A novel regularized approach for functional data clustering: An application to milking kinetics in dairy goats

Christophe Denis, Emilie Lebarbier, C. Lévy-Leduc, Olivier Martin, Laure Sansonnet

A NOVEL REGULARIZED APPROACH FOR FUNCTIONAL DATA CLUSTERING: AN APPLICATION TO MILKING KINETICS IN DAIRY GOATS

C. DENIS, E. LEBARBIER, C. LÉVY-LEDUC, O. MARTIN, AND L. SANSONNET

Abstract. Motivated by an application to the clustering of milking kinetics of dairy goats, we propose in this paper a novel approach for functional data clustering. This issue is of growing interest in precision livestock farming that has been largely based on the development of data acquisition automation and on the development of interpretative tools to capitalize on high-throughput raw data and to generate benchmarks for phenotypic traits. The method that we propose in this paper falls in this context. Our methodology relies on a piecewise linear estimation of curves based on a novel regularized change-point estimation method and on the k-means algorithm applied to a vector of coefficients summarizing the curves. The statistical performance of our method is assessed through numerical experiments and is thoroughly compared with existing ones. Our technique is finally applied to milk emission kinetics data with the aim of a better characterization of inter-animal variability and toward a better understanding of the lactation process.

1. Introduction

Precision livestock farming is a blooming field grounded in the development of sensors providing high throughput data and thus potentially increasing access to valuable information on biological processes. Therefore, developing methods for data analysis and interpretation has become a challenging issue in animal science. Economic performance of dairy goat farming systems is primarily based on milk production and a large amount of farmers working time is spent milking animals, see Marnet et al. (2005). Moreover, with the increasing size of goat herds and the rapid growth of the dairy goat industry, more in-depth information on individual milking performance is necessary. In this context, a better understanding of the variability in milk flow kinetics could for instance help refining selection criteria for breeding programs, simplifying milking workload or controlling udder health. Milk emission kinetics recorded during milking of dairy goats are classically described and classified through synthetic parameters such as milking time, maximum and average milk flow rates, time to reach 500 g/min milk flow, see Romero et al. (2017). In this paper, we explore the possibility of considering milk emission kinetics as a whole function, opening new perspectives to study inter-animal variability.

From a statistical point of view, this issue belongs to the general field of functional data analysis, see Ramsay and Silverman (2005) for a survey on this subject. In the specific functional data clustering framework, several approaches have been proposed by Abraham et al. (2003), Jacques and Preda (2013) and Bouveyron et al. (2015) among others. For a

review on this subject, we refer the reader to Jacques and Preda (2014a) and the references therein. This kind of approaches was extended to deal with multivariate functional data by Jacques and Preda (2014b) who proposed the first model-based clustering algorithm in this multivariate context and more recently by Schmutz et al. (2018).

To deal with the functional clustering of the milking kinetics of goats, some specific features have to be taken into account, see Figure 1 for some examples of such kinetics. We can see from this figure that these curves are nondecreasing and can be split into two parts, namely an increasing linear part and an almost constant one. Inspired by Abraham et al. (2003), we propose in this paper a dimension reduction approach based on a continuous piecewise linear function fit to each curve which boils down to a change-point detection issue which will be crucial in our method.

The problem of detecting change-points in the mean of a signal is largely addressed in the literature. In particular, it is now well known that in (penalized-) maximum likelihood frameworks the Dynamic Programming (DP) algorithm (Bellman (1961); Auger and Lawrence (1989)) and its recent pruned versions Killick et al. (2012); Rigaill (2015); Maidstone et al. (2016) are the only algorithms that retrieve the exact solution very quickly. However, DP can only be used if the contrast to be optimized is additive with respect to the segments, see for example Bai and Perron (2003); Picard et al. (2005); Lavielle (2005). When detecting changes in the slope with a continuity condition, the segments will unavoidably be linked and therefore the additivity condition is not satisfied. This

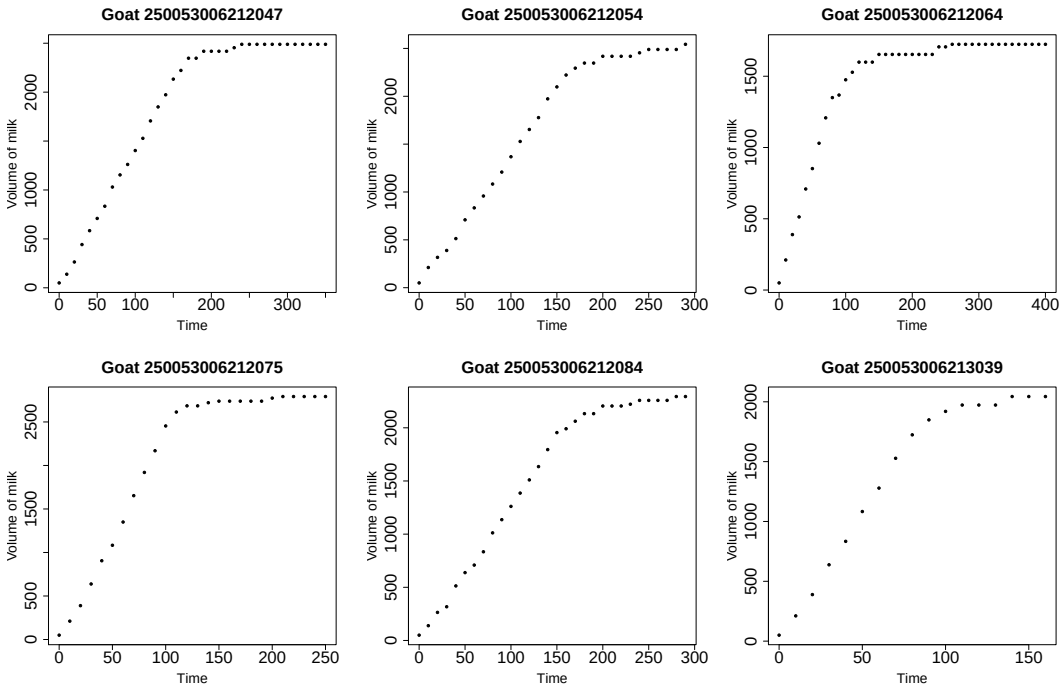

FIGURE 1. Some examples of milking kinetics of goats.

partly explained that this change-point detection problem has not been thoroughly investigated in the literature compared to the simplest detection in the mean problem. Recently, Fearnhead et al. (2019) proposed to extend the PELT algorithm Killick et al. (2012) to this problem. Their idea is to include the penalty in the DP algorithm with a pruning strategy. The penalty they proposed is proportional to the number of change-points up to a penalty constant. However, this penalty constant needs to be chosen in advance, which is not easy in practical situations.

In this paper, we first propose a novel change-point estimation in the slope method combining the trend filtering proposed by Tibshirani (2014) with a (penalized-) maximum likelihood approach which is useful for removing the spurious change-points that may have been proposed by trend filtering. These change-points estimators are then used for devising a new dimension reduction approach: Each curve is summarized by a vector containing the coefficients of its projection onto an order 2 B-spline basis having for knots the obtained change-points and also the change-point locations. Including the change-points both in the features characterizing the curves and in the B-spline knots is the main novelty compared to classical approaches reviewed in Jacques and Preda (2014a).

The paper is organized as follows. The methodology that we propose is decribed in Section 2. The performance of our approach is investigated in Section 3 through numerical experiments. Finally, in Section 4, we apply our method to the data that motivated this study.

2. Methodology

In this section, we describe our novel functional data clustering approach which consists of two steps which can be summarized as follows:

- First step: Piecewise linear estimation of the curves using a novel change-point estimation method based on the trend filtering approach and B-splines.
- Second step: Applying the k-means algorithm to a vector of coefficients summarizing the curves obtained in the first step.

These two steps are further described hereafter.

2.1. First step: Piecewise linear estimation of the curves based on a change-point estimation method.

In the following, we assume that the observations of a given curve $\mathbf{Y} = (Y_1, \ldots, Y_n)$ correspond to a noisy function evaluated at the input points $\mathbf{x} = (x_1, \ldots, x_n)$. In this step, we aim at estimating each curve by a piecewise linear function using a two-stage approach described below.

2.1.1. *First stage: Trend filtering for change-point estimation.*

We use the trend filtering approach proposed by Tibshirani (2014) which consists in fitting to the observations $\mathbf{Y}$ the vector $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_n)$ using a regularized method. More precisely, we use

$$\widehat{\boldsymbol{\beta}}(\lambda) = \text{Argmin}_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\{ \|\mathbf{Y} - \boldsymbol{\beta}\|_2^2 + \lambda \|D^{(2)}\boldsymbol{\beta}\|_1 \right\},$$

where $\|y\|_2^2 = \sum_{i=1}^n y_i^2$, $\|y\|_1 = \sum_{i=1}^n |y_i|$, for $y = (y_1, \ldots, y_n)$, λ is a positive constant which has to be tuned and $D^{(2)}$ is the discrete difference operator of order 2 defined by

$$D^{(2)} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \end{pmatrix}.$$

The final estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$ where $\widehat{\lambda}$ has to be properly chosen. Usually, this parameter is chosen using resampling approaches such as cross-validation or stability selection, see Meinshausen and Bühlmann (2010). From $\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$, we define a set of potential change-point indices as the coordinates where the vector $D^{(2)}\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$ is not equal to zero. However, in change-point estimation frameworks, the performance of such methods may be altered since some change-points may be omitted by subsampling. Moreover, it is well known that such regularization approaches lead to over-segmentation phenomena. Usually, in this case, a DP algorithm is then used on the set of potential change-points obtained with the latter strategy in order to remove the irrelevant ones, see for instance Harchaoui and Lévy-Leduc (2007) and Harchaoui and Lévy-Leduc (2010).

We propose following this strategy: In order to avoid the use of a resampling method, we choose a small enough λ in order to obtain a large enough set of potential change-points. More precisely, we set a maximal number of change-points denoted $K_{\max}$ and choose λ such that among the λ's leading to $K_{\max}$ change-points, $\widehat{\lambda}$ is the one minimizing $\|\mathbf{Y} - \widehat{\boldsymbol{\beta}}(\lambda)\|_2^2$.

Let $(\widehat{n}_1, \ldots, \widehat{n}_{K_{\max}})$ the resulting change-point indices and the associated change-point positions $(\widehat{t}_1, \ldots, \widehat{t}_{K_{\max}}) = (x_{\widehat{n}_1}, \ldots, x_{\widehat{n}_{K_{\max}}})$. For each K in $\{1, \ldots, K_{\max}\}$, we use the DP algorithm to retrieve the K most relevant change-point indices among $\widehat{n}_1, \ldots, \widehat{n}_{K_{\max}}$. DP is thus applied to $Y_{\widehat{n}_1}, \ldots, Y_{\widehat{n}_{K_{\max}}}$ instead of $Y_1, \ldots, Y_n$. Note that a slight modification of the algorithm is considered to make the piecewise linear fit to data continuous. The optimal number of change-points $\widehat{K}$ is then chosen by using the criterion proposed by Lavielle (2005).

2.1.2. *Second stage: Projection onto the B-spline basis having as knots the obtained change-points.* Each curve will then be summarized by a few coefficients corresponding to the coefficients of its projection onto the B-spline basis $(B_{i,2})_{1 \leq i \leq \widehat{K}+2}$ defined as follows, see (Hastie et al., 2009, p. 206) for a review on the subject. Let $\widehat{t}_0 = x_1$ and $\widehat{t}_{\widehat{K}+1} = x_n$. Let us also define the augmented knot sequence τ such that:

$$\tau_1 = \tau_2 = \widehat{t}_0 = x_1,$$

$$\tau_{j+2} = \widehat{t}_j, \ j = 1, \ldots, \widehat{K},$$

$$\tau_{\widehat{K}+3} = \tau_{\widehat{K}+4} = \widehat{t}_{\widehat{K}+1} = x_n,$$

namely,

$$(\tau_1, \ldots, \tau_{\widehat{K}+4}) = (x_1, x_1, \widehat{t}_1, \cdots, \widehat{t}_{\widehat{K}}, x_n, x_n).$$

The ith B-spline function $B_{i,2}$ having τ for knot sequence satisfies:

$$B_{i,2}(u) = \frac{u - \tau_i}{\tau_{i+1} - \tau_i} B_{i,1}(u) + \frac{\tau_{i+2} - u}{\tau_{i+2} - \tau_{i+1}} B_{i+1,1}(u),$$

where

$$B_{i,1}(u) = \begin{cases} 1, \text{ if } \tau_i \leq u < \tau_{i+1} \\ 0, \text{ otherwise} \end{cases}$$

with $i \in \{1, \ldots, \widehat{K} + 2\}$. Thus, each curve is estimated by $\widehat{f}$ defined by:

$$\widehat{f}(u) = \sum_{i=1}^{\widehat{K}+2} \widehat{\theta}_i B_{i,2}(u), \tag{1}$$

where the $\hat{\theta}_i$'s are obtained using a least-square criterion. Hence, the coefficients summarizing each curve is:

$$(\widehat{\theta}_1, \ldots, \widehat{\theta}_{\widehat{K}+2}, \widehat{t}_1, \ldots, \widehat{t}_{\widehat{K}}). \tag{2}$$

2.2. **Second step: Clustering using the k-means algorithm.** In order to obtain a clustering of the curves (milking kinetics), we use the k-means algorithm of Hartigan and Wong (1979) on the scaled summarized coefficients (2) obtained in the previous step. It has to be noticed that the number of change-points $\widehat{K}$ may change from one curve to the other. Thus, we consider summarized coefficients of length $\widehat{K}_M$ corresponding to the largest value of $\widehat{K}$. For kinetics having a number of change-points smaller than $\widehat{K}_M$, we replace the missing $\widehat{t}_k$ and the missing coefficients by 0. Our goal is indeed to propose a strategy which is able to distinguish the curves both thanks to the change-point positions and/or the coefficient values.

The number k of clusters is chosen by using the strategy proposed by Charrad et al. (2014) which consists in using the majority rule that is taking for k the value chosen by the largest number of criteria among 30 indices such as: CH index, Duda index, Pseudot2 index, C index, Hartigan index, ... Further details on these indices can be found in Charrad et al. (2014). Here, we focused on the four following indices: KL index, Hartigan index, SDindex, Ptbiserial index.

3. Numerical experiments

In this section, we investigate the statistical performance of our procedure. The simulation scheme that we used for this investigation is described in Section 3.1. We also propose in Section 3.2 to benchmark our procedure with existing approaches and to assess our change-point estimation approach in Section 3.3.

3.1. **Simulation scheme.** In order to be as close as possible to the data coming from our motivating application, we consider two different models for generating the data that we will refer to as `Model 1` and `Model 2` in the following. For each model, the complete observed data is $(\mathbf{Y}, Z)$, where $\mathbf{Y}$ is in $\mathbb{R}^n$ and corresponds to the observations of an underlying function, which we will specify hereafter, at the input points $\mathbf{x} = (x_i)_{1\leq i\leq n} = (10(i-1))_{1\leq i\leq n}$ with $n = 51$. Z denotes the label of $\mathbf{Y}$ which takes its value in $\mathcal{Z} = \{1, 2, 3, 4\}$. Moreover, for each $z \in \mathcal{Z}$, the associated cluster $\mathcal{C}_z$ is characterized by a number of change-points K_z, a vector of change-points t^z, and a vector of parameters $\theta^z \in \mathbb{R}^{K_z+1}$. Hence, each model is defined by a set of parameters $\{K_z, t^z, \theta^z : z \in \mathcal{Z}\}$. The values of the parameters associated to each model are reported in Tables 1 and 2. Note that for each model, the clusters are distinguishable by both the change points and the parameters.

For each model, the vector $(\mathbf{Y}, Z)$ is simulated according to the following procedure:

(a) The label Z is drawn from a uniform distribution on $\mathcal{Z}$;
(b) We generate $\tilde{t}^Z = t^Z + \mathcal{U}$, such that $\mathcal{U} = (U, \ldots, U)$, where U is a uniformly distributed random variable on $\{-30, -20, 10, 0, 10, 20, 30\}$;
(c) We generate $\tilde{\theta}^Z = \theta^Z + \mathcal{V}$, such that $\mathcal{V} = (V, \ldots, V)$, where V is a uniformly distributed random variable on $[-200, 200]$;
(d) Then, we consider the sequences $(\tilde{t}^Z_0, \ldots, \tilde{t}^Z_{K_Z+1}) = (0, \tilde{t}^Z, 500)$, $(\tilde{\theta}^Z_0, \ldots, \tilde{\theta}^Z_{K_Z+1}) = (0, \tilde{\theta}^Z)$, and define for $x \in [\tilde{t}^Z_j, \tilde{t}^Z_{j+1}]$, and $j \in \{0, \ldots, K_Z\}$

$$f_{\tilde{t}^Z, \tilde{\theta}^Z}(x) = (\tilde{\theta}^Z_{j+1} - \tilde{\theta}^Z_j)\frac{x - \tilde{t}^Z_j}{\tilde{t}^Z_{j+1} - \tilde{t}^Z_j} + \tilde{\theta}^Z_j; \tag{3}$$

TABLE 1. Set of parameters for `Model 1`.

`Model 1`			
z	K_z	t^z	θ^z
1	2	$(150, 250)$	$(1600, 1900, 2000)$
2	2	$(150, 300)$	$(1400, 1800, 2200)$
3	4	$(100, 200, 300, 400)$	$(300, 1500, 1700, 2000, 2200)$
4	3	$(50, 150, 300)$	$(200, 1300, 1800, 2100)$

TABLE 2. Set of parameters for `Model 2`.

`Model 2`			
z	K_z	t^z	θ^z
1	2	$(150, 250)$	$(1600, 1900, 2000)$
2	2	$(150, 300)$	$(1400, 1800, 2200)$
3	4	$(100, 200, 300, 400)$	$(300, 1500, 1700, 2000, 2200)$
4	3	$(150, 250, 300)$	$(200, 700, 1000, 1600)$

(e) Finally, we define $\mathbf{Y}$ such that, for $i \in \{1, \dots, n\}$,

$$Y_i = f_{\tilde{t}^Z, \tilde{\theta}^Z}(x_i) + \varepsilon_i, \tag{4}$$

where the ε_i's are i.i.d $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma \in \{1, 5\}$.

Note that the function f defined in (3) can be seen as another way of writing (1).

Figure 2 displays some observations generated using the above simulation scheme for each model and for each σ. We can see from this figure that the clustering problem associated to `Model 1` seems to be the most difficult. In `Model 1`, the clusters are indeed completely mixed whereas in `Model 2` Cluster $\mathcal{C}_4$ is well separated from the others. Observe also that the data that is generated has the same behavior as the data coming from our motivating application: They are nondecreasing and piecewise linear constant with a small additive noise, see Figure 1.

3.2. **Statistical performance.** Following the simulation scheme described in Section 3.1, the performance of our procedure is assessed for each model, each σ and is compared with two different clustering methods: the k-means algorithm applied to the raw data $\mathbf{Y}$ and the FunFEM procedure described in Bouveyron et al. (2015) and available in the `R` package `FunFEM`. The latter method is dedicated to the clustering of functional data and is based on a functional mixture model. All the methods are compared thanks to the Adjusted Rand Index (ARI) defined in Hubert and Arabie (1985) which is often used for clustering validation. It is indeed a measure of agreement between two partitions. Note that the number of clusters k in the k-means algorithm is chosen using the same strategy as the one that we considered in our approach. As far as `FunFEM` is concerned, we used the default parameters.

For each model and for each σ in $\{1, 5\}$, we repeat independently 100 times the following steps:

(a) We simulate a sample $\mathcal{D}_N = \{(\mathbf{Y}^1, Z^1) \dots (\mathbf{Y}^N, Z^N)\}$ of size $N = 100$ according to the scheme described in Section 3.1;
(b) We apply each method to $\mathcal{D}_N$;
(c) Based on the obtained clustering, we compute the ARI.

The results are displayed in Figure 3 with $K_{\max} = 10$. We can see from this figure that our method outperforms the other ones in all cases except for Model 2 with $\sigma = 5$ where the performance of our method is on a par with the one of FunFEM. Note that applying the k-means to a relevant summary measure of $\mathbf{Y}$ significantly improves the clustering performance. Moreover, we observe that when σ increases, the performance of our approach is slightly altered since the change-points are more difficult to locate accurately, see Section 3.3.

3.3. **Assessment of our change-point estimation procedure.** We provide the following numerical experiments for assessing the change-point estimation stage of our method. We used the parameters associated to Cluster 3 of Model 1, see Table 1. We repeat 100 times

(a) We simulate $\mathbf{Y}$ according to Equation (4) with $\sigma \in \{1, 5\}$;

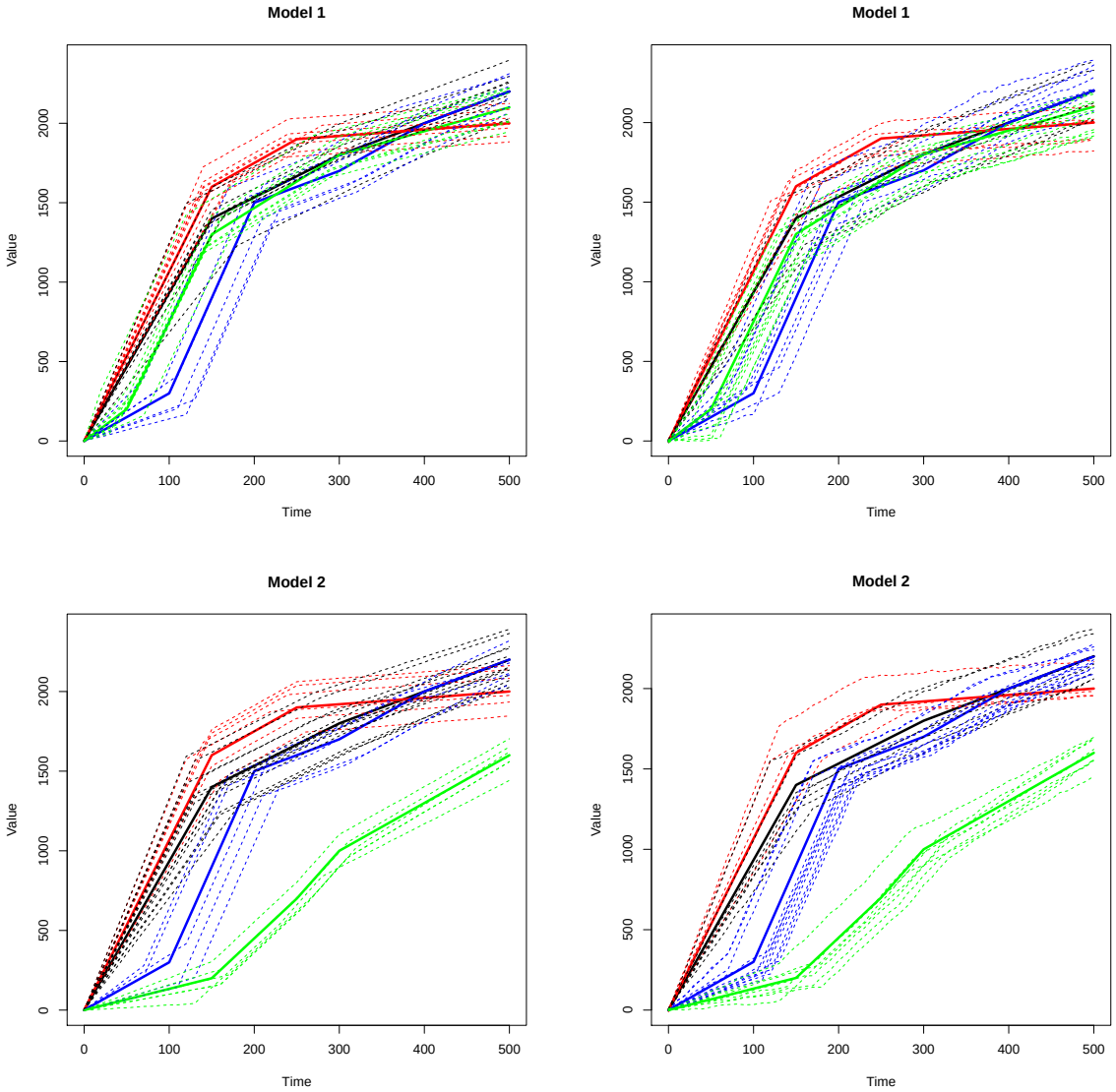

FIGURE 2. Examples of observations generated from `Model 1` (top) and `Model 2` (bottom) for $\sigma = 1$ (left) and $\sigma = 5$ (right). The curves belonging to Cluster 1 (resp. 2, 3, 4) are displayed in red (resp. black, blue and green). The solid lines display the representative curves of each cluster f_{t^z,θ^z} and the dashed ones are some examples of the corresponding $\mathbf{Y}$.

(b) We estimate the change-points according to the procedure described in the first stage of the first step in Section 2.

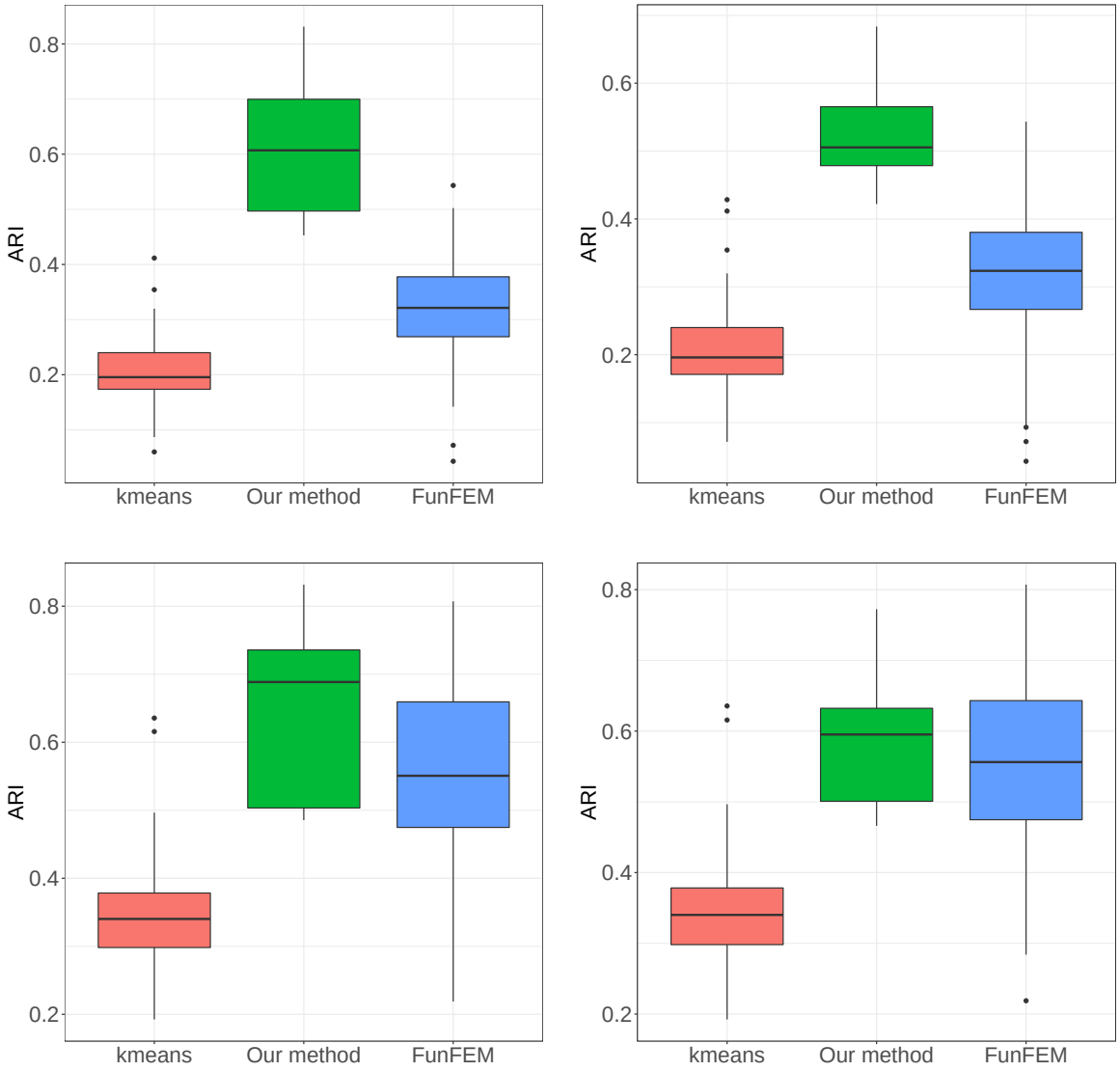

FIGURE 3. Boxplots of the ARI for `Model 1` (top) and `Model 2` (bottom) for $\sigma = 1$ (left) and $\sigma = 5$ (right).

Some examples of $\mathbf{Y}$ for the two values of σ are displayed in Figure 4. We can see from this figure that the change-points located at 300 and 400 are more difficult to detect than the others. It is all the more true when $\sigma = 5$.

Figure 5 displays the frequency of the number of times where each position has been estimated as a change-point. We can see that the change-points are all retrieved and that no spurious change-points are provided when $\sigma = 1$. In the case where $\sigma = 5$, although the

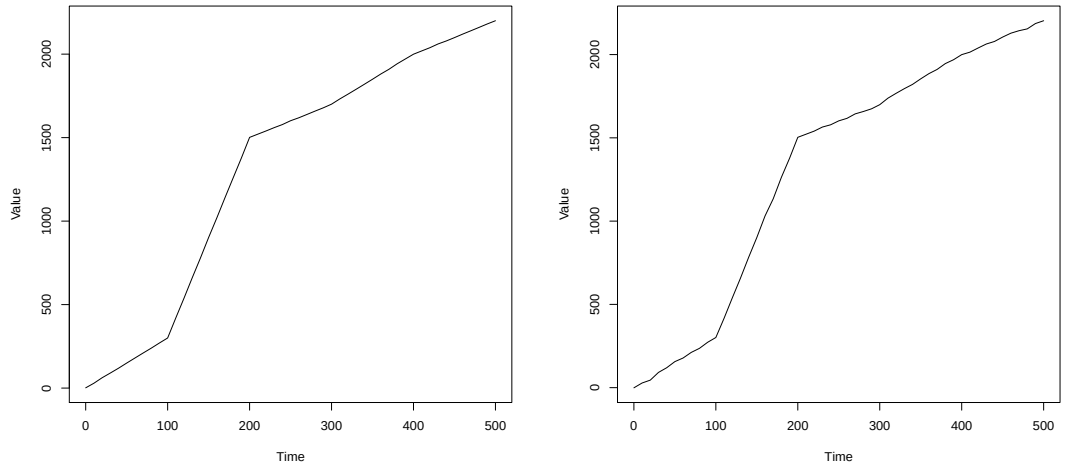

FIGURE 4. Examples of $\mathbf{Y}$ belonging to Cluster 3 of Model 1 for $\sigma = 1$ (left) and $\sigma = 5$ (right).

positions of the true change-points are retrieved most of the time, some additional spurious change-points are also selected with a very low frequency.

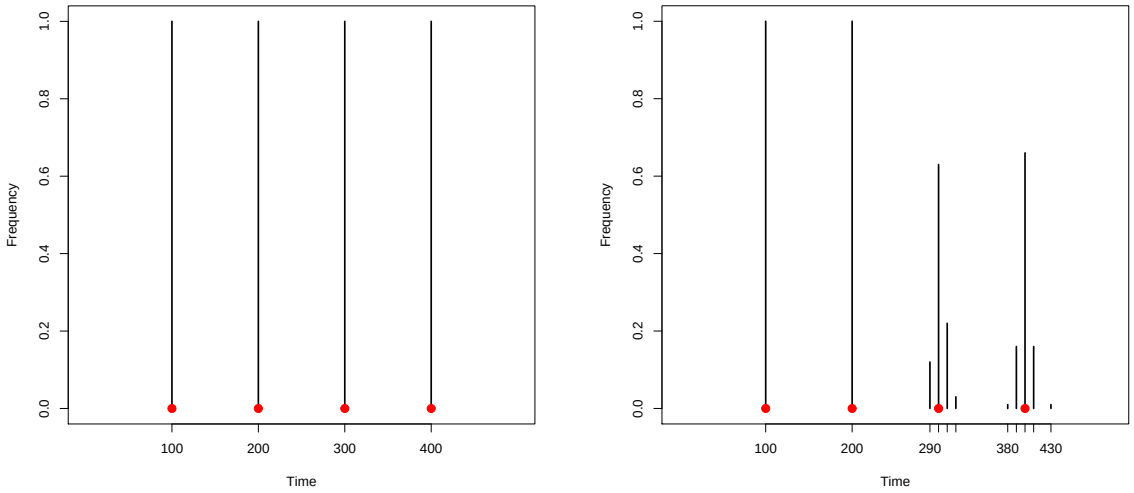

FIGURE 5. Change-point estimation frequencies for $\sigma = 1$ (left) and $\sigma = 5$ (right). The true change-point positions are denoted with red plain circles.

4. Application

In this section, we apply the methodology described in Section 2 to milking kinetics of dairy goats coming from the experimental herd of the research unit Systemic Modelling Applied to Ruminants (Paris, France).

4.1. **Data description.** The data set contains 100470 milking kinetics of goats of two different breeds: "Alpine" and "Saanen". All these kinetics are morning milking kinetics and several kinetics are available for each goat. The kinetics can also be separated according to parity which corresponds to the lactation rank *i.e.* to the number of times a goat has given birth and started a new lactation. In the considered dataset, there are in particular 276 (resp. 191) goats for which we have their milking kinetics for Parity 1 (resp. 2).

4.2. **Kinetics clustering.** First, note that based on the shapes of the milking kinetics of this data set, the parameter $K_{\max}$ defined in the first stage of the first step in Section 2 was set to 2. We obtained three clusters containing 57498, 36757 and 6215 kinetics, respectively. Some examples of kinetics belonging to Clusters 1, 2 and 3 are displayed in Figures 6, 7 and 8, respectively. The average of the kinetics estimations obtained within each cluster is displayed in Figure 9. We can observe that the three clusters can be distinguished in terms of quantity of milk production: Cluster 1 has the lowest production, Cluster 3 the highest and Cluster 2 is between them.

Another difference between the three clusters is the number and the positions of changes. The number of changes in the kinetics of Cluster 1 and 2 is mainly one contrary to Cluster 3 where this number is always equal to two. Figure 10 displays the histogram of the change-point positions for Clusters 1 and 2. We can observe that the change-point having the highest frequency is not located at the same position for these two clusters. Interestingly, our methodology was able to distinguish these two clusters thanks to the change-point position which illustrates the potential of our methodology to extract synthetic traits from raw data.

In practice, such a clustering may be very useful in the precision farming context to refine selection criteria for breeding programs, to simplify milking workload or to control udder health. Thanks to the clustering results, we should be able to define a milking profile for each goat. Moreover, we propose in the next section to characterize dairy goats belonging to a given parity.

4.3. **Parity characterization.** In order to go further into this analysis, we tried to characterize the parities 1 and 2 in terms of the proportion of kinetics of type 1, 2 or 3 according to the clustering previously obtained. We thus created for each goat belonging to a given parity a vector of proportions corresponding to its belonging frequency to each Cluster 1, 2 or 3. For each parity, the goats are clustered using the k-means algorithm applied to the vectors of proportions. The results are displayed in Figures 11 and 12 for Parities 1 and 2, respectively. The number of groups is selected using the method described in Section 2.2: We found 6 (resp. 5) groups for Parity 1 (resp. 2). We can notice that there is one goat which produces a large quantity of milk compared to the others for both parities. In

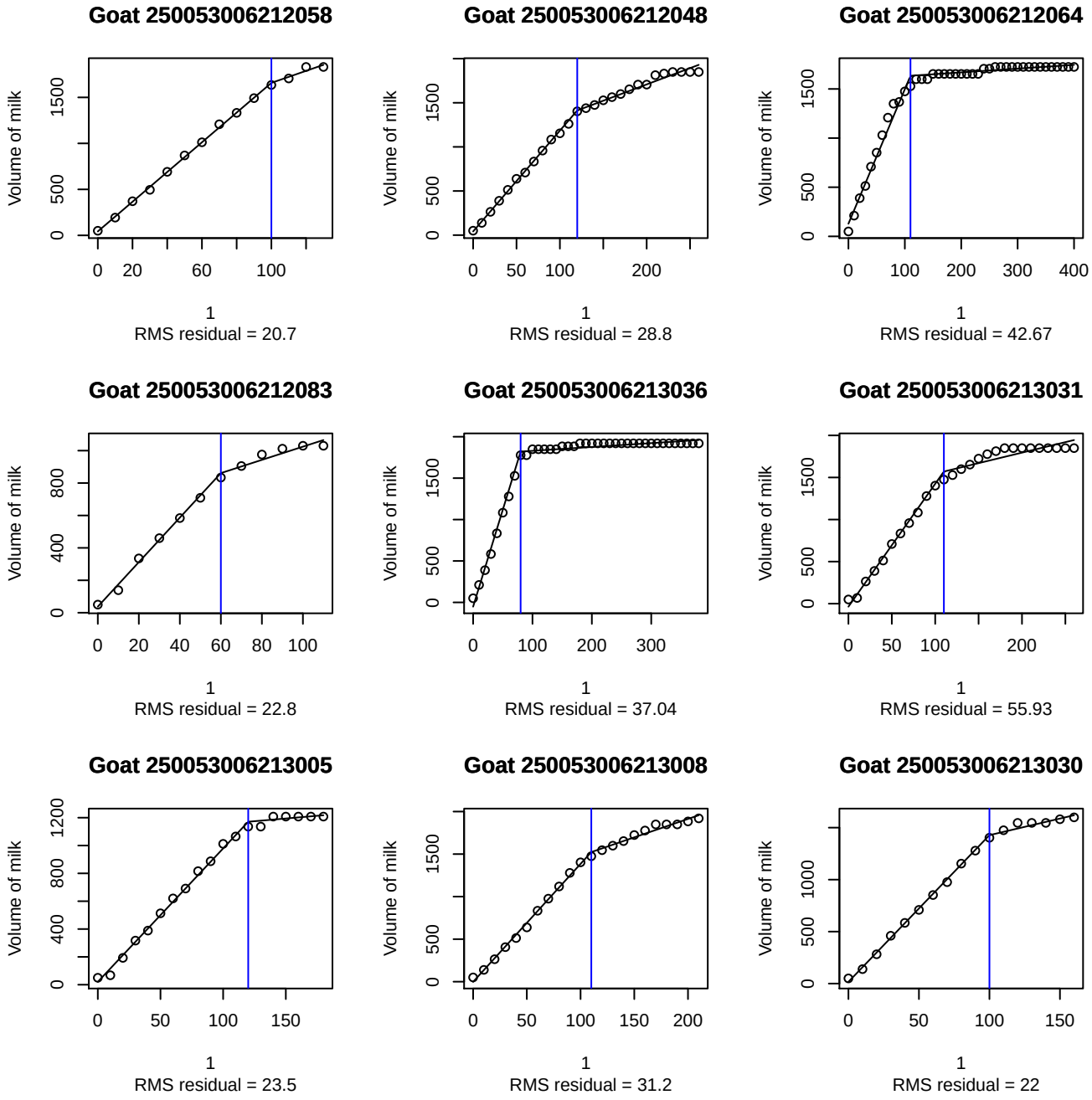

FIGURE 6. Some examples of milking kinetics belonging to Cluster 1. The data are displayed with 'o', the straight lines correspond to the piecewise linear fit obtained thanks to our method and the vertical line corresponds to the position of the change-point.

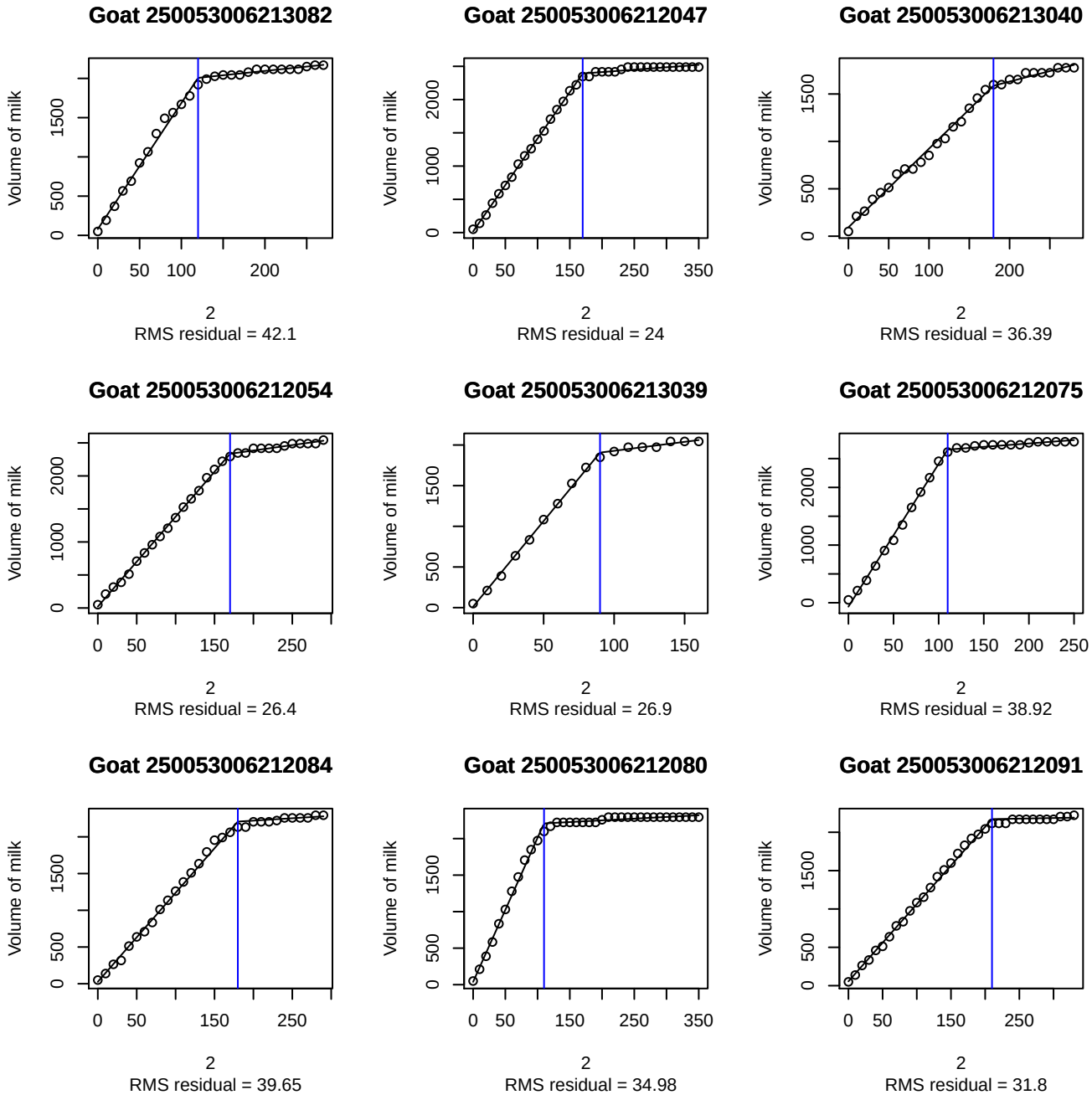

FIGURE 7. Some examples of milking kinetics belonging to Cluster 2. The data are displayed with 'o', the straight lines correspond to the piecewise linear fit obtained thanks to our method and the vertical line corresponds to the position of the change-point.

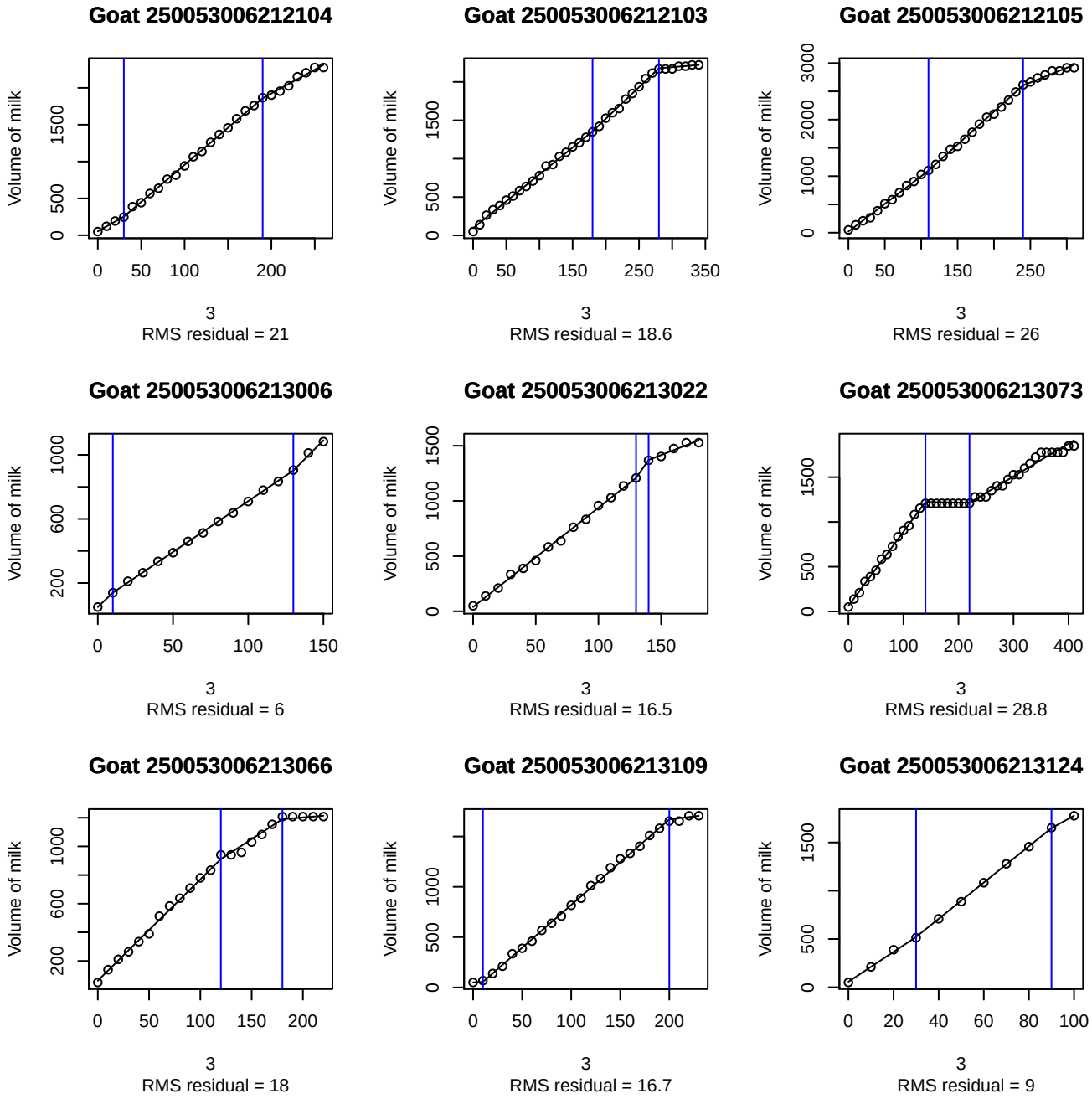

FIGURE 8. Some examples of milking kinetics belonging to Cluster 3. The data are displayed with 'o', the straight lines correspond to the piecewise linear fit obtained thanks to our method and the vertical line corresponds to the position of the change-points.

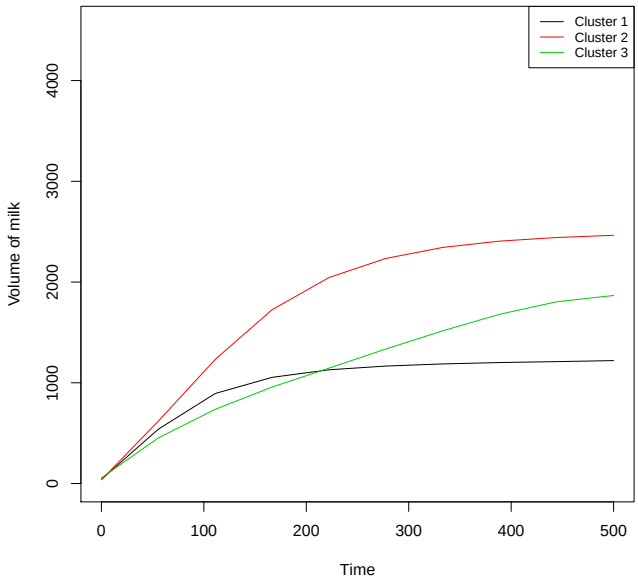

FIGURE 9. Kinetics average obtained within each of the three clusters.

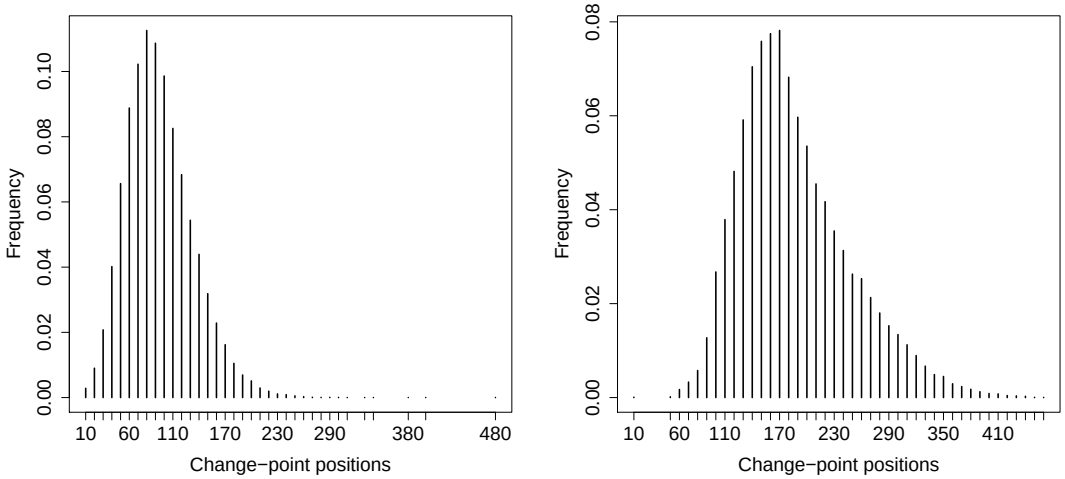

FIGURE 10. Histograms of the change-point positions for Cluster 1 (left) and Cluster 2 (right).

Parity 1, 80% of its milking kinetics belong to Cluster 2 and only 20% to Cluster 1. In Parity 2, 100% of its milking kinetics belong to Cluster 2.

We also observe from Figures 11 and 12 that in both parities, the belonging frequency of the milking kinetics to Cluster 2 is between 50% and 70%. In Parity 2, there is one group (in red) for which the proportion of milking kinetics belonging to Cluster 2 is very high (around 65%) and the proportions of milking kinetics belonging to Cluster 1 and Cluster 3 are very low (around 25% and 13%, respectively). For the other groups the proportions of milking kinetics belonging to Cluster 1 are higher. In Parity 1, the behavior is a little bit different in the sense that the majority of goats have a high proportion of milking kinetics belonging to Cluster 2 (around 65%) and a low proportion of milking kinetics belonging to Cluster 1 (around 25%). Such results may be interesting in the context of precision breeding since they could help to forecast the production of milk at the different parities.

Further analysis should be perfomed in the future to study how evolve the cluster belonging along the lactation course lasting around 150 days in goats. The daily milk yield of a goat for a given parity follows indeed a typical triphasic shape (respectively increasing, plateau and decreasing phase), each daily milk yield being the sum of the total milk produced during each milking (respectively morning and afternoon milking). Being able to link a particular shape at the milking kinetics scale with one at the lactation scale could open perspectives to better characterize individual goats and thus propose options for individual milking management.

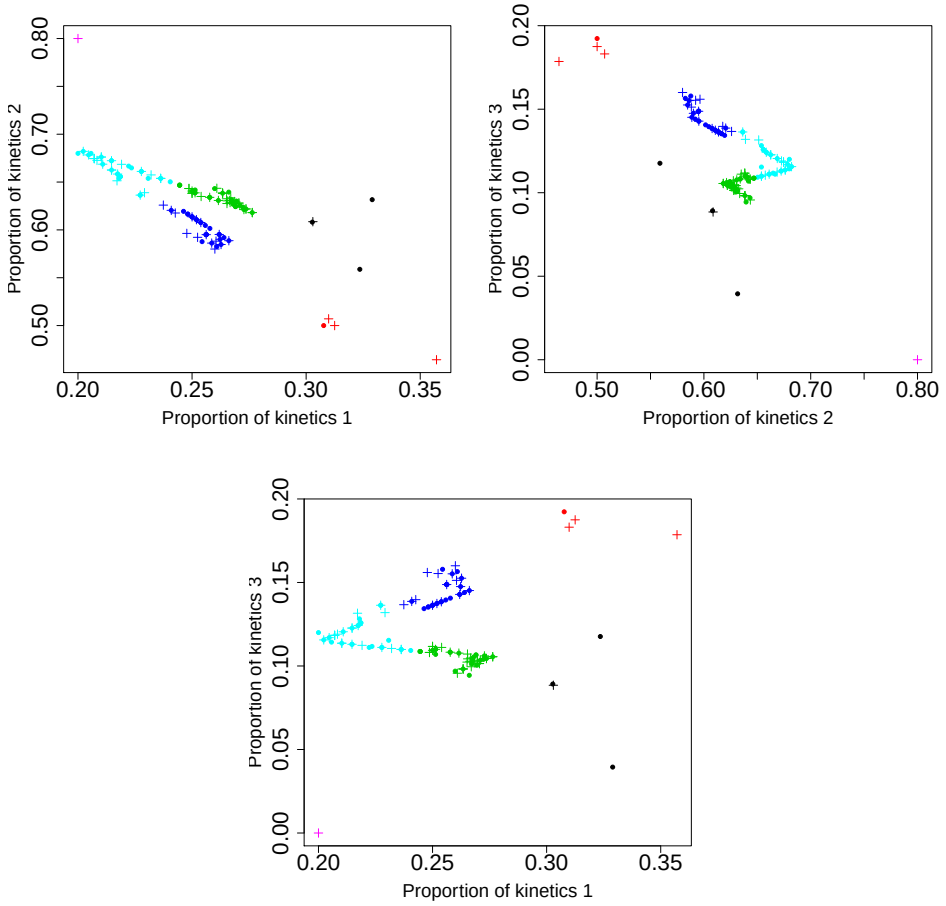

FIGURE 11. Clustering obtained for goats in Parity 1 (6 clusters) displayed on the plane having for axes the proportion of kinetics belonging to Clusters 1 and 2 (top left), 2 and 3 (top right), 1 and 3 (bottom). The Saanen (resp. Alpine) goats are displayed with '•' (resp. '+').

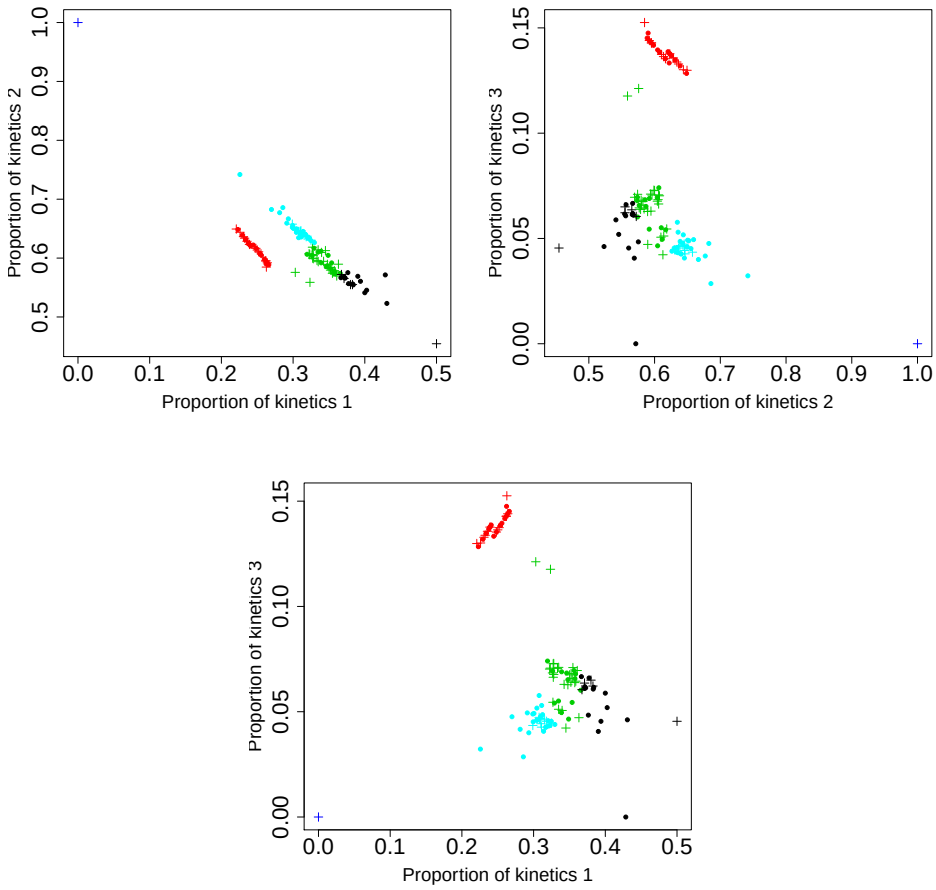

FIGURE 12. Clustering obtained for goats in Parity 2 (5 clusters) displayed on the plane having for axes the proportion of kinetics belonging to Clusters 1 and 2 (top left), 2 and 3 (top right), 1 and 3 (bottom). The Saanen (resp. Alpine) goats are displayed with '•' (resp. '+').

References

Abraham, C., P. A. Cornillon, E. Matzner-Løber, and N. Molinari (2003). Unsupervised curve clustering using b-splines. Scandinavian Journal of Statistics 30(3), 581–595.

Auger, I. and C. Lawrence (1989). Algorithms for the optimal identification of segments neighborhoods. Bull Math Biol 51, 39–54.

Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. J. Appl. Econ. 18, 1–22.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. Commun. ACM 4(6), 284–.

Bouveyron, C., E. Côme, and J. Jacques (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. Ann. Appl. Stat. 9(4), 1726–1760.

Charrad, M., N. Ghazzali, V. Boiteau, and A. Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, Articles 61(6), 1–36.

Fearnhead, P., R. Maidstone, and A. Letchford (2019). Detecting changes in slope with an l0 penalty. Journal of Computational and Graphical Statistics 28(2), 265–275.

Harchaoui, Z. and C. Lévy-Leduc (2007). Catching change-points with lasso. In NIPS, Volume 617, pp. 624.

Harchaoui, Z. and C. Lévy-Leduc (2010). Multiple change-point estimation with a total variation penalty. Journal of the American Statistical Association 105(492), 1480–1493.

Hartigan, J. and M. Wong (1979). Algorithm AS 136: A K-means clustering algorithm. Applied Statistics, 100–108.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning: data mining, inference and prediction (2 ed.). Springer.

Hubert, L. and P. Arabie (1985). Comparing partitions. Journal of Classification 2(1), 193–218.

Jacques, J. and C. Preda (2013). Funclust: a curves clustering method using functional random variables density approximation. Neurocomputing 112, 164–171.

Jacques, J. and C. Preda (2014a). Functional data clustering: A survey. Adv. Data Anal. Classif. 8(3), 231–255.

Jacques, J. and C. Preda (2014b). Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis 71, 92 – 106.

Killick, R., P. Fearnhead, and I. Eckley (2012). Optimal detection of changepoints with a linear computational cost. J. Amer. Statist. Assoc. 107(500), 1590–1598.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. Signal Processing 85(8), 1501 – 1510.

Maidstone, R., T. Hocking, G. Rigaill, and P. Fearnhead (2016). On optimal multiple changepoint algorithms for large data. Statistics and Computing, 1–15.

Marnet, P. G., P. Billon, E. Sinapsis, P. Da Ponte, and E. Manfredi (2005). Machine milking ability in goats: Genetic variability and physiological basis of milk flow rate. 10, pp. 15–24. ICAR Technical Series, Rome, Italy.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4), 417–473.

Picard, F., S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin (2005). A statistical approach for array CGH data analysis. BMC Bioinformatics 6(27), 1.

Ramsay, J. and B. Silverman (2005). Functional data analysis.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations in 1 to Kmax changes. J. SFDS 156(4), 180–205.

Romero, G., R. Panzalis, and P. Ruegg (2017). Relationship of goat milk flow emission variables with milking routine, milking parameters, milking machine characteristics and goat physiology. Animal 11(11), 2070–2075.

Schmutz, A., J. Jacques, C. Bouveyron, L. Cheze, and P. Martin (2018). Clustering multivariate functional data in group-specific functional subspaces.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. Ann. Statist. 42(1), 285–323.

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France and LAMA - Université Paris-Est - Marne-la-Vallée, 77420 Champs-sur-Marne, France
E-mail address: christophe.denis@u-pem.fr

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France
E-mail address: emilie.lebarbier@agroparistech.fr

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France
E-mail address: celine.levy-leduc@agroparistech.fr

UMR Modélisation Systémique Appliquée aux Ruminants, INRA, AgroParisTech, Université Paris-Saclay, 75005, Paris, France
E-mail address: olivier.martin@agroparistech.fr

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France
E-mail address: laure.sansonnet@agroparistech.fr