TEST (2016) 25:35–40
DOI 10.1007/s11749-015-0476-9

DISCUSSION

Comments on: Probability enhanced effective dimension reduction for classifying sparse functional data

Manuel Febrero-Bande[1]

Published online: 25 January 2016

1 Introduction

The authors deal with the interesting problem of classifying sparse functional data. This task has been treated extensively in the literature from different points of view, most of them trying to extend ideas from the multivariate setting. Being $\mathcal{X}_i$ the trajectories and Y_i the labels taking values on $\{-1, 1\}$ (or on $\{0, 1\}$ depending whether we are interested on discriminant functions or on probabilities of belonging to a certain group), the techniques are based on the two conditional distributions: $Y|\mathcal{X}$ or $\mathcal{X}|Y$. The former is the most popular alternative based on estimating $\mathbb{E}[Y/\mathcal{X}]$ directly or using a transformation of the covariate. For instance, the use of a logistic regression after representing the trajectories in a fixed (Fourier, B-spline, wavelets) or data-driven (PC, PLS) basis is the approach employed by James (2002), Escabias et al. (2005, 2007), Cardot and Sarda (2005), Müller and Stadtmüller (2005), Leng and Müller (2006), Preda et al. (2007) or Müller and Yao (2008) among others. The works by Ferraty and Vieu (2003) or Febrero-Bande and González-Manteiga (2013) follow the ideas from the literature of the nonparametric regression framework allowing to extend the classification techniques to non-Hilbert spaces using norms and distances between trajectories. The third alternative is the transformation of the information of the functional covariate $\mathcal{X}$ using, for instance, notions of depth like in Cuevas et al. (2007) or Li et al. (2012). The first example in functional data of the use of the conditional distribution $\mathcal{X}|Y$ can be found in James and Hastie (2001) as the extension of classical linear

This comment refers to the invited paper available at: doi:10.1007/s11749-015-0470-2.

✉ Manuel Febrero-Bande
manuel.febrero@usc.es

[1] Department of Statistics and Operations Research, Faculty of Mathematics, University of Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela, Spain

discriminant analysis. Again, taking into account the difference between populations, the paper by Delaigle and Hall (2012) employs a centroid classifier. For the particular case of gaussian distributions, the work by Baíllo and Cuevas (2008) contains some theoretical remarks. An interesting review of several classification techniques using both approaches can be found in Baillo et al. (2010).

This paper also follows this second stream. The main idea is to estimate $\mathbb{E}[\mathcal{X}|Y \in I_s]$ recovering a discriminant direction between populations but using the partition scheme provided by $p(\mathcal{X} = p(Y = 1|\mathcal{X}))$ which is not estimated but sequentially split. This involves several issues that will be discussed in the following sections jointly with the other key of the paper: sparsity.

2 Sparsity

The difficulty of dealing with sparsity comes from the way of reconstructing the loss of information. Without dense grids, it is almost impossible to know what happens in the neighborhood of every $t \in T$ and this is an important issue because it determines the shape of the covariance matrix Σ (and so, the eigenfunctions). In this paper, the alternative chosen is to estimate Σ through a local linear estimator. But, this approach has, at least, two drawbacks. First of all, there is no guarantee that the matrix $\hat{\Sigma}$ will be a positive definite one (see, for instance, the introduction of Wu and Pourahmadi 2003). Second, the optimal choice of the bandwidth h_2 (and possibly of the kernel) is an open issue (and even more the use of leave-one-curve-out cross-validation). A naive alternative could be to represent every trajectory in a fixed basis selected *ad hoc*. Of course, both solutions are imperfect and a discussion about the gain/loss of every procedure when fitted to a particular data set it would be interesting. For instance, in Fig. 2 the curves of spinal bone density for Hispanic males and females are represented. Looking at the trajectories for females, there are three with values below 0.7 in the interval [9, 12]. Due to sparsity, these trajectories are not continued after 13, and this has a clear impact in the procedure. Clearly, the mean in the first interval will be biased downward affecting to the computation of the U_{ij} and so, having its impact in the estimation of Σ. The same happens with the male trajectory with values over 1.4. Despite the fact of considering the same covariance matrix for both populations, an open question is about the influence of these apparently strange curves in the classification procedure. It seems quite difficult to detect these curves as outliers (again due to sparsity). Therefore, the presence of outliers jointly with certain types of sparsity can lead every technique to a big fail and an interesting question is what is the degree of sparsity/outlyingness that a procedure can handle and how to correct it.

But, the main challenge with sparsity is how to discover whether the pattern of the sparsity is related with the process $\mathcal{X}$ or with the groups. This seems to be the case of the third real example in the paper where the curves of the second group (dead within 10 years) have the propensity of being unobserved at the end of the interval by obvious reasons. Certainly, this will cause a bias in the procedure. But, the lack of observations can occur in any place along the interval and the challenge is how to avoid its effect in our procedure. In this case, it is obvious that the estimate of the

covariance matrix for the points at the end of the interval is biased to the elements of the first group and so, any classification procedure based on this matrix cannot be able to distinguish between groups using only the last values of the interval. But, what is better? Must we design a procedure that impartially avoids the effect of sparsity or another one that exploits the hidden pattern in sparsity? I am afraid that there is no an easy answer and probably, a tailored solution must be constructed for every particular example. In some sense, this is the same debate as in censored data related with the type of censoring and its effect on the estimates.

3 Probability-enhanced functional cumulative slicing

The validity of the procedure is based on three assumptions. The second one is the most critical from the point of view of the functional data analysis because it establishes that the generating process $\mathcal{X}$ has, in essence, a finite dimension K. This occurs when the data are generated, for example, using the first K elements of a basis but seems quite restrictive in a general framework. The third assumption is of technical nature but uses all the eigenvalues in a double sum which seems a bit surprising. Depending on the decay rate of the eigenvalues, that quantity could be arbitrarily large. A discussion focusing on the practical issues of both assumptions would be welcome.

More shocking is the minor assumption that the observed data are generated as $U_{ij} = \mathcal{X}_i(t_{ij}) + \epsilon_{ij}$ being ϵ_{ij} an i.i.d. measurement error. The reason is that such an error process $\epsilon(t)$ (i.e. $\mathbb{E}[\epsilon(t)] = 0$, $\mathrm{Var}(\epsilon(t)) = \sigma_\epsilon^2$ and $\mathrm{Cov}(\epsilon(t), \epsilon(t')) = 0, \forall t \neq t'$) does not belong to $L^2(T)$ because it is not continuous. This assumption is critical when the grid becomes more and more dense and the minimum distance between discretization points goes to zero. Perhaps it is not so important for a fixed design of the grid, considering that the dependence vanishes for distances lower than that minimum but it is strange from the theoretical point of view and, probably, unnecessary.

4 Simulations and data examples

The application of the procedure involves the selection of various parameters by cross-validation. The list is quite large including λ in WSVM, the dimension K, the truncation s_n, and the bandwidths h_1 and h_2. Despite the fact that the computational time should be large, no clues are provided about how good are the selectors for the approximations involved. And after a lot of computational work, the results in Table 1 are not so impressive as desired. The gain after a such complex procedure is less than 0.80 % between the PEFCS method and the FPCA one. In fact, comparing the numbers in Table 1 with other simple alternatives in the dense case (using the same scenario proposed in the paper), the results obtained (with 1000 Monte Carlo runs) are comparable or slightly better (see Table 1). The two alternatives considered here are both based on logistic regression models: functional generalized linear models (see, for instance, James 2002; Escabias et al. 2007; Cardot and Sarda 2005; Müller and Stadtmüller 2005) and functional generalized spectral additive models (see Müller and Yao 2008). In both cases, the first five principal components were used (no optimization proce-

Table 1 Simulation results comparable to Table 1 in the dense case with functional generalized linear models (FGLM) and functional generalized spectral additive models (FGSAM)

Dense case	FGLM	FGSAM	Best of PEFCS
Model I	11.3	12.6	12.1
Model II	16.2	17.1	16.3
Model III	12.0	14.9	14.2
Model IV	3.7	4.5	6.46

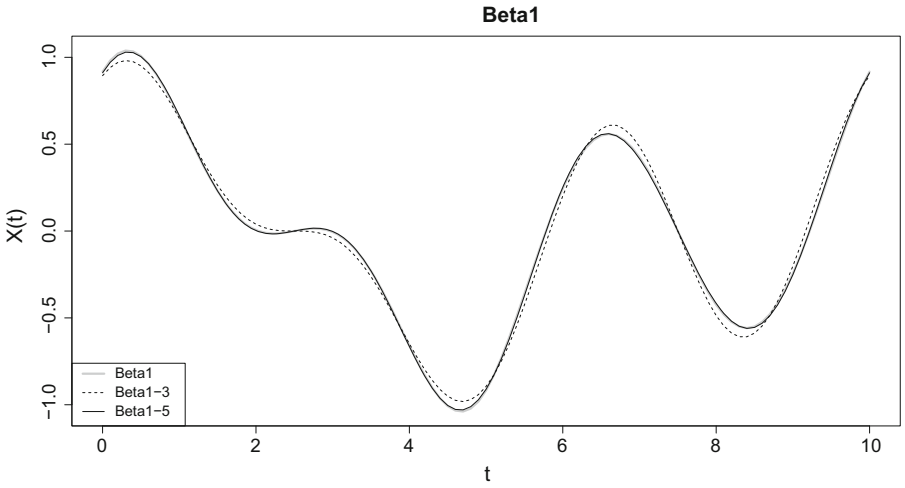

Fig. 1 Comparison between β_1 and different approximations with the first three and five elements

dure was considered for the number of components) and the options by default were employed.

The final message joining both simulation studies seems to be the triumph of the simplicity. Taking into account that the FGLM can be considered a particular case of the FGSAM, it seems shocking that the classification error provided by FGLM were always better than for FGSAM. But, as in Hand (2006), not always a more sophisticated method can obtain better results than a simpler one. Something similar happens in the original Table 1 where the results using QDA are not always better than using LDA. Probably, the examples chosen for the simulation are not complex enough to emphasize the advantages of the new procedure and, in a more simpler scenario, the classical methods obtain better results. As an example, the parameter β_1 is defined as a sum of the first 50 eigenfunctions of X and, in principle, any technique based on principal components has certain advantages even though the huge number of eigenfunctions. But, as can be seen in Fig. 1, the approximation considering only the first five elements is almost undistinguishable from the complete one and so, all the information needed for the classification task is included in the first five eigenfunctions.

The parameter β_2 is also a simple function and probably, there is no enough complexity in the simulations to show the ability of the new procedure to catch the space of directions that better separates both groups or at least, the space generated by these

directions is quite similar to the generated by the first principal eigenfunctions. Note that the procedure is quite high consuming (by the need of computing the optimal of several parameters) and some other alternatives also high demanding like, for instance boosting, could improve the results. As a final comment, there is no information about the structural dimension K in the examples. It is said that the structural dimension has been correctly identified but what is this number? (1 for models I and II and 2 for models III and IV?). In that sense, what was the s_n parameter or the number of components employed when using FPCA? If the number of principal components is only chosen to explain the variability of X, this selection could not be the right one for classification. In that sense, an approach based on PLS (see for instance, Preda et al. 2007) could be more adequate. The situation under sparsity is less clear. The procedure is based on the computation of the covariance matrix that, depending on the type of sparsity, could be very different from the theoretical one affecting not only the PEFCS method but also the FPCA one. In fact, it seems that there is nothing in PEFCS procedure specifically designed to correct sparsity. In essence, the proposed method is adapted to the sparsity scenario but there are doubts about how the different types of sparsity could affect the procedure. This is not explored in the simulation study and the data examples are not the best ones to check this effect. The classification error for the first data example (Berkeley growth study) depends on how much of the last part of the curve is included in the procedure. Using only the data over 14-year-old, a classification error about 8 % is obtained (using FKNN) whereas when using only the curves up to 14-year-old, the error raises to a 25 %. This means that the second half of the growth curves are more informative for classification and so, if the sparsity affects how this part is represented the classification error will raise. On the other hand, if only the data in the first part is affected by sparsity, probably the classification will be unaffected. Therefore, how the sparsity affects the classification error it is not a clear matter. The last two data examples seem to be biased by sparsity: the first one contains curves with just a few observations and located far away from the other ones. These curves affect the estimation of the mean and the covariance matrix which is central for the procedure. The curves for died patients in the cirrhosis data seem to be right-censored affecting again the computation of the covariance matrix. Also, in this example, the results of QDA seem to be erroneous. Taking into account that LDA can be considered a particular case of QDA, the differences between both classifiers in Table 4 are excessively large. How can be this explained? Is it an aside effect of the sparsity?

Therefore, as a final conclusion, the procedure presented in this paper has interesting ideas but also a couple of drawbacks that must be solved in order to achieve that this classification technique can be considered as a standard among practitioners. The main one is the complexity of the method with a lot of parameters that must be jointly optimized and the aside effect of being high time consuming. A second drawback is how to interpret the results obtained for the procedure. Is it possible to know anything about the classification rule through $\hat{m}(t, \pi)$ or $\hat{\Delta}(s, t)$? Typically, the success of a classification technique lies on three pillars: interpretability, precision, and speed of execution. Typically, the precision is considered the most important one although the most popular methods usually sacrifice a little bit in the ability for prediction to gain useful insights and/or fastness. The procedure proposed here seems to go in

the opposite direction showing a small gain in prediction respect to the others but with the high cost of the complexity. The challenge for the authors in the near future is to make its procedure accessible, solving the practical issues providing clues and rules for the choice of parameters, and distributing the procedure worldwide through a friendly environment like `R` or `MatLab`. Finally, from a general perspective, there are a lot of classification methods in FDA without a clear guide about its applicability, on the contrary of their counterparts in the multivariate framework. Surely, a deep comparative study is needed to determine the strength and weakness of every particular proposal compared with the others.

Acknowledgments This research has been partially funded by Project MTM2013-41383-P from Ministerio de Ciencia e Innovación, Spain and by Project 2013-PG064 from Xunta de Galicia, Spain.

References

Baíllo A, Cuevas A (2008) Supervised functional classification: a theoretical remark and some comparisons. arXiv:0806.2831v1

Baillo A, Cuevas A, Fraiman R (2010) Classification methods for functional data. In: The Oxford handbook of functional data analysis. Oxford University Press, Oxford, pp 259–297

Cardot H, Sarda P (2005) Estimation in generalized linear models for functional data via penalized likelihood. J Multivar Anal 92(1):24–41

Cuevas A, Febrero M, Fraiman R (2007) Robust estimation and classification for functional data via projection-based depth notions. Comput Stat 22(3):481–496

Delaigle A, Hall P (2012) Achieving near perfect classification for functional data. J R Stat Soc Ser B (Stat Methodol) 74(2):267–286

Escabias M, Aguilera A, Valderrama M (2005) Modeling environmental data by functional principal component logistic regression. Environmetrics 16(1):95–107

Escabias M, Aguilera A, Valderrama M (2007) Functional PLS logit regression model. Comput Stat Data Anal 51(10):4891–4902

Febrero-Bande M, González-Manteiga W (2013) Generalized additive models for functional data. TEST 22(2):278–292

Ferraty F, Vieu P (2003) Curves discrimination: a nonparametric functional approach. Comput Stat Data Anal 44(1):161–173

Hand DJ et al (2006) Classifier technology and the illusion of progress. Stat Sci 21(1):1–14

James G (2002) Generalized linear models with functional predictors. J R Stat Soc Ser B (Stat Methodol) 64(3):411–432

James GM, Hastie TJ (2001) Functional linear discriminant analysis for irregularly sampled curves. J R Stat Soc Ser B (Stat Methodol) 63(3):533–550

Leng X, Müller HG (2006) Classification using functional data analysis for temporal gene expression data. Bioinformatics 22(1):68–76

Li J, Cuesta-Albertos JA, Liu RY (2012) DD-classifier: nonparametric classification procedure based on DD-plot. J Am Stat Assoc 107(498):737–753

Müller H, Stadtmüller U (2005) Generalized functional linear models. Ann Stat 33(2):774–805

Müller H, Yao F (2008) Functional additive models. J Am Stat Assoc 103(484):1534–1544

Preda C, Saporta G, Lévéder C (2007) PLS classification of functional data. Comput Stat 22(2):223–235

Wu WB, Pourahmadi M (2003) Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika 90(4):831–844