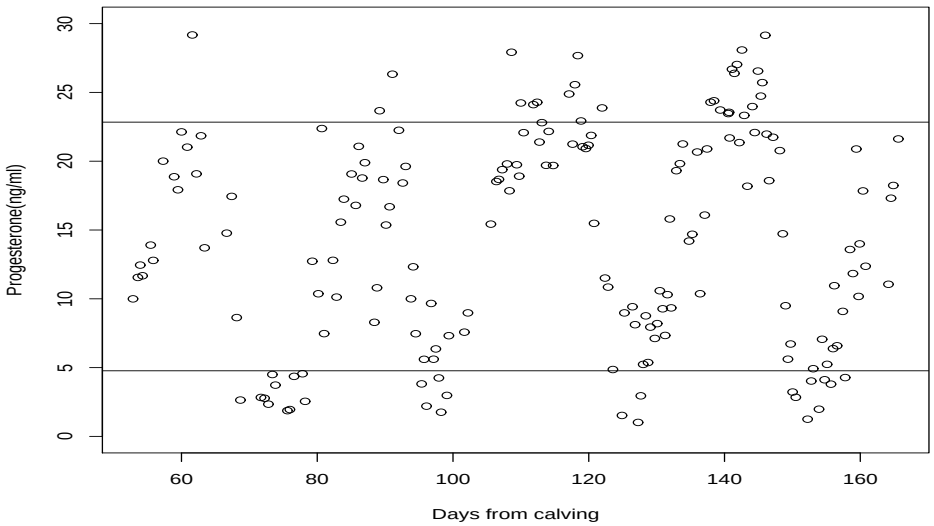

Figure 3: In this plot the 15 percent and the 85 percent quantile of all measurements are indicated by the horizontal lines. These two quantiles are used to create crude estimates of the parameters μ_q and ω_q^2, $q = 1, 2, 3, 4$.

The procedure for estimating the residual variances are carried out for fixed values of the parameters α_q and β_q of the waiting time distributions and vice versa. Therefore an iterative procedure is used for the simultaneous estimation where each of the two sets of parameters $(\alpha_q, \beta_q, q = 1, 2, 3, 4)$

Table 1: Estimates of those parameters of the cyclic model that are the same for all cow-lactations.

q	μ_q (ng/ml)	ω_q (ng/ml)	$E(W_q)$ (days)	$\sqrt{V(W_q)}$ (days)
1	3.122	1.170	5.12	2.27
2			8.25	1.81
3	20.929	2.074	6.82	2.80
4			2.74	0.98

and (σ^2 specific for every cow-lactation) are updated one at a time. To speed up the computations we take the unit of time to be one day during the estimation of the parameters. Furthermore, the maximal length of the waiting time is set to $M = 12$ days. The estimates for the four gamma distributions defining the waiting times are given in Table 1 summarized as means (α/β) and standard deviations ($\sqrt{\alpha/\beta^2}$).

The result of estimating the residual variances is summarized in a histogram of the standard deviations shown in Figure 4.

6.3.2 Prediction of oestrus

For each cow-lactation in our data the day (but not the exact hour) at which an artificial insemination resulted in a confirmed pregnancy is known. The filter was run with a 6-hour interval between the updates. In Figure 5 the development of the filter probability of being in the low stage, P(S=1), is shown for nine cow-lactations. In each plot a vertical line is drawn at noon on the day of succesful artificial insemination.

Knowing the day for the confirmed succesful inseminations we have the possibility of evaluating how well our model can predict the time point when a cow enters oestrus. A possible way of constructing an alarm telling the farmer that a cow is about to go into oestrus is to say that when the probability of being in the low stage increases to a certain level the alarm should go off. For most of the cow-lactations the alarm goes off more than once because we observe more than one cycle for most cows. In this case we take the time of alarm t_{predict} to be the time point closest to the stipulated time of confirmed succesful insemination t_{ins}, which we in all cases define to be at noon. The difference $t_{\text{ins}} - t_{\text{predict}}$ indicates how much time in advance the farmer is given to observe the cow in detail. For 7 out of the 112 cow-

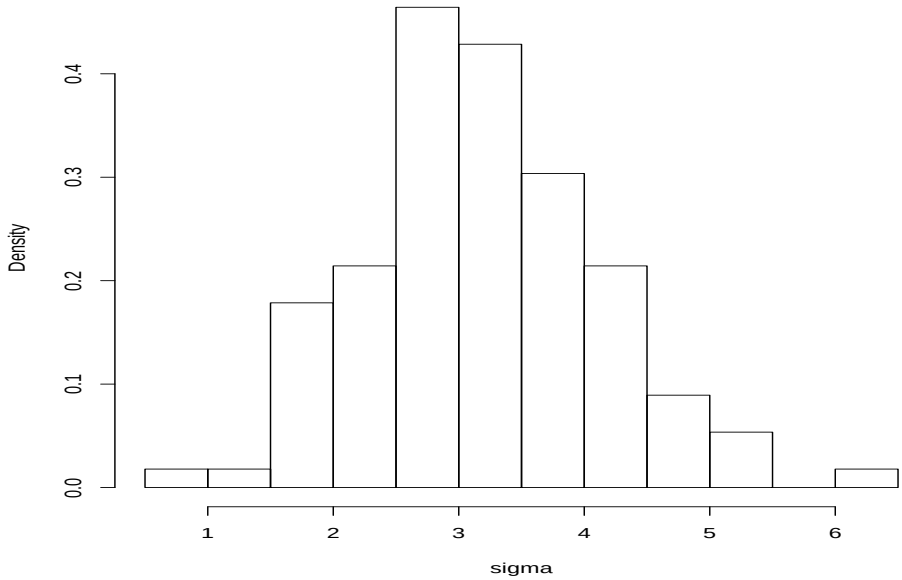

Figure 4: Histogram of the estimates of the cow-lactation specific standard deviation σ.

lactations, the time point for insemination was placed outside the time range of the observations (which in Figure 1 is outside the two vertical lines) and therefore no t_{predict}-value was found. With a threshold probability of 0.5 of $P(S = 1)$ we observed that the this probability did not exceed the threshold in the cycle including the time of insemination for 3 of the remaining 105 cow-lactations. This leave 102 values of $t_{\text{ins}} - t_{\text{predict}}$. In Figure 6 a histogram of these values are shown. The observed mean and standard deviation of this sample was 1.431 days and 1.556 respectively meaning that on average the alarm will tell the farmer to look for signs of oestrus a little less than one and a half day before oestrus actually occurs. In 88 of the 102 cases the alarm went off prior to the actual insemination of these cows.

All analysis was performed using R (R Development Core Team, 2008). The speed of running the whole filter is proportional to the cube of the number of updates per day. If for a cow-lactation the filter is updated each hour with $M = 288$ hours (12 days) an update takes approximately 200 seconds on a standard laptop.

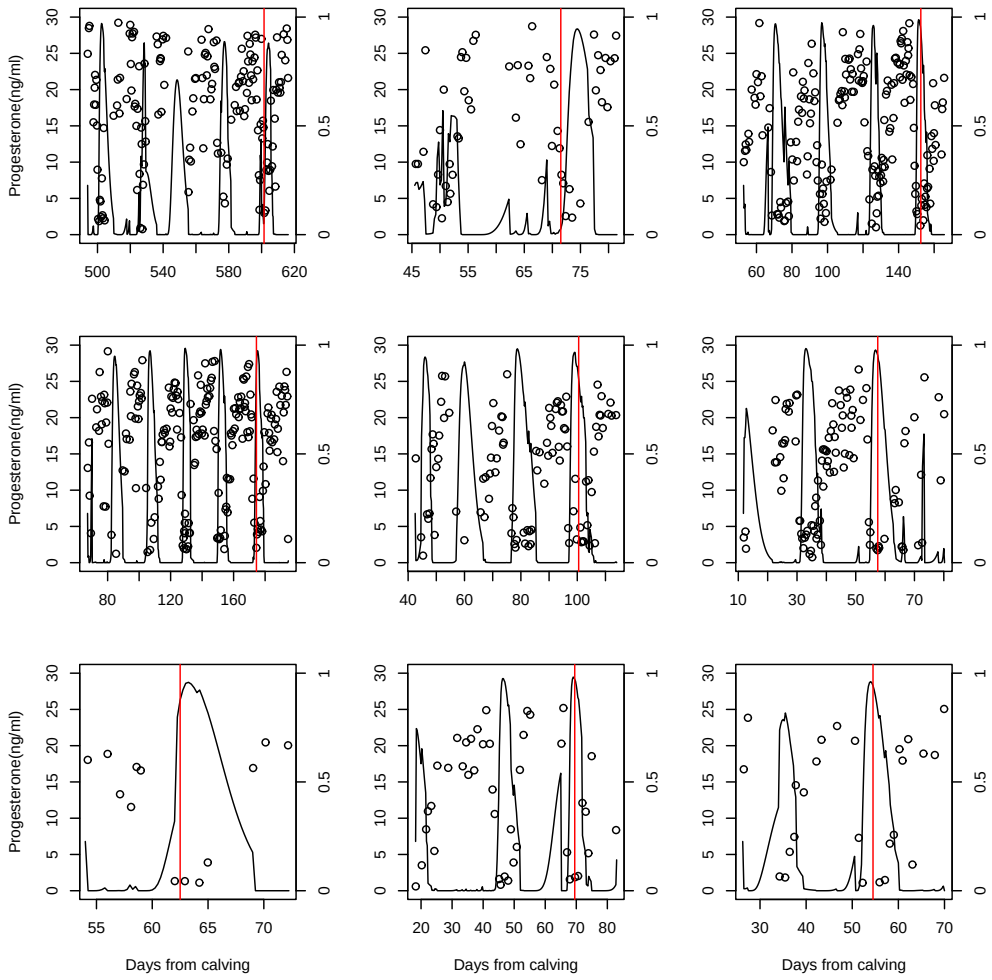

Figure 5: Filter probabilities (full drawn curve) of being in the low stage plotted against time for nine cow-lactations. The probability scale is shown on the right vertical axis. The vertical line shows the time of an insemination resulting in a confirmed pregnancy.

7 Conclusions

The first objective of this study was to develop a state space model with a corresponding Kalman filter to model data with a cyclic nature. This has been done as described in Section 3 and Section 4. Furthermore in Section 5 we discussed techniques for estimation of some of the parameters in the model.

In Section 6 we analysed the progesterone data using the state space

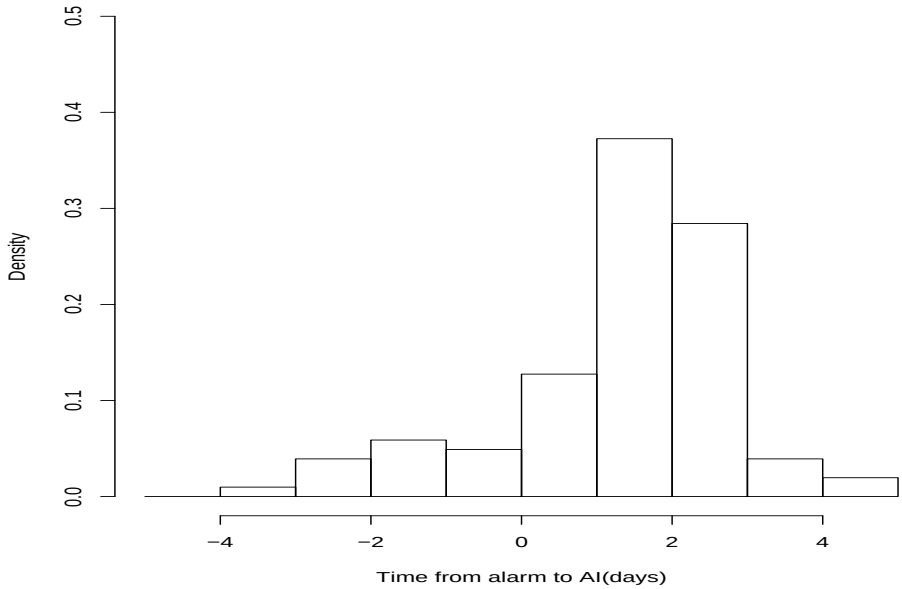

Figure 6: Histogram of the time from alarm goes off to known time for artificial insemination $t_{\text{ins}} - t_{\text{predict}}$. The observed mean of this sample is 1.431 meaning that on average the alarm will tell the farmer to look for signs of oestrus a little less than one and a half day before oestrus actually occurs.

model developed in this study. We discussed how the model was able to provide an alarm for oestrus in cows. Because the aim was to evalute the use of progesterone for detection of oestrus, the time difference $t_{\text{ins}} - t_{\text{predict}}$ has to be positive so that the farmer is told to look for signs of oestrus before oestrus occurs. To be an efficient mechanism for detection of oestrus the variance of $t_{\text{ins}} - t_{\text{predict}}$ must be as small as possible. In Section 6 we found the mean of $t_{\text{ins}} - t_{\text{predict}}$ to be positive fulfilling the first requirement of a possible alarm. To judge if the corresponding variance is small enough, for this study to prove the usefulness of progesterone in oestrus detection, two issues with the data need to be mentioned. Firstly, only the day for succesful insemination is given for each cow-lactation. Secondly, the time point of the succesful insemination is not the optimal measure to evaluate an alarm. We would rather wish to know the time at which an insemination has the highest probability of being succesful. In biological terms, this is related to the time of ovulation, which has been shown to be a rather variable time interval after the onset of oestrus. Onset of oestrus can not be measured by progesterone. The confirmed inseminations in our data are spread around this time of highest probability of success with some variation. Both of these issues contribute to a certain variance that no alarm based on any

progesterone model can remove.

8 Acknowledgements

We gratefully acknowledge the contribution of the farm staff at the Danish Cattle Research Centre for providing the data used in this study, which was partly funded by the Danish Ministry of Food, Agriculture and Fisheries and the Danish Cattle Association through the Biosens project. Further, the first author thanks Paul Fearnhead for inspiration in the process of constructing the cyclic model and the corresponding filter.

A Stationary distribution of (R_t, S_t, N_t)

Lemma 1. *The stationary distribution of* $(R_t - t, S_t, N_t - t)$ *is*

$$\pi(j, q, l) = \frac{W_q(l - j)}{\sum_r \nu(r)}.$$

Proof. With

$$P(j, q, l|j', q', l') = P(R_{t+1} = j, S_{t+1} = q, N_{t+1} = l|R_t = j', S_t = q', N_t = l')$$

we must show that

$$\sum_{j',q',l'} \pi(j', q', l')P(j, q, l|j', q', l') = \pi(j, q, l),$$

for all j, q and l. We split the proof in two cases. Firstly we consider the case of no changepoint at time t which means $j < -1$. Then by (2) $P(j, q, l|j', q', l') = 0$ unless $j' = j + 1$, $l' = l + 1$ and $q' = q$. Also by (2) we find that

$$\pi(j + 1, q, l + 1)P(j, q, l|j + 1, q, l + 1) = \pi(j, q, l).$$

In the other case where $j = -1$ corresponding to a changepoint at time t we have that $P(j, q, l|j', q', l') = 0$ unless $q' = q - 1(\text{mod } m)$ and $l' = 0$. Here we find that

$$\begin{aligned}
&\sum_{j'} \pi(j', q - 1(\text{mod } m), 0)P(-1, q, l|j', q - 1(\text{mod } m), 0) \\
&\quad = \sum_{j'} \frac{W_{q-1(\text{mod } m)}(0 - j')}{\sum_r \nu(r)} W_q(l + 1) \\
&\quad = \frac{1}{\sum_r \nu(r)} W_q(l - j),
\end{aligned}$$

to complete the proof of the lemma. □

B Mathematical description of updating equations

In section 4 we presented an approximate filter for the state variables of our model. Here we give a detailed derivation of the filter recursions. That is, we

derive $p(R_{t+1}=j, S_t=q, N_{t+1}=l, a_{t+1}, b_{t+1}|y^{t+1}) = p_{t+1}(j,q,l,a_{t+1},b_{t+1})$ from p_t, the transition density and the likelihood of y_t^{t+1}, using standard updating formulas. Because we use the approximation (7), at each time point t we have to update the quantities $p_t(j,q,l)$, $\mu_t(j,q,l)$ and $\Sigma_t(j,q,l)$ for all j,q and l. Therefore we now assume that all the quantities are known at time t and in the following we will prove that the updating equations given in Section 4 are valid.

We first consider the case where $j < t$, that is, there is no change of the stage at time t. In this case according to (2), $R_t = R_{t+1}$, $S_t = S_{t+1}$ and $N_t = N_{t+1}$. Then by (5), $a_t = a_{t+1}$ and $b_t = b_{t+1}$. Therefore only one transition is possible and we get directly

$$\begin{aligned}
&p_{t+1}(j,q,l,a,b)\\
&\quad = c_{t+1}(y^{t+1})p_t(j,q,l)\phi(a,b;\ \mu_t(j,q,l),\Sigma_t(j,q,l))\\
&\qquad \times \prod_{(t)} \phi\Big(y_i;\ a\frac{l-s_i}{l-j} + b\frac{s_i-j}{l-j}, \sigma_q^2\Big)\\
&\quad = p_t(j,q,l)\frac{\phi(0;\ \mu_t(j,q,l),\Sigma_t(j,q,l))}{\phi(0;\ \tilde{\mu}_t(j,q,l),\tilde{\Sigma}_t(j,q,l))}\\
&\qquad \times \phi(a,b;\ \tilde{\mu}_t(j,q,l),\tilde{\Sigma}_t(j,q,l)) \prod_{(t)} \phi(y_i;\ 0,\sigma_q^2), \qquad (16)
\end{aligned}$$

with $\tilde{\Sigma}_t(j,q,l)^{-1}$ and $\tilde{\mu}_t(j,q,l)$ given by the right hand sides of (9) and (10). The normalizing constant $c_{t+1}(y^{t+1})$ is $p(y^t)/p(y^{t+1})$. Integrating (16) with respect to (a,b) the $\phi(a,b;\cdot)$ term disapear, and the formula (8) for $p_{t+1}(j,q,l)$ is obtained. Next dividing (16) by (8) we see that the filtering distribution of (a_{t+1}, b_{t+1}) is the normal distribution with mean $\tilde{\mu}_t(j,q,l)$ and variance $\tilde{\Sigma}_t(j,q,l)$ which proves (9) and (10).

We next consider the case where $j = t$, that is, there is a change of the stage at time t. In this case $N_t = t$, R_t can be any value $j' < t$, $S_t = q' = q-1 (\mathrm{mod}\ m)$, and $b_t = a_{t+1} = a$. Letting $\tilde{q} = q+1 (\mathrm{mod}\ m)$, using the transistion density (2) we find

$$\begin{aligned}
p_{t+1}(t,q,l,a,b) &= c_{t+1}(y^{t+1})W_q(l-t)\sum_{j'<t} p_t(j',q',t)\\
&\quad \times \int_{a'} \phi(a',a;\ \mu_t(j',q',t),\Sigma_t(j',q',t))\phi(b;\ \mu_{\tilde{q}},\omega_{\tilde{q}}^2)\\
&\quad \times \prod_{(t)} \phi(y_i;\ a\frac{l-s_i}{l-t} + b\frac{s_i-t}{l-t}, \sigma_q^2)
\end{aligned}$$

$$
\begin{aligned}
&= W_q(l-t)\sum_{j'<t} p_t(j',q',t)\phi(a;\ \mu_t(j',q',t)_2, \Sigma_t(j',q',t)_{22}) \\
&\quad \times \phi(b;\ \mu_{\tilde{q}}, \omega_{\tilde{q}}^2)\prod_{(t)} \phi(y_i;\ a\frac{l-s_i}{l-t} + b\frac{s_i-t}{l-t}, \sigma_q^2) \\
&= W_q(l-t)\prod_{(t)} \phi(y_i;\ 0, \sigma_q^2)\sum_{j'<t} p_t(j',q',t) \\
&\quad \times \frac{\phi(0;\ \mu_t(j',q',t)_2, \Sigma_t(j',q',t)_{22})\phi(0;\ \mu_{\tilde{q}}, \omega_{\tilde{q}}^2)}{\phi(0;\ \bar{\mu}_t(j',q',t,l), \bar{\Sigma}_t(j',q',t,l))} \\
&\quad \times \phi(a,b;\ \bar{\mu}_t(j',q',t,l), \bar{\Sigma}_t(j',q',t,l)),
\end{aligned} \tag{17}
$$

with

$$
\begin{aligned}
\bar{\Sigma}_t(j',q',t,l)^{-1} &= \begin{pmatrix} \Sigma_t(j',q',t)_{22}^{-1} & 0 \\ 0 & (\omega_{\tilde{q}}^2)^{-1} \end{pmatrix} \\
&\quad + \frac{1}{\sigma_q^2}\sum_{(t)} \begin{pmatrix} [(l-s_i)/(l-t)]^2 & (s_i-t)(l-s_i)/(l-t)^2 \\ (s_i-t)(l-s_i)/(l-t)^2 & [(s_i-t)/(l-t)]^2 \end{pmatrix}
\end{aligned} \tag{18}
$$

and

$$
\begin{aligned}
&\bar{\Sigma}_t(j',q',t,l)^{-1}\bar{\mu}_t(j',q',t,l) \\
&\quad = \begin{pmatrix} \mu_t(j',q',t)_2/\Sigma_t(j',q',t)_{22} \\ \mu_{\tilde{q}}/\omega_{\tilde{q}}^2 \end{pmatrix} + \frac{1}{\sigma_q^2}\sum_{(t)} \begin{pmatrix} y_i(l-s_i)/(l-t) \\ y_i(s_i-t)/(l-t) \end{pmatrix}
\end{aligned} \tag{19}
$$

Integrating (17) with respect to (a,b) the $\phi(a,b;\cdot)$ term disappear, and the formula (11) is obtained. Dividing (17) by (11) we see that the density of (a_{t+1}, b_{t+1}) is

$$
\sum_{j'<t} \alpha_t(j',q',t,l)\phi(a,b;\bar{\mu}_t(j',q',t,l), \bar{\Sigma}_t(j',q',t,l)), \tag{20}
$$

where

$$
\alpha_t(j',q',t,l) = \frac{\gamma_t(j',q',t,l)}{\sum_{\tilde{j}<t}\gamma_t(\tilde{j},q',t,l)}
$$

with

$$
\gamma_t(j',q',t,l) = p_t(j',q',t)\frac{\phi(0;\ \mu_t(j',q',t)_2, \Sigma_t(j',q',t)_{22})}{\phi(0;\ \bar{\mu}_t(j',q',t,l), \bar{\Sigma}_t(j',q',t,l))}.
$$

We approximate the Gaussian mixture in (20) by a single Gaussian density with the same mean and variance. This gives the formulas (12) and (13).

C Estimation of waiting time distributions

When we estimate the waiting time probabilities $W_q(l)$ as described in Section 5.2 we need to calculate $E(\sum_1^k \psi_i | y^k)$ iteratively. Now let $\psi_i = \psi(z_i, z_{i-1})$ be a general estimation function where $z_i = (R_i, S_i, N_i)$. Using the approximation

$$p(z_{k+1}, y_{k+1} | z_1^k, y_1^k) \approx p(z_{k+1}, y_{k+1} | z_k, y_1^k)$$

we have the following updating rule

$$E\Big(\sum_1^{k+1} \psi_i | z_{k+1}, y_1^{k+1}\Big) = \sum_{z_k} \Big(E(\sum_1^k \psi_i | z_k, y_1^k) + \psi_{k+1}\Big) \frac{p(z_{k+1}, y_{k+1} | z_k, y_1^k) p(z_k | y_1^k)}{\sum_{\tilde{z}_k} p(z_{k+1}, y_{k+1} | \tilde{z}_k, y_1^k) p(\tilde{z}_k | y_1^k)}.$$

We next specialize this formula. First we consider the case $j = R_{k+1} < k$ for which we find

$$E\Big(\sum_1^{k+1} \psi_i | (j,q,l), y_1^{k+1}\Big) = E\Big(\sum_1^k \psi_i | (j,q,l), y_1^k\Big) + \psi((j,q,l),(j,q,l)).$$

For the case $j = k$ we use $q' = q - 1 (\mathrm{mod}\ m)$ and get

$$E\Big(\sum_1^{k+1} \psi_i | (k,q,l), y_1^{k+1}\Big) = \frac{\sum_{j'<k} \{E(\sum_1^k \psi_i | (j',q',k), y_1^k) + \psi((k,q,l),(j',q',k))\} \gamma_t(j',q',t,l)}{\sum_{j'<k} \gamma_t(j',q',t,l)}.$$

Finally, we find

$$E\Big(\sum_1^{k+1} \psi_i | y_1^{k+1}\Big) = \sum_{k,q,l} p_{k+1}(k,q,l) E\Big(\sum_1^{k+1} \psi_i | (k,q,l), y_1^{k+1}\Big).$$

Bibliography

Cavalieri, J., V. E. Eagles, M. Ryan, and K. L. Macmillan (2001). Comparison of four methods for detection of oestrus and diseases in dairy cattle based on time series analysis combined with a kalman filter. *Computers and Electronics in Agriculture 17*, 399–407.

Cavalieri, J., V. E. Eagles, M. Ryan, and K. L. Macmillan (2003). Comparison of four methods for detection of oestrus in dairy cows with resynchronised oestrus cycles. *Aust. Vet. J. 81*, 422–25.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B 39(1)*, 1–38.

Dobson, H., S. L. Walker, M. J. Morris, J. E. Routly, and R. F. Smith (2008). Why is it getting more difficult to successfully ai dairy cows? *Animal in press.*

Elashoff, M. and L. Ryan (2004). An EM algorithm for estimating equations. *J. Comput. Graph. Statist. 13*, 48–65.

Elton, C. and M. Nicholson (1942). The ten-year cycle in numbers of the lynx in canada. *J. Animal Ecol. 11*, 215–244.

Fearnhead, P. and Z. Liu (2007). Online inference for multiple changepoint problems. *J. Roy. Statist. Soc. B 69*, 589–605.

Fearnhead, P. and D. Vasileiou (2008). Bayesian analysis of isochores. *Preprint available from* `www.maths.lancs.ac.uk/~fearnhea/publications`.

Friggens, N. C., M. Bjerring, C. Ridder, S. Højsgaard, and T. Larsen (2008). Improved detection of reproductive status in dairy cows using milk progesterone measurements. *To appear in Reprod. Domestic Anim.*

Fulkerson, W. J., G. J. Sawyer, and I. Crothers (1983). The accuracy of several aids in detecting oestrus in cattle. *Appl. Anim. Ethol. 10*, 199–208.

Heyde, C. and R. Morton (1996). Quasi-likelihood and generalizing the EM-algorithm. *J. Roy. Statist. Soc. B 58*, 317–327.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika 41*, 100–115.

Peters, A. R. and P. J. H. Ball (1995). *Reproduction in cattle*. Blackwell Science, Oxford, UK.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Roelofs, J., F. V. Eerdenburg, W. Hazeleger, N. Soede, and B. Kemp (2006). Relationship between progesterone concentrations in milk and blood and time of ovulation in dairy cattle. *Animal Reproduction Science 91*, 337–343.

Roelofs, J., E. Graat, E. Mullaart, N. Soede, W. Voskamp-Harkema, and B. Kemp (2006). Effects of insemination-ovulation interval on fertilization rates and embryo characteristics in dairy cattle. *Theriogenology 66*, 2173–2181.

Rosen, O., W. Jiang, and M. Tanner (2000). Mixtures of marginal models. *Biometrika 87*, 391–404.

Xu, Z. Z., D. J. McKnight, R. Vishwanath, C. J. Pitt, and L. J. Burton (1998). Estrus detection using radiotelemetry or visual observation and tail painting for dairy cows on pasture. *J. Dairy. Sci. 81*, 2890–2896.

Paper

C

Hansen, J.V., and Jensen, J.L. (2008).
Asymptotics for estimating equations in hidden Markov models.

Submitted to *Statistica Sinica*.
To appear as Thiele Research Report, Department of Mathematical Sciences, University of Aarhus.

ASYMPTOTICS FOR ESTIMATING EQUATIONS IN HIDDEN MARKOV MODELS

Jørgen Vinsløv Hansen and Jens Ledet Jensen

University of Aarhus

Abstract

Results on asymptotic normality for the maximum likelihood estimate in hidden Markov models are extended in two directions. The stationarity assumption is relaxed, which allows for a covariate process influencing the hidden Markov process. Furthermore a class of estimating equations is considered instead of the maximum likelihood estimate. The basic ingredients are mixing properties of the process and a general central limit theorem for weakly dependent variables. The results are illustrated with a cyclic model for the progesterone concentration in cowmilk.

Key words and phrases: Cyclic model, Estimating equation, Mixing properties, Progesterone concentration.

1 Introduction

Unless simulation based methods are used inference in hidden Markov models is based on the asymptotic normality of the parameter estimates. For the case of a finite state space for both the hidden variable x and the observed variable y, asymptotic normality for the maximum likelihood estimate was established in the pioneering paper of Baum and Petrie (1966). More than thirty years elapsed until this result was generalised to a general state space for the observed variable y by Bickel, Ritov and Rydén (1998), and still further generalized to a

non-discrete state space for the hidden variable x by Jensen and Petersen (1999). In these papers stationarity is a crucial assumption. The log likelihood is a sum where the individual terms are the log densitites of y_i given the past $y_1, \ldots, y_{i-1}$. These are replaced by the similar expressions conditioned instead on the infinite past $\ldots, y_{-1}, y_0, \ldots, y_{i-1}$. A martingale central limit theorem is then used to establish asymptotic normality of the score function. In this paper we use a different approach that allows us to consider nonhomogeneous processes and to consider alternatives to the maximum likelihood estimates. To illustrate the scope of the setup we briefly describe an example from evolutionary biology.

Example 1. Let $v(t) = (v_1(t), \ldots, v_n(t))$ be a sequence of letters from the alphabet $\{A, G, C, T\}$ of nucleotides at time t. The sequence at time $t = 0$ is fixed and known. Time is discrete. The process is observed at time $t = T$, but not observed at the times $t = 1, 2, \ldots, T-1$ in between. The sequence $v(t)$ evolves according to a Markov chain with transition probabilitites of the form

$$p(v(t+1)|v(t)) = \prod_{i=1}^{n} h(v_i(t+1)|v_{i-1}(t+1), v_i(t), v_{i+1}(t)),$$

for some transition probability h. This formalizes a time discretized version of a model where the probability of a change of a nucleotide $v_i(t)$ depends on the two neighbouring nucleotides. Let now $x_i = (v_i(1), v_i(2), \ldots, v_i(T))$ be the complete history for nucleotide i. It can be seen that the conditional distribution of x_i given $x_1, \ldots, x_{i-1}$ depends on (x_{i-2}, x_{i-1}) only. We thus have a second order hidden Markov model where the observed variable is $y_i = v_i(T)$. The underlying Markov structure is inhomogeneous due to the fixed initial sequence $v(0)$.

Asymptotic normality for a class of estimating equations, in the setting af evolutionary models for DNA, has been treated in Jensen (2005). In that paper both the state space of the hidden variable x and the observed variable y is finite. Here we extend the results in Jensen (2005) to a setup akin that of Jensen and Petersen (1999) with a general state space for the observed variable and a general state space for the hidden variable. Nonhomogeneity is introduced through a covariate. We base the asymptotic normality of the "score function" directly on the mixing properties of the process, using a central limit theorem extracted from Götze and Hipp (1983). Although the state space is general the conditions imposed effectively restricts the space to be compact.

In section 2 we describe the setup and results in detail and define the class of estimating equations that we consider. In section 3 we illustrate the results for a hidden cyclic model used to describe the progesterone concentration in cowmilk. The proofs of the results are split into three sections. In section 4 we study the mixing properties of the process and use these in section 5 to derive a central limit theorem for the "score function". Finally, in section 6 we derive the uniform convergence of the "observed information".

2 Assumptions and results

We consider an observed process $y_1, \ldots, y_n$ controlled by an unobserved Markov process $\{x_i\}$. Conditionally on the x-process the y_is are independent. Both the observed y_i and the unobserved x_i may be influenced by a covariate z_i, making the process inhomogeneous. The transition density of the Markov process is $p_\theta(x_i|x_{i-1}; z_i)$, where $p_\theta(\tilde{x}|x; z)$ is a density in $\tilde{x}$ with respect to a probability measure μ on the state space for the hidden variable. The conditional density of y_i is $p_\theta(y_i|x_i; z_i)$, where $p_\theta(y|x; z)$ is a density in y with respect to a measure ν. Both these densities are parametrized by the d-dimensional parameter θ. We split the assumptions into two parts, one part concerned with the process itself, Conditions 2 and 3 below, and another part concerned with the estimating function used, Condition 5 below.

Condition 2 ensures mixing of the underlying Markov chain. In order to allow for the possibility that in a single step the Markov chain can reach only a subset of the state space, we use the m_0–step transition probabilities in the condition. This transition density depends on several z_i's, but in order not to overburden the notation we write simply z instead. We can start by establishing exponential mixing of the m_0–step chain $\{x_{jm_0}\}$, and from this trivially obtain mixing of the original chain $\{x_j\}$. To avoid complicated notation we consider in the proofs the case with $m_0 = 1$. In the setting of a DNA sequence as in Jensen (2005) the two-step transition probabilities will suffice, whereas in the setting of a process with a cyclic nature as described in section 3 higher order transitions may be needed. Point (i) of Condition 3 limits the influence of the hidden variable on the observed variable. This condition is needed when studying the mixing properties of the hidden chain conditioned on the observed y–process. Point (ii) of Condition 3 limits the conditional score function based on (x_i, y_i) given x_{i-1}. The true value of the parameter is θ_0.

Condition 2. *There exists $\delta_0 > 0$, a positive integer m_0, and constants $0 < \tau < \sigma < \infty$, such that*

$$\tau \leq p_\theta(x_{m_0}|x_0; z) \leq \sigma \quad \textit{for all } (x_0, x_{m_0}, z) \textit{ and all } |\theta - \theta_0| \leq \delta_0.$$

To state the next condition we introduce some notation. Likelihood quantities for the chain (x_i, y_i) are denoted by ω as follows

$$\omega_i(\theta) = \log[p_\theta(x_i|x_{i-1}; z_i)p_\theta(y_i|x_i; z_i)] \quad \text{and} \quad \omega_i^r(\theta) = \frac{\partial}{\partial \theta_r}\omega_i(\theta).$$

With δ_0, τ, and σ from Condition 2 define

$$\xi(y) = \sup_{x_1, x_2, z, |\theta - \theta_0| \leq \delta_0} \frac{p_\theta(y|x_1; z)}{p_\theta(y|x_2; z)}, \qquad \rho(y) = 1 - \tau^2/(\sigma\xi(y)),$$

and

$$\beta_1 = \inf_{x,z} \int p_{\theta_0}(y|x;z)/\xi(y)\nu(dy).$$

Condition 3. *Let δ_0 be as in Condition 2.*

i) Assume that $\xi(y) < \infty$ for all y and that $\beta_1 > 0$.

ii) Assume that there exists a function $h_0(y)$ with

$$c_0 = \sup_{x,z} \int h_0(y) p_{\theta_0}(y|x;z)\nu(dy) < \infty,$$

such that for all $r = 1, \ldots, d$ and all i,

$$\sup_{x_{i-1},x_i,z_i,|\theta-\theta_0|\leq\delta_0} |\omega_i^r(\theta)| \leq h_0(y_i).$$

The second part of the conditions relates to the estimating equation. Let $\psi(\theta, \bar{x}, y; z)$ be a function of the parameter θ, a triple $\bar{x}$ of consequtive states, an observed variable y and covariates z. Let $\psi_i(\theta) = \psi(\theta, \bar{x}_i, y_i; z)$, where $\bar{x}_i = (x_{i-1}, x_i, x_{i+1})$. We think of $\sum_{i=1}^n \psi_i(\theta) = 0$ as an estimating equation had both x and y being observed. Having observed y only we use the estimating equation

$$\sum_{i=1}^n E_\theta[\psi_i(\theta)|(1,n)] = 0, \tag{1}$$

where $E_\theta(\cdot|(1,n))$ is the conditional mean given $y_1, \ldots, y_n$. The coordinates of $\psi_i(\theta)$ are denoted by $\psi_i^r(\theta)$, $r = 1, \ldots, d$ and the derivatives of these are $\psi_i^{rs}(\theta) = \frac{\partial}{\partial\theta_s}\psi_i^r(\theta)$. In Appendix I a recursive formula for evaluating the estimating function on the left hand side of (1) is given. To solve (1) one often uses an EM-type algorithm, that is, $\sum_{i=1}^n E_\theta[\psi_i(\tilde{\theta})|(1,n)] = 0$ is solved with respect to $\tilde{\theta}$, and this defines a new value improving on the old value θ. This *EEE*-algorithm (Expectation–Estimating–Equation) has been considered in Heyde and Morton (1996), Rosen, Jiang and Tanner (2000) and Elashoff and Ryan (2004).

The "observed information" in this setting, that is, the derivative of the left hand side of (1), is given by

$$\begin{aligned} J_n(\theta) &= -\frac{\partial}{\partial\theta} E_\theta\Big[\sum_{i=1}^n \psi_i(\theta)|(1,n)\Big] \\ &= -E_\theta\Big[\sum_{i=1}^n \frac{\partial}{\partial\theta}\psi_i(\theta)|(1,n)\Big] - V_\theta\Big[\sum_{i=1}^n \psi_i(\theta), \sum_{i=1}^n \frac{\partial}{\partial\theta}\omega_i(\theta)|(1,n)\Big]. \end{aligned}$$

This formula corresponds to the formula in Louis (1982) for the maximum likelihood equation. A derivation can be found in Jensen (2005).

Before stating the second part of the conditions it is convenient to introduce a notation for a class of functions satisfying suitable conditions.

Definition 4. *Consider for each i a function $a_i(\theta)$ which also depends on $(\bar{x}_i, y_i, z)$. We say that these functions belong to class G_k if there exist $\delta_0 > 0$, a function $a_0(y)$ and a finite constant $c_0^k(a)$ such that*

$$|a_i(\theta)| \leq a_0(y_i) \quad \text{for all } (\bar{x}_i, z) \text{ and all } |\theta - \theta_0| \leq \delta_0,$$

and

$$\sup_{x,z} \int a_0(y)^k p_{\theta_0}(y|x;z)\nu(dy) \leq c_0^k(a).$$

If, furthermore, there exist a function $a_1(y)$ and finite constants $c_1(a), c_1^m(a)$ such that

$$|a_i(\theta) - a_i(\theta_0)| \leq |\theta - \theta_0| a_1(y_i) \quad \text{for all } (\bar{x}_i, z) \text{ and all } |\theta - \theta_0| \leq \delta_0,$$

and

$$\sup_{x,z} \int a_1(y) p_{\theta_0}(y|x;z)\nu(dy) \leq c_1(a),$$

$$\sup_{x,z} \int a_0(y)^m h_0(y) p_{\theta_0}(y|x;z)\nu(dy) \leq c_1^m(a),$$

we say that the set of functions belong to class $G_{k,m}$.

Condition 5. *i) For all $r = 1, \ldots, d$ the set of functions $\{\psi_i^r(\theta)\}$ belongs to class G_3 and $E_\theta \psi_i(\theta) = 0$.*

ii) For all $r, s = 1, \ldots, d$ the set of functions $\{\psi_i^{rs}(\theta)\}$ belongs to class $G_{1,2}$.

iii) For all $r = 1, \ldots, d$ the set of functions $\{\omega_i^r(\theta)\}$ belongs to class $G_{1,2}$.

We now formulate the results of this paper.

Theorem 6. *Assumme that Condition 2, Condition 3(i), and Condition 5(i) hold. Define $S_n = \sum_{i=1}^n E_{\theta_0}(\psi_i(\theta_0)|(1,n)/\sqrt{n}$ and assume that the covariates $\{z_i\}$ are such that the variance of S_n converges to a positive definite limit. Then a central limit theorem holds for the normalized sum S_n.*

Theorem 7. *Assume that Condition 2, 3, and 5 hold. Let $\delta_n \to 0$ for $n \to \infty$. Then*

$$E_{\theta_0}\Big\{ \sup_{|\theta - \theta_0| \leq \delta_n} \frac{1}{n} \big| J_n(\theta) - J_n(\theta_0) \big| \Big\} \to 0.$$

Corollary 8. *Assume that Condition 2, 3, and 5 hold. Assume that the covariates $\{z_i\}$ are such that the variance of $S_n = \sum_{i=1}^n E_{\theta_0}(\psi_i(\theta_0)|(1,n)/\sqrt{n}$ converges to a positive definite limit $V(\theta_0)$, and also $\frac{1}{n} J_n(\theta_0)$ converges to a positive definite limit $I(\theta_0)$. Then there exists a sequence $\hat{\theta}_n$ solving the estimating equation such that $\hat{\theta} \to \theta_0$ in probability and $\sqrt{n}(\hat{\theta} - \theta_0)$ has a limiting normal distribution with mean zero and variance $I(\theta_0)^{-1} V(\theta_0) I(\theta_0)^{-1}$.*

3 Example: cyclic model

In Hansen (2008) a cyclic hidden Markov model is described for the progesterone concentration in cowmilk. The observed process y_j is the measured progesterone concentration in the milk at each milking. The underlying dynamic is described by a stage i_j, a level v_j giving the mean of the observed process, a slope s_j which defines the increase in the level v_j, and a waiting time r_j until the next change of the stage. The stage describes a cyclic nature where $i = 1$ corresponds to a low stage, this is followed by an increasing stage $i = 2$, followed next by a high stage $i = 3$, and ending in a decreasing stage $i = 4$. Below, when $i = 4$ the sum $i+1$ means the stage 1. The process is controlled by two transition probabilities, $p(r|i;\gamma)$ which is the probability of a new waiting time r at a point in time where the stage changes from $i-1$ to i and which depends on a parameter γ, and $p(s|r,v,i)$ which is the probability of a new slope s at a point in time where the stage changes from $i-1$ to i, the present level is v, and the new waiting time is r. Formally, the Markov structure for the hidden variable $x_j = (i_j, r_j, v_j, s_j)$ is given by

$$\left.\begin{array}{l} i_{j+1} = i_j \\ r_{j+1} = r_j - 1 \\ v_{j+1} = v_j + s_j \\ s_{j+1} = s_j \end{array}\right\} \quad r_j > 1,$$

$$\left.\begin{array}{l} i_{j+1} = i_j + 1 \\ r_{j+1} \sim p(\cdot|i_{j+1};\gamma) \\ v_{j+1} = v_j + s_j \\ s_{j+1} \sim p(\cdot|r_{j+1}, v_{j+1}, i_{j+1}) \end{array}\right\} \quad r_j = 1.$$

Conditionally on the hidden state the observed variable y_j is normally distributed with mean v_j and variance σ^2.

We consider γ and σ^2 to be cow specific parameters with γ allowing for variation in the mean cycle length from cow to cow, and with σ^2 allowing for varying degree of fit of the hidden model. Finally, we consider the case where s_j and v_j belong to compact sets and r_j belongs to a finite set (this is slightly different from the setup in Hansen, 2008).

The full likelihood, having observed both x_j and y_j, $j = 1, \ldots, n$, and conditioning on x_0, leads to the likelihood equations

$$\sum_{j=1}^{n} \big[(y_j - v_j)^2 - \sigma^2\big] = 0,$$

$$\sum_{j=1}^{n} \big[\frac{d}{d\gamma} \log p(r_j|i_j;\gamma)\big] 1(i_{j-1} \neq i_j) = 0.$$

We replace the first of these equations with one giving a more robust estimate of σ. Thus we use instead

$$\sum_{j=1}^{n}\big[|y_j - v_j| - \sigma\sqrt{\frac{2}{\pi}}\big] = 0.$$

In relation to our general setup we thus have $\theta = (\sigma, \gamma)$ and

$$\psi_j^1 = |y_j - v_j| - \sigma\sqrt{\frac{2}{\pi}}, \qquad \psi_j^2 = \big[\frac{d}{d\gamma}\log p(r_j|i_j;\gamma)\big]1(i_{j-1} \neq i_j).$$

The derivatives of these with respect to σ and γ are

$$\begin{aligned} \psi_j^{11} &= -\sqrt{\frac{2}{\pi}}, & \psi_j^{12} &= 0, \\ \psi_j^{21} &= 0, & \psi_j^{22} &= \big[\frac{d^2}{d\gamma^2}\log p(r_j|i_j;\gamma)\big]1(i_{j-1} \neq i_j). \end{aligned}$$

Furthermore, we have

$$\omega_j = \begin{cases} \log\big[\varphi(y_j; v_j, \sigma)\big] & i_j = i_{j-1}, \\ \log\big[p(r_j|i_j;\gamma)p(s_j|r_j, v_j, i_j)\varphi(y_j; v_j, \sigma)\big] & i_j \neq i_{j-1}, \end{cases}$$

where $\varphi(y; v, \sigma)$ is the density of a normal distribution with mean v and variance σ^2.

The derivatives of ω_j with respect to σ and γ are

$$\omega_j^1 = \frac{1}{\sigma^3}(y_j - v_j)^2 - \frac{1}{\sigma}, \quad \omega_j^2 = \frac{d}{d\gamma}\log p(r_j|i_j;\gamma)\big]1(i_{j-1} \neq i_j).$$

Condition 2 will hold under mild conditions on the transition densities due to the compactness of the state space. We do not discuss this further here. Since the state space is bounded we make the assumption that the first three derivatives of $p(\cdot|i;\gamma)$ are bounded. For condition 3 i) we find the bound $\xi(y) \leq \exp(b_0 + b_1|y|)$ for suitable constants b_0 and b_1, due to the finiteness of the level v. Then clearly also $\beta_1 > 0$. For condition 3 ii) we can take $h_0(y) = b_0 + b_1|y| + b_2y^2$ for suitable constants b_0, b_1, b_2. Similarly in condition 5 i) we can use a bound on the form $a_0(y) = b_0 + b_1|y|$. For condition 5 ii) $a_0(y)$ and $a_1(y)$ can be taken as constants. And, finally, for condition 5 iii) both $a_0(y)$ and $a_1(y)$ can be bounded by $b_0 + b_1|y| + b_2y^2$ for suitable constants b_0, b_1, b_2.

In conclusion we see that the standard asymptotic results hold for the estimates in a cyclic model as described here and considered (with slight modifications) in Hansen (2008).

4 Mixing

As a first step in the proof of the main results we study the mixing properties of the process. We use throughout Condition 2 with $m_0 = 1$. Our results hold for all $|\theta - \theta_0| \leq \delta_0$, and we skip θ in the notation below. First we state bounds on the transition densities for the hidden chain conditioned on the observed y–process. The lemma has been proved in Jensen and Petersen (1999).

Lemma 9. *Assume Condition 2. Conditioned on the y–process $\{x_n\}$ constitute a Markov chain with*

$$\frac{\tau^2}{\sigma\xi(y_s)} \leq p(x_s|x_{s-1}, x_{s+1}, y; z) \leq \frac{\sigma^2\xi(y_s)}{\tau}.$$

For the original Markov chain (not conditioned on y) we have trivially from Condition 2 that

$$\frac{\tau^2}{\sigma} \leq p(x_{s+1}|x_s, x_{s+2}; z) \leq \frac{\sigma^2}{\tau}. \tag{2}$$

For easy reference we state here a Lemma from Jensen and Petersen (1999) that will be used repeatedly.

Lemma 10. *Assume that ν_1 and ν_2 are dominated by μ and $\nu_1(\mathcal{X}) = \nu_2(\mathcal{X})$. Then for any real valued measurable function h on $\mathcal{X}$ we have*

$$\Big|\int_{\mathcal{X}} h d\nu_1 - \int_{\mathcal{X}} h d\nu_2\Big| \leq \{\sup_x h(x) - \inf_x h(x)\}\{\nu_1(S^+) - \nu_2(S^+)\},$$

where $S^+ = \{d\nu_1/d\mu - d\nu_2/d\mu > 0\}$.

To establish mixing results for both the original hidden Markov chain and for the chain conditioned on the y–process we consider a general Markov chain $\{x_s\}$ satisfying

$$\tau_s \leq p(x_s|x_{s-1}, x_{s+1}) \leq \sigma_s. \tag{3}$$

with $0 < \tau_s < \sigma_s < \infty$ for all s. We start with a result on one-sided and two-sided mixing. To make the notation more transparent we let u_r, for a lower case letter u, denote $x_r = u$, and let A_s, for a upper case letter A, denote $x_s \in A$.

Lemma 11. *Assume (3). Let $r < s < t$ and let $\rho_j = 1 - \tau_j$. Then*

$$\sup_u P(A_s|u_r) - \inf_v P(A_s|v_r) \leq \prod_{j=r+1}^{s} \rho_j,$$

and

$$\sup_{a,b} P(A_s|a_r, b_t) - \inf_{u,v} P(A_s|u_r, v_t) \leq \prod_{j=r+1}^{s} \rho_j + \prod_{j=s}^{t-1} \rho_j.$$

Proof. The proof of the one-sided case is given in Jensen and Petersen (1999) based on Doob (1953, page 198). In Jensen(2005) a similar proof for the two-sided case is indicated. We give here the details of this proof.

Let $r < s < t$. Define, for a fixed set A and a fixed state w, $D(r) = \max_u P(A_s|u_r, w_t)$, $d(r) = \min_u P(A_s|u_r, w_t)$, and $S_r = \{x : p(x_r = x|u_{r-1}, w_t) > p(x_r = x|v_{r-1}, w_t)\}$. Using Lemma 10 in the first inequality below we find

$$\begin{aligned}
&D(r-1) - d(r-1) \\
&\quad = \max_{u,v} \left[P(A_s|u_{r-1}, w_t) - P(A_s|v_{r-1}, w_t)\right] \\
&\quad = \max_{u,v} \int P(A_s|\alpha_r, w_t) \left[p(\alpha_r|u_{r-1}, w_t) - p(\alpha_r|v_{r-1}, w_t)\right] \mu(d\alpha) \\
&\quad \leq (D(r) - d(r)) \max_{u,v} \left[P(S_r|u_{r-1}, w_t) - P(S_r|v_{r-1}, w_t)\right] \\
&\quad \leq (D(r) - d(r)) \max_{u,v} \left[1 - P(S_r^c|u_{r-1}, w_t) - P(S_r|v_{r-1}, w_t)\right] \\
&\quad \leq (D(r) - d(r))\big(1 - \tau_r\big) \\
&\quad = (D(r) - d(r))\rho_r,
\end{aligned}$$

where we used the bound

$$p(x_r|u_{r-1}, w_t) = \int p(x_r|u_{r-1}, v_{r+1}) p(v_{r+1}|u_{r-1}, w_t) \mu(dv) \geq \tau_r.$$

Iterating, we obtain

$$\max_{u,v} |P(A_s|u_r, w_t) - P(A_s|v_r, w_t)| \leq \prod_{j=r+1}^{s} \rho_j,$$

A similar argument gives

$$\max_{u,v} |P(A_s|w_r, u_t) - P(A_s|w_r, v_t)| \leq \prod_{j=s}^{t-1} \rho_j.$$

Combining the two latter bounds lead to

$$\begin{aligned}
&\max_{a,b,u,v} |P(A_s|a_r, b_t) - P(A_s|u_r, v_t)| \\
&\quad \leq |P(A_s|a_r, b_t) - P(A_s|u_r, b_t)| + |P(A_s|u_r, b_t) - P(A_s|u_r, v_t)| \\
&\quad \leq \prod_{j=r+1}^{s} \rho_j + \prod_{j=s}^{t-1} \rho_j. \qquad (4)
\end{aligned}$$

□

Lemma 12. *Assume Condition 2. Define $\rho = 1 - \tau^2/\sigma$. For the y-process we have mixing as in Lemma 11 with $\rho_j \equiv \rho$.*

Proof. For the original Markov chain $\{X_n\}$ we have the bounds in Lemma 11 with $\rho_j \equiv \rho$. Letting y_r^j denote $y_r = y^j$ and similarly with x_r^j, we find by using Lemma 10 twice

$$\begin{aligned}
&P(y_s \in A|y_r^1, y_t^1; z) - P(y_s \in A|y_r^2, y_t^2; z) \\
&\quad = \iint P(y_s \in A|x_s; z)p(x_s|x_r, x_t; z)\mu(dx_s) \\
&\qquad \times [p(d(x_r, x_t)|y_r^1, y_t^1; z) - p(d(x_r, x_t)|y_r^2, y_t^2; z)] \\
&\quad \leq \sup_{x_r^1, x_t^1, x_r^2, x_t^2} \Big[\int P(y_s \in A|x_s; z)p(x_s|x_r^1, x_t^1; z)\mu(dx_s) \\
&\qquad - \int P(y_s \in A|x_s; z)p(x_s|x_r^2, x_t^2; z)\mu(dx_s) \Big] \\
&\quad \leq \sup_{x_r^1, x_t^1, x_r^2, x_t^2, B} \big[p(x_s \in B|x_r^1, x_t^1; z) - p(x_s \in B|x_r^2, x_t^2; z) \big] \\
&\quad \leq \rho^{s-r} + \rho^{t-s}.
\end{aligned}$$

□

5 Central limit theorem

In this section we prove Theorem 6. First some notation. Mean values and probabilities are with respect to the true measure corresponding to $\theta = \theta_0$. We do not show θ_0 in the notation. The conditional mean given $(y_s, y_{s+1}, \ldots, y_t)$ is denoted by $E(\cdot|(s,t))$. If, furthermore, we condition on x_s and x_t we use the notation $E(\cdot|[s,t])$. The expression $\prod_{j=s(-u)}^{t} c_j$ is a short hand notation for the expression $\prod_{j=s}^{u-1} c_j + \prod_{j=u+1}^{t} c_j$.

From Götze and Hipp (1983), which deals with Edgeworth expansion, we can extract a central limit theorem suitable for our purpose. We have already seen in Lemma 12 that the observed process is exponentially fast mixing. If w_i is a sequence of random variables with uniformly bounded third absolute moment a central limit theorem holds for the normalized sum under two additional assumptions. The first condition is the standard assumption that the variance of the normalized sum converge. The second condition says that each w_i can for each m be approximated by a function of y_{i-m}^{i+m} introducing an error that is exponentially small in m. To handle this last requirement we have the following lemma.

Lemma 13. *Assume Condition 2 and 3. Let a_i be a function of $(\bar{x}_i, y_i, z)$. Assume that the set $\{a_i\}$ belongs to class G_1. Then*

$$E\left|E\left(a_i\middle|(1,n)\right) - E\left(a_i\middle|(i-l, i+l)\right)\right| \leq 4c_0^1(a)(1 - \tau^2\beta_1/\sigma)^{l-1},$$

where $i-l$ is replaced by 1 when $i-l<1$ and, similarly, $i+l$ is replaced by n when $i+l>n$.

Proof. For the case $i-l \geq 1$ and $i+l \leq n$, one finds using Lemma 10 and Lemma 11 with $\rho_j = \rho(y_j) = 1-\tau^2/(\sigma\xi(y_j))$ (see Lemma 9) that

$$\begin{aligned}
&\left|E\left(a_i\middle|(1,n)\right) - E\left(a_i\middle|(i-l,i+l)\right)\right| \\
&\quad = \left|\int E\left(a_i\middle|[i-l,i+l]\right)\left\{P\left(d(x_{i-l},x_{i+l})\middle|(1,n)\right)\right.\right. \\
&\qquad\qquad \left.\left. - P\left(d(x_{i-l},x_{i+l})\middle|(i-l,i+l)\right)\right\}\right| \\
&\quad \leq 2a_0(y_i) \max_{A,a,b,u,v} |P(\bar{x}_i \in A|a_{i-l},b_{i+l},y;z) - P(\bar{x}_i \in A|u_{i-l},v_{i+l},y;z)| \\
&\quad \leq 2a_0(y_i) \prod_{j=i-l+1(-i)}^{i+l-1} \rho(y_j). \qquad (5)
\end{aligned}$$

To bound the mean of this we condition on the x–process and use the conditional independence of the y's given the x's,

$$\begin{aligned}
&E\left|E\left(a_i\middle|(1,n)\right) - E\left(a_i\middle|(i-l,i+l)\right)\right| \\
&\quad \leq 2E\Big\{E(a_0(y_i)|x_i) \prod_{j=i-l+1(-i)}^{i+l-1} E(\rho(y_j)|x_j)\Big\} \\
&\quad \leq 2c_0^1(a) \prod_{j=i-l+1(-i)}^{i+l-1} \Big(1-\frac{\tau^2}{\sigma}\beta_1\Big) \\
&\quad = 4c_0^1(a)(1-\tau^2\beta_1/\sigma)^{l-1},
\end{aligned}$$

where we have used Assumption 2 and 5. The two cases $i-k<1$ and $i+k>n$ are treated similarly using one-sided mixing. □

Proof of Theorem 6. Since $\{\psi_i^r\}$ are assumed to be of class G_3 the third absolute moments are uniformly bounded. Furthermore, since $G_3 \subseteq G_1$ we can use Lemma 13 with a_i replaced by $\psi_i^r(\theta_0)$. The central limit theorem extracted from Götze and Hipp (1983) is then applicable. □

6 Uniform convergence of "observed information"

As a final step we prove here Theorem 7. In particular then we work under Condition 2.

To show uniform convergence of $\frac{1}{n}J_n(\theta)$ we need to bound the difference between conditional mean values evaluated under θ and under θ_0.

Lemma 14. *Let b^u be a funtion of $\bar{x}_u$ with $|b^u| \leq 1$. Let $s \leq u-2$ and let $t \geq u+2$. For $|\theta - \theta_0| \leq \delta_0$ we have*

$$|E_\theta(b^u|[s,t]) - E_{\theta_0}(b^u|[s,t])| \leq 2d|\theta - \theta_0| \sum_{i=s+1}^{t} h_0(y_i).$$

Proof. This lemma corresponds to Lemma 5 in Jensen (2005) with sums replaced by integrals. The representation of the conditional density of $\bar{x}_u$ given $[s,t]$ is in our case

$$\frac{\int \prod_{i=s+1}^{t} \omega_i(\theta) \prod_{i=s+1}^{u-2} \mu(dx_i) \prod_{i=u+2}^{t} \mu(dx_i)}{\int \prod_{i=s+1}^{t} \omega_i(\theta) \prod_{i=s+1}^{t} \mu(dx_i)},$$

with $\omega_i(\theta) = p_\theta(x_i|x_{i-1};z_i)p_\theta(y_i|x_i;z_i)$. An interchange of differentiation and integration is possible since the derivative of the integrand is bounded. The details of the proof can be seen in Jensen (2005). □

Lemma 15. *Let b^u be a function of $\bar{x}_u$ with $|b^u| \leq 1$. For $|\theta - \theta_0| \leq \delta_0$ and any integer $l \geq 1$ we have*

$$|E_\theta(b^u|(1,n)) - E_{\theta_0}(b^u|(1,n))| \leq 2d|\theta - \theta_0| \sum_{i=u-l+1}^{u+l} h_0(y_i) + 4 \prod_{j=u-l+1(-u)}^{u+l-1} \rho(y_j).$$

Proof. We can replace $E_\theta(b^u|(1,n))$ by $E_\theta(b^u|[u-l,u+l])$ with an error less than

$$\sup_{x_{u-l},x_{u+l}} E_\theta(b^u|(u-l,u+l),x_{u-l},x_{u+l}) - \inf_{x_{u-l},x_{u+l}} E_\theta(b^u|(u-l,u+l),x_{u-l},x_{u+l}).$$

Combining Lemma 11 and Lemma 10 this gives the bound $2\prod_{j=u-l+1(-u)}^{u+l-1} \rho(y_j)$. We use this for both E_θ and for E_{θ_0}. Finally we use the bound from Lemma 14 for $E_\theta(b^u|[u-l,u+l]) - E_{\theta_0}(b^u|[u-l,u+l])$. □

Lemma 16. *Let the functions $a_i(\theta)$ belong to class $G_{1,1}$ and let $\delta_n \to 0$ for $n \to \infty$. Then*

$$\lim_{n\to\infty} E_{\theta_0} \sup_{|\theta-\theta_0|\leq\delta_n} \Big| \frac{1}{n} \sum_{i=1}^{n} \big\{ E_\theta(a_i(\theta)|(1,n)) - E_{\theta_0}(a_i(\theta_0)|(1,n)) \big\} \Big| = 0$$

Proof. We can replace $E_\theta(a_i(\theta)|(1,n))$ by $E_\theta(a_i(\theta_0)|(1,n))$ with an error bounded by $\delta_n a_1(y_i)$. Next, from Lemma 15, we can replace $E_\theta(a_i(\theta_0)|(1,n))$ with $E_{\theta_0}(a_i(\theta_0)|(1,n))$. Adding together the error terms we need to consider

$$E_{\theta_0}\left\{ \frac{1}{n} \sum_{u=1}^{n} \Big[\delta_n a_1(y_u) + a_0(y_u) \Big(2d\delta_n \sum_{i=u-l+1}^{u+l} h_0(y_i) + 4 \prod_{j=u-l+1(-u)}^{u+l-1} \rho(y_j) \Big) \Big] \right\}.$$

Conditioning first on the hidden process this gives the bound

$$\delta_n c_1(a) + 2d\delta_n\big[c_1^1(a) + 2lc_0^1(a)c_0\big] + 8c_0^1(a)(1-\tau^2\beta_1/\sigma)^{l-1}.$$

If we take $l = \delta_n^{-1/2}$ the last expression tends to zero for $n \to \infty$. □

Lemma 17. *Let the functions $a_i(\theta)$ and $b_j(\theta)$ belong to the class $G_{1,2}$. Then there exist constants q_1, q_2, q_3 such that for any integer $l \geq 1$*

$$E_{\theta_0} \sup_{|\theta-\theta_0|\leq\delta} |V_\theta(a_u(\theta), b_v(\theta)|(1,n)) - V_{\theta_0}(a_u(\theta_0), b_v(\theta_0)|(1,n))|$$
$$\leq d\delta\big[q_1 + q_2(|v-u|+6l)\big] + q_3(1-\tau^2\beta_1/\sigma)^{l-1}.$$

Proof. Let $u \leq v$. The difference between the covariances can be written as the sum of the two terms

$$E_\theta(a_u(\theta)b_v(\theta)|(1,n)) - E_{\theta_0}(a_u(\theta_0)b_v(\theta_0)|(1,n))$$

and

$$E_\theta(a_u(\theta)|(1,n))E_\theta(b_v(\theta)|(1,n)) - E_{\theta_0}(a_u(\theta_0)|(1,n))E_{\theta_0}(b_v(\theta_0)|(1,n))$$
$$= E_\theta(a_u(\theta)|(1,n))\{E_\theta(b_v(\theta)|(1,n)) - E_{\theta_0}(b_v(\theta_0)|(1,n))\}$$
$$+ \{E_\theta(a_u(\theta)|(1,n)) - E_{\theta_0}(a_u(\theta_0)|(1,n))\}E_{\theta_0}(b_v(\theta_0)|(1,n)).$$

For each of these terms we apply Lemma 15. For the first term this gives the bound

$$a_0(y_u)b_0(y_v)\Big\{2d\delta \sum_{i=u-l+1}^{v+l} h_0(y_i) + 4 \prod_{j=u-l+1(-(u:v))}^{v+l-1} \rho(y_j)\Big\}$$

for $|\theta - \theta_0| \leq \delta$. For the second term the bound becomes

$$a_0(y_u)b_0(y_v)\Big\{2d\delta \sum_{i=v-l+1}^{v+l} h_0(y_i) + 4 \prod_{j=v-l+1(-v)}^{v+l-1} \rho(y_j)$$
$$+ 2d\delta \sum_{i=u-l+1}^{u+l} h_0(y_i) + 4 \prod_{j=u-l+1(-u)}^{u+l-1} \rho(y_j)\Big\}.$$

We next bound the mean of the sum of these two terms by first bounding the conditional mean given the hidden process $\{x_i\}$. For the case $u \neq v$ we get the bound in the lemma with

$$q_1 = 4(c_1^1(a)c_0^1(b) + c_1^1(b)c_0^1(a)), \quad q_2 = 2c_0^1(a)c_0^1(b)c_0, \quad q_3 = 24c_0^1(a)c_0^1(b),$$

and for the case $u = v$ we get bound in the lemma with

$$q_1 = 6\sqrt{c_1^2(a)c_1^2(b)}, \quad q_2 = 2\sqrt{c_0^2(a)c_0^2(b)}c_0, \quad q_3 = 24\sqrt{c_0^2(a)c_0^2(b)}.$$

We use here that $\int a_0(y)b_0(y)p_{\theta_0}(y|x;z)\nu(dy)$ is bounded by $\sqrt{c_0^2(a)c_0^2(b)}$ and, similarly, $\int a_0(y)b_0(y)h_0(y)p_{\theta_0}(y|x;z)\nu(dy)$ is bounded by $\sqrt{c_1^2(a)c_1^2(b)}$. □

Lemma 18. *Let the assumptions be as in Lemma 17. Let $\delta_n \to 0$ for $n \to \infty$. Then*

$$\lim_{n\to\infty} E_{\theta_0}\left\{ \sup_{|\theta-\theta_0|\le\delta_n} \Big| \frac{1}{n} \sum_{u,v=1}^{n} \big\{ V_\theta(a_u(\theta), b_v(\theta)|(1,n)) - V_{\theta_0}(a_u(\theta_0), b_v(\theta_0)|(1,n)) \big\} \Big| \right\}$$
$$= 0$$

Proof. The mixing result in Lemma 11 for the hidden process conditioned on the observed process gives (for the case $v > u$)

$$|V_\theta(a_u(\theta), b_v(\theta)|(1,n))| \le 4a_0(y_u)b_0(y_v) \prod_{i=u+2}^{v-2} \rho(y_i),$$

see Ibragimov and Linnik (1971, Theorem 17.2.1). Taking the mean of this, by first evaluating the conditional mean given the hidden process, gives the bound

$$4c_0^1(a)c_0^1(b)(1-\tau^2\beta_1/\sigma)^{|v-u|-3}. \tag{6}$$

Consider now a fixed u and the sum over v of the difference between the two covariances. We split this sum into terms with $|u-v| > l$ and terms with $|u-v| \le l$. For the first set we use the bound in (6) for each covariance, and for the second set we use the bound from Lemma 17. This gives the bound

$$\frac{16c_0^1(a)c_0^1(b)}{\tau^2\beta_1/\sigma}(1-\tau^2\beta_1/\sigma)^{l-3} + d\delta_n\big[(2l+1)q_1 + q_2(l(l+1)+6l(2l+1))\big]$$
$$+ q_3(2l+1)(1-\tau^2\beta_1/\sigma)^{l-1}.$$

Taking $l = \delta_n^{-1/4}$ this bound tends to zero as $\delta_n^{1/2}$ and the lemma has been proved. □

Proof of Theorem 7. The theorem follows directly from Lemma 18. □

Appendix I: Recursions

Let us write the traditional recursive filter for the hidden Markov process in terms of the joint density $p(x_k, y_1^k)$ of the state x_k at time k and the observations $y_1^k = (y_1, y_2, \ldots, y_k)$. We skip the covariates $\{z_i\}$ from the notation here. The recursion takes the form

$$p(x_{k+1}, y_1^{k+1}) = p(y_{k+1}|x_{k+1}) \int p(x_k, y_1^k) p(x_{k+1}|x_k) \mu(dx_k).$$

We next state a similar recursion for the estimating function on the left hand side of (1). Define $a_k(x_k) = E\big(\sum_{i=1}^{k-1} \psi_i | x_k, y_1^k\big)$, where ψ_i is a function of y_i and $\bar{x}_i = (x_{i-1}, x_i, x_{i+1})$. We then have

$$\begin{aligned}
a_{k+1}(x_{k+1}) &= E\Big(\sum_{i=1}^{k-1} \psi_i + \psi_k | x_{k+1}, y_1^{k+1}\Big) \\
&= \int \Big\{ E\Big(\sum_{i=1}^{k-1} \psi_i | x_k, x_{k+1}, y_1^{k+1}\Big) + E(\psi_k | x_k, x_{k+1}, y_1^{k+1}) \Big\} \\
&\qquad \times p(x_k | x_{k+1}, y_1^{k+1}) \mu(dx_k) \\
&= \int \Big\{ a_k(x_k) + \int \psi_k p(x_{k-1} | x_k, x_{k+1}, y_1^{k+1}) \mu(dx_{k-1}) \Big\} \\
&\qquad \times \frac{p(x_k, y_1^k) p(x_{k+1} | x_k) p(y_{k+1} | x_{k+1})}{p(x_{k+1}, y_1^{k+1})} \mu(dx_k) \\
&= \int \Big\{ a_k(x_k) + \int \psi_k \frac{p(x_{k-1}, y_1^{k-1}) p(x_k | x_{k-1}) p(y_k | x_k)}{p(x_k, y_1^k)} \mu(dx_{k-1}) \Big\} \\
&\qquad \times \frac{p(x_k, y_1^k) p(x_{k+1} | x_k) p(y_{k+1} | x_{k+1})}{p(x_{k+1}, y_1^{k+1})} \mu(dx_k) \\
&= \int \Big\{ a_k(x_k) p(x_k, y_1^k) + \int \psi_k p(x_{k-1}, y_1^{k-1}) p(x_k | x_{k-1}) p(y_k | x_k) \mu(dx_{k-1}) \Big\} \\
&\qquad \times \frac{p(x_{k+1} | x_k) p(y_{k+1} | x_{k+1})}{p(x_{k+1}, y_1^{k+1})} \mu(dx_k).
\end{aligned}$$

A similar calculation gives that the estimating function is

$$\begin{aligned}
E\Big(\sum_{i=1}^{n} \psi_i | y_1^n\Big) = \Big\{ \int p(x_n, y_1^n) \mu(dx_n) \Big\}^{-1} \int \Big\{ a_n(x_n) p(x_n, y_1^n) + \iint \psi_n \\
\times p(x_{n-1}, y_1^{n-1}) p(x_n | x_{n-1}) p(y_n | x_n) p(x_{n+1} | x_n) \mu(dx_{n-1}) \mu(dx_{n+1}) \Big\} \mu(dx_n).
\end{aligned}$$

For the special case where ψ_i depends on y_i and (x_{i-1}, x_i) only, we define instead $\tilde{a}_k(x_k) = E\big(\sum_{i=1}^{k} \psi_i | x_k, y_1^k\big)$. The recursion becomes

$$\tilde{a}_{k+1}(x_{k+1}) = \int \big\{ \tilde{a}_k(x_k) + \psi_k \big\} \frac{p(x_k, y_1^k) p(x_{k+1} | x_k) p(y_{k+1} | x_{k+1})}{p(x_{k+1}, y_1^{k+1})} \mu(dx_k).$$

References

Baum, L.E. and Petrie, T.P. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **37**, 1554–1563.

Bickel, P.J., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26**, 1614–1635.

Doob, J.L. (1953). *Stochastic Processes*. Wiley.

Elashoff, M. and Ryan, L. (2004). An em algorithm for estimating equations. *J. Comput. Graph. Statist.*, **13**, 485–465.

Götze, F. and Hipp, C. (1983). Asymptotic expansions for sums of weakly dependent random vectors. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, **64**, 211–240.

Hansen, J.V., Jensen J.L., Friggens, N.C. and Højsgaard, S. (2008). A state space model exhibiting a cyclic structure with an application to progesterone concentration in cow milk.

Heyde, C. and Morton, R. (1996). Quasi-likilhood and generalizing the em algortihm. *J. Roy. Statist. Soc. B*, **58**, 317–327.

Ibragimov, I.A. and Linnik, Yu.V. (1971). *Independent and stationary sequences of random variables*. Wolters-Noordhoff Series of Monographs and Textbooks on Pure and Applied Mathematics.

Jensen, J.L. (2005). Context dependent DNA evolutionary models. Research Report No. 458, Department of Theoretical Statistics, University of Aarhus.

Jensen, J.L. and Petersen, N.V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist*, **27**, 514–535.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **44**, 226–233.

Rosen, O., Jiang, W. and Tanner, M. (2000). Mixtures of marginal models. *Biometrika*, **87**, 391–404.

Department of Genetics and Biotechnology and Department of Mathematical Sciences, University of Aarhus, DK-8000 Aarhus C, Denmark.

E-mail: `jvhansen@imf.au.dk`

Department of Mathematical Sciences, University of Aarhus, DK-8000 Aarhus C, Denmark.

E-mail: `jlj@imf.au.dk`