Unsupervised Curve Clustering using B-Splines

C. ABRAHAM
ENSA-INRA Montpellier

P. A. CORNILLON
Université Rennes II

ERIC MATZNER-LØBER
Université Rennes II

NICOLAS MOLINARI
Universitat Montpellier I

ABSTRACT. Data in many different fields come to practitioners through a process naturally described as functional. Although data are gathered as finite vector and may contain measurement errors, the functional form have to be taken into account. We propose a clustering procedure of such data emphasizing the functional nature of the objects. The new clustering method consists of two stages: fitting the functional data by B-splines and partitioning the estimated model coefficients using a *k*-means algorithm. Strong consistency of the clustering method is proved and a real-world example from food industry is given.

Key words: B-splines, clustering, epi-convergence, functional data, *k*-means, partitioning

1. Introduction

Most data collected by practitioners and scientists in many fields, e.g. biology, meteorology, and industry, are functional. Such examples include growth curves, evolution of temperature and evolution of pH in a food industry process. Ramsay & Silverman (1997) presented several techniques for analysing such data, e.g. principal components analysis, linear modelling, canonical correlation analysis, and proposed some challenges for the future, including asymptotic results for functional data analytic methods. This paper is motivated by two questions: given a sample of n functional data, can one propose a segmentation procedure leading to homogeneous classes and can some asymptotic results as strong consistency be proven for that procedure?

An obvious answer could be the following: consider the measurements of n curves as vectors and use a straightforward clustering algorithm (Hartigan, 1975; Diday *et al.*, 1983). However, there are many reasons against doing so. If the index sets are not exactly the same for the n curves, that technique cannot be employed. If numerous measurements are available, the obvious answer leads to computational problems, especially when complex algorithms such as the hypervolume clustering algorithm (Hartigan & Wong, 1979) are used. Moreover, in the presence of measurement errors, direct clustering methods do not take advantage of the functional structure.

B 7 7 0 0 1 1 2 B
Journal No. Manuscript No.
Dispatch: 14.2.03 Journal: SJOS
Author Received: No. of pages: 15

Thus, it is advantageous to partition the functional data keeping the functional structure. In order to keep the structure of the index set, which usually represents time, it seems natural first to fit the curves by linear or non-linear parametric models and then to partition the model coefficients. Linear models are often too restrictive to fit the underlying phenomenon, and non-linear parametric models are so numerous that finding a relevant parametric model is already a source of problems. In this paper, we propose to fit the functional data by B-splines and then partition the model coefficients using a k-means procedure.

The paper is organized as follows. The next section introduces our working example. In section 3, we present the method: section 3.1 is devoted to the presentation of the B-spline functions and smoothing; section 3.2 to the k-means procedure. In section 4, we prove the strong consistency of our method. Section 5 is devoted to a real example: from a set of n curves measuring the evolution of a cheese product's pH, we construct an appropriate partition leading to homogeneous classes. Each cluster is represented by its centre. We discuss our procedure in the last section.

2. Acidification process in cheese-making

The production of cooked and pressed type cheese such as Comté or Emmental including maturing takes several months. The first process stage that produces young cheese takes about 1 day depending on the cheese, whereas the second stage takes months. The first stage is divided into several steps: milk maturation, coagulation, draining, pressing and salting. The processing of the milk into cheese is characterized by a great number of state variables, evolving under the action of miscellaneous factors. Very few variables are being measured and usually cheese makers focus their attention to the evolution of pH as the acidification plays a key role to achieve a good quality product. The evolution of pH is measured using a pH sensor and stored in a computer. The partitioning of these acidification curves gives an insight into the quality of cheese without having to wait for months and using subjective appraisement of some sensorial criteria on matured cheese.

The data set consists of $n = 148$ observations of pH evolution between 5800 and 70,000 s. Each observational unit (curve) i consists of m_i measurements $\{y_j^i\}_{j=1}^{m_i}$ along different sampling points $\left\{x_j^i\right\}_{j=1}^{m_i}$; each m_i is around 224. Figure 1 represents five of these observations; they are chosen in order to keep the figure clear and to give a representative sample of raw data.

All the 148 observations have not been measured at the same time, thus a direct application of clustering methods to the data is not possible. Another important thing is to take into account the functional aspect of acidification. Thus it seems natural to fit a curve through each observational unit based on measurements.

As we want to design a procedure which can be extended to numerous problems of curve clustering, we do not use a classical parameterization using a sigmoid-type curve (Diday *et al.*, 1983). But as the cheese manufacturer wants a fast procedure on a personal computer, we have to use a parameterization with a few number of coefficients but flexible enough to be applied to a variety of problems. Intensive computing methods such as standard neural network approaches (Muller & Hébrail, 1996; Bock, 1998) do not seem to be suitable for our problem.

Moreover, measurements of the curves G^i are done with errors which can be thought as added to the underlying smooth phenomenon of acidification. The model can be written as:

$$y_j^i = G^i(x_j^i) + \varepsilon_j^i, \tag{1}$$

where ε_j^i are independent random errors. The procedure will have to remove this noisy part to focus on the smooth interesting part: the acidification process. Thus, in conclusion we have to

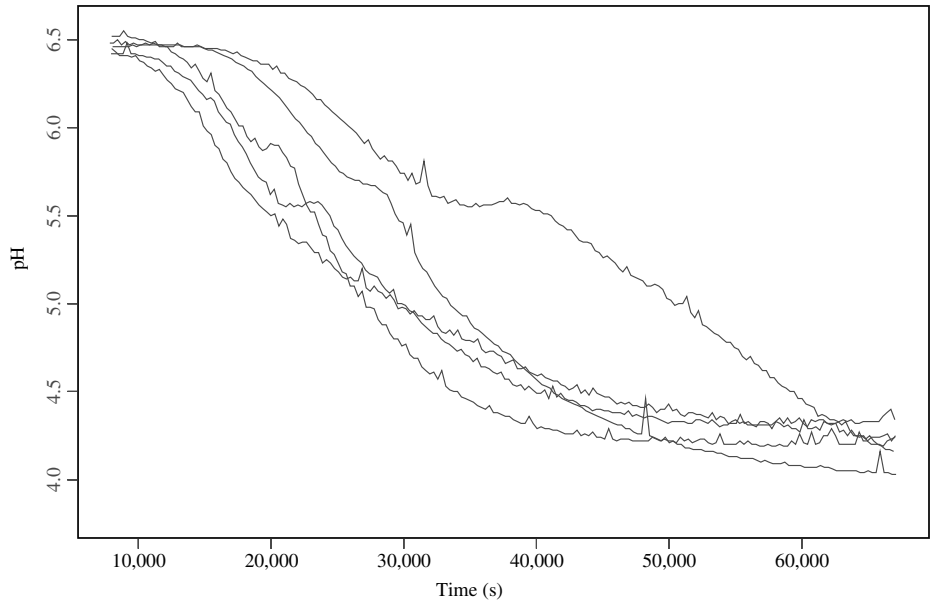

Fig. 1. Original pH evolution for five observations.

(i) summarize each curve by a few coefficients which capture the smooth part of the acidification (with enough flexibility), (ii) partition these coefficients. B-splines fitting seems to be perfectly adequate for the first step and the partitioning is done using a k-means algorithm which is implemented in numerous software.

3. Description of the method

3.1. Curve parameterization

We first fit each observation $\{x_j^i, y_j^i\}_{j=1}^{m_i}$ by a regression spline function in order to estimate G^i. Piecewise polynomials or splines extend the advantages of polynomials to include greater flexibility on estimated functions. Basic references are De Boor (1978) and Schumaker (1981). To make the paper self-contained, we recall some essential background.

Let $x \in [a, b]$ and let $(\xi_0 =)a < \xi_1 < \xi_2 < \cdots < \xi_K < b(=\xi_{K+1})$ be a subdivision by K distinct points on $[a, b]$; these points are called the 'knots'. The spline function $s(x)$ is a polynomial of degree d (or order $d+1$) on any interval $[\xi_{i-1}, \xi_i]$, and has $d-1$ continuous derivatives on the open interval (a, b). For a fixed sequence of knots $\xi = (\xi_1, \xi_2, \ldots, \xi_K)$, the set of such splines is a linear space of functions with $K + d + 1$ free parameters. A useful basis $(B_1, \ldots, B_{K+d+1})$, for this linear space is given by Schoenberg's B-splines, or Basic-splines (Curry & Schoenberg, 1966). We can write a spline as

$$s(x, \beta) = \sum_{l=1}^{K+d+1} \beta_l B_l(x),$$

where $\beta = (\beta_1, \ldots, \beta_{K+d+1})'$ is the vector of spline coefficients and $'$ denotes transposition. A simple expression of B-splines involving classical functions such as exponential, polynomial or logarithm is not possible but computation of their value at a given point is easy with numerous softwares.

A linear combination of (say) third-degree B-splines gives a smooth curve. B-splines are very attractive as basis functions for univariate regression. Using spline functions in a regression model allows the investigation of non-linear effects with continuous covariates. In particular, B-spline basis functions are very appropriate in this case because of the fact that they are numerically well-conditioned, and also because they are locally sensitive to data (De Boor, 1978). Two splines belonging to the same approximating space (same degree $d = 3$, same knots $\xi = (25{,}000,\ 50,\ 000)'$) are drawn in Fig. 2. This clearly shows the flexibility of splines.

With fixed knots, the least-squares spline approximation is equivalent to a linear problem. Once one can compute the B-splines themselves, their application is not more difficult than polynomial regression. Let $(x_j^i,\ y_j^i)_{j=1,\ldots,m_i}$ be a regression type data set of m_i measurements of the curve G^i ranging over $[a,\ b] \times \mathbb{R}$. Denote by $B^i = \{B_l(x_j^i)\}_{j=1,\ldots,m_i}^{l=1,\ldots,K+d+1}$ the corresponding $m_i \times \{K + d + 1\}$ matrix of sampled basis functions, and by y^i the vector $(y_1^i,\ \ldots,\ y_{m_i}^i)'$. It is a straightforward linear least-squares problem to fit the data by splines. We suppose that $B'B$ is non-singular. The spline coefficients are estimated by

$$\hat{\beta}^i := \arg\min_{\beta^i} \frac{1}{m_i} \sum_{j=1}^{m_i} (y_j^i - s(x_j^i, \beta^i))^2 = [(B^i)'B^i]^{-1}(B^i)'y^i, \tag{2}$$

where $[B'B]^{-1}$ is the inverse of $B'B$. Then, $\hat{s}^i(x) := s(x, \hat{\beta}^i)$ estimates $G^i(x)$. The set of curves $\{G^1,\ \ldots,\ G^n\}$ is summarized by $\{\hat{\beta}^1,\ \ldots,\ \hat{\beta}^n\}$, a set of vectors of $\mathbb{R}^{K+d+1}$. As we use the same degree and vector of knots, the same basis functions $(B_1, \ldots, B_{K+d+1})$ are used for the n curves. Thus, each coordinate $\hat{\beta}^i$ has the same meaning for each curve G^i.

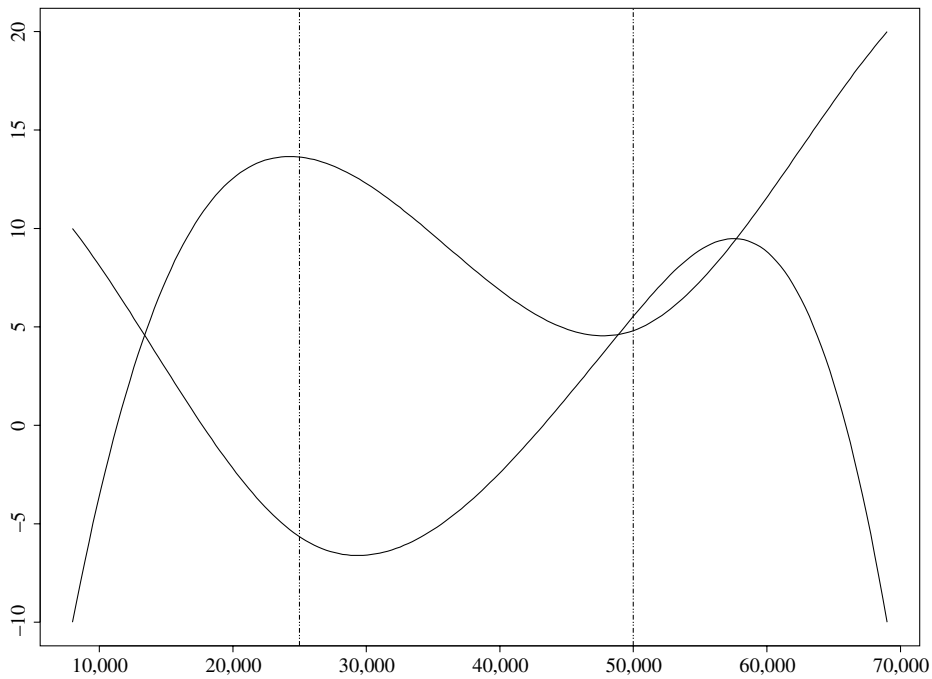

Fig. 2. Two B-splines of degree 3, with two interior knots at $\xi = (25{,}000,\ 50{,}000)'$. Each knot is figured by a vertical dotted line.

3.2. Clustering

In the previous section, we have summarized each curve G^i by its decomposition in the B-splines basis. Thus to partition the n curves into k clusters, where k is given, we just need to partition their coefficients $\hat{\boldsymbol{\beta}}^i \in \mathbb{R}^{K+d+1}$. From the set $\{\hat{\boldsymbol{\beta}}^1, \ldots, \hat{\boldsymbol{\beta}}^n\}$ we want to construct an appropriate partition with homogeneous classes and class representatives $z=\{c^1,\ldots,c^k\}$, where each c^i belongs to $\mathbb{R}^{K+d+1}$. Classical methods such as k-means can be suitably used for this purpose (Hartigan, 1975; Ripley, 1996).

In the following, we describe the k-means clustering procedure. The problem is to choose $z = \{c^1,\ldots,c^k\}$ which minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \min_{c\in z} \|\hat{\boldsymbol{\beta}}^i - c\|^2,$$

where $\|\cdot\|$ denotes the usual Euclidean norm. Note that this problem is equivalent to looking for a partition $\{C^1, \ldots, C^k\}$ of $\{\hat{\boldsymbol{\beta}}^1, \ldots, \hat{\boldsymbol{\beta}}^n\}$ in k classes such that

$$\frac{1}{n}\sum_{j=1}^{k} \sum_{\hat{\boldsymbol{\beta}}^i \in C^j} \|\hat{\boldsymbol{\beta}}^i - c^j\|^2$$

attains its minimum, where c^i is the centre of C^i.

In step 1 of the algorithm, we need initial guesses for the centres of the clusters. In step 2, each $\hat{\boldsymbol{\beta}}^i$ is classified using the actual centres. In step 3, using the result of the previous step, each cluster centre c^j is recomputed as the mean of the $\hat{\boldsymbol{\beta}}^i$ assigned to cluster j. In step 4, if the centres of the clusters are the same then the algorithm stops; otherwise it goes to step 2.

There is no procedure which actually guarantees that the global minimum will be reached. The chosen k-means algorithm finds a stationary point, that is, a solution such that there is no single switch of an observation from one cluster to another cluster that will decrease the function to minimize (see Hartigan & Wong, 1979, for details).

3.3. Outlines of the method

In conclusion, our procedure of unsupervised curves clustering follows the generic strategy. (i) The user chooses the approximating space by deciding the degree d and K interior knots ξ. (ii) For each curve i the matrix B^i (the B-spline values at sampled points $\{x_j^i\}_{j=1}^{m_i}$) is computed. Recall that functions for computing B-spline values at given points are implemented in numerous statistics software. (iii) Then by equation (1) the B-spline coefficients $\hat{\boldsymbol{\beta}}^i$ are calculated. At that stage each curve i is summarized by $K+d+1$ coefficients. (iv) The n vectors $\hat{\boldsymbol{\beta}}^i$ of $\mathbb{R}^{K+d+1}$ are clustered by a k-means algorithm.

4. Strong consistency of *k*-means clustering

This section examines the asymptotic behaviour of the set z of the k centres of the clusters. First, we prove the consistency of the procedure without measurement errors and examine the practical consequences of this proposition. Secondly, we prove the strong consistency with errors; this theorem requires additional assumptions which are explained on a practical ground. Before that, we need to introduce some notation. Let v be a positive measure on $[a, b]$ (its use will be specified later) and L^2 the usual Hilbert space of functions f from $[a, b]$ into $\mathbb{R}$ such that

$$\|f\| = \left(\int_a^b f(x)^2 v(dx) \right)^{1/2} < \infty.$$

Let $(G^n)_n$ be a sequence of independent identical distributed (i.i.d) random functions from a probability space $(\Omega, \mathscr{A}, P)$ into $(L^2, \mathscr{B})$ where $\mathscr{B}$ is the Borel σ-field. For every $f \in L^2$, let $\Pi(f)$ be the vector of coordinates of the orthogonal projection from L^2 onto the vector subspace $\mathscr{S}$ generated by the B-spline basis $(B_1, \ldots, B_{K+d+1})$, which is the approximating subspace. Thus, $\Pi(f)$ is the unique $\beta \in \mathbb{R}^{K+d+1}$ such that

$$\inf_{\beta \in \mathbb{R}^{K+d+1}} \|f - s(\cdot, \beta)\| = \|f - s(\cdot, \Pi(f))\|.$$

Let $\mathscr{B}_{\mathbb{R}^{K+d+1}}$ and μ denote, respectively, the Borel σ-field of $\mathbb{R}^{K+d+1}$ and the image measure of P induced by Π. As Π is continuous $(\mathbb{R}^{K+d+1}, \mathscr{B}_{\mathbb{R}^{K+d+1}}, \mu)$ is a probability space. The sequence $(G^1, G^2, \ldots, G^n)$ induces a sequence $\underline{\beta}^n = (\beta^1, \beta^2, \ldots, \beta^n)$ of i.i.d random vectors $\beta^i = \Pi(G^i)$ in $\mathbb{R}^{K+d+1}$.

The k-means procedure associates to each $\underline{\beta}^n$ a centre $z = \{c^1, \ldots, c^k\}$ $\mathbb{R}^{K+d+1}$ such that

$$u_n(\underline{\beta}^n, z) := \frac{1}{n} \sum_{i=1}^n \min_{c \in z} \|\beta^i - c\|^2$$

is minimized. The following proposition 1 asserts the consistency of this procedure. In order to keep similar notations as in Lemaire (1983), let

$$F = \{z \subset \mathbb{R}^{K+d+1} \,|\, \text{card } z \leq k\},$$

$$u(\beta, z) = \min_{c \in z} \|\beta - c\|^2 \quad \text{and} \quad u_n(\underline{\beta}^n, z) = \frac{1}{n} \sum_{i=1}^n u(\beta^i, z),$$

for all $\beta \in \mathbb{R}^{K+d+1}$, $c \in \mathbb{R}^{K+d+1}$ and $z \in F$. Let $(M_n)_n$ be any increasing sequence of convex and compact subsets of $\mathbb{R}^{K+d+1}$ such that $\mathbb{R}^{K+d+1} = \bigcup M_n$ and let $(z^n)_n$ be a sequence of minimizers of $u_n(\underline{\beta}^n, \cdot)$ with the constraint $z^n \subset M_n$:

$$u_n(\underline{\beta}^n, z^n) = \inf_{z \subset M_n} u_n(\underline{\beta}^n, z).$$

Proposition 1 shows that this sequence is strongly consistent. Furthermore, the limit is a minimizer of

$$u(z) = \int_{\mathbb{R}^{K+d+1}} u(\beta, z) \mu(d\beta).$$

Recall that $(z^n)_n$ is a sequence of sets with k elements of $\mathbb{R}^{K+d+1}$. The convergence of $(z^n)_n$ is taken with respect to the Hausdorff metric h. This metric h is defined for compact subsets A and B of $\mathbb{R}^{K+d+1}$ by: $h(A, B) < \delta$ if and only if every point of A is within distance δ of at least one point of B, and vice versa. Let $\mathscr{B}_F$ denotes the Borel σ field of F derived from the Hausdorff metric. We need the following technical assumption:

(A1) $$(\inf\{u(z) \,|\, z \in F\} < \inf\{u(z) \,|\, z \in F, \text{ card } z < k\}.$$

Assumption A1 is less restrictive than the one in Pollard (1981) as was pointed out by Lemaire (1983) and is needed in the proof that all the minimizers of $u(\cdot)$ belong to a compact set.

Proposition 1
Under (A1), the (unique) minimizer $z^\star$ of u exists and there also exists a unique sequence of measurable functions z^n from $(\Omega, \mathcal{A}, P)$ into $(F, \mathcal{B}_F)$ such that $z^n(\omega) \subset M_n$ for all $\omega \in \Omega$ and

$$u_n(\underline{\beta}^n, z^n) = \inf_{z \subset M_n} u_n(\underline{\beta}^n, z) \quad a.s.$$

Furthermore, this sequence $(z^n)_n$ is strongly consistent to $z^\star$: $\lim_n h(z^n, z^\star) = 0$ a.s.

The proof of this proposition is given in appendix 2. Proposition 1 shows that if we choose an approximating space, and if we are able to calculate perfectly the projection on this subspace, then our procedure is stable as we get more and more curves: the centres of the clusters, z^n, converge to an unique cluster set $z^\star$.

This proposition is not sufficient for our problem because, in practice, one does not observe the whole curve G^i but only the curve at some points x_j^i for $j = 1, \ldots, m_i$. Furthermore, because of measurement errors, it is more realistic to allow that the practitioner observes $G^i(x_j^i)$ with an error ε_j^i. As a consequence, we take the model given in (1) which we recall here:

$$y_j^i = G^i(x_j^i) + \varepsilon_j^i,$$

where ε_j^i are i.i.d random variables with $\mathbb{E}\varepsilon_j^i = 0$ and $\mathbb{E}(\varepsilon_j^i)^2 = \sigma^2$. Thus, for each curve G^i, the data $x^i = (x_1^i, \ldots, x_{m_i}^i)'$ and $y^i = (y_1^i, \ldots, y_{m_i}^i)'$ induces an estimated spline $\hat{s}^i(\cdot) = s(\cdot, \hat{\beta}^i)$ as described in Section 3.1. Let

$$u_n(\underline{\hat{\beta}}^n, z) = \frac{1}{n}\sum_{i=1}^{n} u(\hat{\beta}^i, z)$$

where $\underline{\hat{\beta}}^n = (\hat{\beta}^1, \ldots, \hat{\beta}^n)$. As above, let $\hat{z}^n \in F$ be the minimizer of $u_n(\underline{\hat{\beta}}^n, \cdot)$ on the compact M_n. Clearly, the probability distribution of $\hat{\beta}^i$ depends on the point sequence x^i. If all the sequences x^i are identical, then $\hat{\beta}^i$ for $i = 1, \ldots, n$ are i.i.d random vectors of $\mathbb{R}^{K+d+1}$. Thus using arguments similar to those used in the proof of proposition 1, we can prove the strong consistency of the sequence $(\hat{z}^n)_n$. Nevertheless, most of the time the sequences x^i are not identical. For that reason, we suppose that every sequence x^i is defined as the first m_i elements of an infinite sequence in $[a, b]$. So, if $m = \min\{m_1, \ldots, m_n\}$ goes to infinity, it can be proved that all the $\hat{\beta}^i$ are close to $\beta^i = \Pi(G^i)$ or that the sequence $u_n(\underline{\hat{\beta}}^n, \cdot) \to u_n(\underline{\beta}^n, \cdot)$ a.s. when $m \to \infty$ as soon as each empirical distribution associated with x^i has the same limit v. Then, we have to prove the convergence of a sequence of minimizers $\hat{z}^n$ of $u_n(\underline{\hat{\beta}}^n, \cdot)$ to z^n, which is a minimizer of $u_n(\beta^n, \cdot)$. Finally, the consistency of $\hat{z}^n$ when m and n go to infinity is deduced from the consistency of z^n.

To prove the strong consistency, we need some additional assumptions on the design in order to get the same information on all the curves, at least when the number of measurements goes to infinity. As in Van de Geer (2000), we suppose that the design is random.

(A2) $x_1^i, \ldots, x_{m_i}^i$ are i.i.d with probability distribution v. The functions $B_1, \ldots, B_{K+d+1}$ of the B-spline basis are linearly independent on the support of v.

We need also a technical assumption on the space of functions of the unknown curves G^i.

(A3) The unknown curves G^i, $1 \leq i \leq n$ belong to the space $\mathcal{G}$ of continuous finite functions with bounded variations on $[a, b]$.

Obviously, we need some assumption on the errors.

(A4) The errors are i.i.d. with zero mean and finite variance σ^2. They are also independent of curves and independent of the design.

Theorem 1
If all the assumptions (A1)–(A4) are satisfied, for every n, if m is sufficiently large, the set $\operatorname{argmin}_{z \subset M_n} u_n(\underline{\hat{\beta}}^n, \cdot)$ *is nonempty. For all* $\omega \in \Omega$, *let* $\hat{z}_m^n(\omega)$ *be a minimizer of* $u_n(\underline{\hat{\beta}}^n(\omega), \cdot)$ *with the constraint that* $\hat{z}_m^n(\omega) \subset M_n$, *then* $\lim_n \lim_m h(\hat{z}_m^n, z^\star) = 0$ *a.s.*

This theorem expresses that, if n is sufficiently large, there exists m sufficiently large so that $\hat{z}_m^n$ is arbitrarily close to $z^\star$ which is the unique minimizer of u. Thus, even with measurement errors, and sampling points which can vary from curve to curve, our procedure is stable.

The technical assumption (A3) ensures that the space of functions G^i is not too large. As we measure pH the underlying G^i are obviously bounded by 14. In a more general framework, as this procedure is designed to be used on computers, functions have to be bounded. Moreover, as we are interested in a smooth underlying function G^i, the assumption of bounded variation is not restrictive.

5. Modelling and clustering the acidification process

Recall that we want to segment a data set of $n = 148$ curves of acidification. The main goal of this study is to get a better knowledge of the process of cheese-making through the evolution of pH. The cheese manufacturer has chosen $k = 3$ clusters. In order to implement our method we have to choose an approximating space by choosing the degree of B-splines and the set ξ of interior knots. Usual choice for degree is less or equal to 3. According to the number of sampled points per curve (224) we can afford degree 3 to get the maximum of flexibility. The set ξ of interior knots has to be suitably chosen according to the localization of user's interest: if all the sampling intervals were of equal importance, then equidistant knots will be appropriated. Usually only parts of the intervals are known to be important. Thus, interior knots have to be spread along these intervals. According to the number of sampled points and the prior knowledge of manufacturer we choose B-splines of degree 3 and 7 interior knots $\xi = (10{,}000, 14{,}000, 18{,}000, 22{,}000, 28{,}000, 40{,}000, 55{,}000)'$. Spline estimations for the same five chosen curves given in Fig. 2 are shown in Fig. 3.

We can notice that the regression on the B-splines basis capture easily the shape of the curves, even when they are not as regular as a sigmoid. We use the S-Plus k-means function to partition the 148 vectors $\hat{\beta}$ into three groups. The corresponding curve clustering is presented in Fig. 4.

It is easily seen (Fig. 5) that there are three types of curves: those quickly acidifying, those leading to final pH higher than the others, and the intermediate cluster which has a slower acidification. The cluster 3, with high final pH, usually leads to bad matured cheese compared with the two other clusters. The cluster 1, with low final pH and fast acidification, usually leads to good quality products. The cluster 2 corresponds to an intermediate quality.

6. Discussion

This paper has presented a new method for partitioning functional data, keeping in mind the functional form of the observations. The proposed method achieves this goal with a parameterization of function using B-splines. As pointed out in previous sections, we are interested in an acidification process which is a continuous and smooth phenomenon. Using splines allows to satisfy this description of the process. Moreover B-splines summarize each curve by a few coefficients, improving stability and rapidity of the methods. Other splines such as smoothing splines are often used for fitting non-linear curves. The most widely used splines are the cubic (smoothing) ones. Recall that calculating cubic splines for the observed measurements $\{y_j^i\}$ at time $\{x_j^i\}$ leads usually to the minimization,

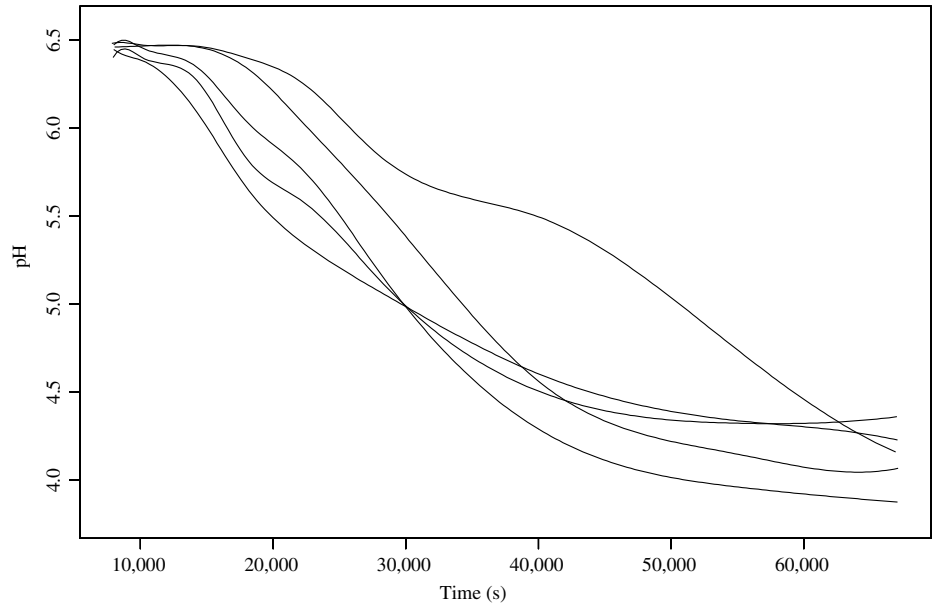

Fig. 3. Spline estimates of the pH evolution for the observations presented in Fig. 1.

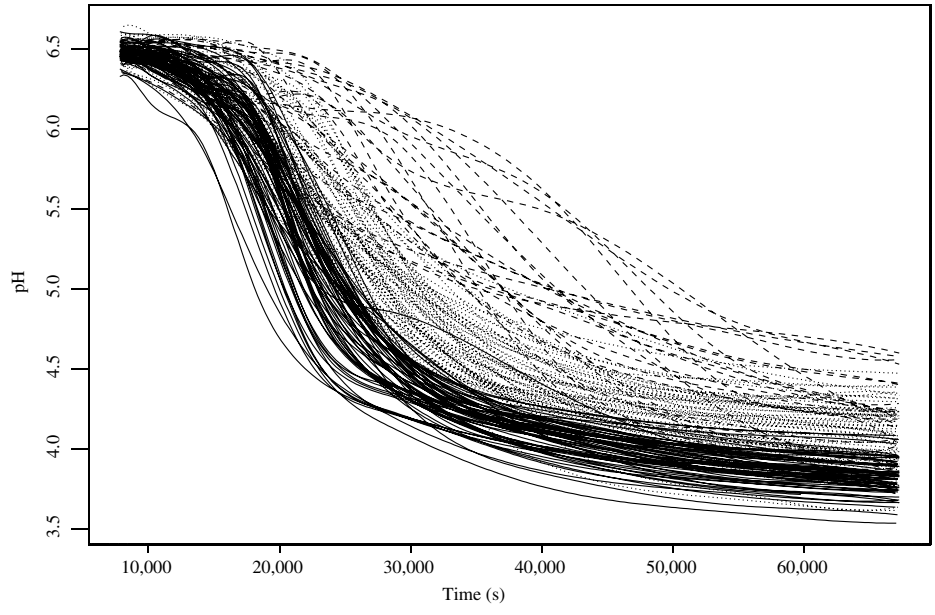

Fig. 4. Results of the clustering procedure.

in the space of functions with continuous second derivative, of the following objective function:

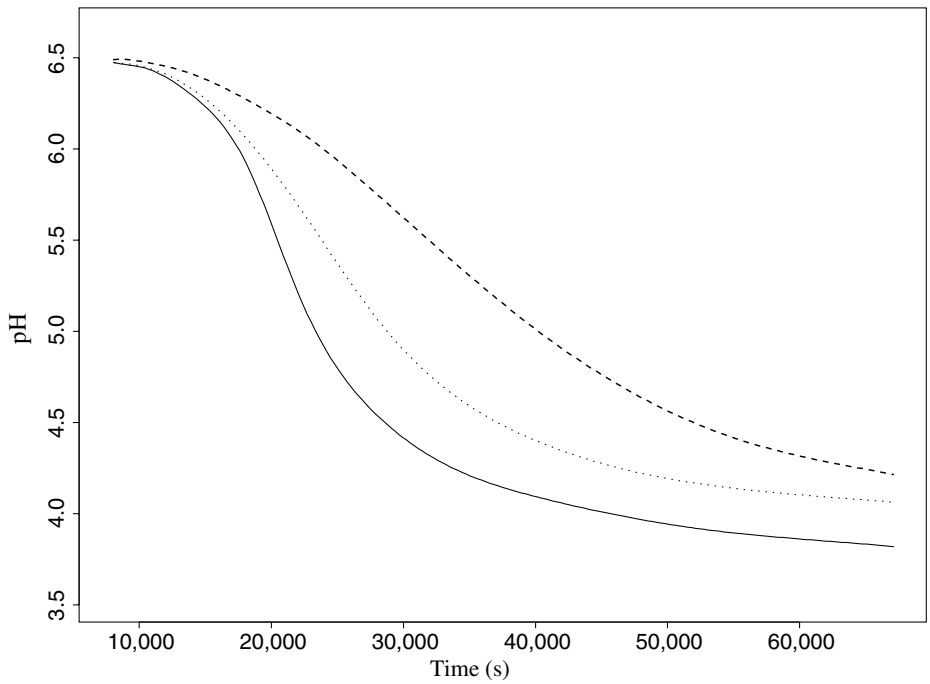

Fig. 5. Centres of the clusters.

$$\sum_{i=1}^{n}(y_j^i - f^i(x_j^i))^2 - \lambda \int_a^b \left(\{f^i\}^{(2)}(x)\right)^2 dx,$$

where $\{f^i\}^{(2)}$ is the second derivative of f^i and $\lambda > 0$ is the positive, fixed smoothing parameter. This criterion corresponds to the least-squares term augmented by a penalty term for lack of smoothness. The minimum of the objective function is obtained with polynomial splines of degree 3 with knots at each measurement x_j^i (see for instance Wegman and Wright, 1983). In order to partition the curves we need parameters which summarize a major part of the information contained within the raw data. These parameters must have the same meaning, that is a common basis for these polynomial splines is needed. That leads to at least as many knots as there are different sampling points $\{x_j^i\}$ minus 2. In the simple case of the same m sampling points $\{x_j\}$ for each curve, we have knots at each interior design points of $[x_1, x_m]$, that is $\xi = (x_2, \ldots, x_{m-1})$. In this case, cubic splines lead to $m + 2$ parameters for each curve which is usually much larger than $K + 4$. For instance, in our working example, if the measurement times are the same for all the curves, we would have $m + 2 = 226$ parameters with cubic splines compared with $K + 4 = 11$ for B-splines. In our context, using B-splines directly is far more appropriate than smoothing (cubic) splines.

Why using the B-splines instead of classical polynomial regression or non-linear regression? B-splines with only a few coefficients can capture a lot of different shapes. This is an important feature as we want to analyse patterns of acidification. B-splines have a local support, that is the domain of abscissa (time), where a B-spline is non-null, is a very small domain. Thus, outliers or small changes in data in one part of the time domain do not affect other parts of the domain. This kind of robustness is a key point of our method. As we only use coefficients,

their estimation should be robust, and this requirement is achieved by using B-spline regression. Polynomial regression does not have this property and small changes of the data can dramatically affect the coefficients.

Another advantage of this method is the weight that can be used in k-means (Diday *et al*., 1983). The B-splines have local support, thus every coefficient represents part of the time domain. Weighting each coefficient, and thus part of the time domain, can emphasize a selected time period, which is important for the user. In cheese-making, we cannot see any practical reason for selecting a particular time period, but it may be different for other users.

We proved the consistency of our method in theorem 1. This theorem is stated in a general way. Thus if we had used other basis functions than B-splines such as Fourier series, polynomial bases, or wavelet bases, the results of the theorem still hold, as soon as the assumptions are satisfied. This theorem allows to say that this method is stable: if we increase the information through more accurate sampling (i.e. shorter sampling intervals) and through additional observational units (curves), then the proposed procedure will converge to a stable minimum. For instance, if we sample a mixture of k curves which comes randomly with error, then the proposed method will find cluster centres which converge to the projection of these k curves on the chosen approximating space.

Finally recall that assumption (A3) is a technical assumption which is fulfilled in our application and in all practical applications. But one can weaken this assumption: working with the entropy of the class of functions (proposition 2) can lead to a broader class of functions.

Acknowledgement

The authors express their gratitude to the two anonymous referees, the associate editor and the editor whose comments greatly improved this paper.

References

Bock, H. H. (1998). Clustering and neural networks. *Advances in data science and classification* (eds A. Rizzi, M. Vichi & H. H. Bock), 265–277. Springer, Berlin.

Curry, H. B. & Schoenberg, I. J. (1966). On Polya frequency functions. IV: The fundamental splines and their limits. *J. Anal. Math.* **17**, 71–107.

De Boor, C. (1978). *A practical guide to splines*. Springer-Verlag, New York.

Diday, E., Lemaire, J, Pouget, J. & Testu, F. (1983). *Eléments d'analyse de données*. Dunod, Paris.

Hartigan, J. A. (1975). *Clustering algorithms*. Wiley, New York.

Hartigan, J. A. & Wong, M. A. (1979). A k-means clustering algorithm. *J. Appl. Statist.* **28**, 100–108.

Lemaire, J. (1983). Propriétés asymptotiques en classification. *Statistiques et analyse des données* **8**, 41–58.

Muller, C. & Hébrail, G. (1996). Le courboscope: un outil pour visualiser plusieurs milliers de courbes. *Proceedings from the XXIX^{e} journées de statistique*, 600–601. ASU, Carcassonne.

Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9**, 135–140.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer, New York.

Ramsay, J. & Silverman, B. (1997). *Functional data analysis*. Springer, Berlin.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press, Cambridge.

Rockafellar, R. T. & Wets, R. J.-B. (1998). *Variational analysis*. Springer, Berlin.

Schumaker, L. L. (1981). *Spline functions: basic theory*. Wiley, New York.

Van de Geer, S. (1987). A new approach to least-squares estimation, with applications. *Ann. Statist.* **15**, 587–602.

Van de Geer, S. (2000). *Empirical processes in M-estimation*. Cambridge University Press, Cambridge.

Van der Vaart, A. W. & Wellner J. A. (1996). *Weak convergence and empirical processes with applications to statistics*. Springer, New York.

Wegman, E. J. & Wright, I. W. (1983). Splines in statistics. *J. Amer. Statist. Assoc.* **78**, 351–365.

Received October 2000, in final form September 2002

Eric Matzner-Løber, UFR Sciences Sociales, Université Haute Bretagne, 6 avenue G. Berger, 35 043 Rennes Cedex, France.
E-mail: eml@ensai.fr

Appendix

We introduce first the following notation for all the demonstrations given in this section: denote by $\mathcal{S}$, the finite dimensional vector space of real functions generated by the B-spline basis $(B_1, \ldots, B_{K+d+1})$; let $p = K + d + 1$ the dimension of approximating space, where K is the number of interior knots, and d the degree of splines.

Appendix 1: Strong consistency of k-means clustering without error

Proof. We first need to prove that the minimizers of $u_n(\underline{\beta}^n, \cdot)$ and u are unique if they exist. Consider the function ϕ from F into $(\mathbb{R}^p)^k$ such that $\phi(z)$ is an ordered vector of $(\mathbb{R}^p)^k$ composed with the elements c_i of $z = \{c_1, \ldots, c_k\}$. The order considered here can be the lexicographical one. Let us also consider the function ψ from $(\mathbb{R}^p)^k$ onto F such that $\psi(c_1, \ldots, c_k) = \{c_1, \ldots, c_k\}$. Obviously, we have $\psi \circ \phi(z) = z$. Let us define a scalar multiple by $\lambda \in \mathbb{R}$, and a sum of elements $z = \{c_1, \ldots, c_k\}$ and $z' = \{c'_1, \ldots, c'_k\}$ of F by

$$\lambda z := \psi(\lambda\phi(z)) = \{\lambda c_1, \ldots, \lambda c_k\},$$
$$z + z' := \psi(\phi(z) + \phi(z')).$$

Note that λz and $z + z'$ also are elements of F. Take $z = \{c_1, \ldots, c_k\}$ and $z' = \{c_1', \ldots, c_k'\}$ in F and suppose that, with no loss of generality, $\phi(z) = (c_1, \ldots, c_k)$ and $\phi(z') = (c'_1, \ldots, c'_k)$. In other words this means that $c_1 \leq c_2 \leq \cdots \leq c_k$ and $c'_1 \leq c'_2 \leq \cdots \leq c'_k$.

$$\begin{aligned} u(\beta, \lambda z + (1-\lambda)z') &= \min_{c \in \lambda z + (1-\lambda)z'} \|\beta - c\|^2 \\ &= \min_{1 \leq i \leq k} \|\beta - (\lambda c_i + (1-\lambda)c'_i)\|^2 \\ &< \min_{1 \leq i \leq k} \lambda\|\beta - c_i\|^2 + (1-\lambda)\|\beta - c'_i\|^2 \\ &\leq \lambda u(\beta, z) + (1-\lambda)u(\beta, z'). \end{aligned}$$

Thus, $u(\beta, \cdot)$ is strictly convex. Consequently, $u_n(\underline{\beta}^n, \cdot)$ and u are also strictly convex. Using the fact that any strictly convex function has only one unique minimizer, the minimizers of $u_n(\underline{\beta}^n, \cdot)$ and u are thus unique if they exist.

The existence of minimizers of u and $u_n(\underline{\beta}^n, \cdot)$ restricted to M_n comes from proposition 7, theorems 1 and 2 of Lemaire (1983). All the assumptions which are required in the paper of Lemaire are clearly satisfied in our context because $\|\beta - c\|^2$ is non-negative.

Appendix 2: Proof of theorem 1

The proof of consistency of $\hat{z}_n$ needs the following proposition. In the sequel, we will omit the dependence on i for simplicity of notation. Thus, m observations $y_1 = G(x_1) + \varepsilon_1, \ldots, y_m = G(x_m) + \varepsilon_m$ are measured with errors $\varepsilon_1, \ldots, \varepsilon_m$ at design points $x_1, \ldots, x_m$. In proposition 2, $G \in \mathcal{G}$ is a non-random curve and $\hat{\boldsymbol{\beta}}$ is the associated estimator of the spline coefficients defined in section 3.

Proposition 2

Under (A2) and (A3), $\hat{\boldsymbol{\beta}}$ converges strongly to $\boldsymbol{\beta} = \Pi(G)$ when $m \to \infty$ uniformly over the space $\mathcal{G}$. Consequently, for almost all $\omega \in \Omega$ and all $G \in \mathcal{G}$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \to 0$ when $m \to \infty$.

Let P_ε denote the distribution of the error and recall that v is the distribution of x_j. We follow the notation of Van de Geer (1987) and write $\|\cdot\|$ and $\|\cdot\|_m$ for the L^2 norm associated with the joint distribution $v \times P_\varepsilon$ and the empirical distribution function based on $(x_1, \varepsilon_1), \ldots, (x_m, \varepsilon_m)$, respectively. Let denote y the function $y(x, \varepsilon) = g(x) + \varepsilon$. Thus,

$$\|g\|^2 = \int g^2(x)v(dx),$$

$$\|y - s(\cdot, \beta)\|^2 = \int (g(x) + \varepsilon - s(x, \beta))^2 v(dx)P_\varepsilon(d\varepsilon) = \|g - s(\cdot, \beta)\|^2 + \|\varepsilon\|^2,$$

$$\|g\|_m^2 = \frac{1}{m}\sum_{j=1}^{m} g^2(x_j)$$

$$\|y - s(\cdot, \beta)\|_m^2 = \frac{1}{m}\sum_{j=1}^{m}(g(x_j) + \varepsilon_j - s(x_j, \beta))^2 = \|\varepsilon - (g - s(\cdot, \beta))\|_m^2.$$

We shall prove first the following uniform strong law of large numbers on the space $\mathcal{F} = \{y - s(\cdot, \alpha), g \in \mathcal{G}, s \in \mathcal{S}\}$ where $\mathcal{S}$ and $\mathcal{G}$ are defined in section 4:

$$\sup_{g \in \mathcal{G}, s(\cdot,\alpha) \in \mathcal{S}} \left| \|y - s(\cdot, \alpha)\|_m^2 - \|y - s(\cdot, \alpha)\|^2 \right| \to 0 \text{ almost surely.} \tag{3}$$

To prove this, we need to apply lemma 2.1 of Van de Geer (1987). First note that the supremum is measurable according to problem 3 of Pollard (1984, p. 38). Secondly, the envelope condition of this lemma, that is $\sup_{g \in \mathcal{G},\ s \in \mathcal{S}} |g - s(\cdot, \alpha)| \in L^2(v)$, is obviously fulfilled using (A3). Thirdly, let us prove that the entropy condition is fulfilled as well. Let v_m be the empirical distribution function based on $x_1, \ldots, x_m$. Assumption (A3) ensures that the entropy condition, $\log N_2(\delta, v_m, \mathcal{G})/m \to_P 0$, is fulfilled (Van de Geer, 1987), where $N_2(\delta, v_m, \mathcal{G})$ denotes the covering number of space $\mathcal{G}$ for the usual $L^2(v_m)$ distance $d(f, g) = \|f - g\|_m^2$. $\mathcal{S}$ is a finite vector space, thus the class of graph of functions from $\mathcal{S}$ have polynomial discrimination and $N_2(\delta, v_m, \mathcal{S})/m$ is less than a fixed polynomial in δ^{-1}. We refer to Pollard (1984, pp. 27–30) for the definition of the graph of a real-valued function, for the definition of the polynomial discrimination and for the proof of the polynomial discrimination of the class of graphs of functions of a finite dimensional vector space of real functions. Other exposition of empirical processes and entropy can be found in Van der Vaart & Wellner (1996) or Van de Geer (2000). By the definition of $\mathcal{F}$, we can bound for every probability measure Q, $\log N_2(\delta, Q, \mathcal{F})$ by $\log N_2(\delta/2, Q, \mathcal{G}) + \log N_2(\delta/2, Q, \mathcal{S})$ and thus the entropy condition $\log N_2(\delta, v_m, \mathcal{F})/m \to_P 0$ is fulfilled.

Fix $\eta > 0$. Using (3), for almost every $\omega \in \Omega$, for all $g \in \mathcal{G}$, there exists an integer N such that $\forall m > N$, $\sup_{s(\cdot,\beta) \in \mathcal{S}} \left| \|y - s(\cdot, \beta)\|_m^2 - \|y - s(\cdot, \beta)\|^2 \right| < \eta/2$. Let $\hat{\boldsymbol{\beta}}$ stand for the least-

squares estimate of $\Pi(g)$ using the observations of y_j at design points. Using this last inequality and following the same lines as Van de Geer (1987) we have for m sufficiently large:

$$\begin{aligned}\|\varepsilon\|^2 + \|g - s(\cdot, \hat{\beta})\|^2 &= \|y - s(\cdot, \hat{\beta})\|^2 \\ &\leq \|y - s(\cdot, \hat{\beta})\|_m^2 + \frac{\eta}{2} \\ &\leq \|y - s(\cdot, \Pi(g))\|_m^2 + \frac{\eta}{2} \\ &\leq \|g + \varepsilon - s(\cdot, \Pi(g))\|^2 + \eta \\ &= \|\varepsilon\|^2 + \|g - s(\cdot, \Pi(G))\|^2 + \eta,\end{aligned}$$

cancelling out $\|\varepsilon\|^2$ and using Pythagorean theorem we have

$$\|s(\cdot, \Pi(g)) - s(\cdot, \hat{\beta})\|^2 \leq \eta.$$

As the functions $B_1, \ldots, B_{K+d+1}$ are linearly independent on the support of v (A2), $N(\beta) = \|s(\cdot, \beta)\|^2$ is a norm and thus the proof is completed.

Proof of theorem 1. From proposition 2, for almost all $\omega \in \Omega$, and for all $G \in \mathcal{G}$ the sequence $u_n(\underline{\hat{\beta}}^n, \cdot) \to u_n(\underline{\beta}^n, \cdot)$ when $m \to \infty$ (recall that $\underline{\beta}^n = (\beta^1, \ldots, \beta^n)$). Then, we have to prove the convergence of a sequence of minimizers $\hat{z}^n$ of $u_n(\underline{\hat{\beta}}^n, \cdot)$ to z^n which is a minimizer of $u_n(\underline{\beta}^n, \cdot)$.

Let us recall some definitions and results of variational analysis. Let $(g_m)_m$ be a sequence of functions from $\mathbb{R}^k$ into $(-\infty, +\infty]$. This sequence is eventually level bounded if, for every $\alpha \in \mathbb{R}$, there exist a compact K and an integer M such that

$$\bigcup_{m \geq M} \{t \in \mathbb{R}^k \mid g_m(t) \leq \alpha\} \subset K.$$

A function g is the epi-limit of $(g_m)_m$ if at each point $t \in \mathbb{R}^k$

$$\begin{cases} \liminf g_m(t_m) \geq g(t) & \text{for every sequence } t_m \to t, \\ \limsup g_m(t_m) \leq g(t) & \text{for some sequence } t_m \to t. \end{cases}$$

A function g from $\mathbb{R}^k$ into $(-\infty, +\infty]$ is lower semicontinuous if the level sets $\{t \in \mathbb{R}^d \mid g(t) \leq \alpha\}$ are all closed. Finally, it is proper if $g(t) < \infty$ for at least one t. Let us state the main theorem of convergence in minimization (Rockafellar & Wets, 1998, p. 266): if the sequence $(g_m)_m$ is eventually level-bounded and epi-converges to g with g_m and g lower semicontinuous and proper, then, for m sufficiently large, the sets argmin g_m are non-empty and are all included in a same compact set. Furthermore, if $t_m \in$ argmin g_m and if t is a cluster point of $(t_m)_m$ (i.e. there exists a subsequence of $(t_m)_m$ with limit t), then $t \in$ argmin g.

In order to apply the above theorem, we have to define several functions. Let ψ and Φ_n be the following functions

$$\begin{aligned}\psi : (\mathbb{R}^p)^k &\longrightarrow F \\ c &\longrightarrow \{c_1, \ldots, c_k\}\end{aligned}$$

where $c = (c_1, \ldots, c_k)'$ and

$$\begin{aligned}\Phi_n : (\mathbb{R}^p)^n \times (\mathbb{R}^p)^k &\longrightarrow (-\infty, +\infty] \\ (\underline{\beta}^n, c) &\longrightarrow u_n(\underline{\beta}^n, \psi(c)) + I_{K_n}(\psi(c))\end{aligned}$$

where $K_n = \{z \in F \mid z \subset M_n\}$ and $I_{K_n}(z) = 0$ if $z \in K_n$ and $I_{K_n}(z) = \infty$ if $z \notin K_n$. Clearly, finding the set $\operatorname{argmin}_{z \subset M_n} u_n(\beta^n, z)$ is equivalent to finding the set $\operatorname{argmin}_{c \in (\mathbb{R}^p)^k} \Phi_n(\beta^n, c)$. Finally, for all $\omega \in \Omega$, let us define the functions g_m and g by

$$\begin{aligned} g_m(\omega, \cdot) : (\mathbb{R}^p)^k &\longrightarrow (-\infty, +\infty] \\ c &\longrightarrow \Phi_n(\underline{\hat{\beta}}^n(\omega), c), \\ g(\omega, \cdot) : (\mathbb{R}^p)^k &\longrightarrow (-\infty, +\infty] \\ c &\longrightarrow \Phi_n(\underline{\beta}^n(\omega), c). \end{aligned}$$

Recall that $\underline{\hat{\beta}}^n(\omega)$ depends on m although this dependence is not explicitly written in order to avoid messy notations. Now, we can apply the theorem of convergence in minimization.

Take $\omega \in \Omega$ such that $\underline{\hat{\beta}}^n(\omega) \to \underline{\beta}^n(\omega)$ when $m \to \infty$ (which is verified for almost all ω and for all $G \in \mathcal{G}$ by proposition 2). Thus for all $G \in \mathcal{G}$, for all $c \in (\mathbb{R}^p)^k$ and all sequences $(c_m)_m$ such that $c_m \to c$, by the continuity of u_n and ψ, we have

$$\lim_{m \to \infty} g_m(\omega, c_m) = g(\omega, c).$$

Thus $g(\omega, \cdot)$ is clearly the epi-limit of $g_m(\omega, \cdot)$. As $g(\omega, \cdot)$ and $g_m(\omega, \cdot)$ take finite values on K_n, they are proper. As $\psi^{-1}(K_n) = \prod_{i=1}^k M_n$ and

$$\{c \in (\mathbb{R}^p)^k \mid g_m(\omega, c) \leq \alpha\} \subset \psi^{-1}(K_n)$$

for all $\alpha \in \mathbb{R}$, it follows that $g_m(\omega, \cdot)$ is eventually level bounded. We next prove that $g_m(\omega, \cdot)$ and $g(\omega, \cdot)$ are lower semicontinuous.

By the continuity of u_n and ψ, the set

$$\{c \in (\mathbb{R}^p)^k \mid g_m(\omega, c) \leq \alpha\} = \{c \in (\mathbb{R}^p)^k \mid u_n(\underline{\hat{\beta}}^n(\omega), \psi(c)) \leq \alpha\} \cap \psi^{-1}(K_n)$$

is closed and so $g_m(\omega, \cdot)$ is lower semicontinuous. We can use similar arguments to prove that $g(\omega, \cdot)$ is also lower semicontinuous. Then, by the theorem of convergence in minimization, we conclude that, for m sufficiently large, $\operatorname{argmin} g_m(\omega, \cdot)$ is non-empty, and every cluster point of any sequence $c_m(\omega)$ of minimizer of $g_m(\omega, \cdot)$ is a minimizer of $g(\omega, \cdot)$.

Note that

$$\operatorname{argmin}_{z \subset M_n} u_n(\underline{\hat{\beta}}^n(\omega), z) = \psi(\operatorname{argmin} g_m(\omega, \cdot)). \tag{4}$$

Clearly for m large, $\operatorname{argmin}_{z \subset M_n} u_n(\underline{\hat{\beta}}^n(\omega), z)$ is non-empty. Now, take $(\hat{z}_m^n(\omega))_m$ a sequence of minimizers of $u_n(\underline{\hat{\beta}}^n(\omega), \cdot)$ included in M_n. There exists a sequence $(c_m(\omega))_m$ such that $\hat{z}_m^n(\omega) = \psi(c_m(\omega))$ and $c_m(\omega)$ is a minimizer of $g_m(\omega, \cdot)$. Let $z^n(\omega)$ be the unique minimizer of $u_n(\beta^n(\omega))$. Using an equation similar to equation (4), the set

$$\operatorname{argmin} g(\omega, \cdot) = \psi^{-1}(z^n(\omega))$$

is a finite set and $(c_m(\omega))_m$ has a finite number of cluster points. Then, for all $\varepsilon > 0$, there exists M such that, for all $m > M$, $c_m(\omega)$ is close to a cluster point so that $h(\hat{z}_m^n(\omega), z^n(\omega)) < \varepsilon$. In other words, $\hat{z}_m^n(\omega) \to_m z^n(\omega)$.

By proposition 1, we conclude that

$$\lim_n \lim_m h(\hat{z}_m^n(\omega), z^\star) = 0,$$

for almost all ω.

Author Query Form

Journal: SJOS

Article: 00_112

Dear Author,
During the preparation of your manuscript for publication, the questions listed below have arisen. The numbers pertain to the numbers in the margin of the proof. Please attend to these matters and return the form with this proof.

Many thanks for your assistance.

Query reference	Query	Remarks
Q1	**de Boor 1978 has been changed to De Boor 1978 so that this citation matches the list**	