Sequential Predictions of Menstrual Cycle Lengths

Paola Bortot[1], Bruno Scarpa[2] and Guido Masarotto[3]

[1] University of Bologna, Italy — E–mail: bortot@stat.unibo.it
[2] University of Pavia, Italy — E–mail: bruno.scarpa@unipv.it
[3] University of Padova, Italy — E–mail: guido@sirio.stat.unipd.it

The problem

Motivation

- Forecasting the length of the menstrual cycle and of its phases is a problem of greatest importance in Natural Family Planning and infertility management.
- The development of a simple and effective method for ovulation prediction would increase the efficacy of the rhythm method of contraception.
- The accuracy of some procedures (e.g. postcoital tests) and the success of some therapeutic measures are dependent upon adequate timing of ovulation.

Goals

- The aim of this work is to develop a statistical approach to the problem.
- Consider a particular woman and let y_t be the length of her t-th menstrual cycle ($t = 1, 2, \ldots$).
  Our objective is the estimation of the predictive distribution

$$F_{t+1}(x) = P\{y_{t+1} \leq x | y_1, \ldots, y_t\}$$

  in its entirety, since the non–negligible intra–woman variability means interval forecasts are more appropriate than point predictions (Marshall, 1965).
- By using an analogous approach we aim to forecast other menstrual parameters, e.g. the hypothermic phase in the basal body temperature.

The data

- The data were collected at the Catholic Marriage Advisory Council of England and Wales, a centre which provides counselling and educational service, free of charge and irrespective of race, nationality or religious affiliation.
- The information recorded (Miolo et al. 1993) comes from a sample of 1798 women, each providing a sequence of at least 6 consecutive cycles, leading to a total of 36641 cycles.
- The longest recorded sequence of consecutive cycles for any woman comprises 109 measurements.

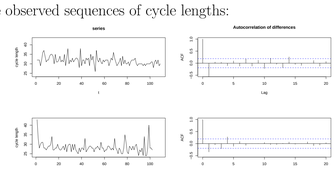
The model

- In looking for a statistical model suitable to describe the observed cycle lengths we have to take into account that there is variability both **within women** and **between women**:
  – a woman's cycle lengths are not constant over time;
  – different women often have different mean cycle lengths and different variabilities around these means.
- The basic ideas to model these two sources of variation can be described as follows:
  1. To explain the intra–woman variation we propose a parametric model, i.e.
  $$P\{y_{t+1} \leq x | y_1, \ldots, y_t, \theta\} = G_{t+1}(x | y_1, \ldots, y_t, \theta)$$
  where G_{t+1} is fully specified and θ is a vector of unknown parameters.
  2. To describe the inter–woman variability we allow θ to vary across women according to a probability distribution $p(\theta|\zeta)$.
  3. Finally, we cast the problem in a Bayesian framework by specifying a prior distribution for ζ.

A model for an individual woman

Empirical evidence

- Plots of y_t against t and the related autocorrelation of first differences highlight some features of the observed sequences of cycle lengths:



  1. Observed cycle lengths are discrete with only a limited number of values. (*Data are recorded in days.*)
  2. A slow downward time trend is generally observed for sequences covering many years. (*The mean cycle length can vary over a long period of time since it depends on the woman's age, see Vollman, 1977.*)
  3. A negative one-lag autocorrelation on first differences is present for many women. (*Women may misunderstand the signals for the end of a cycle, anticipating it, then add to the next cycle the days belonging to the previous one.*)
  4. Some observations are very different from the others and can be considered as outliers. (*Some cycles can have a peculiar biological explanation, such as non detectable early loss...*)
  5. After allowing for the above effects the process remains noisy.

Modelling intra–woman variability

- How to account for the above features within a statistical model?
- We adopt a **state–space formulation**
  1. The observed cycle length y_t can be seen as a discrete measure of the true **unobserved** cycle length z_t, which is a continuous variable:
  $$y_t = [z_t].$$
  2. The true unobserved cycle lengths are generated by the following state-space model:
  $$z_t = m_t + \psi_t + \varepsilon_t,$$
  where m_t, ψ_t and ε_t are defined as follows.
  – Assuming that trajectories are locally constant, we adopt for m_t the random walk model
  $$m_t = m_{t-1} + \eta_t,$$
  where η_t is a sequence of i.i.d. $\mathcal{N}(0, \sigma^2_\eta)$ variables.
  – After allowing for a possible trend, to enable a negative one-lag autocorrelation, we specify for ψ_t the model
  $$\psi_t = \rho\psi_{t-1} + \nu_t,$$
  where ν_t is a sequence of i.i.d. $\mathcal{N}(0, \sigma^2_\nu)$ variables.
  – Accounting for the possible contamination of a small fraction of outliers, we assume ε_t to be a sequence of i.i.d. random variables following a Normal mixture distribution with components $\mathcal{N}(0, \sigma^2_\varepsilon)$ and $\mathcal{N}(0, \alpha\sigma^2_\varepsilon)$, and mixing proportions $(1 - \pi)$ and π, respectively.

The intra-woman variability model

- The complete model is

$$\begin{aligned} y_t &= [z_t] \\ z_t &= m_t + \psi_t + \varepsilon_t \\ m_t &= m_{t-1} + \eta_t \\ \psi_t &= \rho\psi_{t-1} + \nu_t \end{aligned}$$

  with

$$\begin{aligned} \varepsilon_t &\sim (1-\pi)\mathcal{N}(0, \sigma^2_\varepsilon) + \pi\mathcal{N}(0, \alpha\sigma^2_\varepsilon) \text{ i.i.d.}, \\ \eta_t &\sim \mathcal{N}(0, \sigma^2_\eta) \text{ i.i.d.}, \qquad \nu_t \sim \mathcal{N}(0, \sigma^2_\nu) \text{ i.i.d.} \end{aligned}$$

Specification of prior distributions

- The vector of model parameters is $\theta = (\sigma^2_\varepsilon, \pi, \alpha, \sigma^2_\eta, \sigma^2_\nu, \rho)$. We assume a priori that the components of θ are mutually independent with the following distributions

$$\begin{aligned} \pi &\sim \text{Beta}(1, 10), \\ \alpha^{-1} &\sim \text{Gamma}(5, 10), \\ \sigma^{-2}_\varepsilon &\sim \text{Gamma}(0.1^3, 0.1^3), \\ \sigma^{-2}_\eta &\sim \text{Gamma}(0.1^3, 0.1^3), \\ \sigma^{-2}_\nu &\sim \text{Gamma}(0.1^3, 0.1^3), \\ \rho &\sim \mathcal{N}(0, 100). \end{aligned}$$

- Consequently, flat priors are specified for σ^{-2}_ε, σ^{-2}_η and ρ. For π and α^{-1} the prior distributions are such that the largest mass lies on $(0, 0.5)$ and $(0, 1)$, respectively, thus guaranteeing identifiability.

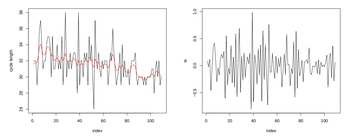
Inference

- Inference is made via MCMC.
- We augmented the parameter space to include the unobserved cycle lengths z_t and the latent Bernoulli variables identifying outliers. Conditional on this augmented vector of unknown parameters, the proposed model can be reduced to a linear Gaussian state–space model.
- The multi–move Gibbs Sampler algorithm of Shephard (1994) and Carter and Kohn (1996) can then be applied to draw inference.
- Given the relatively large number of unknown variances, to simplify estimation we imposed the constraint
$$\sigma_\nu / \sigma_\varepsilon = 0.5.$$

Results for an individual woman

Woman I

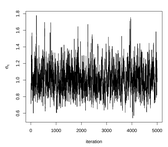
- We show results obtained by fitting the model to the longest observed sequence of consecutive cycle lengths from an individual woman (109 observations).
- Estimated m_t process superimposed on the plot of the observed cycle lengths and estimated ψ_t process:



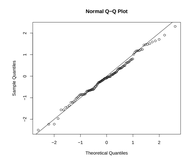
- Summary of model parameter estimates:

	σ_η	σ_ε	ρ	π	α
median	1.01	1.50	-0.27	0.07	3.24
95% credibility interval	(0.72, 1.36)	(0.98, 1.92)	(-0.75, 0.71)	(0.01, 0.24)	(1.2, 11.0)

- Good performance of the MCMC algorithm. As an example, we show the MCMC output (5000 iterations) for σ_η:

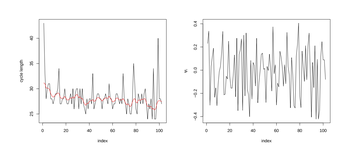


- The model assumptions seem valid. Below we show the Normal qqplot for the model residuals:



Woman II

- We also show inference results for a sequence of consecutive cycle lengths from a second woman (102 observations).
- Estimated m_t process superimposed on the plot of the observed cycle lengths and estimated ψ_t process:



	σ_η	σ_ε	ρ	π	α
median	0.98	1.40	0.14	0.10	13.4
95% credibility interval	(0.58, 1.42)	(0.98, 1.88)	(-0.66, 0.90)	(0.03, 0.21)	(5.5, 30.1)

- Main differences between the two women occur in the estimates of ρ and α. However, in the broader cohort we observed differences also in the estimates of σ_ε.

A model for all women

- Most women have relatively short sequences of cycle lengths, which precludes the possibility of individual estimation.
- A possible solution is a hierarchical model that allows a transfer of information across women and the estimation of population parameters for prediction on women not included in the survey.
- The formulation of the hierarchical model follows directly from the intra–woman model, by allowing θ to vary across women according to a probability distribution $p(\theta|\zeta)$.
- The analysis of some sequences suggests that a simplification is possible: the biggest variation across women is observed in the estimates of ρ (autocorrelation coefficient), α (ratio between outlier and non–outlier variances) and σ_ε (the residual error variance).
- Thus, we allow random effects only on ρ, α and σ_ε, treating the other parameters as fixed across women.
- We assume

$$\rho \sim \mathcal{N}(\mu, \sigma^2), \;\; \alpha^{-1} \sim \text{Gamma}(a_1, b_1) \text{ and } \sigma^{-2}_\varepsilon \sim \text{Gamma}(a_2, b_2)$$

  with prior specification

$$\begin{aligned} \mu &\sim \mathcal{N}(0, 100), \qquad \sigma^{-2} \sim \text{Gamma}(0.1^3, 0.1^3) \\ a_1 &\sim \text{Gamma}(0.1^3, 0.1^3), \qquad b_1 \sim \text{Gamma}(0.1^3, 0.1^3) \\ a_2 &\sim \text{Gamma}(0.1^3, 0.1^3), \qquad b_2 \sim \text{Gamma}(0.1^3, 0.1^3) \end{aligned}$$

- Inference is carried out by an extension of Gibbs sampling with block state update (Shephard, 1994; Carter and Kohn, 1996)

Results

- Not available yet! Wait!

Future work

Model estimation and applications

- Estimation of the hierarchical model from the total cycle lengths.
- Estimation of the hierarchical model from the pre– and post–ovular phase lengths.
- Application to prediction of probability of conception under given intercourse behaviour.

Model extension

- Inclusion of covariates (woman's age, temperature, mucus types, ...).

Inference extensions

- Parameter estimation using Monte Carlo sequential procedures, in order to develop a dynamic algorithm for prediction.

References

- Carter C.K., Kohn R. (1996) Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika*, **83**, 589–601.
- Marshall J. (1965) Predicting length of the menstrual cycle. *Lancet*, **30**, 263–265.
- Miolo L., Colombo B., Marshall J. (1993) A data base for biometric research on changes in basal body temperature in the menstrual cycle. *Statistica*, **LII**, 563–572.
- Shephard N. (1994) Partial Non-Gaussian State Space. *Biometrika*, **81**, 115–131.
- Vollman R.F. (1977) *The menstrual Cycle*, Friedman, London.