LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Riccardo De Bin, Bruno Scarpa

Non-parametric Bayesian modeling of cervical mucus symptom

Technical Report Number 170, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Non-parametric Bayesian modeling of cervical mucus symptom

Riccardo De Bin* Bruno Scarpa†

September 4, 2014

Abstract

The analysis of the cervical mucus symptom is useful to identify the period of maximum fertility of a woman. In this paper we analyze the daily evolution of the cervical mucus symptom during the menstrual cycle, based on the data collected in two prospective studies, in which the mucus symptom is treated as an ordinal variable. To produce our statistical model, we follow a non-parametric Bayesian approach. In particular, we use the idea of non-parametric mixtures of rounded continuous kernels, recently proposed in literature to deal with categorical functional data. Fitting the model, we identify the typical pattern of the mucus symptom during the menstrual cycle, i.e. a slow increase of the fertility until the ovulation and, in the aftermath, a steep decrease to a situation less favorable for the fecundation. From the results, it is possible to extract useful information to predict the beginning of the most fertile period and, in case, to identify possible physio-pathological conditions. As a by-product of our analysis, we are able to group the menstrual cycles based on the differences in the daily evolution of the cervical mucus symptom. This division may help in the identification of cycles with particular characteristics.

Keywords: categorical functional data, cervical mucus symptom, non-parametric Bayes, woman fertility.

*Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität of Munich, Germany

†Department of Statistical Sciences, University of Padova, Italy

Corresponding author: Riccardo De Bin, Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität of Munich, Marchioninistr. 15, 81377 Munich (Germany). E-mail: debin@ibe.med.uni-muenchen.de

1 Introduction

The study of women fertility is relevant in many scientific fields, such as biology, demography or medicine. In particular, much attention has been devoted to the investigation of the women menstrual cycles, whose analysis may provide useful information to identify, among other things, the period of maximal fertility of a woman. This issue is strongly relevant for those couples which desire to conceive a child or, on the contrary, to avoid conception by periodic abstinence. In this regard, the simple knowledge of the calendar day is often not sufficient [1]. For this reason, the researchers are interested in additional measurements, such as basal body temperature or cervical mucus symptom, that can provide further information on the daily fertility and may help to identify the days in which the probability of conception is high. The woman menstrual cycle, indeed, is governed by an intense hormonal activity, which determines its characteristics and phases. One visible effect of this hormonal activity is the different appearance and/or sensation of the cervical mucus, whose daily observation allows to identify in a non-invasive way the different phases of the woman menstrual cycle. Being related to the ability of human spermatozoa to move in the cervix [1], the characteristics of the cervical mucus are strongly relevant in the context of women fertility. Here we focus on the cervical mucus symptom and on its daily evolution during the menstrual cycle, in order to derive a statistical model that can support medical investigations and give new insights on the fertility issue.

With this goal in mind, we take advantage of two multi-center studies of daily fecundability co-ordinated by Bernardo Colombo in the Department of Statistical Sciences of the University of Padova [2, 3]. In these prospective studies, women are asked to collect information related to the consistency and the appearance of their mucus secretion, which are later categorized on an ordinal scale [2, 3]. The data are collected daily throughout the entire menstrual cycle, and the mucus symptom peak day is identified for each cycle. This day, in particular, is defined as the last day with best quality mucus, in terms of sensation or appearance [2] and it is important to define the different phases of the menstrual cycle. The mucus symptom peak day, indeed, comes immediately before the ovulation, and, therefore, may be used to determine the end of the fertile window [4, 5].

From a statistical point of view, the daily observed mucus score can be seen as a categorical function of time. The categorical nature of the functional data raises challenging issues: other studies, in the same context, do not consider

this characteristic in order to simplify the problem, avoiding to work with a categorical outcome. For example, Dunson and Colombo develop a model to estimate the trajectory of the mucus consistency as a function of the time considering the observations as instances of a continuous variable [6]. In particular, in their study, they fit a hierarchical Bayesian model to estimate this trajectory. Our aim is to avoid this simplification and to analyze the trajectory of the cervical mucus symptom preserving its categorical nature. We also avoid to define a parametric model, preferring a non-parametric strategy. In order to pursue our goal, an excellent starting point is represented by a recent paper by Canale and Dunson [7]. Following a non-parametric Bayesian approach, they introduce a technique to model counts data based on non-parametric mixtures of rounded continuous kernels. In addition to theoretical results, they develop an efficient Gibbs sampler for the posterior computation. A later contribution from the same authors [8] provides an extension to cope with count functional data. Here, we adapt these methods in order to tackle ordinal categorical functional data, which are characterized by a finite support. Our adaptation allows us to model the trajectories of the mucus symptom in the women menstrual cycles.

In the non-parametric Bayesian framework, an important aspect of the model building process is the choice of the prior on the space of the mixing distributions of the rounded continuous kernels. As in the two aforementioned papers by Canale and Dunson, here we implement a Dirichlet process [9]. Thanks to its clustering property, as a byproduct of the analysis the trajectories of the cervical mucus symptom are clustered in groups. From a medical point of view, this may be useful to identify possible trajectories with peculiar characteristics, which it is worth studying more deeply.

The present paper is organized as follows: first we introduce the two datasets in Section 2, then we formalize the problem and describe the methods in Section 3. In particular, the statistical model used to analyze the data is presented at the end of this section. The results are shown in Section 4 for both the datasets. Finally, a brief discussion is presented in Section 5.

2 Daily fecundability data

In this section we report some information about the two datasets used in the paper. Both are related to studies on women fertility and contain information

about the cervical mucus symptom during the menstrual cycles. The former dataset, which is more recent and contains a larger number of observations, is our dataset of interest. We refer to it as "Italian database". We also extend our analyses to a second dataset, here denoted as Paris data, to investigate the results of our method on the same data used by Dunson and Colombo [6].

In order to be included in the study, in both cases the women must meet some specific criteria, involving their relationship status (they must be married or in a stable relationship), their age (only women between 18 and 40 at the entry are considered) and several other characteristics, which refer to their individual status (they must have at least one menses after cessation of breastfeeding, delivery or miscarriage and they must not be taking any hormonal medication or drug affecting fertility) and to their relation with the partner (couples mixing incidences of unprotected and protected intercourses are excluded). Moreover, the partner cannot be permanently infertile and both must be free from illness causing sub-fertility.

In both studies, each woman is asked to record some data in each day of her menstrual cycle. Here we are interested in the cervical mucus symptom, which is classified on an ordinal scale based on characteristics such as its appearance and the sensation that it produces. In the studies, the menstrual cycle is defined as the interval in days between two consecutive beginnings of vaginal bleeding, and it is characterized by a mucus peak. This information is fundamental for our analysis, since we define the time with respect to the mucus peak day. For this reason we use only the information of those cycles (82.9% of the total) whose mucus peak day is recorded.

2.1 Italian database

This dataset is presented in details by Colombo and colleagues in a 2006 paper [3]. Data are collected by four Italian centers providing natural family planning services, namely Milan, Parma, Saluzzo and Rome, and concern 2755 menstrual cycles of 193 women. The average age of the women is 28.3 years (standard deviation 3.77), while the average age of their partner is 31.2 years (SD = 4.58). Among the women, 128 (52.9%) had at least one past pregnancy, while the percentage of women with past use of hormonal contraception is 16.1% (39 subjects). Due to the absence of information on the mucus peak day, in our analysis we discard 472 cycles, keeping the information about 2283 menstrual cycles (82.9% of the total). For further information about the women and their

cycles, we refer to the original paper [3]. We report, instead, the table with the description of classes used to categorize the cervical mucus symptom (Table 1). In particular, the scale is ordinal and it ranges from 1 to 5, where 1 represents the least fertile status and 5 the situation most favorable for the conception.

Class	Sensation	Appearance
1	No sensation or dry sensation	No mucus, nor any insubstantial discharge
2	No longer dry sensation	No substantial discharge, nor any noticeable mucus
3	Damp sensation	Thick, creamy, whitish, yellowish, sticky, elastic mucus
4	Wet, slippery sensation	–
5	Wet, slippery sensation	Clear, stringy (or stretchy), fluid, watery mucus, blood trails

Table 1: Italian database: classification of the cervical mucus symptom as described in the original study [3]. In our paper, values equal to 0 (no information) and outliers are treated as missing values.

2.2 Paris data

The data for the second example come from a study involving eight centers around Europe, namely Verona, Milan, Lugano, Paris, Düsserdolf, London and Brussels [2]. As in the aforementioned study by Dunson and Colombo [6], we consider only the observations collected in the Paris center. They involve 104 women for a total of 787 cycles. The average age of these women is 29.3 years (standard deviation 4.52), while their partners have an average age of 31.4 years (SD = 5.42). Moreover, 76 out of 104 women (73.1%) had at least one pregnancy, while 38 out of 104 (36.5) have a record of past use of hormonal contraception. Also for this sample, due to the absence of information about the mucus peak day, in our study we consider only a part of the available cycles (576, the 74% of the total). As for the previous dataset, we refer to the original paper [2] to have more information about women and menstrual cycles' characteristics, and we report only the description of the codification of the cervical mucus symptom (Table 2). The scale used to measure the mucus is again ordinal, but, differently from the Italian database, it has only 4 categories. As in the previous case, the larger the number is, the more favorable to conception is the mucus.

Class	Sensation	Appearance
1	Dry, rough and itchy feeling or nothing felt	Nothing seen, no mucus
2	Damp feeling	Nothing seen, no mucus
3	Damp feeling	Mucus is thick, creamy, whitish, yellowish, not stretchy/elastic, sticky
4	Wet, slippery, smooth feeling	Mucus is transparent, like raw egg white, stretchy/elastic, liquid, watery, reddish (with some blood)

Table 2: Paris data: the classification of the cervical mucus symptom provided in the original study [2]. In our paper, values equal to 0 (no information) and outliers are treated as missing values.

3 Non-parametric Bayesian modeling of cervical mucus symptom trajectory

3.1 Data characteristics and formalization of the problem

Let us denote with $y(t)$ the random process which describes the classification of the cervical mucus symptom in a menstrual cycle. More extensively, let $y_c(t)$ indicate the modality of the cervical mucus symptom for the c-th cycle at the t-th day (understood as the t-th day from the mucus peak day), where c is a natural number assuming values from 1 to C. Here, C denotes the total number of cycles analyzed, i.e., $C = 2283$ for the first study, $C = 576$ for the second one.

About t, here it is an integer ranging between -11 and 8. It denotes the day of the cycle with respect to the peak of the mucus symptom. Following a common strategy in the literature [6, 10, 11], indeed, we center the time around this day, which can be seen as the last day of the fertile period. The choice of the 20 days interval $[-11; 8]$, already used in this context by Dunson and Colombo [6], is made to focus on the period of maximum fertility, and it represents a good compromise between the need to include as much (relevant) information as possible, and the need to prevent the inclusion of too many missing values (in particular in those days which are far from the mucus peak). The total length of the observed menstrual cycles varies a lot within each study, from a minimum of 13 days to a maximum of 87 days.

Finally, $y(t)$ can assume values in $\{1, 2, \ldots, K\}$, where K, the number of categories used to codify the cervical mucus symptom, varies in the two studies

(5 in the Italian database, 4 in the Paris data). In both cases, however, K represents, due to the ordinal nature of the categories, the highest value in terms of favorable condition for the fecundation.

3.2 Model derivation

Following the theoretical results provided by Canale and Dunson [7], we consider $y(t)$ as the transformation of an underlying continuous function $y^*(t)$ through a threshold mapping function h. Due to the flexibility of the non-parametric Bayesian method, the values of the thresholds in h can be fixed arbitrarily. As mentioned in the introduction, Canale and Dunson developed their method for functional count data, i.e., for $y(t)$ with infinite support. Here, instead, $y(t)$ can assume only a finite number of values ($\{1, 2, \ldots, K\}$) and therefore we need to adapt the existing method introducing $K + 1$ thresholds, with the first and the last thresholds being respectively $-\infty$ and ∞ in order to allow $y^*(t)$ to assume values in $\mathcal{R}$. In our model, for $j = 1, \ldots, K$ the map function h is defined as

$$h(x) = j \text{ if } a_j \leq x < a_{j+1},$$

where the thresholds are $a_0 = -\infty, a_1 = 2, \ldots, a_{K-1} = K$ and $a_K = \infty$. Since $y(t) = h(y^*(t))$, our model is based on $y^*(t)$, which is defined as

$$y^*(t) = f(t) + \epsilon(t).$$

where $\epsilon(t) \sim N(0, \tau^{-1})$ denotes the error term. Here τ represents the precision terms, and has the usual prior $p(\tau) \propto \tau^{-1}$. We model the functional part $f(t)$ using a basis spline function (hereafter, B-spline), i.e.,

$$f(t) = B\theta,$$

where B denotes the B-spline basis, associated with a large number (50) of knots, and θ represents the basis coefficients. Taking advantage of the properties of the Bayesian non-parametric approach, we leave the prior of θ, let say P, unspecified. Therefore, P is totally unknown and flexible, representing the class of all the possible priors for θ. In our implementation, it follows a Dirichlet process. In formula

$$P \sim DP(\alpha, P_0),$$

where $\alpha > 0$ is a concentration parameter characterizing prior precision and clustering, while P_0 has density $p_0(\theta|\lambda) \propto \exp\{-0.5\lambda\theta^T D^\top D\theta\}$. Here, $\lambda \sim Ga(\nu/2, \rho\nu/2)$, $\rho \sim Ga(1,1)$ and D is a second order difference matrix,

$$D = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & -2 & 1 \end{pmatrix}.$$

In this paper, we denote with $Ga(\psi_1, \psi_2)$ a Gamma distribution with shape parameter ψ_1 and rate parameter ψ_2.

The implementation of this B-splines-based methodology follows the results of Jullion and Lambert [12], including the choice of the matrix D. Also $p_0(\theta|\lambda)$, here chosen as density for P_0 for computational convenience, is defined in their paper. Finally, to be more general, we also consider a prior for the concentration parameter of the Dirichlet process, namely $\alpha \sim Ga(1,1)$.

In a more general formulation, here

$$f(y^*; P) = \int N_T(y^*; B\theta, \mathrm{diag}(\tau^{-1}))dP(\theta),$$

with $P \sim DP(\alpha, P_0)$.

3.3 Posterior computation

For posterior computation, we use an adaption of the Gibbs sampling algorithm presented by Canale and Dunson in their paper [7], with the modifications made to take into account the B-spline technique and the multivariate nature of the problem. The algorithm is based on the blocked Gibbs sampler of Ishwaran and James [13], and takes advantage of the stick-breaking representation of the Dirichlet process. The steps are:

step 1: generate each $y_c^*(t)$ from the conditional posteriors, i.e.,

1. generate $u_c(t)$ from a uniform distribution between $\Phi(a_{y_c(t)}; B\theta_{S_c}, \tau^{-1})$ and $\Phi(a_{y_c(t)+1}; B\theta_{S_c}, \tau^{-1})$;
2. let $y_c^*(t) = \Phi^{-1}(u_c(t); B\theta_{S_c}, \tau^{-1})$;

step 2: update S_c from its multinomial conditional posterior with

$$Pr(S_c = h|-) = \frac{\pi_h p(y_c|B, \theta_h, \tau)}{\sum_{i=1}^{C} \pi_i p(y_c|B, \theta_i, \tau)},$$

where $p(j|B, \theta_h, \tau) = \Pi(a_{j+1}|B\theta_h, \text{diag}(\tau^{-1})) - \Pi(a_j|B\theta_h, \text{diag}(\tau^{-1}))$;

step 3: update the stick-breaking weights using

$$V_h \sim \text{Beta}\left(1 + n_h, \alpha + \sum_{i=h+1}^{C} n_i\right),$$

with n_j denoting the number of processes belonging to group j;

step 4: update the parameters from their conditional posterior,

$$\begin{aligned}
\theta_{S_c}|\tau, \lambda &\sim N(\tau B^\top y_{S_c} V_c), V_c) \\
\tau|\theta &\sim Ga(tC/2, R/2) \\
\lambda|\nu, \rho &\sim Ga(\nu/2 + C\,\text{rank}\{D^\top D\}/2, \rho\nu/2 + U/2) \\
\rho|a, b, \nu, \lambda &\sim Ga(a + \nu/2, b + \nu\lambda/2),
\end{aligned}$$

where $V_c = (\tau n_c B^T B + \lambda D^\top D)^{-1}$, $R = \sum_{i=1}^{C}(y_i - B\theta_i)^T(y_i - B\theta_i)$ and $U = \sum_{i=1}^{C} \theta_i^T \theta_i$. For the update of α, instead, we use the result of Escobar and West [14].

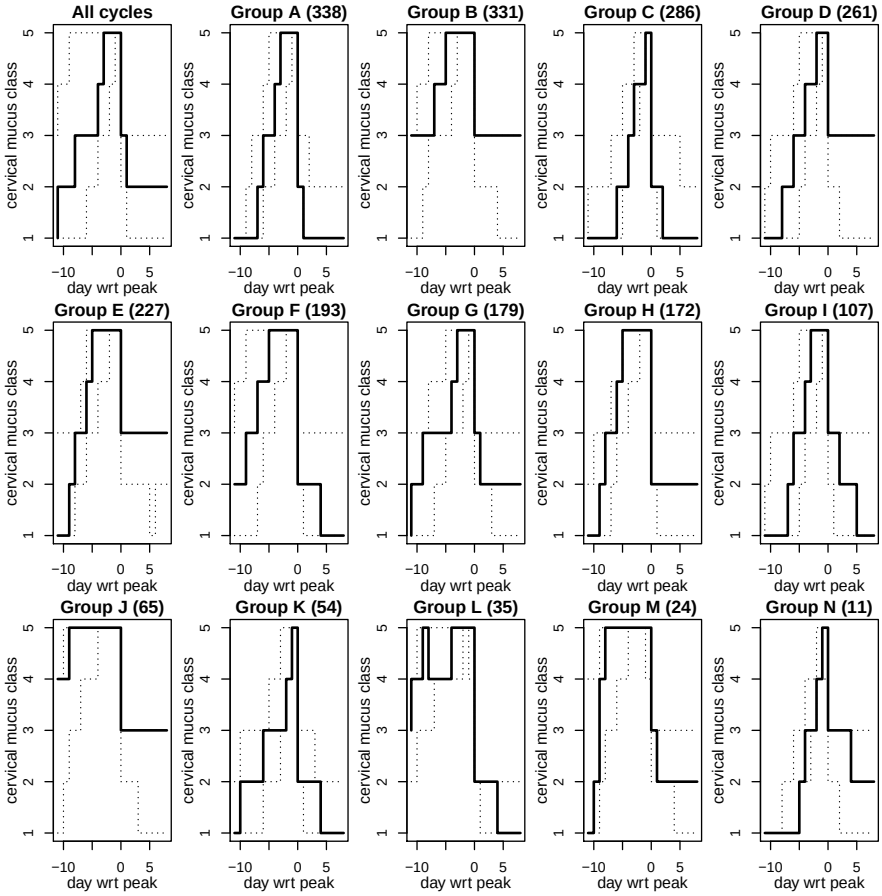
4 Results

4.1 Italian database

4.1.1 Estimate of the trajectories

We perform a MCMC algorithm to estimate the trajectories of the cervical mucus symptom. After an initial burn-in of 3 000 iterations, the algorithm is run 8 000 times. A graphical inspection of the parameter estimates (not shown) seems to confirm their convergence. The final estimation of the trajectory for each cycle corresponds to the most frequent instance among those obtained in the MCMC process. In other words, for each cycle we look at the 8 000 estimates and we select that which occurs with the highest frequency (modal trajectory). In this way, we can bypass the issue of the "label switching" that affects the

Figure 1: Italian dataset: estimation of the mucus symptom trajectories.



approaches based on the Dirichlet process. All the cycles which share the same modal trajectory are then clustered in the same group. It is worth noting that this split does not perfectly correspond to that induced by the Dirichlet process, but it is strictly related: the modal trajectory of each cycle, indeed, is probably the modal trajectory of the group identified by the Dirichlet process to which the cycle belong most of the times. Here, the number of clusters obtained with our procedure and the number of clusters defined by the Dirichlet process are equal, namely 14.

The results are shown in Figure 1. To give an idea of the variability of the estimates, we include a 90% functional interval based on the notion of band

depth for curve, as presented by Lòpez-Pintado and Romo [15, 16]. Very briefly, the band depth is the proportion of times in which, for each point of the support, a curve is inside a band delimited by the minimum and maximum values of all the subsets (of a pre-specified cardinality) that can be formed with the other curves. For example, using subsets of dimension 2 (as in our paper),

$$\begin{aligned}\mathrm{BD2}(y_i) = \quad & \binom{n_g B}{2}^{-1} \sum_{y_j, y_k \in S} I(\min\{y_j(1), y_k(1)\} \leq y_i(1) \leq \max\{y_j(1), y_k(1)\}, \ldots, \\ & , \ldots, \min\{y_j(20), y_k(20)\} \leq y_i(20) \leq \max\{y_j(20), y_k(20)\})\end{aligned}$$

where S indicates the set of all the possible subsets of dimension 2 obtained with the available curves (in our case B, the number of MCMC iterations, times n_g, the cardinality of the cluster) and $I(\cdot)$ denotes the indicator function.

The band depth allows us to introduce an order in the curves, because it is inversely proportional to their "distance" from the median curve. We can then follow an approach similar to that used by Sun and Genton [17] to construct their "functional box-plots". Here, we select the 90% of the curves with the largest band depth, and, for each point of the support (i.e., for t in $[-11; 8]$), we compute the maximum and the minimum values of these curves. The curve with values equal to the maxima represents the upper limit, while that with the minima the lower limit. Please note that this interval is not a proper credibility interval for the modal curve, but it provides a useful indication of the variability of the estimated curves insides the groups.

Moreover, for matter of comparison, we also report the results for all the cycles, without taking into account the split in groups. Since in this case it makes no sense to report the modal curve, we draw the curves with maximum band depth, which is related to the median curve [17]. The 90% interval, instead, is computed as for the groups, considering all the 8000×2283 curves obtained in the MCMC process.

From the analysis of Figure 1, we can clearly identify the well-known pattern of the evolution of the fertility during the woman cycle. Starting from a situation of scarce fertility, characterized by a dry sensation and no-mucus (class 1), all the curves show the progression through the intermediate statuses (class 2, 3 and 4), until the situation of maximal fertility (class 5) is reached. The length of this high-fertility window differs from group to group, but then all the trajectories show a fast drop to less fertile classes (2 or 3) in the day immediately after the peak.

Although the pattern described above is common, we also note some differ-

ences among the trajectories of the mucus symptom for the different groups. We have underlined in the introduction that this clustering, obtained as by-product of the estimation process by our non-parametric Bayesian approach, may provide useful information about the cycles. For this reason, in the following we briefly analyze the results from this point of view.

4.1.2 Description of the groups.

Among the groups, the trajectories of the mucus symptom differ to each other for characteristics such as the length of the high-fertility window (from 1 day for groups C, K, N to the 9 days for group J), the shape of the increasing process (for example, very steep for group M, relatively gentle for group K) and of the decreasing process (for example, for group A the less fertile class, 1, is reached two days after the peak, while for several groups this does not happen within the 8 days after the peak considered in our analysis). It is worth noting, however, that the right tails of the trajectories (decreasing process) are steeper than the left tails (increasing process) for all the groups: for example, the mucus symptom always drops at least to class 3 the day after the peak, while to reach class 5 it always passes through class 4.

Group L shows a modal trajectory with an unexpected shape, in which we can recognize two peaks. If we investigate the observed cycles included in this group, we can note that they are characterized by two peculiarities: an early presence of fertile mucus symptom (classes 4 and 5) and a very fast drop to the least fertile classes (1 and 2) after the peak. The latter characteristic, in particular, differentiates the observed cycles of this group from those of groups M and J, which also have an early presence of fertile mucus symptom. More interestingly, we find the "two peaks shape" of the modal curve directly in part of the observed cycles: therefore, it may be worth investigating the menstrual cycles clustered in this group in order to identify possible physio-pathological situations.

4.1.3 Prediction of the beginning of the most fertile window.

From a practical point of view, these results may drive to some suggestions for those couples which want/do not want to have children: in particular, the experience of a mucus symptom with characteristics belonging to class 4 is a strong signal for the fast approaching of the period of maximum fertility. In all the groups, indeed, class 4 occurs immediately before the start of the high-

fertility window, namely one or two days before it (i.e., the mucus symptom does not last long in class 4, but goes quickly to class 5). If we want to predict the beginning of the high-fertility window based on the mucus symptom of class 3, instead, we must be more careful: on the one hand, in large part of the groups, the step from class 2 to class 3 happens 3 or 4 days before the first instance of class 5; on the other hand, there are counter-examples, like groups G and K, where the mucus symptom stays longer in this class, and the high-fertility window is reached only after 6 or 7 days. Moreover, we cannot really say anything for group B and L: in the former case, the step from class 2 to class 3 is likely before day -11, in the latter, the situation is not clear due to the peculiar behavior of the trajectory.

4.1.4 Inclusion frequencies.

The curve used to estimate the mucus symptom trajectory for each cycle is the modal curve, namely the most selected curve among the 8 000 estimates obtained in the MCMC process. It may be interesting to analyze the selection frequencies of these modal curves. The results are reported in Table 3, summarized by group. Please note that in total there are 5^{20} possible trajectories, and therefore the chance to select randomly a curve is very small (around 10^{-14}). Since in our study we perform only 8 000 MCMC iterations, the situation of maximum entropy is here represented by 8000 different estimates, each with selection frequency 0.000125. As we can see from Table 3, the selection frequencies for the modal curves obtained in our analysis are far from this case, with a highest value of 0.706.

As we expected, the most numerous groups tend to be characterized by high selection frequencies. The large number of cycles available within the group, indeed, makes the estimate of the group trajectory robust, in the sense that small perturbations in the group composition, derived from possible inclusion/exclusion of some cycles in the specific MCMC iteration, do not change the estimate of the trajectory. Conversely, in groups of small size, the inclusion/exclusion of one or few cycles may modify the estimate, generating different curves with relatively small selection frequency. This is the case of group L, which is characterized by the lowest average selection frequencies for the modal curves. If we look at the 8 000 estimates for the cycles belonging to this group, indeed, we find curves slightly different from the modal one. In particular, for the second and third most selected curves the class 1 is reached, in the right

Group	Minimum	1^{st} quart.	Median	Mean	3^{rd} quart.	Maximum
A	0.096	0.269	0.400	0.373	0.487	0.528
B	0.068	0.431	0.597	0.532	0.666	0.677
C	0.060	0.243	0.317	0.287	0.349	0.356
D	0.048	0.248	0.462	0.415	0.587	0.650
E	0.156	0.363	0.458	0.455	0.572	0.706
F	0.045	0.251	0.360	0.338	0.433	0.466
G	0.028	0.153	0.202	0.191	0.230	0.244
H	0.078	0.191	0.235	0.231	0.280	0.318
I	0.085	0.145	0.182	0.178	0.214	0.243
J	0.050	0.302	0.416	0.354	0.465	0.465
K	0.057	0.092	0.107	0.102	0.112	0.119
L	0.038	0.037	0.038	0.037	0.038	0.038
M	0.090	0.132	0.162	0.152	0.173	0.180
N	0.048	0.050	0.052	0.052	0.055	0.055

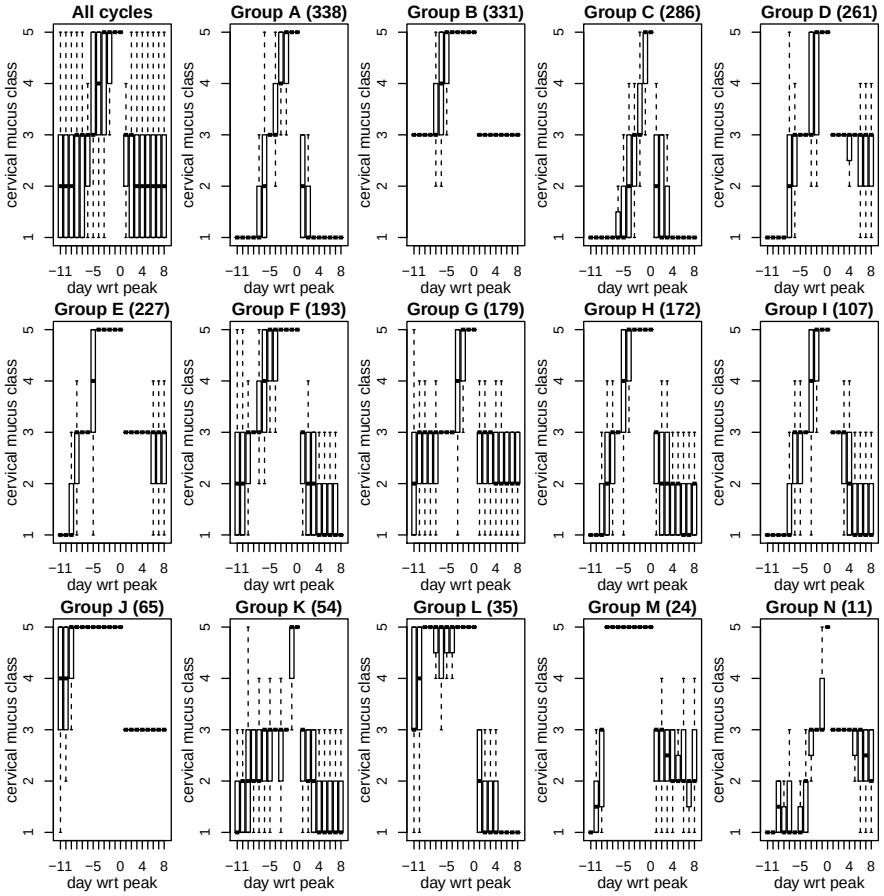
Table 3: Italian dataset: selection frequencies for the modal trajectories per group

tails, one day earlier or later than in the modal curve. Surprisingly, the differences do not involve the "two peak shape" described above, which seems to be common to all the most selected estimates for the cycles belonging to this group.

The size itself, however, cannot explain alone the average selection frequencies of the modal curves of the groups. For example, we note that the selection frequencies of the modal trajectories of groups J are generally larger than those of group C, despite that the former group contains 65 cycles against the 286 of the latter. It is possible, indeed, that smaller groups can result more stable because they contain observations very similar to each other. This is the case of group J: if we look at the box-plots computed day by day (i.e., not as functional data) on the observed cycles for each group (Figure 2), we note a very strong homogeneity among the observations belonging to this group.

With the aim of detecting possible problematic cycles, it may be worth deepening the analysis on the modal trajectories and their selection frequencies. In particular, the trajectories with low selection frequency are highly relevant, especially if they belong to groups with a large average selection frequency. Here, as an example, we investigate the cycles with absolute minimum selection frequency (0.028, group G) and that with minimum selection frequency in the group D (for which the ratio minimum/average selection frequency is minimum). The former is

Figure 2: Italian dataset: box-plots computed day by day on the observed cycles for each groups. Outliers are not reported.



t	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
y	2	3	3	5	5	5	5	2	2	2	5	5	2	2	3	2	2	2	2	2

The trajectory of the cervical mucus symptom for this cycle is peculiar, with class 5 already reached at day -8 and an unexpected drop to class 2 between days -4 and -2. Due to these peculiarities, it is hard for our algorithm to assign the cycle to a specific group. It is easy to imagine that this cycle "jumps" often among the different groups detected by the algorithm. As a consequence, very different trajectories are estimates for this cycle during the MCMC process, each of them very few times (low selection frequency). As a curiosity, we note that the second and third most selected estimates among the 8 000 trajectories obtained in the MCMC process have a shape similar to the modal trajectory of group L.

The cycle of group D whose modal curve has minimum selection frequency, instead, is

t	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
y	na	na	na	na	na	na	na	na	3	3	5	5	3	3	3	3	na	na	2	3

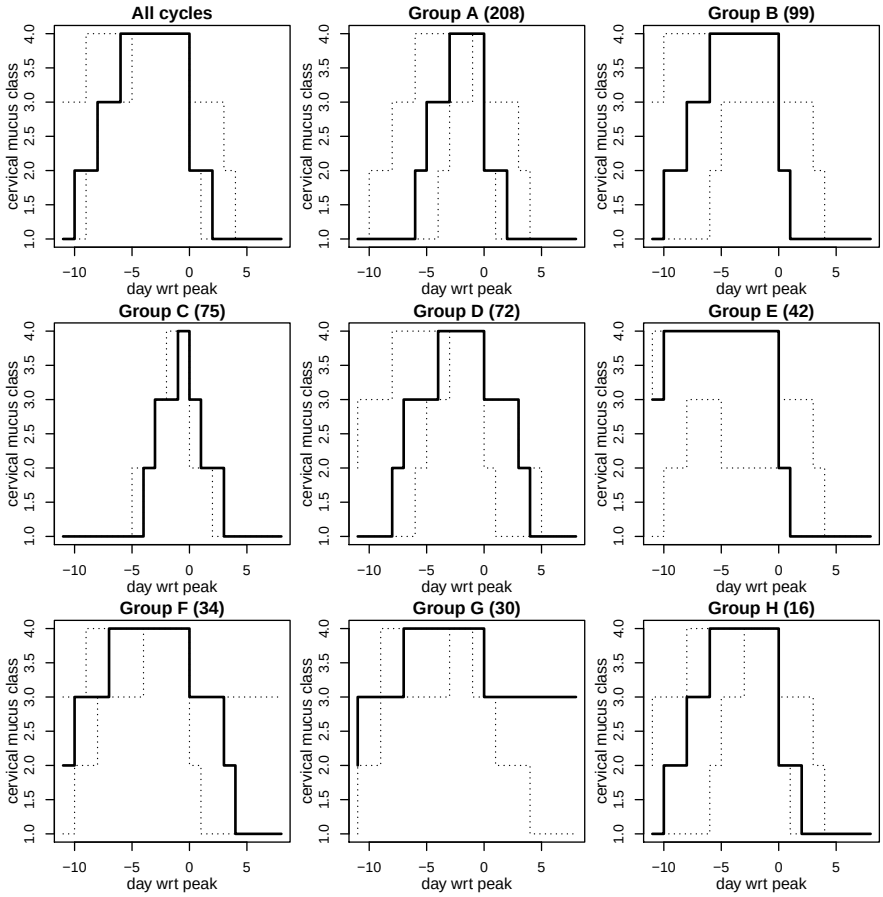
In this case the issue is related to the numerous missing values in the cycle, which also make complicated to assign the cycle to a specific group. It is worth noting that a moderate presence of missing values does not affect the stability of the estimates, and a comparison between number of missing values and selection frequencies of the modal trajectories does not show any strong correspondence. However, when the number of missing values is large (approximately more than 1/3 of the days), the problems become severe. The last cycle analyzed is an example.

4.2 Paris data

We replicate the analysis using the same data used in the aforementioned paper by Dunson and Colombo, introduced as "Paris data" in Section 2. Again, we perform a MCMC algorithm, discarding the first iterations (here 4 000) as burn-in, and then iterate the MCMC process other 8 000 times. Again, the convergence of the parameters is checked by graphical inspection. Figure 3 report the results of our analysis, in which the estimated trajectory for each cycle is again chosen by selecting the mode of all the different estimates obtained in the MCMC algorithm.

In this dataset, partially due to the smaller sample size ($C = 576$) and the smaller number of possible classes for the mucus symptom ($K = 4$), we identify fewer clusters (groups). The characteristics of the groups are similar to those

Figure 3: Paris data: estimation of the mucus symptom trajectories.



Group	Minimum	1^{st} quart.	Median	Mean	3^{rd} quart.	Maximum
A	0.173	0.484	0.611	0.552	0.665	0.681
B	0.023	0.218	0.265	0.247	0.287	0.291
C	0.017	0.284	0.315	0.299	0.325	0.329
D	0.097	0.287	0.396	0.352	0.432	0.456
E	0.056	0.158	0.201	0.179	0.213	0.216
F	0.060	0.090	0.097	0.093	0.100	0.103
G	0.071	0.133	0.148	0.136	0.148	0.148
H	0.074	0.143	0.155	0.145	0.165	0.181

Table 4: Italian dataset: selection frequencies for the modal trajectories per group

obtained in the Italian database: there is a gentle slope between the less fertile status (mucus symptom class equal to 1) to the high-fertility window (class 4), and usually a steeper decrease starting from the day immediately after the peak. In this dataset, however, there is an example of complete symmetric trajectory (group D). Moreover, also in group C we do not see a clear difference between the left and the right tails of the trajectory. However, the two most numerous groups, which jointly contain more than half cycles, show a very steep decrease from class 4 to class 2 the day after the peak, and from class 2 to the less fertile class after one (group B) or two (group A) additional days. Also for group E, the status of minimum fertility is only two days after the peak. In this dataset the most unusual group is group G, in which are included all the cycles for which the mucus symptom is of class 3 during the eight days which follow the peak.

From a practical point of view, the different scale of measurement allows us to use only the class 3 as possible criterion to predict the beginning of the high-fertility window: the trajectories last on this status in almost all the cases two days, with the exception of groups F and G, for which it lasts 3 days. More seriously, for group E we cannot see, in the time interval considered in our analysis, how many days the mucus symptom is of class 3. Nevertheless, this group is relatively small (42 cycles), and, therefore, we may consider 2/3 days after the appearance of the cervical mucus symptom of class 3 as a good point of reference for a couple to stop/intensify the sexual intercourse according to their desire to avoid/have a conception.

We replicate the analysis on the selection frequencies of the modal curves for this example as well. The results, reported in Table 4, show a situation not dissimilar to that seen in the previous dataset. In this case, we have only one group with a large size (group A), for which the inclusion frequencies of

the modal curves are relatively high. In this case, the homogeneity of the observations within the groups are similar (see also Figure 4), and therefore the group average inclusion frequency is almost proportional to the group size.

We again investigate the cycles related to the smallest selection frequencies, obtaining results similar to those seen in the previous example. The modal curve with minimum selection frequency,

t	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
y	na	na	na	na	na	4	4	1	1	1	1	4	1	1	1	na	na	na	na	na

and that related to the smallest ratio minimum/average selection frequency,

t	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
y	na	na	na	4	4	4	4	4	3	1	1	1	1	1	1	1	1	1	1	1

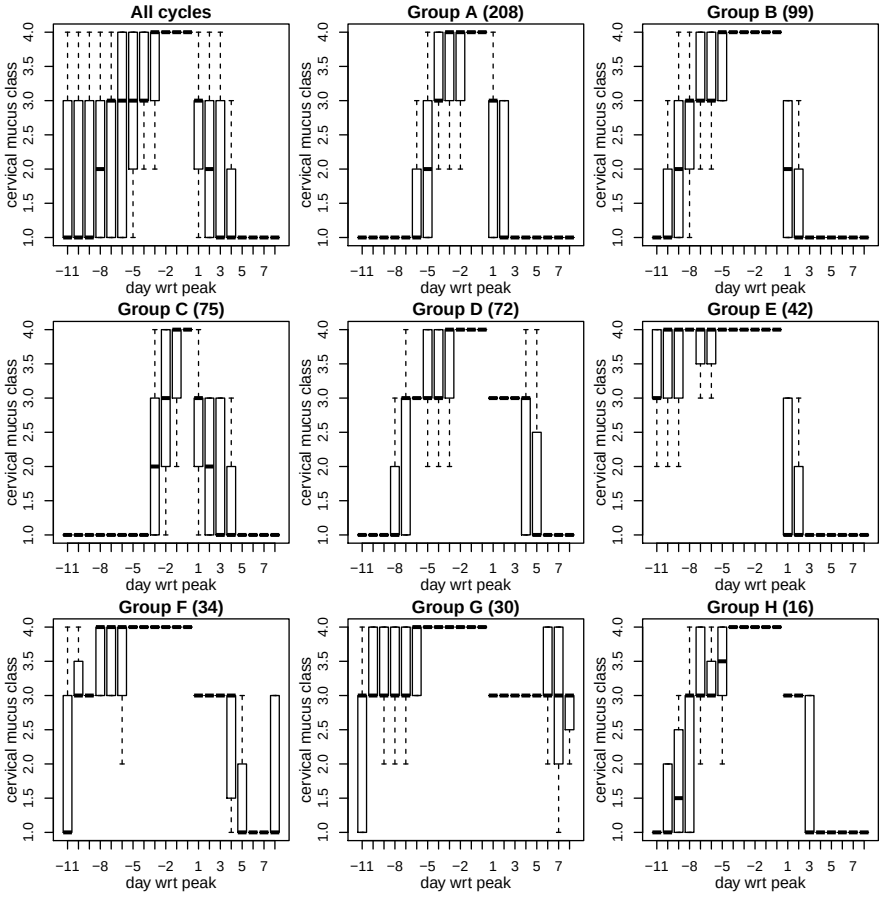
either contains several missing values or shows a peculiar shape. In particular, in latter case, it seems that the peak day have been registered incorrectly, because the mucus symptom has class 1 at day 0 and the last day with class 4 is $t = -4$.

As a final remark, we note that our results are in line with those provided by the aforementioned paper by Dunson and Colombo. In particular, both in their and in our study, it is possible to identify the well-known trajectory of the cervical mucus symptom, which increases gently from a status of low fertility to a status of high fertility, and decreases steeply after the mucus peak day.

5 Discussion

In this paper we have seen that the non-parametric Bayesian approach is suitable to treat categorical functional data. In particular, it is a good tool to investigate the evolution of the cervical mucus symptom during the woman menstrual cycle. We have seen how it can be profitably used to estimate the trajectory of the mucus symptom, to predict the beginning of the most fertile period and, in case, to identify possible physio-pathological situations. With respect to the latter issue, we have seen that our approach based on the computation of the modal curves may provide, through the analysis of the selection frequencies, a useful tool to identify cycles with peculiar characteristic. Moreover, the issue related with the "label switching" connected with the use of a Dirichlet process is avoided. On the other hand, our approach may be less accurate than others, because the estimates of the trajectories are based only on a fraction of the MCMC iterations. In the examples, we have seen that this fraction may be very large, up to 0.702, which means that the estimates probably correspond to those computed with any other technique, but also very small, which implies

Figure 4: Paris data: box-plots computed day by day on the observed cycles for each groups. Outliers are not reported.



that the estimates of the trajectories may be influenced by the randomness of the MCMC process. We have seen, however, that those cycles (observations) for which the modal curves have low selection frequencies are problematic (too many missing values) or may need further investigations (peculiar trajectories).

In the paper we have stressed the flexibility of the non-parametric Bayesian approach. Here we have considered a simple hierarchical Bayesian model treating each cycle independently, but more complex situation may be investigated, for example taking into account the fact that multiple cycles belong to the same woman. Besides the flexibility of the model, an interesting and useful aspect of this approach is related to the clustering property of the Dirichlet process. We have seen how the split in clusters can provide useful information and allow more specific estimates for the observed cycles (we have an estimate for each group and not only a general one). On the negative side, we note that the statistical model is not the easiest to be explained, especially to non-statisticians. Another weakness is the large computational effort necessary to obtain the results, in particular for the burn-in phase of the MCMC process. For theoretical reasons, indeed, during the burn-in it is not possible to take advantage of all the possibilities offered by the modern computational techniques, for example the parallel computation. Nevertheless, the computational power is rapidly increasing year by year, and this issue may rapidly become irrelevant.

Acknowledgments

RDB was partially financed by the University of Padova under the project *"Modelli bayesiani non parametrici per l'analisi di dati funzionali: teoria ed applicazioni in biometria e marketing"* and by grant BO3139/4-1 from the German Science Foundation (DFG). The authors wish to thank Anne-Laure Boulesteix and Antonio Canale for their advices and suggestions.

The authors declare that there is no conflict of interest.

References

1. Wood JW. *Dynamics of Human Reproduction: Biology, Biometry, Demography.* Aldine de Gruyter, New York, 1994.

2. Colombo B, Masarotto G. Daily fecundability: first results from a new data base. *Demographic Research* 2000; **3**:5.

3. Colombo B, Mion A, Passarin K, Scarpa B. Cervical mucus symptom and daily fecundability: first results from a new database. *Statistical Methods in Medical Research* 2006; **15**:161–180.

4. Guida M, Tommaselli GA, Palomba S, Pellicano M, Moccia G, Di Carlo C, Nappi C. Efficacy of methods for determining ovulation in a natural family planning program. *Fertility and Sterility* 1999; **72**:900–904.

5. Ecochard R, Boehringer H, Rabilloud M, Marret H. Chronological aspects of ultrasonic, hormonal, and other indirect indices of ovulation. *BJOG: An International Journal of Obstetrics & Gynaecology* 2001; **108**:822–829.

6. Dunson D, Colombo B. Bayesian modeling of markers of day-specific fertility. *Journal of the American Statistical Association* 2003; **98**:28–37.

7. Canale A, Dunson D. Bayesian kernel mixtures for counts. *Journal of the American Statistical Association* 2011; **106**:1528–1539.

8. Canale A, Dunson D. Nonparametric Bayes modelling of count processes. *Biometrika* 2013; **100**:801–816.

9. Ferguson TS. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1973; **1**:209–230.

10. Dunson D, Weinberg C, Baird D, Kesner J, Wilcox A. Assessing human fertility using several markers of ovulation. *Statistics in Medicine* 2001; **20**:965–978.

11. Bigelow J, Dunson D. Bayesian adaptive regression splines for hierarchical data. *Biometrics* 2007; **63**:724–732.

12. Jullion A, Lambert P. Robust specification of the roughness penalty prior distribution in spatially adaptive bayesian p-splines models. *Computational Statistics & Data Analysis* 2007; **51**:2542–2558.

13. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 2001; **96**:161–173.

14. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 1995; **90**:577–588.

15. López-Pintado S, Romo J. Depth-based inference for functional data. *Computational Statistics & Data Analysis* 2007; **51**:4957–4968.

16. López-Pintado S, Romo J. On the concept of depth for functional data. *Journal of the American Statistical Association* 2009; **104**:718–734.

17. Sun Y, Genton MG. Functional boxplots. *Journal of Computational and Graphical Statistics* 2011; **20**:4957–4968.