# Basal body temperature curves and fitting with periodic smoothing splines

Ingrid Odlén
ingrid.odlen@fysik.lth.se

February 26, 2019

Master's thesis work carried out at
the Centre for Mathematical Sciences, Lund University.

Supervisors: Sara Maad Sasane, sara.maad_sasane@math.lth.se

Examiner: Niels Christian Overgaard,
niels_christian.overgaard@math.lth.se

**Abstract**

Fertility tracking phone apps are widely used by women today, in order to both achieve or prevent conception. Due to the health risks of hormone based contraceptions, many women opt for non-invasive methods, such as fertility tracking. Most of the algorithms for tracking fertility use parameters such as the basal body temperature (BBT), cycle length and days of menstruation.
The widely used "coverline" method is based on daily charting of the BBT in order to identify a sudden temperature increase following ovulation. This method can only find ovulation after it has occurred and has been used as a contraception method for centuries. Our goal is to use curve fitting of the basal body temperature in order to model fertility during the whole cycle, and hence predict future ovulation. The method uses cubic periodic smoothing splines on a data set consisting of daily temperatures from 1151 women. The data is anonymized and was provided by the fertility tracking company Natural Cycles.
We optimize the parameters of the smoothing spline algorithm through cross validation and statistical analysis with the given data. We then correlate the periodicity of the fitted curve with the day of ovulation, through the use of data containing results from ovulation tests. We finally use the splines to model future cycles and predict ovulation. We conclude that this method could be used for increasing accuracy in existing fertility tracking algorithms.

# Acknowledgements

I would like to thank Sara Maad Sasane for supervising the project. I would also like to thank Natural Cycles for providing anonymized data as well as supervising, by Jonathan Bull, Data Scientist, Natural Cycles.

# Contents

# Chapter 1
# Introduction

There are numerous methods that can be used for studying female fertility in order to increase the chances for or avoid conception. These methods include looking at hormonal levels, cervical mucus or basal body temperature.

In this project we focus on the use of basal body temperature (BBT) to study female fertility. This is done through curve fitting the BBT curve by the use of cubic periodic smoothing splines, in order to estimate the day of ovulation. This method is particularly suitable due to its noise cancelling capabilities. We also test the method's capabilities for predicting future ovulation.

In order to do this we use anonymized data of a total of 1151 persons, provided by the fertility tracking company Natural Cycles.

We start by describing the chosen curve fitting method for the purpose, which is the cubic periodic smoothing splines. We then briefly explain the theory regarding fertility that is needed to follow and understand the report. Natural Cycles has also provided a brief description of their fertility tracking algorithm that their application is based on. We then describe our method, divided in five main parts:

1. Curve fitting of the basal body temperature curve through the use of cubic periodic smoothing splines,
2. Estimating the values of the smoothing spline constants by the use of cross validation and data,
3. Estimating the day of ovulation from the curve fit based on temperature data for a specific person,
4. Predicting the day of ovulation from the curve fit based on previous temperature data for a specific person,

5. Statistical study of the algorithm.

We end by discussing the efficiency of the chosen method, its possible uses and areas for improvement. We want to emphasize that our method is not built on Natural Cycles's algorithm.

# Chapter 2
# Curve fitting

When given data points, curve fitting is frequently used to get an understanding of the data. There are numerous different methods to be used but most of them require some knowledge about an underlying mathematical model. Curve fitting can be done either by interpolating (when the curve passes through the points) or by regression (when the curve comes close to the points).

## 2.1 Least Squares

One of the simplest ways to fit a curve is through the linear least squares method, e.g. by finding the coefficients $k$ and $m$ such that

$$F(k,m) := \sum_{i=1}^{n} (y_i - kx_i - m)^2 \tag{2.1}$$

is minimized. The data points are given by $(x_i, y_i)$ where $i = 1, ..., n$. Differentiating $F(k,m)$ with respect to $k$ and $m$ respectively gives

$$\begin{cases} \dfrac{\partial F(k,m)}{\partial k} = \displaystyle\sum_{i=1}^{n} 2(y_i - kx_i - m) \cdot (-x_i) \\ \dfrac{\partial F(k,m)}{\partial m} = \displaystyle\sum_{i=1}^{n} 2(y_i - kx_i - m) \cdot (-1). \end{cases} \tag{2.2}$$

Since $F$ is a convex function it is minimized when $\dfrac{\partial F(k,m)}{\partial m}$ and $\dfrac{\partial F(k,m)}{\partial m}$ are equal to zero, which gives

$$\begin{cases} -\sum_{i=1}^{n} x_i y_i + k \sum_{i=1}^{n} x_i^2 + m \sum_{i=1}^{n} x_i = 0 \\ -\sum_{i=1}^{n} y_i + k \sum_{i=1}^{n} x_i + m \cdot n = 0. \end{cases} \tag{2.3}$$

We rewrite (2.3) in matrix form:

$$\begin{bmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{bmatrix} \cdot \begin{bmatrix} k \\ m \end{bmatrix} = \begin{bmatrix} \overline{xy} \\ \overline{y} \end{bmatrix} \tag{2.4}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$, $\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$ and $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
The coefficients $k$ and $m$ are therefore given by

$$\begin{bmatrix} k \\ m \end{bmatrix} = \begin{bmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \overline{xy} \\ \overline{y} \end{bmatrix}, \tag{2.5}$$

which is solvable if $\overline{x^2} - \overline{x}^2 \neq 0$.

## 2.2 Smoothing splines

A spline is a function of piecewise polynomials ("Smooth Polynomial Lines Interpolating Numerical Estimates"). Its name originates from a drawing tool used in ship building, as seen in Figure 2.1. The spline S is defined piecewise for $k$ subintervals.

The description of smoothing splines that we will cover in this section is partially based on the work of Massimo Zanetti [11] as well as the article "Surface fitting with boundary data" [8].

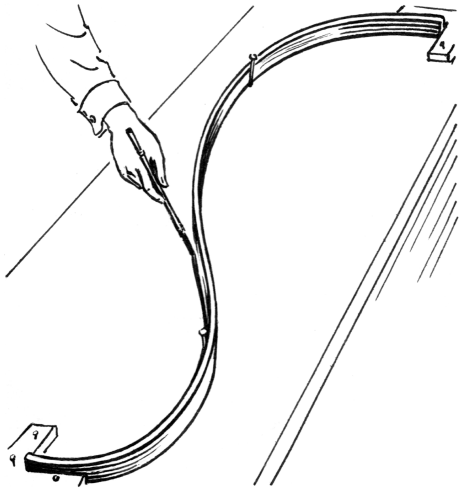

**Figure 2.1:** A wooden spline. [Pearson Scott Foresman (Public domain), via Wikimedia Commons]

When there are uncertainties, noise or error in the data points, it is useful to use a method that doesn't force the fitted curve to go through data points. Instead, we will use a method that tries to minimize the "energy" of the curve, while coming close to the data points. Let $(x_i, y_i)$ be the data points, $i = 1, \ldots, n$, and $a = x_1 < x_2, ... < x_n < b$. We want to find a function $y(x)$ that minimizes

$$p \sum_{i=1}^{n} (y_i - y(x_i))^2 + (1-p) \int_a^b y''(x)^2 \, dx \tag{2.6}$$

among all $y \in X$, where

$$X := \{ \, y \in C^2([a,b]) \mid y(a) = y(b), \;\; y'(a) = y'(b), \;\; y''(a) = y''(b) \, \}. \tag{2.7}$$

The parameter $p \in (0,1)$ is the interpolating parameter. Note that when $p$ is close to 1 the spline function will almost go through the data points and when $p$ is close to 0 the function will almost be a straight line.

Note that the functions $y \in X$ can be extended to a $C^2$ periodic function on $\mathbb{R}$. We will now introduce the arbitrary fixed function $v \in X$.

We assume that (2.6) is minimal when $y = y_*$, and study the expression when $y = y_* + \epsilon v$ for $\epsilon \in \mathbb{R}$.
When studying (2.6) for an abitrary $\epsilon$ we get

$$F(\epsilon) := I(y_* + \epsilon h) = p \sum_{i=1}^{n} (y_i - y_*(x_i) - \epsilon(x_i))^2 + (1-p) \int_a^b (y_*''(x) + \epsilon v''(x))^2 dx, \tag{2.8}$$

which is a real-valued function of one variable. The minimum for $F(\epsilon)$ occurs when $\epsilon = 0$, and hence $F'(0) = 0$, which gives

$$2p \sum_{i=0}^{n} (y_i - y_*(x_i)) \, v(x_i) + 2(1-p) \int_a^b y_*''(x) v''(x) dx = 0 \tag{2.9}$$

for every $v \in X$. Since $v$ is an arbitrary function, we can choose it to be 0 on the whole interval $[a, b]$, except on the $[x_i, x_{i+1}]$, for a fixed $i$. For such $v$ we have

$$v(x_i) = v'(x_i) = v(x_{i+1}) = v'(x_{i+1}) = 0, \tag{2.10}$$

giving

$$\int_{x_i}^{x_{i+1}} y_*''(x) v''(x) dx = 0. \tag{2.11}$$

The lemma of calculus of variations [4] for higher order derivatives states that if a set of continuous functions $f_0, f_1, \ldots, f_n$ on an interval $(c, d)$ satisfies the equality

$$\int_c^d \Big( f_0(x)v(x) + f_1(x)v'(x) + ... + f_n(x)v^{(n)}(x) \Big) \, dx = 0 \tag{2.12}$$

for all $C^\infty$ functions that are compactly supported on $(c, d)$, then there exist continuously differentiable functions $u_0, u_1, ..., u_{n-1}$ on $(c, d)$ such that

$$f_0 = u_0', \;\; f_1 = u_0 + u_1', \;\; f_{n-1} = u_{n-2} + u_{n-1}', \;\; f_n = u_{n-1}. \tag{2.13}$$

Using the lemma for $n = 2$, $f_0 = f_1 = 0$ and $f_2 = y_*''$ on the interval $[x_i, x_{i+1}]$, we obtain

$$u_0' = 0,\ u_1' = -u_0,\ y_*'' = u_1, \tag{2.14}$$

which implies that $y_*''$ is a polynomial of at most degree 1 on $[x_i, x_{i+1}]$. Integrating this $y_*''$ twice, we get a third degree polynomial on $[x_i, x_{i+1}]$, which we denote by

$$y_*(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3. \tag{2.15}$$

Since $y_* \in X$, the function values and its derivatives up to order 2 have to coincide at the spline knots and can be extended to a $C^2$ periodic function on $\mathbb{R}$. This gives the following set of equations

$$\begin{cases} a_{i+1} = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 \\ b_{i+1} = b_i + 2c_i h_i + 3d_i h_i^2 \\ c_{i+1} = c_i + 3d_i h_i \end{cases} \tag{2.16}$$

where $h_i = x_{i+1} - x_i$ and we use the convention that $i + 1 = 1$ if $i = n$. Since we always use a constant step size $h$, we can write $h_i = h$ for all $i$.

The above set of equations can be rewritten as

$$\begin{cases} b_i = \frac{a_{i+1} - a_i}{h} - c_i h - d_i h^2 \\ c_i = \frac{b_{i+1} - b_i}{2h} - \frac{3}{2} d_i h \\ d_i = \frac{c_{i+1} - c_i}{3h}. \end{cases} \tag{2.17}$$

By combining the equations in (2.17) we obtain the system of equations

$$\begin{cases} h b_i = a_{i+1} - a_i - c_i h^2 - d_i h^3 \\ h c_{i-1} + 4c_i + h c_{i+1} = \frac{3}{h}\left(a_i - \frac{a_{i+1}}{2} + a_{i+2}\right) \\ 3h d_i = c_{i+1} - c_i. \end{cases} \tag{2.18}$$

Let

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix},\ \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix},\ \mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix},\ \mathbf{d} = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}. \tag{2.19}$$

We can rewrite (2.17) to

$$\begin{cases} h\mathbf{b} = D\mathbf{a} - h^2\mathbf{c} - h^3\mathbf{d} \\ S\mathbf{c} = 3V\mathbf{a} \\ 3h\mathbf{d} = D\mathbf{c}, \end{cases} \tag{2.20}$$

where

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 1 \\ 0 & -1 & 1 & & & 0 \\ & & & & & \\ \vdots & & & \ddots & & \vdots \\ & & & & & \\ & & & & & 0 \\ 0 & & & & -1 & 1 \\ 1 & 0 & & \dots & 0 & -1 \end{bmatrix}, \tag{2.21}$$

$$S = h \begin{bmatrix} 4 & 1 & 0 & \dots & 0 & 1 \\ 1 & 4 & 1 & & & 0 \\ & & & & & \\ \vdots & & & \ddots & & \vdots \\ & & & & & \\ & & & & & 0 \\ 0 & & & & 1 & 4 & 1 \\ 1 & 0 & & \dots & 0 & 1 & 4 \end{bmatrix} \tag{2.22}$$

and

$$V = \frac{1}{h} \begin{bmatrix} -2 & 1 & 0 & \dots & 0 & 1 \\ 1 & -2 & 1 & & & 0 \\ & & & & & \\ \vdots & & & \ddots & & \vdots \\ & & & & & \\ & & & & & 0 \\ 0 & & & & 1 & -2 & 1 \\ 1 & 0 & & \dots & 0 & 1 & -2 \end{bmatrix}. \tag{2.23}$$

We want to show that $S$ is positive definite, or in other words that it satisfies $\mathbf{x}^T S\mathbf{x} > 0$ for all vectors $\mathbf{x} \neq 0$. We can rewrite $S$ as the sum of the matrices $S_1 + S_2 + ... + S_n$ as follows

$$S = h\cdot \begin{bmatrix} 2 & & & & \\ & 2 & & & \\ & & \ddots & & \\ & & & 2 & \\ & & & & 2 \end{bmatrix} + h\cdot \begin{bmatrix} 1 & 1 & 0 & & & \\ 1 & 1 & 0 & & & \\ 0 & 0 & 0 & & & \\ & & & \ddots & & \\ & & & & 0 & 0 \\ & & & & 0 & 0 \end{bmatrix} + h\cdot \begin{bmatrix} 0 & 0 & 0 & & & \\ 0 & 1 & 1 & & & \\ 0 & 1 & 1 & & & \\ & & & \ddots & & \\ & & & & 0 & 0 \\ & & & & 0 & 0 \end{bmatrix} +$$

$$+ \quad \dots \quad + h \cdot \begin{bmatrix} 0 & 0 & & & & & \\ 0 & 0 & & & & & \\ & & \ddots & & & & \\ & & & 0 & 0 & 0 \\ & & & 0 & 1 & 1 \\ & & & 0 & 1 & 1 \end{bmatrix} + h \cdot \begin{bmatrix} 1 & 0 & & & & & 1 \\ 0 & 0 & & & & & \\ & & \ddots & & & & \\ & & & & & 0 & 0 \\ & & & & & 0 & 1 \\ 1 & & & & & 0 & 1 \end{bmatrix}. \tag{2.24}$$

We see that all matrices are positive semidefinite. This means for a matrix $M$ that $\mathbf{x}^T M\mathbf{x} \geq 0$ for all vectors $\mathbf{x}$. Since the first matrix is diagonal with positive diagonal elements, it is positive definite. Therefore $\mathbf{x}^T(S_1 + S_2 + S_3 + ... + S_n)\mathbf{x} > 0$ for each vector $\mathbf{x} \neq 0$, so the matrix $S$ is therefore positive definite.

Recall that a positive definite matrix is always invertible with a positive definite inverse, and hence $S^{-1}$ and is positive definite.

We can now rewrite (2.20) as

$$\begin{cases} \mathbf{b} = \dfrac{1}{h} D\mathbf{a} - h\mathbf{c} - h^2\mathbf{d} \\ \mathbf{c} = 3S^{-1}V\mathbf{a} \\ \mathbf{d} = \dfrac{1}{3h} D\mathbf{c}. \end{cases} \tag{2.25}$$

Since $\mathbf{b}, \mathbf{c}$ and $\mathbf{d}$ depend on $\mathbf{a}$, also y depends on $\mathbf{a}$, and when substituting (2.15) into (2.6) we obtain

$$G(\mathbf{a}) := F(y) = p \sum_{i=1}^{n} w_i(a_i - y_i)^2 + (1-p) \int_a^b y_*''(x)^2 dx. \tag{2.26}$$

The integral can be rewritten as

$$\int_a^b y_*''(x)^2 dx = \sum_{i=1}^{n-1} \left( \int_0^h (2c_i + 6d_i(x - x_i) \right)^2 dx = \sum_{i=1}^{n} 4(c_i^2 h + 3c_i d_i h^2 + 3d_i^2 h^3). \tag{2.27}$$

Substituting (2.25) into (2.27) gives

$$\int_a^b y_*''(x)^2 dx = \frac{4h}{3} \sum_{i=1}^{n} (c_i^2 + c_i c_{i+1} + c_{i+1}^2) = \frac{2h}{3} \mathbf{c}^T S\mathbf{c} = 6h\mathbf{a}^T V^T (S^{-1}) V\mathbf{a}. \tag{2.28}$$

The function $G(\mathbf{a})$ can now be written as

$$G(\mathbf{a}) = p(\mathbf{a} - \mathbf{y})^T W(\mathbf{a} - \mathbf{y}) + 6(1-p)\mathbf{a}^T V^T S^{-1} V\mathbf{a} = \mathbf{a}^T U\mathbf{a} - \mathbf{v}^T\mathbf{a} + r, \tag{2.29}$$

where $W = diag(w_i)$ and

$$\begin{cases} U = 2pW + 12(1-p)V^T S^{-1} V \\ \mathbf{v} = 2pW\mathbf{y} \\ r = p\mathbf{y}^T W\mathbf{y}. \end{cases} \tag{2.30}$$

The weight matrix $W$ contains only non-negative values and is therefore positive semi-definite. By multiplying $V^T S^{-1} V$ with an arbitrary vector $\mathbf{x}$, we get

$$\mathbf{x}^T V^T S^{-1} V \mathbf{x} = \mathbf{y}^T S^{-1} \mathbf{y}. \tag{2.31}$$

Since $S^{-1}$ is positive definite, $\mathbf{y}^T S^{-1} \mathbf{y} > 0$ for all $\mathbf{y} \neq 0$. Hence

$$\mathbf{x}^T (V^T S^{-1} V) \mathbf{x} \geq 0 \tag{2.32}$$

for all $\mathbf{x} \neq 0$. $U$ is the sum of two positive definite matrices. Since the kernels of the two sums in $U$ don't intersect, $U$ is positive definite, and therefore invertible.

Due to convexity, minimizing $G(\mathbf{a})$ is equivalent to solving

$$G'(\mathbf{a}) = 0, \tag{2.33}$$

and so solving the minimization problem (2.6) is equivalent to finding the unique solution to

$$U\mathbf{a} = \mathbf{v}. \tag{2.34}$$

# Chapter 3
# Cross Validation

We have now found the spline function and will, from this section, write $y$ or $y_p$, when we need to emphasize the dependence of the smoothing parameter $p$, instead of $y_*$.
In order to determine the interpolating parameter $p$, we will use a leave-one-out cross validation method. This method consists of creating $n$ new spline functions $y_p^i$, where the data point $x_i$ has been removed, for $i = 1, \ldots, n$. For each $i$, we calculate the spline function estimate at the point $x_i$ for a given $p$, which is $y_p^i(x_i)$. For each $i$, we introduce $\delta_p^i$ as

$$\delta_p^i = (y_i - y_p^i(x_i))^2 \tag{3.1}$$

which is the square of the vertical distance between the data value $y_i$ and the estimated value $y_p^i(x_i)$.
The suitable interpolating parameter $p$ is then found by minimizing the sum of $\delta_p^i$ times the weight $w_i$, for $i = 1, \ldots, n$, giving the estimator

$$CV(p) = \frac{1}{n} \sum_{i=1}^{n} w_i \delta_p^i. \tag{3.2}$$

See [5] and [10], for more information.

# Chapter 4

# Fertility tracking

Fertility is referred to as the ability to produce a child. A woman is considered as fertile when ovulation has occurred [12]. The length of the fertility window, which denotes the days when sexual intercourse can lead to conception, depends on the lifetime of the egg, of approximately one day, and the survival of the sperm, which is on average three days but can be up to three weeks [1].

## 4.1 The menstrual cycle

A menstrual cycle consists of two stages, the follicular phase and the luteal phase.

The follicular phase is the phase prior to ovulation in which the ovum matures. The length of this phase usually varies between 10 and 16 days [7]. The luteal phase follows ovulation and has, for most women, a relatively constant length of 14 days.

The length of a menstrual cycle is defined by the number of days between the first day of menstruation of one cycle and the first day of menstruation of the upcoming cycle [7]. The menstrual cycle length is on average 28 days, with a wide variation both individually and over time [9]. During the menstrual cycle, hormone levels such as the estrogen, progesterone and the luteinizing hormone levels vary, see Figure 4.1.

The luteinizing hormone, also known as luthropin, drastically increases prior to ovulation and this is known as the "LH-surge". This LH-surge occurs on average one day prior to ovulation [7].

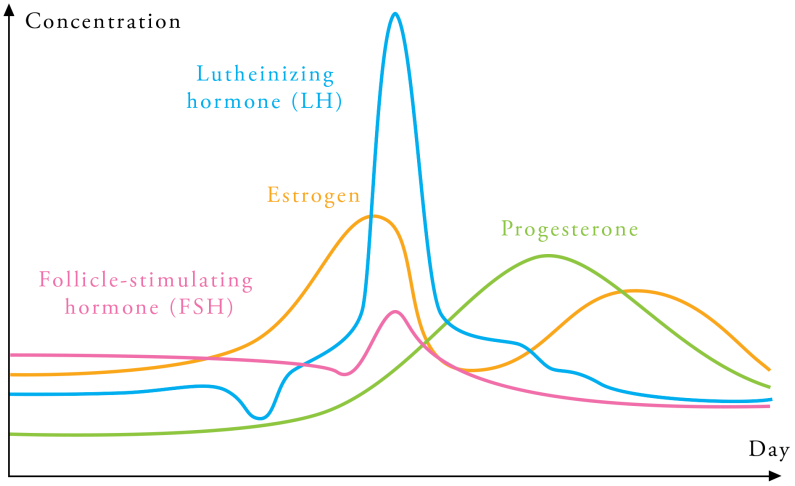


**Figure 4.1:** Concentration of hormone levels during a menstrual cycle

## 4.2 The basal body temperature

There are different factors that are looked at in order to identify a woman's fertility window. This can for example be cervical mucus, hormone levels or basal body temperature. The hormone levels are commonly measured through ovulation tests, where the LH levels are measured. This helps identifying the LH-surge [7].

The basal body temperature is the lowest temperature attained during rest. Traditionally, Natural Family Planning (NFP), is based on identifying a sudden increase in basal body temperature (BBT), usually three consecutive days of temperature above the past six days before the probability of sex leading to conception approaches zero. When using this method for avoiding pregnancy, a couple should abstains from intercourse from the first day of menstruation until the third day after the BBT has risen [2].

This method is the so called "coverline" method.

The basal body temperature increases when the estrogen levels decrease. The BBT increases with the rising progesterone levels, occurring right after ovulation [6].

## 4.3 The Algorithm of Natural Cycles

Natural Cycles uses an algorithm to estimate fertility prior to ovulation based on basal body temperature measurements. It also optionally uses the information provided by LH-tests. It has similarities to the coverline method described in Section 4.2, which requires the average BBT over the last three days to be higher than both the woman's follicular phase average and her cover line (the average temperature of all prior data points from the cycle) and consistent with her luteal phase average. If a positive LH-test is recorded, fewer high temperatures are required since the LH-test gives extra confidence provided that ovulation has occurred. When ovulation is determines to have occured, the algorithm

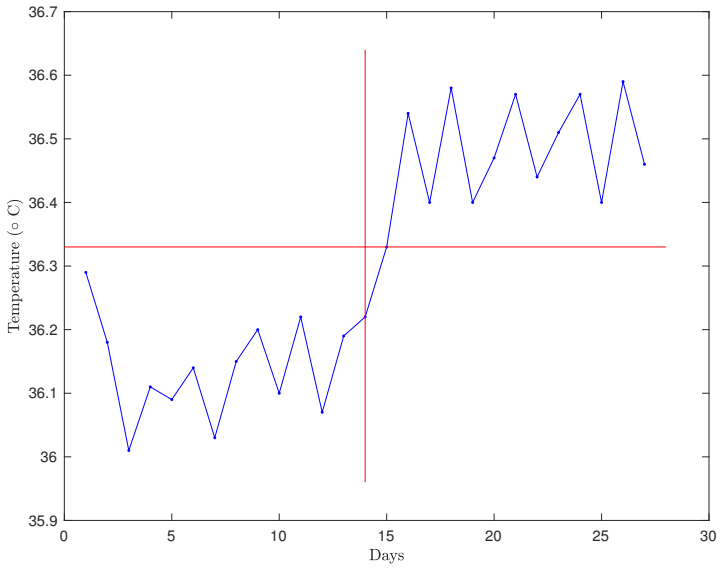


**Figure 4.2:** Coverline method based on basal body temperature measured at the same time for 29 consecutive days of a menstrual cycle

selects the day within the ovulation window with the highest ovulation probability based on comparisons of each temperature to the phase averages and on several heuristic factors. The algorithm is capable of handling missing data points and errors, which are identified through exclusion of anomalous temperatures and performing independent statistical tests [3].

# Chapter 5
# Method

## 5.1 Provided data

The anonymized basal body temperature data, provided by Natural Cycles, is described in the following table. Since the smoothing spline algorithm uses data from previous cycles, samples containing only one cycle are excluded from the data set. This leaves us with data from 1151 persons, each assigned a sample number $k$, and a total of 3866 menstrual cycles. When building the algorithm and choosing the optimal parameters, half of the data set was put aside, leaving us with data from 575 persons and a total of 1970 cycles.

| Notation | Description |
|---|---|
| $ncy_k$ | Number of cycles for the sample number $k$. |
| $x_{i,j,k}$ | Day $i$ of cycle number $j$ and sample number $k$. |
| $y_{i,j,k}$ | Temperature for day $i$ of cycle number $j$ and sample number $k$. |
| $w_{i,j,k}$ | Weight for day $i$ of cycle number $j$ and sample number $k$. |
| $a_{j,k}$ | Cycle length for cycle number $j$ and sample number $k$. |
| $z_{NC\,j,k}$ | NC estimation of ovulation day for cycle number $j$ and sample number $k$. |
| $z_{LH\,j,k}$ | Day of first positive LH-test for cycle number $j$ and sample number $k$. |
| $g_{i,j,k}$ | Fertility color code for day $i$, cycle number $j$ and sample number $k$ . |

The weight is $w_{i,j,k} = 1$ for days with correct temperature measurement and $w_{i,j,k} = 0$ for abnormal or diverging measurements. The fertility color coding is either $g_{i,j,k} = 1$ (fertile), $g_{i,j,k} = 2$ (close to fertile) or $g_{i,j,k} = 3$ (infertile).

The provided data consists of samples with two or more cycles with no missing data points and a weight of 1 for each day of the cycles.

## 5.2 Modeling the basal body temperature curve

The shape of four basal body temperature curves for two consecutive menstrual cycles can be seen in Figure 5.1, and for three consecutive cycles in 5.2. In order to find a pattern in the menstrual cycle for each person, we build an algorithm using the cubic spline method described in Section 2.2. The algorithm finds the optimal spline function $y$ using the whole given set of menstrual cycles as the period length for our smoothing spline.

Figure 5.3 shows the cubic spline smoothing with a high interpolating value $p$ ($p = 0.9$) for four different basal body temperature curves, containing two consecutive cycles. Similarly, Figure 5.4 shows the curve fitting for BBT curves with three consecutive cycles. We will instead choose a lower value and get a curve which resembles a sine function, see Figures 5.5 and 5.6.

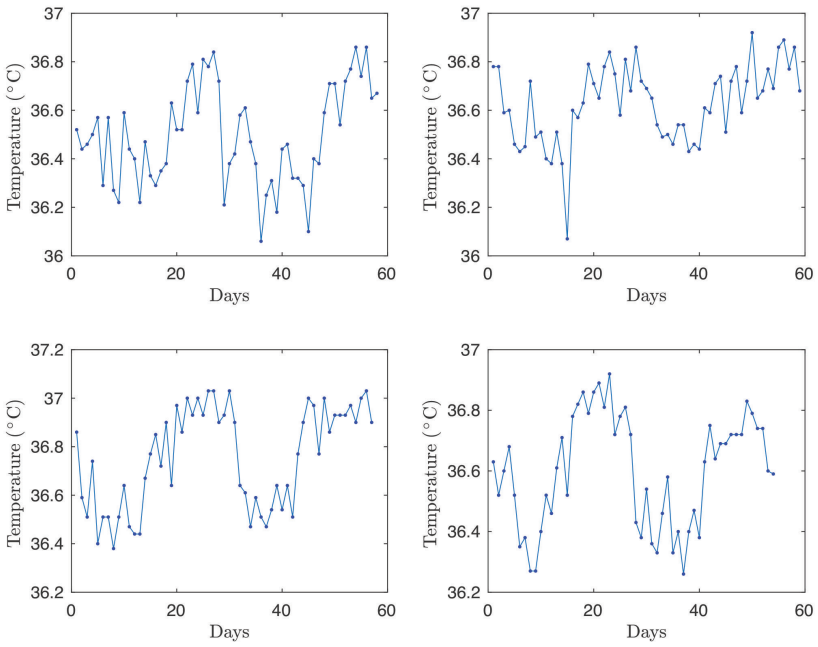


**Figure 5.1:** The basal body temperature for four different persons with two menstrual cycles each.

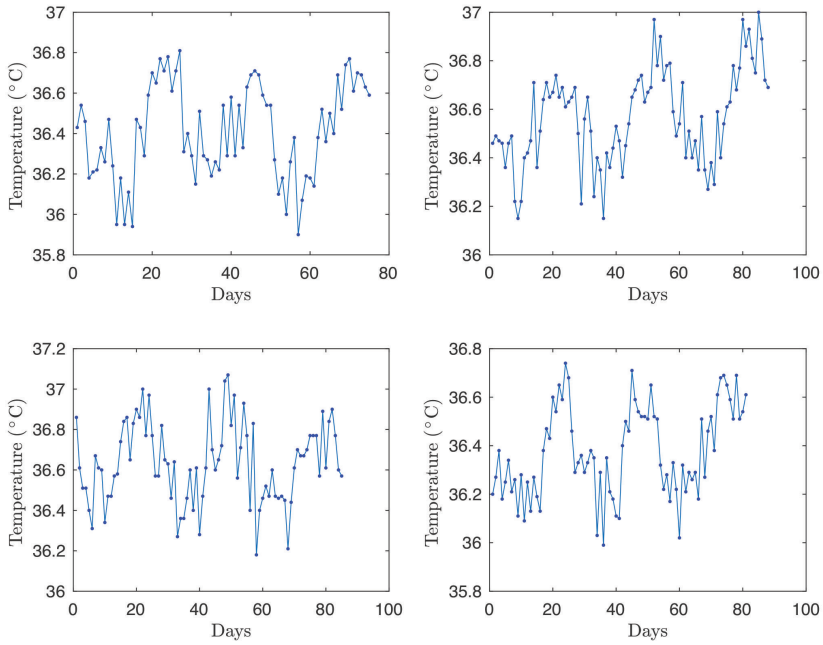


**Figure 5.2:** The basal body temperature for four different persons with three menstrual cycles each.

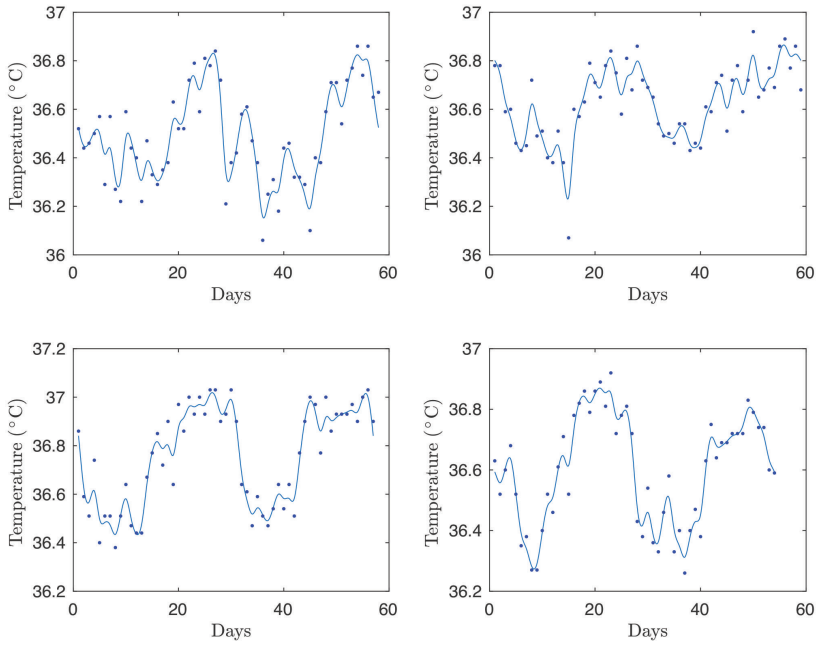


**Figure 5.3:** The basal body temperature for four different persons with two menstrual cycles each and the cubic periodic smoothing splines curve with the interpolating parameter $p = 0.9$

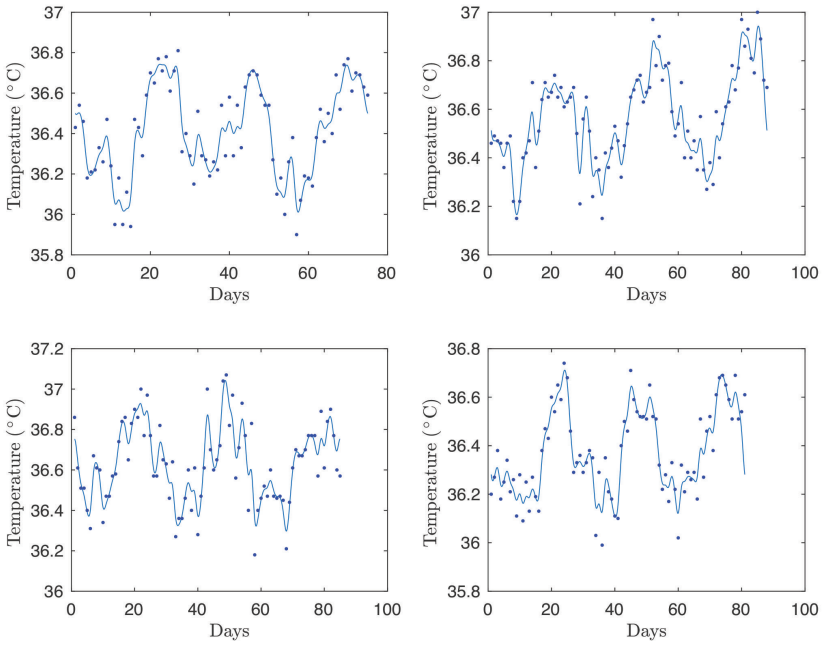


**Figure 5.4:** The basal body temperature for four different persons with three menstrual cycles each and the cubic periodic smoothing splines curve with the interpolating parameter $p = 0.9$

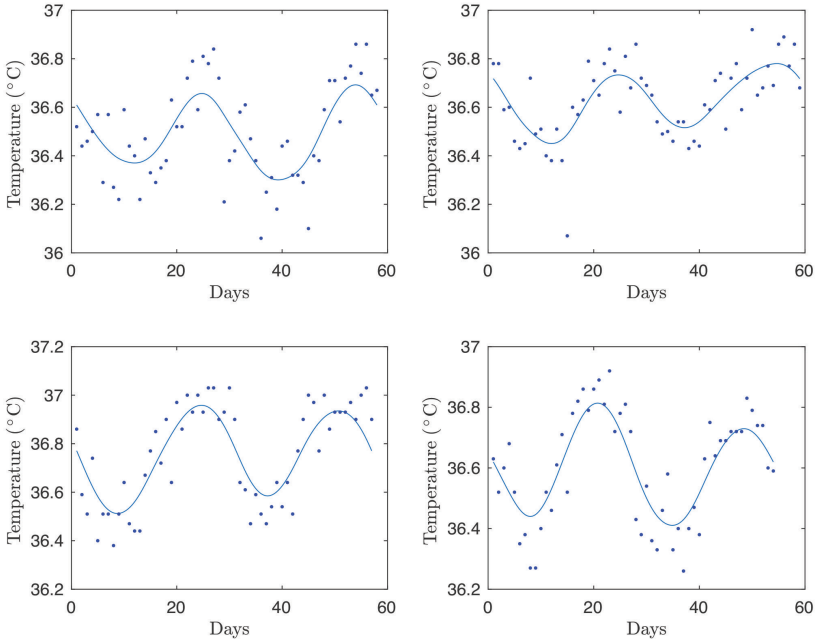


**Figure 5.5:** The basal body temperature for four different persons with two menstrual cycles each and the cubic periodic smoothing splines curve with the interpolating parameter $p = 0.01$

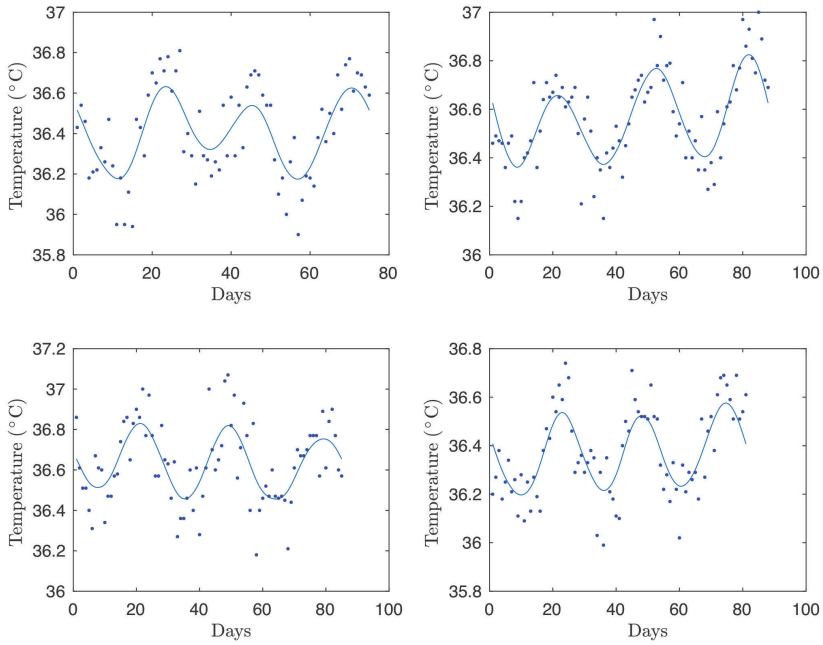


**Figure 5.6:** The basal body temperature for four different persons with three menstrual cycles each and the cubic periodic smoothing splines curve with the interpolating parameter $p = 0.01$

The input data of the smoothing spline algorithm is

| Notation | Description |
|---|---|
| $[x_{i,1,k}, ..., x_{i,ncy,k}]$ | Vector containing the days $i$ of cycle number $j$ and sample number $k$. |
| $[y_{i,1,k}, ..., y_{i,ncy,k}]$ | Temperature vector for day $i$ of cycle number $j$ and sample number $k$. |
| $[w_{i,1,k}, ..., wi, ncy, k]$ | Weight vector for day $i$ of cycle number $j$ and sample number $k$. |
| $[a_{1,k}, ..., a_{ncy,k}]$ | Cycle length vector for cycle number $j$ and sample number $k$. |
| $p$ | interpolating parameter |

## 5.3 Choosing the parameter $p$

We use the cross validation method as described in Section 3, to determine the interpolating parameter $p$.

Another method we use for determining $p$ is based on using the smoothing spline function for a period consisting of a whole set of $i$ values and $k$ cycles and a low interpolating parameter $p$ which gives a sine like curve. The interpolating parameter is determined by minimizing the difference between the estimated day of ovulation and the day given by the LH data (see Section 5.4).

## 5.4 Ovulation Estimation

The smoothing spline algorithm with a low parameter $p$ gives a sine like curve with a certain periodicity, as seen in Section 5.2. Identifying this periodicity and its relation to the day of ovulation could give a good estimation of the correlation between BBT data and ovulation. The estimation algorithm for the day of ovulation for the $j$ cycles of a sample $k$ with known temperature $y_{i,j,k}$ for each given day $x_{ij,k}$ is using the periodic cubic smoothing splines as described in Section 2.2.

As seen in Chapter 4, ovulation is followed by a significant temperature rise, which occurs between the minimum and the maximum of the sine curve shape of the smoothing function. We determine the formula for the estimated day of ovulation $z$ by calculating the constants $c$ and $d$ in

$$z = x_{min} + c \cdot (x_{max} - x_{min}) + d \tag{5.1}$$

where $x_{min}$ and $x_{max}$ are the days for the minimum and the maximum of the smoothing spline function, respectively, see Figure 5.7. We choose $d$ so that the mean of $z$ for the sample coincides with the mean of $z_{LH}+1$. The "+1"-term is due to the LH-surge occurring on average one day before ovulation, as described in Section 4.1. We choose $c$ so that the standard deviation of the difference $z - z_{LH} - 1$ is as small as possible, for half of the given data set. We then use (5.1) to estimate the day of ovulation for each cycle $j$ of a given sample $k$, giving the output function $[z_{1,j}, ..., z_{ncy,j}]$.

## 5.5 Ovulation prediction

In the previous subsection the cubic periodic smoothing spline function is used to estimate the day of ovulation when the temperature data for the whole sample set is known. In order to use this method to instead predict ovulation, as one would like to do it in a fertility app, we need to build the spline function for the remaining days of the last cycle, where the temperature is not yet known. In order to do this with the provided data, only the temperatures up to $y_{n_{days},j,k}$ are used, leaving us with a set $\{x_{i,ncy,k}, y_{i,ncy,k}\}$ for the last cycle, where $i \in 1, \ldots, n_{days}$.

The prediction algorithm will use the following input data

| Notation | Description |
|---|---|
| $[x_{i,1,k}, ..., x_{i,ncy-1,k}]$ | Vector containing the days $i$ of cycle number $j$ and sample number $k$. |
| $[y_{i,1,k}, ..., y_{i,ncy-1,k}]$ | Temperature vector for day $i$ of cycle number $j$ and sample number $k$. |
| $[w_{i,1,k}, ..., w_{i,ncy-1,k}]$ | Weight vector for day $i$ of cycle number $j$ and sample number $k$. |
| $[a_{1,k}, ..., a_{ncy-1,k}]$ | Cycle length vector for cycle number $j$ and sample number $k$. |
| $p$ | interpolating parameter |

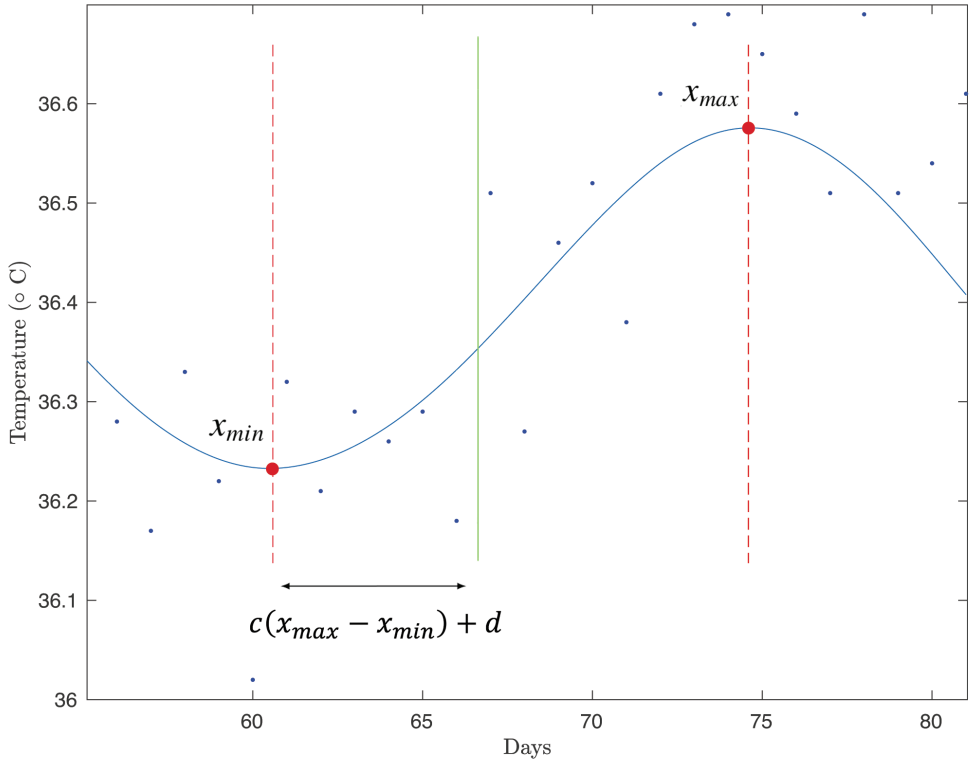


**Figure 5.7:** The basal body temperature data (blue dots), the fitted spline function (blue curve), the day of ovulation ($z_{LH} + 1$) and the parameters $c$ and $d$

The cycle length of the last cycle is estimated by calculating the mean of the previous cycle lengths given sample $k$.

The temperature set for each cycle up to $j = ncy - 1$ is fitted using the smoothing spline algorithm with a high interpolating parameter ($p = 0.99$), building an equidistant $x$-set with a corresponding $y$-set with length $a^*$, used to build an average temperature vector for the unknown days of the last cycle. The built data is described in the following table.

| Notation | Description |
|---|---|
| $y^*_{i,ncy,k}(x_{day})$ | Estimated temperature for the upcoming day $i*$ of the last cycle for sample number $k$ for $x_{day}$ known days of the last cycle. |
| $w^*_{i,ncy,k}(x_{day})$ | Estimated weight for the upcoming day $i*$ of the last cycle for sample number $k$. |
| $a^*_{ncy,k}(x_{day})$ | Estimated cycle length of the last cycle for sample number $k$ for $x_{day}$ known days of the last cycle. |

The ovulation estimation algorithm is finally used on the new data set to predict the day of ovulation of the last cycle. The output data $z_{j,k}$ is the estimation of ovulation for each cycle $j$ for the sample k.

| Notation | Description |
|---|---|
| $y^*_{i,ncy,k}(x_{day})$ | Estimated temperature for the upcoming day $i*$ of the last cycle for sample number $k$ for $x_{day}$ known days of the last cycle. |
| $w^*_{i,ncy,k}(x_{day})$ | Estimated weight for the upcoming day $i*$ of the last cycle for sample number $k$. |
| $a^*_{ncy,k}(x_{day})$ | Estimated cycle length of the last cycle for sample number $k$ for $x_{day}$ known days of the last cycle. |
| $z_k$ | Estimation of the ovulation day for the last cycle for sample number $k$. |

# Chapter 6

# Results and Discussion

## 6.1 Choosing the interpolating parameter

The optimal parameter values for the problem described in Section 5.4 is minimized for $p = 0.003$, $c = 1.1$ and $d = 6$, which are the coefficients we used when predicting the day of ovulation.

The minimum value for the interpolating parameter $p$ found through cross-validation varies from one sample to another, see Figure 6.1. The cross-validation method is therefore not suitable for determining the optimal interpolating parameter $p$ to use on the whole set of samples. Instead we choose to use the interpolating parameter calculated through the method described in Section 2.1. By using a low interpolating parameter, we get a sine like curve which was used for determining the ovulation estimation.

## 6.2 Ovulation Estimation

An example of the ovulation estimation can be seen in Figure 6.2. The graph shows the basal body temperature for all days included in the data set of a given sample together with Natural Cycle's fertility prediction where the green $*$ corresponds to the infertile days and the red $*$ to to the fertile days. Figure 6.2 shows the spline's ability to find an individual periodic pattern based on the basal body temperature. Note that the smoothing spline algorithm does not use the cycle length of each cycle as input data, and does instead use the whole data set length. Despite this, we can clearly see three cycles in the spline function in Figure 6.2.

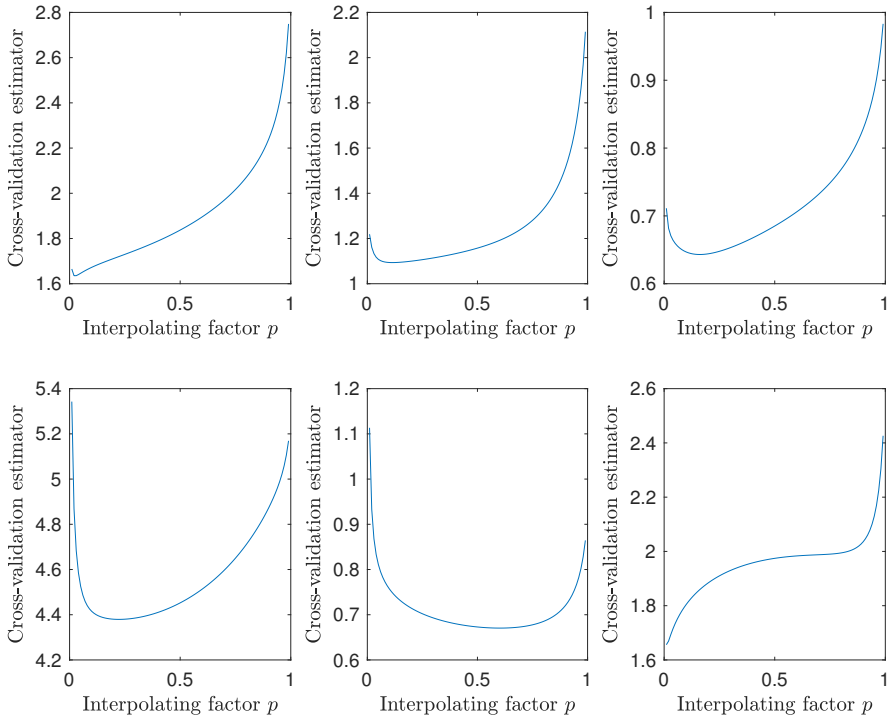


**Figure 6.1:** $CV(p)$ for six randomly chosen samples.

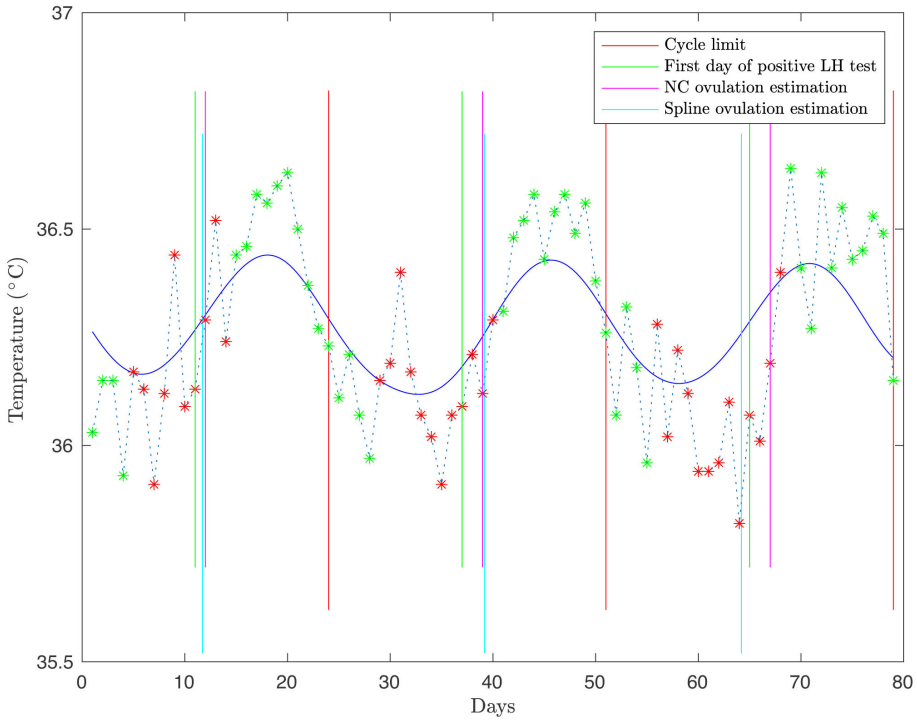


**Figure 6.2:** Ovulation estimation for a three cycle sample.

## 6.3 Ovulation Prediction

In Figure 6.3 we see that the periodic cubic smoothing splines is a suitable method for curve fitting to the basal body temperature and enables the estimation of the fertility window prior to ovulation. The estimation also adapts when given new temperature data. In Figure 6.4 the cycle length of the last cycle differs from the previous one. Since the prediction algorithm estimates the last cycle length as an average of the previous ones, the estimated period of the last cycle strongly differs from the real one. This shows the problem with the method when estimating the cycle length and hence estimating the day of ovulation.

The days shown on the $x$-axis of Figures 6.2 to 6.5 correspond to the real cycle lengths which is why the boundary values seen from the plot lack periodicity. Therefore the method is not suitable in this case.

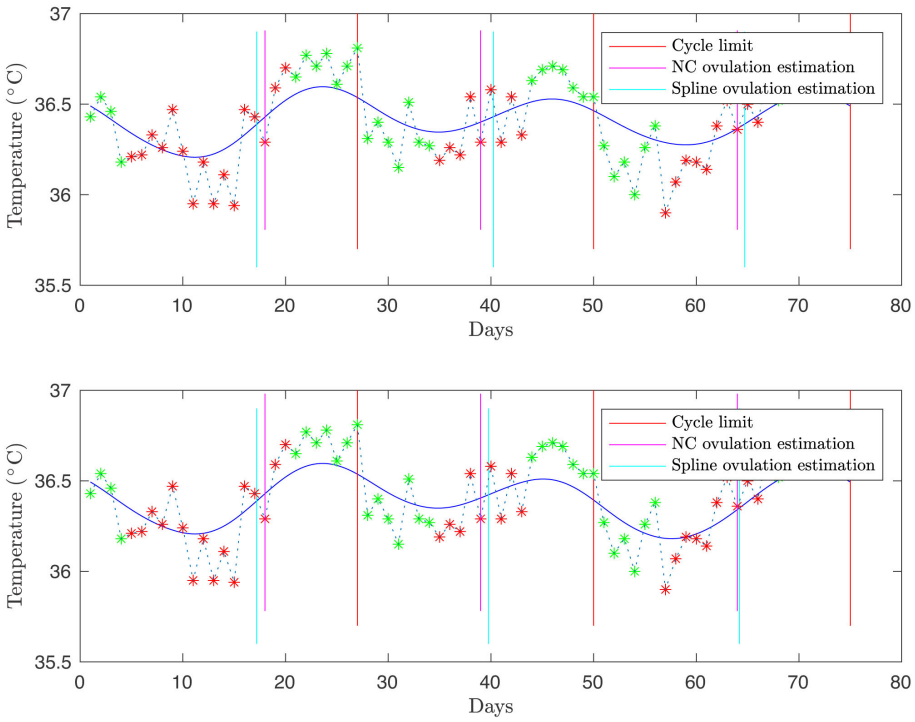


**Figure 6.3:** Ovulation prediction for a two cycle sample with (1) 1 known day of the last cycle (2) 10 known days of the last cycle.

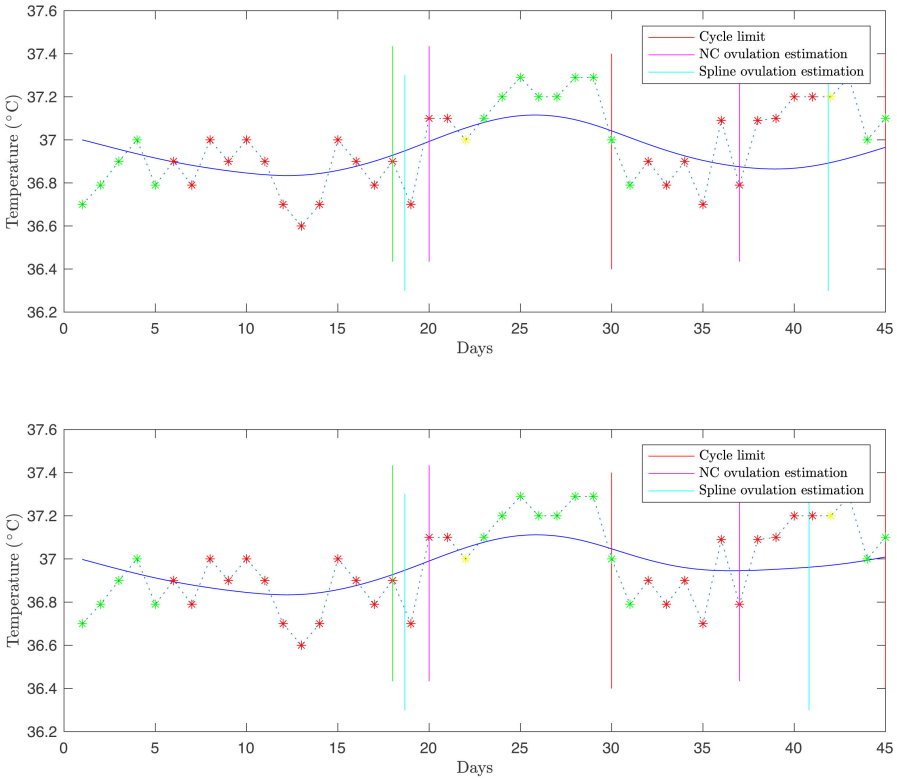


**Figure 6.4:** Ovulation prediction for a two cycle sample with (1) 1 known day of the last cycle (2) 10 known days of the last cycle.

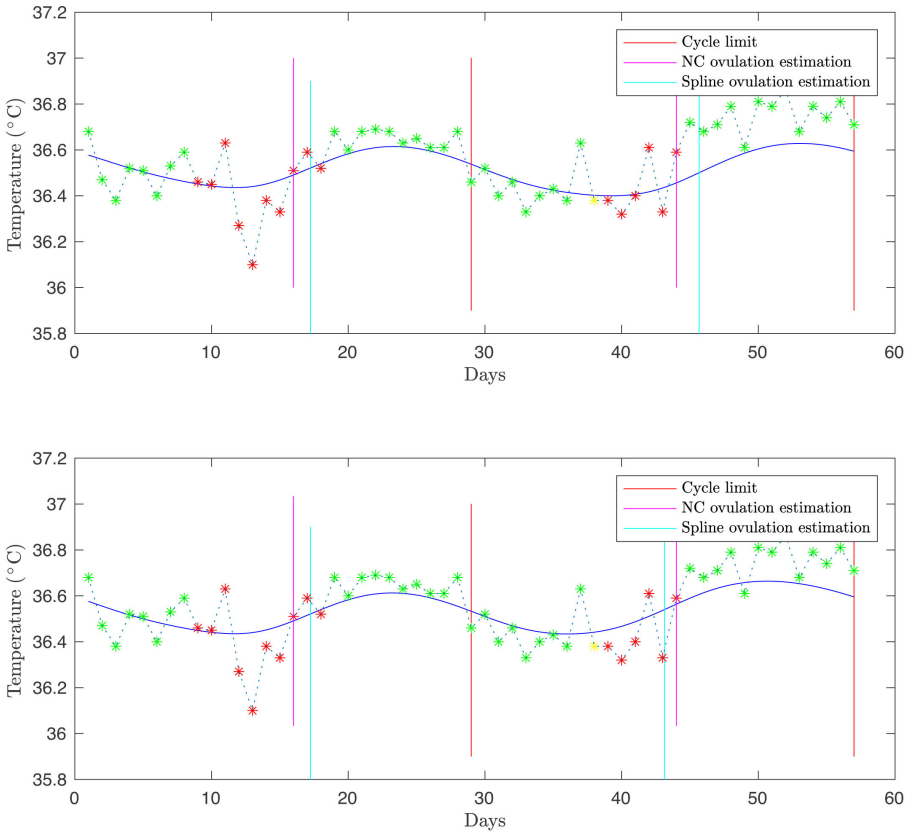


**Figure 6.5:** Ovulation prediction for a two cycle sample with (1) 10 known day of the last cycle (2) 20 known days of the last cycle.

## 6.4 Performance

When looking at the performance of the algorithm we compared out estimations for the whole data set with both the day of ovulation estimated by the algorithm of Natural Cycles and the LH-data. When predicting ovulation when only the temperatures up to the first day of the last cycle are known, our method has a mean difference of 0.6188 with a standard deviation of 2.7335, in comparison to Natural Cycles, where our estimation is slightly prior to theirs. The distribution can be seen in Figure 6.6. When trying to predict ovulation ten days into the last cycle, the mean difference decreases to 0.5908 with a standard deviation of 2.6057, see Figure 6.7.

The difference between our estimation one day into the last cycle and the first day of positive LH-test gives a mean value of 0.8601 with a standard deviation of 2.5698, see Figure 6.8. When the ovulation is estimated ten days into the last cycle, the mean difference is 0.9249 with a standard deviation of 2.1632, see Figure 6.8. Our estimation is on average almost one day after the result from the LH-test. Recall that the first positive LH-test corresponds to the average shift of the LH-surge compared to ovulation, see Section 4.1.

We see that the standard deviation of our prediction algorithm decreases when the

number of known days of the last cycle increases. The comparison of our method with the ovulation tests shows a better accuracy than when compared with the prediction of Natural Cycles. In order to draw any conclusions about the performance of our method compared to other fertility tracking algorithms, we need a greater set of data containing ovulation test results.

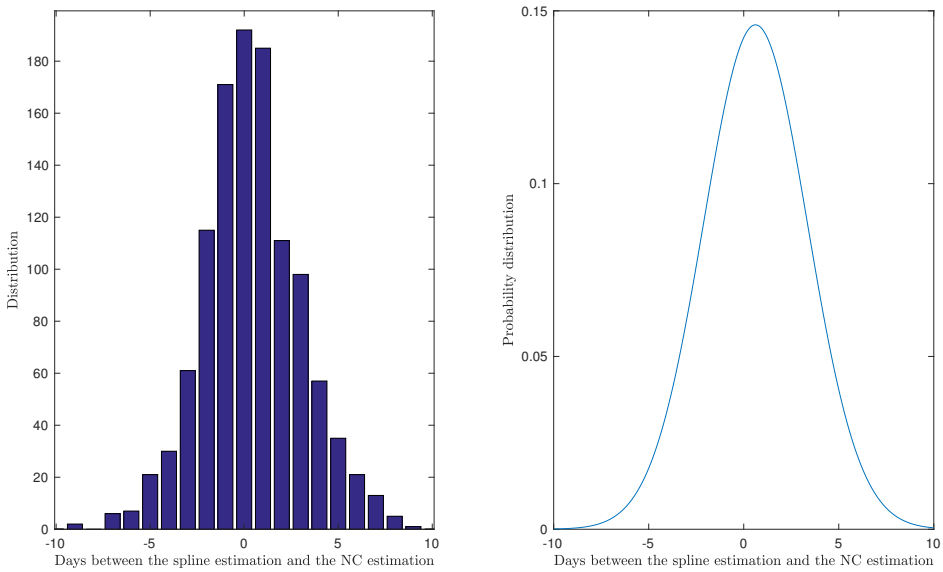


**Figure 6.6:** Distribution and probability distribution of the difference between Natural Cycle's estimation of the ovulation day and the smoothing spline based algorithm of the ovulation day 1 day into the last cycle.

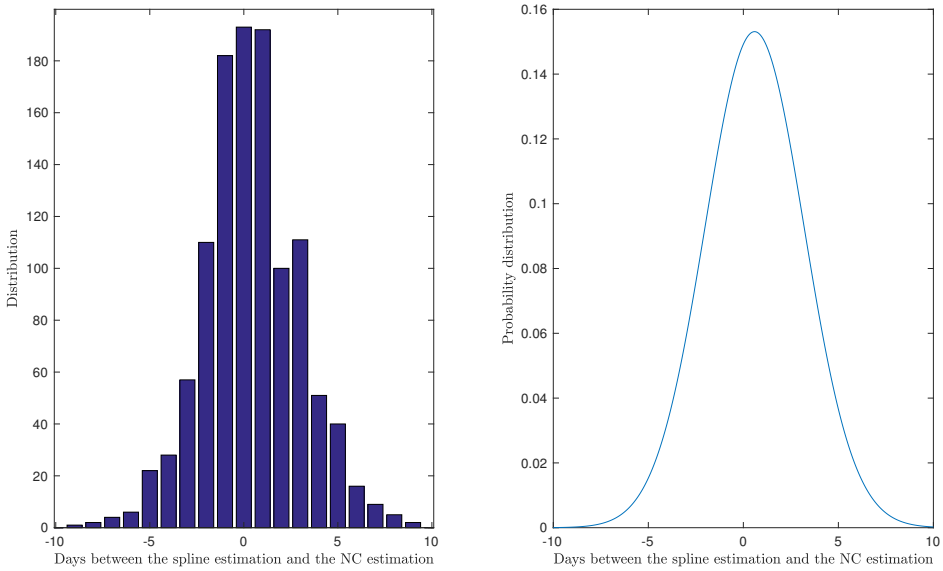


**Figure 6.7:** Distribution and probability distribution of the difference between Natural Cycle's estimation of the ovulation day and the smoothing spline based algorithm of the ovulation day 10 days into the last cycle.

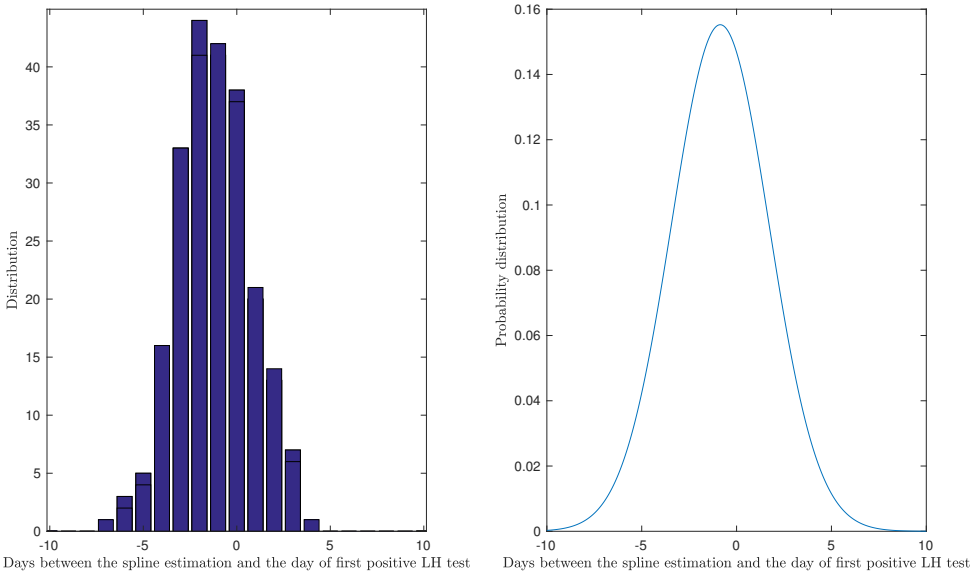


**Figure 6.8:** Distribution and probability distribution of the difference between the first day of positive LH-test and the smoothing spline based algorithm of the ovulation day 1 day into the last cycle

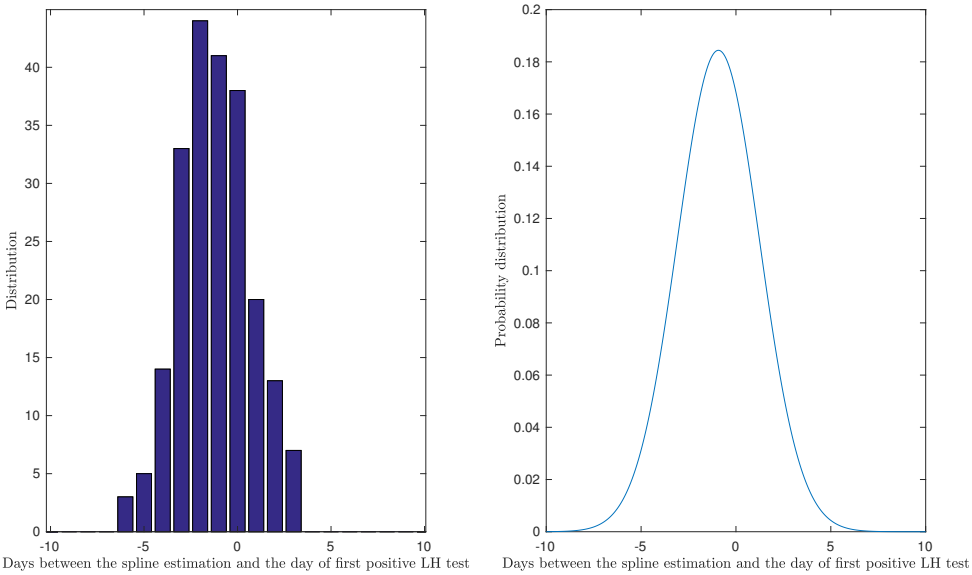


**Figure 6.9:** Distribution and probability distribution of the difference between the first day of positive LH-test and the smoothing spline based algorithm of the ovulation day 10 days into the last cycle

## 6.5 Possible improvements

Using a larger data set could increase our method's accuracy when optimizing the interpolating parameter and the periodic relation to ovulation estimation based on prediction results by comparing to the LH-test results.

Our prediction algorithm could also be improved by not only using previous temperatures when estimating the spline curve of the last cycle. Instead, we could combine previous data for a specific person with data from the whole data set. This would be particularly suitable for cycles with non-regular patterns. Also, using periodic smoothing splines imposes periodic boundary conditions. This forces the first and the last value of the fitted curve to be equal which could, in some cases, lead to inaccuracy in the ovulation prediction. This poses a problem when estimating the cycle length of the last cycle and could be improved by an adaptive method, that estimates the cycle length for each new given data point of the last cycle.

## 6.6 Possible uses

Our algorithm could be used for improving existing fertility tracking applications due to its ability the predict the day of ovulation. If used for preventing pregnancy, an increased accuracy in the ovulation prediction could lower the number of days of abstention from unprotected sexual intercourse. To do this, our algorithm needs to be combined with a coverline-based method.

# Bibliography


[1] Kevin Coward and Dagan Wells. *Textbook of Clinical Embryology*. 2013.

[2] F. Gary. Cunningham. *Williams obstetrics*. 2014.

[3] The European Journal of Contraception E. B. Scherwitzl and Reproductive Health Care. *Identification of the fertile window*.

[4] Magnus R. Hestenes. *Calculus of variations and optimal control theory*. John Wiley, 1966.

[5] Nicoleta Breaz. *The cross-validation method in the smoothing spline regression*. Acta Universitatis Apulensis. Mathematics - Informatics, 2004.

[6] World Health Organization. *Biology of fertility control by periodic abstinence*. 2002.

[7] Chrousos G Dungan K et al. editors. Reed BG, Carr BR. De Groot LJ. *The Normal Menstrual Cycle and the Control of Ovulation.* 2018.

[8] Thomas Strömberg Sara Maad, Martin Clyde and Johan Byström. *Surface fitting with boundary data*. 2004.

[9] Sarah Johnson, Lorrae Marriott and Michael Zinaman. *Can apps and calendar methods predict ovulation with accuracy?* Current Medical Research and Opinion, 2018.

[10] Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[11] Massimo Zanetti. *Periodic cubic smoothing splines as a quadratic minimization problem*.

[12] Zev Rosenwaks and Paul M. Wassarman. *Human Fertility*. Humana Press, 2014.