TEST (2016) 25:52–58
DOI 10.1007/s11749-015-0478-7

DISCUSSION

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

Fang Yao[1] · **Yichao Wu**[2] · **Jialin Zou**[1]

Published online: 25 January 2016

We wish to thank Professors Ana M. Aguilera, Germán Aneiros, Gérard Biau, Jiguo Cao, Manuel Febrero-Bande, Clément Levrard, Yufeng Liu, Yulong Nie, Peijun Sang, Philippe Vieu, Chong Zhang, Hao H. Zhang for their insightful comments and constructive suggestions which were enjoyable to read. In the following, we provide rejoinders to the major points that were raised in the discussions. Our responses are grouped into four themes.

1 Sparsity in functional data and assumptions

The sparsity in functional data stands as a fundamental challenge from both conceptual and practical perspectives, given that the underlying process is considered an infinite-dimensional random variable residing in Hilbert space. Aneiros and Vieu provided a thoughtful discussion on this issue, beginning with a broader view that encompasses two major research areas expanding from the traditional multivariate statistics: high-dimensional data analysis and functional data analysis. In contrast to the discrete nature of high-dimensional problems in which the dimension p can be much larger than the sample size n, functional data are of a continuous nature subject to certain degree of smoothness regularity inherited from nonparametric regression. The analyses of these

This rejoinder refers to the comments available at: doi:10.1007/s11749-015-0471-1; doi:10.1007/s11749-015-0472-0; doi:10.1007/s11749-015-0473-z; doi:10.1007/s11749-015-0474-y; doi:10.1007/s11749-015-0475-x; doi:10.1007/s11749-015-0476-9; doi:10.1007/s11749-015-0477-8.

✉ Fang Yao
fyao@utstat.toronto.edu

[1] Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada

[2] Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC 27695, USA

Table 1 The average classification error with the standard error (in parenthesis) in percentage (%) obtained from 100 Monte Carlo repetitions for an additional model $f(X) = 2^{-1}\exp(2.5\langle\beta_1, X\rangle\cdot\langle\beta_2, X\rangle) - 1$, while all other settings remain unchanged as those in the paper

	Method	LDA	QDA	Centroid	Logistic
Sparse	PEFCS	12.5 (.20)	11.2 (.27)	46.7 (.37)	12.6 (.23)
	FPCA	13.1 (.21)	11.8 (.23)	46.5 (.35)	13.4 (.26)
Dense	PEFCS	12.1 (.25)	10.3 (.28)	46.2 (.42)	12.4 (.23)
	FPCA	13.0 (.21)	11.6 (.23)	46.8 (.37)	13.4 (.26)

two types of data are substantively distinct, resulting in different approaches dealing with potential sparseness, which has enriched the literature of these two areas in the past decades.

The discussion on three types of sparsity by Aneiros and Vieu shreds insights into the various notions of sparsity in functional data. The proposed probability-enhanced functional cumulative slicing (PEFCS) to certain extent crossed the situations (i) and (iii) specified in Aneiros and Vieu. The model assumption on sufficiency of projections of the functional predictor X on K index functions $\{\beta_1, \ldots, \beta_K\}$ serves the purpose of dimension reduction to overcome the curse of dimensionality caused by exponentially decaying small ball probability. This is in line with the semiparametric ideas such as projection pursuit (Chen et al. 2011; Ferraty et al. 2003, 2013; Yao et al. 2015) and partial linear modeling (Aneiros-Pérez and Vieu 2006; Lian 2011), among others. We would like to emphasize that the PEFCS based on sufficiency of linear projections includes (generalized) linear or additive regression as special cases, and is "link-free" compared to multiple index modeling methods. Sang, Nie, and Cao raised a question on the robustness to deviation from the linear sufficiency by considering a model $f(X) = 2^{-1}\exp(2.5\langle\beta_1, X\rangle \cdot \langle\beta_2, X\rangle) - 1$ (scaled to have a reasonable separation between two classes). An additional simulation was conducted keeping other settings unchanged. The results in Table 1 indicate that classifiers based on PEFCS still outperform those based on functional principal component analysis (FPCA), with the quadratic being the winner as expected from the multiplicative relation. However, the centroid classifier breaks down using both methods, which may deserve a further study beyond the scope of this paper. Following Aneiros and Vieu's suggestion, one may consider to explore even less structured models (e.g., fully nonparametric) to accommodate nonlinear model sufficiency, which nevertheless would encounters the curse of dimensionality.

A related issue emerges in the comments by Febrero-Bande as well as Biau and Levrardand on the linearity condition of X (Assumption 2 in the paper). This is made on the mean structure of X for the technical need, but does not imply finite dimensionality of X that is often determined by the covariance structure, e.g., for a Gaussian process. Thanks to the suggestion by Biau and Levrard, a possible remedy for non-elliptically case is to construct a kernel dimension reduction method based on a reproducing kernel Hilbert space embedding of X (Fukumizu et al. 2009). Another technical requirement

(Assumption 3) is a sufficient condition to define a subspace of an L^2 space for the existence of index functions identified by the eigenfunctions of $\Sigma^{-1}\Lambda$. The practical implication is that the correlations among projections decay sufficiently fast. Readers can refer to He et al. (2003) for concrete examples that do or do not satisfy such a convergent double sum condition.

The current paper is concerned with the sparsity in observational grids in conjunction with the linear sufficiency in the context of index models. This type of sparsity is specific to functional data and can obscure the nature of practical problems. The challenges were pointed out by Febrero-Bande, including how to reconstruct loss of information and discover whether the sparsity pattern is related to the process or to the groups. Particularly, for each application, one may seek some tailored solutions that lead to enhancements, which was elaborated in Febrero-Bande's discussion through data examples. We mention a comment on the noisily observed scheme $U_{ij} = X_i(t_{ij}) + \epsilon_{ij}$ with ϵ_{ij} being an i.i.d. measurement error. This is a commonly used observational scheme in nonparametric and functional data literature. Of interest is the unobserved process $X \in L^2(T)$ instead of the noise-prone values U_{ij} that are not demanded to reside in $L^2(T)$. Smoothing techniques are usually employed to recover the underlying process on either subject or population level. It is contrary to assuming fully observed trajectories X_i which can be regarded as an idealization of the noise-contaminated model. Our proposal is intended to adapt to a general missing at random situation in a non-informative manner, while specific data may or may not fully respect this mechanism and tailored solutions are more than welcome to be considered. We do agree with Febrero-Bande that there is a great need of more extensive and deeper comparative studies for functional classification, concerning various aspects, including (but not limited to) models that are "structured" versus "flexible" [for instance, the nonparametric functional regression in the sense of Ferraty and Vieu (2006)], "sparse" versus "dense" designs, methods based on conditional distributions $Y|X$ versus $X|Y$, among others.

2 Relation to partial least squares

The discussion by Aguilera is mainly denoted to relating PEFCS to partial least squares (PLS) that attempts to capture the maximal correlation between the functional predictor and the response. The idea of PLS has been seen advantageous over PCA for regression and classification problems in both multivariate and functional data analysis. In particular, Aguilera et al. (2010), Hall and Delaigle (2012), Preda and Saporta (2005), Reiss and Ogden (2007), have studied PLS for functional linear regression and classification problems via the approach of estimating the conditional distribution of $Y|X$. The main problem with FPCA for regression or classification is to ignore the relationship between the predictor and response, which has been also emphasized in the current paper. The remedy we proposed based on a central subspace method aims at the conditional distribution of $X|Y$ instead of $Y|X$. A natural consideration in the schemes of $Y|X$ is to adopt various classifiers to the projections on the directions obtained by functional PLS, such as GLM-PLS (Escabias et al. 2007), LDA-PLS (Preda and Saporta 2005), Centroid-PLS (Hall and Delaigle 2012). Pointed out by both

Aguilera and Biau and Levrard, expanding PLS to sparsely observed functional data is of considerable interest in not only classification but also more general context of regression problems.

Biau and Levrard provided insightful comments on the approaches based on central subspace and PLS that directly ranking or thresholding PLS scores would likely result in a larger subspace due to choosing redundant directions that strongly correlate with but do not belong to the central subspace. This remark enlightens an advantage of the central subspace paradigm over correlation-based methods. As for PEFCS, a promising alternative in the first step of the initial curve recovery has been suggested by both Biau and Levrard and Aguilera in view of the advantage of PLS over FPCA. Although nearly 100 % of information in X is preserved in our initial recovery by FPCA, the truncated portion might still carry important information for classification in some "not-so-pathological" cases. This suggestion is meaningful and may potentially boost the performance of the current PEFCS method, as the subsequent construction of the EDR space is based on the recovered data. Again this stresses the need of developing PLS approach for regression and classification models in which the predictor process is sparsely observed and contaminated with noise.

3 Implementation and numerical examples

Several discussions have contributed to various aspects of implementation of PEFCS algorithm associated with sparsity, smoothing techniques, as well as computational costs. The estimation of the mean and covariance structures sets the stage for recovering the sparsely observed trajectories and the identification of the effective directions. A valuable comment is made by Aneiros and Vieu on the weighting scheme of measurements from each individual proposed in Li and Hsing (2010). Using the inverse of the number of measurements for each subject as weights when pooling together the data has been carefully studied in Li and Hsing (2010), and it is straightforward to adopt this modification in PEFCS. In sparse setting, Hall et al. (2006) has shown that the current mean and covariance estimation also enjoys the optimal rates of convergence comparable to those in Li and Hsing (2010). As mentioned by Febrero-Bande, the selection of optimal bandwidth of covariance estimation for sparse functional data remains an open issue and deserves a further investigation. Nevertheless, based on our experience and the study in Lin and Carroll (2000), it is suggested to ignore the within-subject correlation in conjunction with cross-validation type search, at least when the measurements are sparsely observed with noise. The issue with the non-negative definiteness of the covariance estimation is less pronounced in our implementation of PACE algorithm, as we followed the suggestion made in Hall et al. (2008) to discard the basis associated with negative eigenvalues. It has been shown theoretically that this amendment leads to asymptotically negligible impact on the resulting covariance estimator.

It is noted in Febrero-Bande's discussion that the simulated predictor processes and index functions may have too simple structures to demonstrate the advantage of the proposed method that are intended for more complex situations. As responded in Sect. 1, it is worth further investigations of extensive and deeper comparative studies

exploring complicated patterns and shapes that warrant the need of advanced methods. In modern scientific activities, such a demand of dealing with increasingly complex data is ever growing, while the current work is an attempt to pioneer along the route. A remark is also made by Febrero-Bande on underlying mechanisms that may lead to various types of sparsity in functional data in light of the discussions on the data examples. The sparsity mechanism is in fact largely unknown to statisticians and has a close connection to the well-researched missing value problems in general. For functional data of infinite-dimensional nature, this becomes a great challenge from both conceptual and methodological considerations. We are still at infancy of identifying such missing links and would like to invite capable hands on this direction of research. In response to Sang, Nie, and Cao, it is a highly nontrivial task to adopt information-based selection criteria, such as the AIC and BIC, as even for standard central subspace methods, this remains as an open issue. To clarify the choices of K and s_n in simulation and data examples, we mention that the selection of both parameters has been based on either testing-sample or cross-validated classification error, i.e., s_n was not chosen from explained variance in X as FPCA. As reported in the paper, the structural dimension K in simulation has been corrected identified for all models ($K = 1$ for models I and II, $K = 2$ for models III and IV). We have the consensus with discussants on the reproducibility/extendability that makes our research useful for real applications and practitioners in not only statistical but also other disciplines. We are currently working on the refinement of the user interface for the PEFCS algorithm in Matlab to make it friendly for public access in near future.

4 Potential extensions

In view of valuable suggestions made by all discussants, we conclude the rejoinder by responding to and commenting on potentially meaningful extensions that the current exposition may have stimulated. Crossing the foregoing sections, we have mentioned several extensions and/or modifications of the PEFCS. In particular, we would like to stress two important extensions that embrace some crucial aspects of the current proposal. The first is due to possible loss of information contained in the correlation between class label and the higher-order eigen-basis when using FPCA to recover the unobserved trajectories from sparsely sampled measurements. It is of considerable interest to develop a version of PLS for the sparsely observed functional data, and will have broader applications in addition to the initial curve recovery in PEFCS. The second extension concerns the linear sufficiency that lays the principle of central subspace approaches as well as multiple index models. A kernel dimension reduction based on a reproducing kernel Hilbert space embedding of X is of importance to reduce this limitation and also remove the linear assumption on $E\langle h, X\rangle$ technically needed for existence of the EDR space.

Another direction outlined by Zhang and Liu and Zhang is to generalize the case of binary labels to multi-class problems for sparse functional data. Both discussions have provided constructive suggestions on how to make such extensions that we truly appreciate. Specifically, Zhang and Liu pointed out that the common approach of using k functions for k classes with a sum-to-zero constraint can be inefficient and subop-

timal. Based on the proposal in Zhang and Liu (2014), Zhang and Liu suggested that the generalization of PEFCS in an angle-based framework that minimizes a penalized surrogate loss for sparse functional data can be feasible and promising. Zhang proposed another idea to develop a model-free estimation for multi-class probabilities by solving a series of multi-class weighted support vector machines (SVM) based on the work in Liu and Shen (2006) and Wu et al. (2010), and again called for proper adoption to sparse functional data. In response to these two proposals, we would like to refresh the motivation of PEFCS for binary response owing to the homogeneity in the central subspace if one applies inverse or cumulative slicing directly on the binary response. It is important to note that, if the class size is large when the multi-category classification faces more challenging, a direct FCS method proposed by Yao et al. (2015) treating a continuous response may be well in place. Thus the difficulty of dealing with multi-class problems for sparse functional data might not be as severe as it appears, which only arises when the class size exceeds 2 to a moderate degree. Liu and Zhang also suggested to employ a more general loss function than the weighted SVM for estimating class probabilities, and the rationale is driven by the transition behavior from soft to hard classifiers that were thoroughly studied in Liu et al. (2011) but remains unknown for functional data classification.

References

Aguilera AM, Escabias M, Preda C, Saporta G (2010) Using basis expansions for estimating functional pls regression: applications with chemometric data. Chemom Intell Lab Syst 104(2):289–305

Aneiros-Pérez G, Vieu P (2006) Semi-functional partial linear regression. Stat Prob Lett 76(11):1102–1110

Chen D, Hall P, Müller HG (2011) Single and multiple index functional regression models with nonparametric link. Ann Stat 39(3):1720–1747

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

Ferraty F, Vieu P (2006) Nonparametric functional data analysis. Springer, New York

Ferraty F, Peuch A, Vieu P (2003) Modèle à indice fonctionnel simple. Comptes Rendus Mathematique 336(12):1025–1028

Ferraty F, Goia A, Salinelli E, Vieu P (2013) Functional projection pursuit regression. Test 22(2):293–320

Fukumizu K, Bach FR, Jordan MI (2009) Kernel dimension reduction in regression. Ann Stat 37:1871–1905

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

Hall P, Müller HG, Wang JL (2006) Properties of principal component methods for functional and longitudinal data analysis. Ann Stat 34(3):1493–1517

Hall P, Müller HG, Yao F (2008) Modeling sparse generalized longitudinal observations with latent gaussian processes. J R Stat Soc Ser B (Statistical Methodology) 70:703–723

He G, Müller HG, Wang JL (2003) Functional canonical analysis for square integrable stochastic processes. J Multivar Anal 85(1):54–77

Li Y, Hsing T (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. Ann Stat 38:3321–3351

Lian H (2011) Functional partial linear model. J Nonparametr Stat 23(1):115–128

Lin X, Carroll RJ (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. J Am Stat Assoc 95(450):520–534

Liu Y, Shen X (2006) Multicategory ψ-learning. J Am Stat Assoc 101:500–509

Liu Y, Zhang HH, Wu Y (2011) Hard or soft classification? Large-margin unified machines. J Am Stat Assoc 106:166–177

Preda C, Saporta G (2005) PLS regression on a stochastic process. Comput Stat Data Anal 48:149–158

Reiss PT, Ogden RT (2007) Functional principal component regression and functional partial least squares. J Am Stat Assoc 102:984–996

Wu Y, Zhang HH, Liu Y (2010) Robust model-free multiclass probability estimation. J Am Stat Assoc 105:424–436

Yao F, Lei E, Wu Y (2015) Effective dimensional reduction for sparse functional data. Biometrika. doi:10.1093/biomet/asv006

Zhang C, Liu Y (2014) Multicategory angle-based large-margin classification. Biometrika 101(3):625–640