Constrained variable clustering for functional data representation

Fabrice Rossi and Yves Lechevallier

Abstract Functional data analysis involves data described by regular functions rather by a finite number of real valued variables. While some robust data analysis methods can be applied directly to the very high dimensional vectors obtained via a fine grid sampling of functional data, all the methods benefit from a prior simplification of the functions that reduces the redundancy induced by the regularity. In this paper we propose to use a variable clustering approach to design a piecewise constant representation of a set of functions. The contiguity constraint induced by the functional nature of the variables leads to an optimal algorithm.

1 Introduction

Functional data [8] appear in applications in which objects to analyse have some form of variability. In spectrometry, for instance, samples are described by spectra: each spectrum is a mapping from wavelengths to e.g., transmittance. Time varying objects offer a more general example: when the characteristics of objects evolve through time, a loss free representation consists in describing these characteristics as functions that map time to values.

In practice, functional data are given as high dimensional vectors (e.g., more than 100 variables) obtained by sampling the functions on a fine grid. For smooth functions (for instance in near infrared spectroscopy), this scheme leads to highly correlated variables. While many data analysis methods can be made robust to this type of problem (see, e.g., [3] for discriminant analysis), all methods benefit from a compression of the data [7] in which relevant and yet easy to interpret features are extracted from the raw functional data.

There are well known standard ways of extracting optimal feature according to a given criterion. For instance in unsupervised problems, the first k principal com-

Projet AxIS, INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France, e-mail: {Fabrice.Rossi,Yves.Lechevallier}@inria.fr

ponents of a dataset give the best linear approximation of the original data in $\mathbb{R}^k$ for the quadratic norm (see [8] for functional principal component analysis (PCA)). In regression problems, the partial least square approach extracts features with maximal correlation with a target variable (see also Sliced Inversion Regression methods [2]). The main drawback of those approaches is that they extract features that are not easy to interpret: while the link between the original features and the new ones are linear, it is seldom sparse; an extracted feature generally depends on many original features.

A different line of thought is followed in the present paper: the goal is to extract features that are easy to interpret in terms of the original variables.

2 Best basis for functional data

Let us consider n functional data, $(s_i)_{1\leq i\leq n}$. Each s_i is a function from $[a,b]$ to $\mathbb{R}$, where $[a,b]$ is a fixed interval common to all functions (more precisely, s_i belongs to $L^2([a,b])$, the set of square integrable functions on $[a,b]$). In terms of functional data, linear feature extraction consists in choosing for each feature a linear operator from $L^2([a,b])$ to $\mathbb{R}$. Equivalently, one can choose a function ϕ from $L^2([a,b])$ and compute $\langle s_i,\phi\rangle_{L^2}=\int_a^b \phi(x)s_i(x)dx$. In an unsupervised context, using e.g., a quadratic error measure, choosing the k best features consists in finding k orthonormal functions $(\phi)_{1\leq i\leq k}$ that minimise the following quantity:

$$\sum_{i=1}^{n}\left\|s_i-\sum_{j=1}^{k}\langle s_i,\phi_k\rangle_{L^2}\phi_k\right\|_{L^2}^2. \tag{1}$$

If the ϕ_k are unconstrained, the optimal basis is given by functional PCA [8]. However, in order for the corresponding feature to be easy to interpret, the ϕ_k should have compact supports, the simple case of $\phi_k=\mathbb{I}_{[u_k,v_k]}$ being the easiest to analyse ($\mathbb{I}_{[u,v]}(x)=1$ when $x\in[u,v]$ and 0 elsewhere).

The problem of choosing an optimal basis among a set of bases has been studied for some time in the wavelet community [1, 10]. In unsupervised context, the best basis is obtained by minimizing the entropy of the features (i.e., of the coordinates of the functions on the basis) in order to enable compression by discarding the less important features. Following [7], [9] proposes a different approach, based on B-splines: a leave-one-out version of Equation (1) is used to select the best B-splines basis. While the orthonormal basis induced by the B-splines does not correspond to compactly supported functions, the dependency between a new feature and the original ones is still localized enough to allow easy interpretation. Nevertheless both approaches have some drawbacks. Wavelet based methods lead to compactly supported basis functions but the basis has to be chosen in a tree structured set of bases. As a consequence, the support of a basis function cannot be any sub-interval of $[a,b]$. The B-spline approach suffers from a similar problem: the approximate sup-

ports have all the same lengths leading either to a poor representation of some local details or to a large number of basis functions.

3 Best basis via constrained clustering

The goal of the present paper is to select an optimal basis using only basis functions of form $\mathbb{I}_{[u,v[}$, without restriction on the possible intervals among sub-interval of $[a,b[$[1]. Let us consider $(\phi_j = \frac{1}{v_j-u_j}\mathbb{I}_{[u_j,v_j[})_{1\leq j\leq k}$ such an orthonormal basis. Orthogonality implies that the $([u_j,v_j[)_{1\leq j\leq k}$ form a partition of $[a,b[$. Moreover, $\langle \phi_j, s_i\rangle = \frac{1}{v_j-u_j}\int_{u_j}^{v_j} s_i(x)dx$, i.e., the feature corresponding to ϕ_j is the mean value of s_i on $[u_j,v_j[$. In other words, $\sum_{j=1}^{k}\langle s_i,\phi_k\rangle_{L^2}\phi_k$ is a piecewise constant approximation of s_i (which is optimal according to the L^2 norm).

In practice, functional data are sampled on a fine grid, i.e., rather than observing the functions $(s_i)_{1\leq i\leq n}$, one gets the vectors $(s_i(t_l))_{1\leq i\leq n,1\leq l\leq m}$ from $\mathbb{R}^m$ (with the $t_l < t_{l+1}$). Then $\langle \phi_j, s_i\rangle$ can be approximated by $\frac{1}{|I_j|}\sum_{l\in I_j} s_i(t_l)$ where I_j is the subset of $\{1,...,m\}$ such that $t_l \in [u_j,v_j[\Leftrightarrow l \in I_j$. Any partition of $([u_j,v_j[)_{1\leq j\leq k}$ of $[a,b[$ corresponds to a partition of $\{1,...,m\}$ in k subsets $(I_j)_{1\leq j\leq k}$ that satisfies an ordering constraint: if r and s belongs to I_j then any integer $t \in [r,s]$ belongs also to I_j. Finding the best basis means for instance minimizing the error measure given by Equation (1) which can be approximated as follows

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l\in I_j}(s_i(t_l) - \frac{1}{|I_j|}\sum_{l\in I_j} s_i(t_l))^2 = \sum_{j=1}^{k} Q(I_j). \tag{2}$$

The second version of the error shows that it corresponds to an additive quality measure of the partition of $\{1,...,m\}$ induced by the $(I_j)_{1\leq j\leq k}$. Therefore, finding the best basis for the sampled functions is equivalent to finding an optimal partition of $\{1,...,m\}$ with some ordering constraints and according to an additive cost function. A suboptimal solution to this problem, based on an ascending hierarchical clustering, is proposed in [4].

However, an optimal solution can be reached in a reasonnable amount of time, as pointed out in [5]: when the quality criterion of a partition is additive and when a total ordering constraint is enforced, a dynamic programming approach leads to the optimal solution. The algorithm is simple and proceeds iteratively by computing $F(j,k)$ as the value of the quality measure (from Equation (2)) of the best partition in k classes of $\{j,...,m\}$:

1. find the best partition in two classes by evaluating $Q(\{1,\ldots,j\})$ and $F(j+1,1) = Q(\{j+1,\ldots,m\})$ for all $1 \leq j \leq m-1$;
2. iterate from $p = 2$ to k:

[1] the exclusion of the right end side of the interval is just a technical trick that prevents cumbersome notations in the rest of the paper.

a. for all $1 \leq j \leq m-p+1$ evaluate $F(j,p)$ by minimizing over l the value of $Q(\{j,\ldots,l\})+F(l+1,p-1)$
b. the best partition with p clusters is the one that realizes $F(1,p)$

The internal loop runs $O(km^2)$ times. It uses the values $Q(\{j,\ldots,l\})$ for all $j \leq l$. Those quantities can be computed prior the search for the optimal partition, using for instance a recursive variance computation formula, leading to a cost in $O(nm^2)$. This algorithm was used to find an optimal basis for a single function in [6].

4 Possible extensions

The previous scheme can be used for any additive quality measure. It is therefore possible to use e.g., a piecewise linear approximation of the functions on a sub-interval rather than a constant approximation. In the case of a regression application, a more interesting solution consists in using as error measure one minus the absolute value of correlation between the feature (the mean of the function on the sub-interval) and the target variable.

In the general case of an arbitrary quality measure Q, there might be no recursive formula for evaluating Q. In this case, the cost of computing the needed quantity might exceed $O(nm^2)$ and reach $O(nm^3)$ or more, depending on the exact definition of Q.

References

1. Coifman, R.R., Wickerhauser, M.V.: Entropy-based algorithms for best basis selection. IEEE Transactions on Information Theory **38**(2), 713–718 (1992)
2. Ferré, L., Yao, A.F.: Functional sliced inverse regression analysis. Statistics **37**(6), 475–488 (2003)
3. Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. Annals of Statistics **23**, 73–102 (1995)
4. Krier, C., Rossi, F., Franois, D., Verleysen, M.: A data-driven functional projection approach for the selection of feature ranges in spectra with ica or cluster analysis. Chemometrics and Intelligent Laboratory Systems **91**(1), 43–53 (2008)
5. Lechevallier, Y.: Classification automatique optimale sous contrainte d'ordre total. Rapport de recherche 200, IRIA (1976)
6. Lechevallier, Y.: Recherche d'une partition optimale sous contrainte d'ordre total. Rapport de recherche RR-1247, INRIA (1990). `http://www.inria.fr/rrrt/rr-1247.html`
7. Olsson, R.J.O., Karlsson, M., Moberg, L.: Compression of first-order spectral data using the b-spline zero compression method. Journal of Chemometrics **10**(5–6), 399–410 (1996)
8. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer Series in Statistics. Springer Verlag (1997)
9. Rossi, F., François, D., Wertz, V., Verleysen, M.: Fast selection of spectral variables with b-spline compression. Chemometrics and Intelligent Laboratory Systems **86**(2), 208–218 (2007)
10. Saito, N., Coifman, R.R.: Local discriminant bases and their applications. Journal of Mathematical Imaging and Vision **5**(4), 337–358 (1995)