MOX-Report No. 47/2018

PCA-based discrimination of partially observed functional data, with an application to Aneurisk65 dataset

Stefanucci, M.; Sangalli, L.M.; Brutti, P.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

PCA–based discrimination of partially observed functional data, with an application to Aneurisk65 dataset

M. Stefanucci[†], L.M. Sangalli[♯], P. Brutti[†]

September 13, 2018

[†] Dipartimento di Scienze Statistiche
Università di Roma - La Sapienza
piazzale Aldo Moro 5, 00185 Roma, Italy
marco.stefanucci@uniroma1.it
pierpaolo.brutti@uniroma1.it

[♯] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
laura.sangalli@polimi.it

Keywords: Functional PCA, Partially Observed Data, Discrimination.

Abstract

Functional data are usually assumed to be observed on a common domain. However, it is often the case that some portion of the functional data is missing for some statistical units, invalidating most of the existing techniques for functional data analysis. The developments of methods able to handle partially observed or incomplete functional data is currently attracting an increasing interest. We here briefly review this literature. We then focus on discrimination based on principal component analysis, and illustrate a few possible methods via simulation studies and an application to the AneuRisk65 dataset. We show that carrying out the analysis over the full domain, where at least one of the functional data is observed, may not be the optimal choice for classification purposes.

1 Introduction

Over the past two decades, functional data analysis has constituted an extremely active area of research and one of the fastest growing fields of modern statistics;

see, e.g., the text books and reviews by Ramsay and Silverman (2005), Ferraty and Vieu (2006), Wang et al. (2016), Kokoszka and Reimherr (2017), and references therein. The interest in this area has been fueled by the explosive growth in the recording of complex and high-dimensional data, exhibiting a functional nature, i.e., representable by means of suitable curves, surfaces or other functions. Functional data are in fact nowadays common in all fields of sciences and engineering, thanks to the development of many devices able to provide images and measures of quantities of interest, captured over time and/or space.

Functional data come as discrete and typically noisy observations of the underlying functional object, measured at different locations in time, space or some other continuum. While the specific observation grid where each functional datum is available may vary across the statistical units, the domain where the data are observed is typically assumed to be the same across units. When this is not the case, the analysis is usually restricted to the intersections of the domains of the data, or some pre–registration procedure is carried out, so that the registered data insist over the same domain. On the other hand, in many application fields, it is common to encounter sets of functional data where the data have missing parts, or equivalently the domains where they are observed varies across statistical units. This setting is referred to as incomplete functional data, or partially observed functional data.

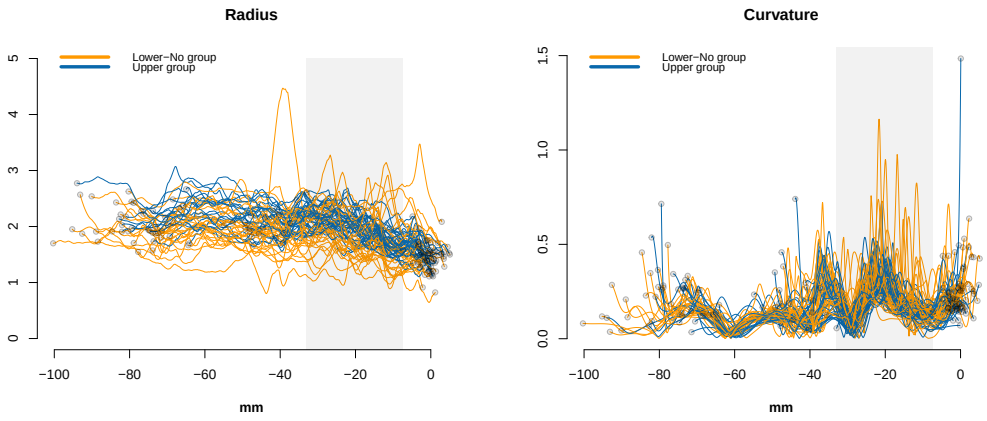

Figure 1: *The Aneurisk65 dataset. Registered radius (left) and curvature (right) of the internal carotid arteries of 65 subjects. The portion of the domain where the data are observed for all subjects is highlighted in light–gray. The circles indicates the starting and ending points for each datum. Two different colors are used for subjects in the Upper group (blue) and subjects in the Lower–No group (orange).*

Figure 1 for instance displays some data from the AneuRisk65 dataset (https://statistics.mox.polimi.it/aneurisk/). These data consist in the profiles of radius (left) and curvature (right) of the internal carotid artery of 65 subjects (see, e.g., Sangalli et al., 2009, 2014b). The data originate from the reconstruc-

tion of three-dimensional angiographic images, taken on subjects suspected to be affected by cerebral aneurysms. The domain where each datum is observed varies across subjects, with longer or shorter portions of the internal carotid artery being observed, depending on where the medical scan has been centered. As highlighted in the figure, there is one portion of the domain where all the data are observed; this corresponds to the (approximately) 3 cm closer to the terminal bifurcation of the artery, that is a point of specific clinical interest; on the other hand, for most subjects, longer portions of the artery are observed, up to more than 10 cm. This incomplete data setting, where there is one portion of the domain where all data are observed, but individual observations are progressively lost when moving from this portion of the domain towards the full domain, is common in functional data coming from medical imaging and from biological studies in general. The analysis of AneuRisk65 data is relevant for the study of the pathology of cerebral aneurysms; in particular, it is relevant to investigate whether the morphology of the internal carotid artery influences aneurysms pathogenesis. The data can be divided into two groups, displayed in orange and blue in the figure, depending on the presence and location of the cerebral aneurysms. In particular, 33 subjects have an aneurysm at or after the terminal bifurcation of the internal carotid artery (Upper group) while the remaining 32 subjects, either have an aneurysm along the internal carotid artery, before the terminal bifurcation, or were found no apparent aneurysm during the angiography (these 32 subjects compose the Lower–No group). Sangalli et al. (2009) present a discriminant analysis between these two groups, based on the scores of the principal components of the radius and curvature profiles; in the latter work, the principal components are computed restricting the attention to the portion of the domain common across subjects. It is however natural to wonder whether these discrimination results may be improved by considering also portions of the domain where not all data are observed.

Unfortunately, most of the nowadays very extensive literature on functional data analysis focuses on the case where all functional data are observed over a common domain, and the vast majority of functional data analysis techniques so far developed is not able to handle this incomplete data framework. The development of methods for partially observed functional data has thus recently started attracting an increasing interest. Classification of functional fragments is discussed in James and Hastie (2001) where an extension of the linear discriminant analysis to the incomplete data framework is proposed. An alternative discrimination technique, based on curves extension, is presented by Delaigle and Hall (2013) and further developed in Delaigle and Hall (2016). Methods for functional Principal Component Analysis (fPCA) of incomplete functional data are described for instance in James et al. (2000), Yao et al. (2005) and in Kraus (2015). Di et al. (2014) extend the technique by Yao et al. (2005) to a multilevel setting, while Liu et al. (2017) employ it to handle spatio–temporal data with gaps. Other works consider partially observed functional data in different applied contexts: Liebl (2013) develops a functional factor model for electricity

spot prices, Goldberg et al. (2014) focus on curve forecasting for call center data, and Gromenko et al. (2017) propose a functional regression model for physical data.

Here, in particular, we shall focus on discrimination based on fPCA scores and consider the case where the incomplete functional data share one common portion of the domain, likewise AneuRisk65 data. The natural usage of techniques for incomplete functional data consists in applying the technique to the whole domain where at least one functional datum is observed. However, this may not be the optimal choice, especially for classification purposes. We will specifically show that, when considering discrimination based on fPCA scores, enlarging the analysis to the whole domain, as well as restricting it to the common domain where all data are observed, may not lead to the best classification results. As illustrated via a simulation study and an application to AneuRisk65 data, the optimal choice often lies between these two extremes. We here suggest to explore different extensions of the domain, ranging from the common domain to the full domain, and select the one that provides the best discrimination result under cross–validation.

Section 2 reviews the techniques for fPCA of incomplete functional data proposed by James et al. (2000), Yao et al. (2005) and Kraus (2015). The same section also generalizes to the incomplete data setting the regularized fPCA technique originally proposed by Huang et al. (2008) in the completely observed data scenario. Section 3 describes the domain extension approach for fPCA–based discrimination. Section 4 illustrates this idea in a simulation study while Section 5 shows the application to AneuRisk65 data. Finally, Section 6 draws some concluding remarks and outlines future directions of possible research.

2 fPCA of partially observed functional data

Assume that n functional data $x_1(t), \ldots, x_n(t)$ are generated from some real-valued random process $X(t)$, with mean $\mu(t)$ and covariance kernel $\Sigma(s,t)$, and that only a discrete and noisy version of each datum is available, i.e., $x_{ij} = x_i(t_{ij}) + \epsilon_{ij}$ for $i \in \{1, \ldots, n\}, j \in \{1, \ldots, m_i\}$, where ϵ_{ij} are measurement errors, with zero mean and finite variance. Consider in particular the case where the observation grids $\{t_{i1}, \ldots, t_{im_i}\}$, with $t_{i1} < \cdots < t_{im_i}$, may differ over the various statistical units, $i = 1, \ldots, n$, and that the domains where they insist, $T_i = [t_{i1}, t_{im_i}] \subset \mathbb{R}$, may as well be different. Standard fPCA assumes the representation

$$x_i(t) = \mu(t) + \sum_{k=1}^{\infty} u_{ik} f_k(t), \quad i \in \{1, \ldots, n\}, \tag{1}$$

where $\mu(t)$ is the mean function, $f_k(t)$ is the $k^{\texttt{th}}$ eigenfunction of the covariance kernel $\Sigma(s,t)$ and u_{ik} is the corresponding score for the $i^{\texttt{th}}$ observation. In practice, only the first K elements of the series are considered. In particular,

when the data are completely observed over a common domain $T = [t_1, t_m]$ and on a common grid $\{t_1, \ldots, t_m\}$, the same for all statistical units, the first K principal components can be estimated performing the eigendecomposition of the empirical covariance matrix $\mathbf{\Sigma}$; the corresponding scores, theoretically defined as $u_{ik} = \int X_i(t) f_k(t)\, \mathrm{d}t$, for each i and k, can be computed by discretizing the integral. When the observation grid differs across the statistical units, but the domain is common to all units, i.e. $T_i = T$, one possibility is to smooth separately each functional datum, and then evaluate each function on a new regular grid, common to all statistical units. Unfortunately, when the data are only partially observed, or observed over different domains T_i, the individual smoothing is not useful for inferring the values of the functions where these are not observed; hence, it is not possible to compute the principal components and associated scores as described above. In this situation, few methodologies try to estimate the scores and the eigenfunctions considering different reformulation of the estimation problem.

James et al. (2000)

Mixed effect models are widely used to handle missing data in longitudinal data analysis. Borrowing from these approaches, James et al. (2000) propose a mixed effect model where the principal component scores are treated as random effects and the mean and principal components are represented via a spline basis. Denote by $\phi(t)$ a spline basis with dimension q. The mean and principal components are then represented as $\mu(t) = \phi(t)^\intercal \mathbf{c}_\mu$ and $\mathbf{f}(t)^\intercal = \phi(t)^\intercal \mathbf{C}$, where $\mathbf{c}_\mu$ and $\mathbf{C}$ are, respectively, a q–dimensional vector of spline coefficients and a $(q \times K)$ matrix of spline coefficients. From equation (1), this leads to the model

$$x_i(t) = \phi(t)^\intercal \mathbf{c}_\mu + \phi(t)^\intercal \mathbf{C}\, \mathbf{u}_i + \zeta_i(t),$$

where the $\mathbf{u}_i$s are assumed to have zero mean and a common variance $\mathbf{\Sigma}_u$, and the $\zeta_i(t)$s are assumed to have zero mean and a constant variance function σ^2. To ensure identifiability of $\mathbf{C}$ and $\mathbf{\Sigma}_u$ the authors restrict the covariance matrix of the $\mathbf{u}_i$s to be diagonal. The fitting procedure is based on maximum likelihood estimation and makes use of the EM algorithm. Once the estimates of the principal components are obtained, the estimates of the scores can be computed through best linear unbiased prediction. The number of basis functions acts as a smoothing parameter that must be carefully selected.

Yao et al. (2005)

Yao et al. (2005) develop an algorithm called PACE (Principal Analysis via Conditional Expectation) that estimates the principal component scores using conditional means. They first estimate the mean $\mu(t)$ and the covariance function $\Sigma(s,t)$ via local linear smoothing of the raw mean vector and covariance matrix obtained from pooled data. An important choice in this context is the

selection of the bandwidths h_μ and h_Σ for the two kernel smoothers. Once the estimates $\widehat{\mu}(t)$ and $\widehat{\Sigma}(s,t)$ of the mean and covariance functions are available, the estimates $\{\widehat{f}_1, \ldots, \widehat{f}_K\}$ of the first K principal components are determined solving the usual discretized eigenvalue–eigenfunction problem, with the associated estimated eigenvalues $\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K\}$. The best prediction for the score vector $\mathbf{u}_i$, associated with the $i^{\texttt{th}}$ observation $\boldsymbol{x}_i = \big(x_{i1}, \ldots, x_{im_i}\big)^{\mathtt{T}}$, is the conditional expectation given $(\boldsymbol{X}_i = \boldsymbol{x}_i)$. Under Gaussian assumptions for the measurement errors ϵ_{ij} and for the scores themselves, this can be shown to be

$$\widehat{u}_{ik} = \widehat{\mathbb{E}}[u_{ik} | \, \boldsymbol{X}_i = \boldsymbol{x}_i] = \widehat{\lambda}_k \, \widehat{\mathbf{f}}_{ik}^{\mathtt{T}} \, \widehat{\boldsymbol{\Sigma}}_i^{-1} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_i),$$

where $\widehat{\mathbf{f}}_{ik} = \big(\widehat{f}_k(t_{i1}), \ldots, \widehat{f}_k(t_{im_i})\big)^{\mathtt{T}}$, $\big[\widehat{\boldsymbol{\Sigma}}_i\big]_{j\ell} = \widehat{\boldsymbol{\Sigma}}(s_{ij}, t_{i\ell})$, and $\widehat{\boldsymbol{\mu}}_i = \big(\widehat{\mu}(t_{i1}), \ldots, \widehat{\mu}(t_{im_i})\big)^{\mathtt{T}}$ are computed on each individual grid T_i.

Huang et al. (2008)

In the case of completely observed functional data, Huang et al. (2008) propose a regularized version of fPCA, that can be be easily generalized to partially observed data, as noted in Lila et al. (2016). This approach relies on a different characterization of the principal components, the so–called best K bases approximation property. Namely, the first K principal components enable the best reconstruction of the signals, in an L^2 sense, among all orthonormal bases of dimension K:

$$\big\{f_k\big\}_{k=1}^K = \underset{\left\{\{\psi_k\}_{k=1}^K : \int \psi_s, \psi_l = \delta_{sl}\right\}}{\operatorname{argmin}} \mathbb{E}\left[\int \left\{X - \mu - \sum_{k=1}^K \left(\int X\psi_k\right)\psi_k\right\}^2\right].$$

Considering only one principal component, the empirical version of the expectation above, for partially observed functional data, is given by $\sum_{i=1}^n \sum_{j=1}^{m_i} \big(x_{ij} - u_i f(t_{ij})\big)^2$. Since the minimization of this quantity involves raw data, a roughness penalty on f is introduced to ensure smoothness of the resulting principal component. In particular, the first principal component and the associated score vector $\mathbf{u} = (u_1, \ldots, u_n)^{\mathtt{T}}$ are estimated solving the following minimization problem:

$$\underset{\mathbf{u}, f}{\operatorname{argmin}} \sum_{i=1}^n \sum_{j=1}^{m_i} \big\{x_{ij} - u_i f(t_{ij})\big\}^2 + \gamma \, \mathbf{u}^{\mathtt{T}}\mathbf{u} \int \big\{f(t)\big\}^2 \, \mathrm{d}t.$$

The smoothing parameter $\gamma > 0$ controls the regularity of the estimated principal component $f(t)$. The term $\mathbf{u}^{\mathtt{T}}\mathbf{u}$ is included to obtain desirable invariance properties (see Huang et al., 2008, for details). Subsequent principal components and the associated score vectors are estimated sequentially solving the same minimization problem, once the contribution to the data of the previously estimated principal components is removed.

Kraus (2015)

Kraus (2015) shows another way to deal with the problem of fPCA in the case of partially observed functional data. The starting point is to estimate the mean function $\mu(t)$ using, for each t, only the available curves at the specific time t, and to estimate $\Sigma(s,t)$ using all complete pairs of functional values at s and t. It is shown that under technical conditions concerning the information provided by the observation grids, these estimators are consistent. The eigenfunctions of the covariance operator can be estimated performing spectral analysis of the complete pairs sample covariance. The missing part of each score is predicted via best linear approximation of its conditional expectation. Using the Riesz representation theorem, the optimization problem can be written as

$$\min_{a_{ik} \in L^2(T_i)} \mathbb{E}\left[\left(u_{ik,mis} - \int_{T_i} a_{ik} x_i\right)^2\right]$$

where $u_{ik,mis}$ is the missing part of the $k^{\texttt{th}}$ score for the $i^{\texttt{th}}$ unit, a_{ik} is an element of $L^2(T_i)$ and x_i is the observed curve. This leads to a linear inverse problem that is regularized, thus involving also in this case the choice of a regularization parameter. Note that this methodology assumes that the data are observed without noise. Moreover, the technique can only deal with data observed over grids that, apart for the starting and ending points, are common across statistical units. This does not create problems in the application to Aneurisk65 data, as these data are preprocessed and evaluated on a common regular grid (see Sangalli et al., 2014b, for details on the preprocessing). In general, a pre-smoothing of each functional data and the re-evaluation on a common grid may be necessary before the technique by Kraus (2015) can be implemented.

3 PCA–based discrimination of partially observed functional data

The four methods for fPCA of incomplete functional data, briefly reviewed in Section 2, are based on different estimation problems and is not clear in advance which one is preferable and in which situation. The first two models rely on parametric assumptions, while the third and the fourth do not. The first three methods involve smoothing, but in different ways: James et al. (2000) use a B–spline basis to represent the mean and principal components, Yao et al. (2005) use a kernel smoothing for the mean and the covariance function, and Huang et al. (2008) smooth the eigenfunctions using a roughness penalty approach. For all the methods, one or more tuning parameters must be selected in some optimal way: the number of B–spline basis in James et al. (2000), the two kernel bandwidths in Yao et al. (2005), the smoothing parameter γ in Huang et al. (2008), and the regularization parameter in Kraus (2015).

In the following sections, we use these methods to perform discrimination of partially observed functional data, where the discrimination is based on the scores of the first K principal components. One natural approach in this sense would be to carry out the analysis over the full domain. However, we show that working on the largest possible domain may not be the optimal choice for classification purposes. On one hand, this may result in imprecise estimates of the principal components, especially of high order, where many data are missing. On the other hand, when the target is classification, considering the total domain may not be useful, when most of the between–group variability is located within the common domain or close to it, or when the missingness is so important that is difficult to distinguish between–group and within–group variability.

We here instead suggest to explore different portions of the domain, moving from the common domain and progressively enlarging towards the full domain. More specifically, we divide the domain where the data are partially observed in L portions, and we thus consider a collection of progressively larger domains I_ℓ for $\ell \in \{0, \ldots, L\}$, with $I_{\ell-1} \subset I_\ell$, where I_0 is the common domain and I_L is the full domain. Figure 2 shows such domains extensions for the AneuRisk65 data. The principal components and their associated scores are then computed over each domain extension I_ℓ, and used for the classification. In particular, in the following sections we consider quadratic discriminant analysis (see, e.g., Izenman, 2009) on the scores of the first K principal components. The optimal number of principal components and the optimal domain extension I_ℓ are selected via cross–validation. For simplicity, the domains I_ℓ are defined by constant enlargements from the common domain to the full domain. Moreover, for illustrative purposes, we here carry out an exhaustive search from I_0 to I_L. Of course, the enlargement step as well as the search could be optimized, if necessary, to decrease the computational cost.

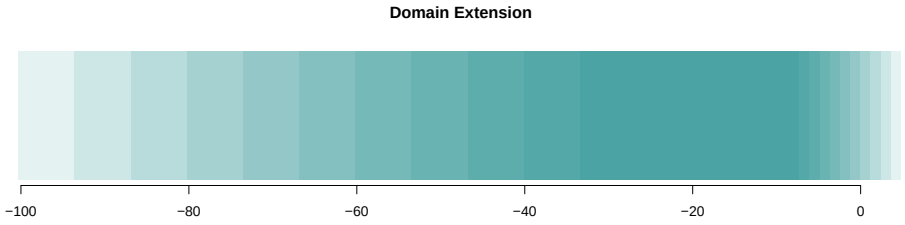

Figure 2: *Visual illustration of the domain extensions for AneuRisk65 data. Moving from the portion of the domain where we have observations for all statistical units, i.e., the common domain, here highlighted by the darkest color, we progressively enlarge the domain by constant steps, until we reach the full domain, where at least one statistical unit is observed, here indicated by the lightest color. The various domain extensions are denoted by progressively lighter shades of color.*

4 Simulations

To illustrate the domain extension approach we carry out a simple simulation study. We generate a set of $n = 100$ functional data over the interval $I_L = [0, 1]$. We then completely retain the data generated over the interval $I_0 = [1/3, 2/3]$, while we censor them over the intervals $I_{\texttt{left}} = [0, 1/3]$ and $I_{\texttt{right}} = [2/3, 1]$, by sampling the starting point of each functional datum uniformly over $I_{\texttt{left}}$, and its ending point uniformly over $I_{\texttt{right}}$. For four statistical units the starting or ending observation points are not sampled but fixed, so that we have one functional datum with starting point in 0, another with starting point in $1/3$, one functional datum with ending point in $2/3$ and another with ending point in 1; this ensures that the full domain is $I_L = [0, 1]$ and the common domain is $I_0 = [1/3, 2/3]$. The data are generated from a cubic B–splines basis with 16 internal knots, corresponding to a total of 20 bases. The position of the spline knots is displayed in Figure 3 by small vertical markers along the x-axis. We generate two groups of functional data, $g \in \{1, 2\}$, composed by 50 curves each, by sampling at each simulation repetition the spline coefficients $\{c_{1,g}, c_{2,g}, \ldots, c_{20,g}\}$ for the two groups from normal distributions with group–specific means, $c_{s,g} \sim N(\mu_{s,g}, \sigma^2)$, for $s \in \{1, \ldots, 20\}$. The means of the first group, $\{\mu_{1,1}, \mu_{2,1}, \ldots, \mu_{20,1}\}$, are set equal to $\{0, 0, 0, 0, 1, 2, 1, 0, -1, 2, 2, -1, 0, 0.5, 1, 0.5, 0, 0, 0, 0\}$; the means of the second group are set to $\{\mu_{1,2}, \mu_{2,2}, \ldots, \mu_{20,2}\} = \{\mu_{20,1}, \mu_{19,1}, \ldots, \mu_{1,1}\}$, thus taking the same values as the first group, but in reverse order. The difference in the means of the spline coefficients constitutes the only structural difference between the two groups. The variance σ^2 of the spline coefficients is the same in both groups and across different coefficients and is set to $\sigma^2 = 0.6$. The n generated curves are evaluated on a regular grid of $p = 150$ over $[0, 1]$ and contaminated by additive, uncorrelated, Gaussian noise, with mean zero and constant variance $\sigma^2_\epsilon = 0.1$. This simulation is repeated 50 times. Different simulation settings are considered in the appendix, changing the amount of noise, the variance of the spline coefficients, the mean values of the coefficients. Figure 3 shows the data sampled in the first simulation repetition. The discrimination between the two groups of data is present both within and outside the common domain, with an important part of the discrimination lying outside the common domain.

Figure 3: *Simulation study 1. Data generated in the first replicate of this simulation. The portion of the domain where the data are not censored is highlighted in light–gray. Two different colors are used for the data in the two groups. Dashed lines represent missing part of the functional data. The small vertical markers along the x-axis indicate the position of the spline knots used for the data generation.*

We thus apply the three methodologies for fPCA of partially observed functional data reviewed in Section 2, over 10 domain extensions, ranging from I_0 to I_L, with constant enlargement steps. The analysis is performed in the `R` environment (R Core Team (2016)). The tuning parameters of each methodology are selected at each simulation replicate by cross–validation. This is carried out separately over each domain extension; the selected tuning parameters can thus differ for different domain extensions. The selection of the optimal number of spline bases in James et al. (2000), implemented in the `R` package `fpca` (Peng and Paul, 2011), is carried out optimizing an approximate cross–validation score. For Yao et al. (2005), implemented in the package `fdapace` (Dai et al., 2017), the optimal bandwidths for the two kernel smoothers are chosen minimizing the leave–one–curve cross validation. For the method based on the extension of Huang et al. (2008) to partially observed data, we implemented a 5–fold cross–validation. The regularization parameter in Kraus (2015), implemented through routines published by the author [1], is selected via generalized cross validation.

[1] available at `http://dx.doi.org/10.1111/rssb.12087`

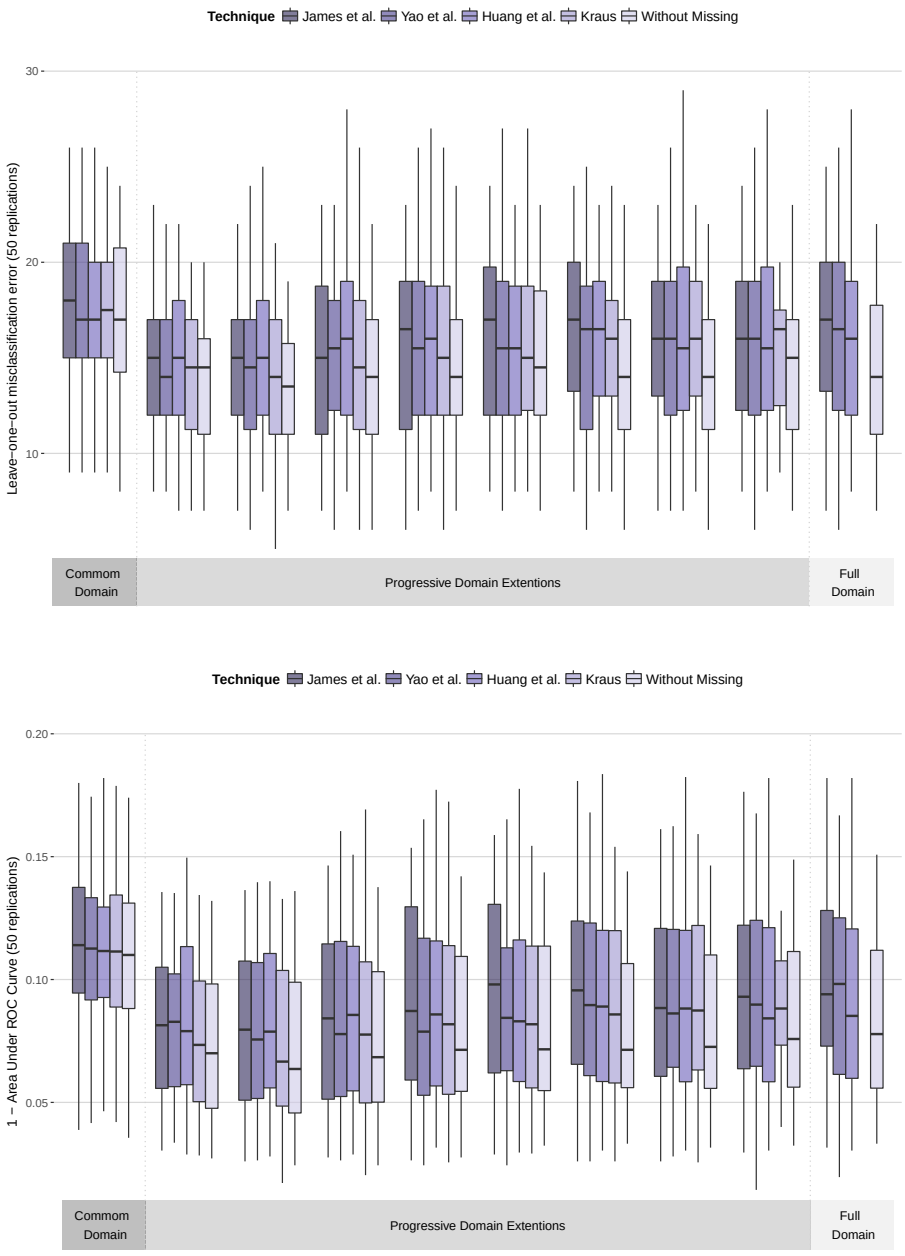

Figure 4: *Simulation study 1. Top: Leave–one–out misclassification error, over the* 50 *simulation replicates, for various domain extensions. Bottom:* 1− *Area Under ROC Curve, over the* 50 *simulation replicates, for various domain extensions.*

A quadratic discriminant analysis on the scores of the first K principal components, with $K \leqslant 5$, is then carried out. In particular, for each replication and

each domain extension, we select the optimal number of principal components scores to be considered for the discrimination via leave–one–out cross–validation, but minimizing in this case the misclassification error. We also apply standard PCA to the fully observed (non–censored) data; the associated misclassification error indicates the best possible classification results achievable in this simulation setting, based on discrimination of the principal component scores, for fully observed data. The top panel of Figure 4 displays the boxplots of the leave-one-out misclassification error, for the various techniques, for various domain extensions. The leave-one-out misclassification error that could be attainable if the uncensored data were available is as well displayed. Note that in the full domain the method by Kraus (2015) is not employable because there are no curves observed jointly at time 0 and 1. For all methods, the misclassification error decreases when we start extending the domain with respect to the common domain, but then progressively increases as we approach the full domain. None of the methods outperforms the other. As an additional measure of the quality of the discrimination we also compute the area under the ROC Curve (see Izenman (2009)); this quantity is bounded between 0 and 1, with the value 1 being attained for perfect classification. In the bottom panel of Figure 4 we show the boxplots of the index (1 - area under the ROC curve), whose minima correspond to the best discrimination. A visual inspection of these boxplots confirms what already commented on the base of the leave-one-out miscassification error: extending the domain with respect to the common domain improves the discrimination between the two groups; on the other hand, larger domain extensions, and in particular the full domain, do not lead to the best discrimination results.

5 Application to AneuRisk65 data

The AneuRisk project (`https://statistics.mox.polimi.it/aneurisk/`) is an interdisciplinary project that involved statisticians and numerical analysts from Politecnico di Milano (Milano, Italy) and Emory University (Atlanta, USA), bioengineers and computer scientists from Istituto Mario Negri (Bergamo, Italy), and medical doctors from Niguarda Hospital and Maggiore Policlinico Hospital (Milano, Italy), with the aim of investigating cerebral aneurysms pathology. This is a very common pathology, totally asymptomatic in the vast majority of cases. Rupture of a cerebral aneurysm is a rare event (affecting one in ten thousand people every year), but unfortunately has associated very high mortality. The origin of the pathology is still largely unknown. One conjecture, investigated by the AneuRisk project, is that aneurysm's pathogenesis may be influenced by the morphology of the hosting vessels, and in particular by the morphology of the internal carotid artery, through the effect that the morphology of the vessel has on the blood fluid–dynamics. The two geometrical quantities that mostly determine the haemodynamics are the radius and the curvature of the

vessel. For this reason, the first studies carried out within the AneuRisk project focused on these two features. Figure 1 shows the profiles of radius (left) and curvature (right) of the internal carotid artery of 65 subjects, pre-processed and registered as described in Sangalli et al. (2009, 2014b,a). As outlined in section 1, the data are divided in 2 groups depending on the presence and location of the aneurysm. Sangalli et al. (2009) carry out a discriminant analysis between these two groups, based on the scores of the principal components of the radius and curvature profiles, computing the principal components by standard fPCA on the portion of the domain common across subjects (the approximately 3cm closer to the terminal bifurcation of the internal carotid artery, as highlighted in Figure 1). The resulting leave–one–out misclassification error amounts to 15 subjects.

Here we test the four methodologies described in the previous sections over various domains extensions; see Figure 2 for the considered domain extensions. Likewise in Sangalli et al. (2009), we consider up to 4 principal components. As for the simulation, the optimal number of principal components is selected for each method and each domain extension via leave-one-out cross validation, minimizing the misclassification error. Figure 5 shows the classification results.

For all considered methods, the domain where the best discrimination is achieved lies between the common domain and the total domain. In this particular application, the approach based on the extension of Huang et al. (2008) to partially observed functional data does the best job, reaching a leave–one–out misclassification error of 9 subjects. Huang et al. (2008) returns the best results also when considering the index based on the area under the ROC curve. Looking at the leave-one-out missclassification error, the best domain extension for this method turns out to be optimal also for the other techniques considered. On the common domain, all methods perform similarly to standard fPCA, with 14, 15 or 16 misclassified units, depending on the method. As highlighted by this figure, the application of the methodologies for partially observed data on the full domain does not lead to any improvement in the discrimination; for discrimination based on James et al. (2000) and Yao et al. (2005), the misclassification error is in fact higher on the full domain than on the common domain, and the index based on the area under the ROC curve is as well worse on the full domain than on the common domain. So, ignoring the domain extension technique would lead to the incorrect conclusion that there is no advantage in including the part of the domain where the data is only partially observed. The estimated principal components over the best domain extension in terms of misclassification error are displayed in Figure 6. The estimates of the principal components returned by the four methods are very similar. The second component for the radius and the first for the curvature have important peaks outside of the common domain (at about -40mm and -38mm, respectively). An important part of the discrimination between the two groups lies here, and for this reason a better classification is possible only when considering the domain extension.

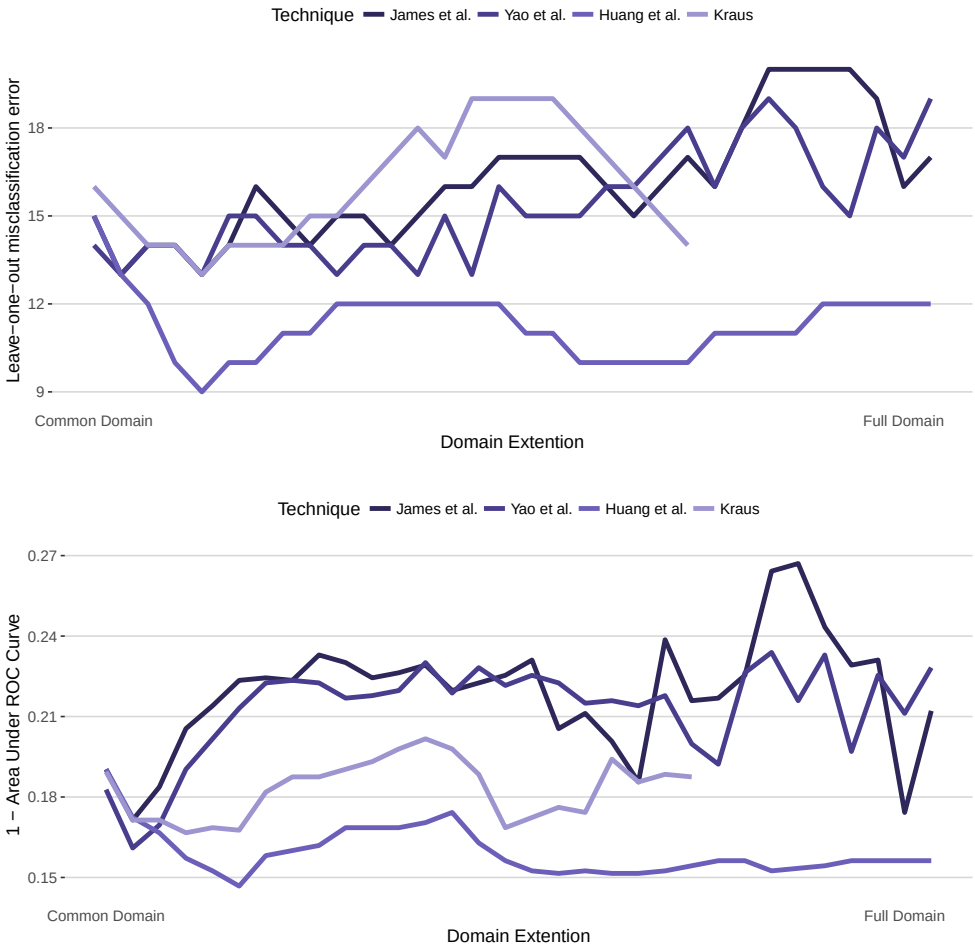

Figure 5: *AneuRisk65 data. Top: leave–one–out misclassification error for various domain extensions. Bottom:* 1− *Area Under ROC Curve for various domain extensions.*

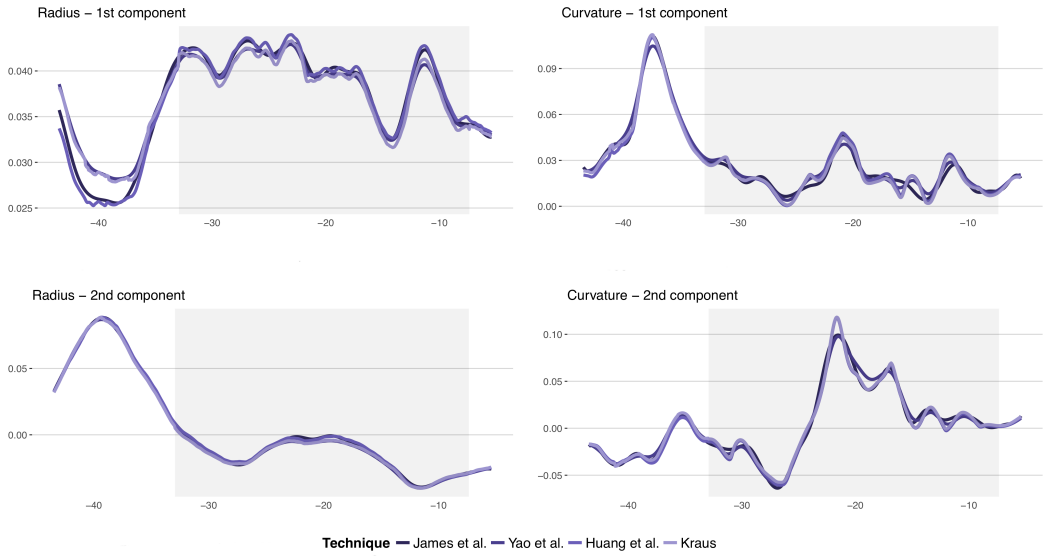

Figure 6: *AneuRisk65 data. Estimates of the principal components provided by the various considered methods on the optimal domain extension. The portion of the domain where the data are observed for all subjects is highlighted in light–gray.*

6 Discussion

As highlighted by the simulation study and the application to AneuRisk65 data, when performing supervised classification of partially observed functional data, considering the full domain where the data are observed may not be optimal. In this illustrated review of PCA–based discrimination of partially observed data, we explored a simple strategy of searching over domain extensions, moving from the common domain where all the data are observed to the full domain where at least some datum is available. An interesting line for future investigation goes towards a more complex and complete search for such optimal domain, where the search is not restricted to progressive extensions of the common domain. In the context of fully observed functional data, a similar idea is explored in Floriello and Vitelli (2017) for unsupervised clustering, and by Pini et al. (2017) for supervised profile monitoring. The domain–selection idea we are here considering differs instead from the approaches explored in Ferraty et al. (2010) and Delaigle et al. (2012), where the search focuses on specific pointwise locations where discrimination between two groups of functional data is optimized.

Acknowledgments. We are grateful to two anonymous referees for their suggestions and constructive comments.

References

Dai, X., Hadjipantelis, P. Z., Ji, H., Mueller, H.-G. and Wang, J.-L. (2017) *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.3.0.

Delaigle, A. and Hall, P. (2013) Classification using censored functional data. *Journal of the American Statistical Association*, **108**, 1269–1283.

— (2016) Approximating fragmented functional data by segments of markov chains. *Biometrika*, **103**, 779–799.

Delaigle, A., Hall, P. and Bathia, N. (2012) Componentwise classification and clustering of functional data. *Biometrika*, **99**, 299–313.

Di, C., Crainiceanu, C. M. and Jank, W. S. (2014) Multilevel sparse functional principal component analysis. *Stat*, **3**, 126–143.

Ferraty, F., Hall, P. and Vieu, P. (2010) Most-predictive design points for functional data predictors. *Biometrika*, **97**, 807–824.

Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer New York.

Floriello, D. and Vitelli, V. (2017) Sparse clustering of functional data. *Journal of Multivariate Analysis*, **154**, 1 – 18.

Goldberg, Y., Ritov, Y. and Mandelbaum, A. (2014) Predicting the continuation of a function with applications to call center data. *Journal of Statistical Planning and Inference*, **147**, 53 – 65.

Gromenko, O., Kokoszka, P. and Sojka, J. (2017) Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Ann. Appl. Stat.*, **11**, 898–918.

Huang, J. Z., Shen, H. and Buja, A. (2008) Functional principal components analysis via penalized rank one approximation. *Electron. J. Statist.*, **2**, 678–695.

Izenman, A. J. (2009) *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer Science & Business Media.

James, G., Hastie, T. and Sugar, C. (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

James, G. M. and Hastie, T. J. (2001) Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 533–550.

Kokoszka, P. and Reimherr, M. (2017) *Introduction to Functional Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC.

Kraus, D. (2015) Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**, 777–801.

Liebl, D. (2013) Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Stat.*, **7**, 1562–1592.

Lila, E., Aston, J. A. D. and Sangalli, L. M. (2016) Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Ann. Appl. Stat.*, **10**, 1854–1879.

Liu, C., Ray, S. and Hooker, G. (2017) Functional principal component analysis of spatially correlated data. *Statistics and Computing*, **27**, 1639–1654.

Peng, J. and Paul, D. (2011) *fpca: Restricted MLE for Functional Principal Components Analysis*. R package version 0.2-1.

Pini, A., Vantini, S., Colosimo, B. M. and Grasso, M. (2017) Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

Ramsay, J. and Silverman, B. (2005) *Functional Data Analysis*. Springer Series in Statistics. Springer.

Sangalli, L. M., Secchi, P. and Vantini, S. (2014a) Analysis of aneurisk65 data: k-mean alignment. *Electron. J. Statist.*, **8**, 1891–1904.

— (2014b) Aneurisk65: A dataset of three-dimensional cerebral vascular geometries. *Electron. J. Statist.*, **8**, 1879–1890.

Sangalli, L. M., Secchi, P., Vantini, S. and Veneziani, A. (2009) A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *Journal of the American Statistical Association*, **104**, 37–48.

Wang, J.-L., Chiou, J.-M. and Mller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

Yao, F., Mller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.

Appendix

We consider three additional simulation studies, where we generate the data as in main simulation study, detailed in Section 4, with the only differences that

- in Simulation study 2 (Figures 7 and 8), the variance of the measurement error is increased to $\sigma_\epsilon^2 = 0.4$;

- in Simulation study 3 (Figures 9 and 10), the variance of the spline coefficients is decreased to $\sigma^2 = 0.3$;

- in Simulation study 4 (Figures 11 and 12), the mean values of the spline coefficients of the first group of functional data are se to $\{\mu_{1,1}, \mu_{2,1}, \ldots, \mu_{20,1}\} = \{0, 0, 0, 0, 1, 2, 1, 0, -1, 1, 1.2, -1, 0, 0.5, 1, 0.5, 0, 0, 0, 0\}$, and the mean values of the spline coefficients of the second group are set to $\{\mu_{1,2}, \mu_{2,2}, \ldots, \mu_{20,2}\} = \{\mu_{20,1}, \mu_{19,1}, \ldots, \mu_{1,1}\}$.

Figures 7, 9 and 11 show the data generated in the first replicates of these simulation studies. We implement the four techniques as detailed in Section 2. Figures 8, 10 and 12 show the boxplots of the misclassification error and area under the ROC curve over the 50 simulation repetitions. Similar comments as those made for the main simulation study hold for all the considered simulation settings: by considering domain extensions it is possible to improve the discrimination results; on the other hand, even though a large part of the separation between the two groups lies outside of the common domain, considering the full domain where at least one of the data is observed leads to sub-optimal results.

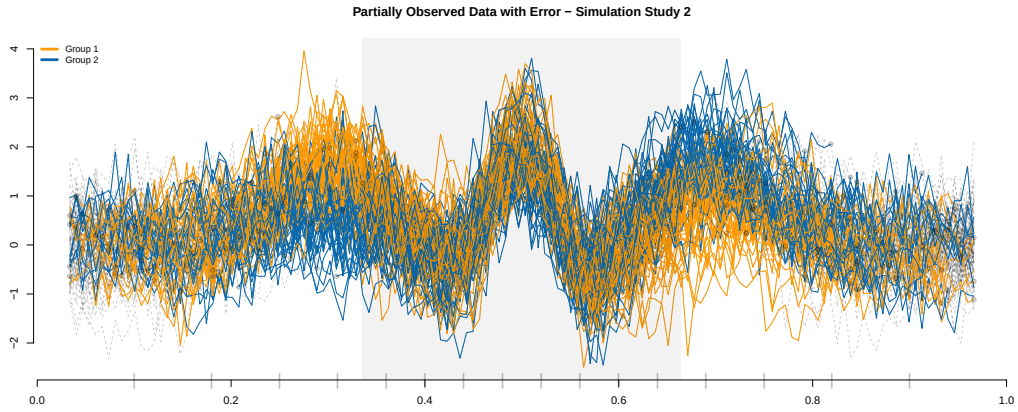

Figure 7: *Simulation study 2. Data generated in the first replicate of this simulation. The portion of the domain where the data are not censored is highlighted in light–gray. Two different colors are used for the data in the two groups. Dashed lines represent missing part of the functional data. The small vertical markers along the x-axis indicate the position of the spline knots used for the data generation.*

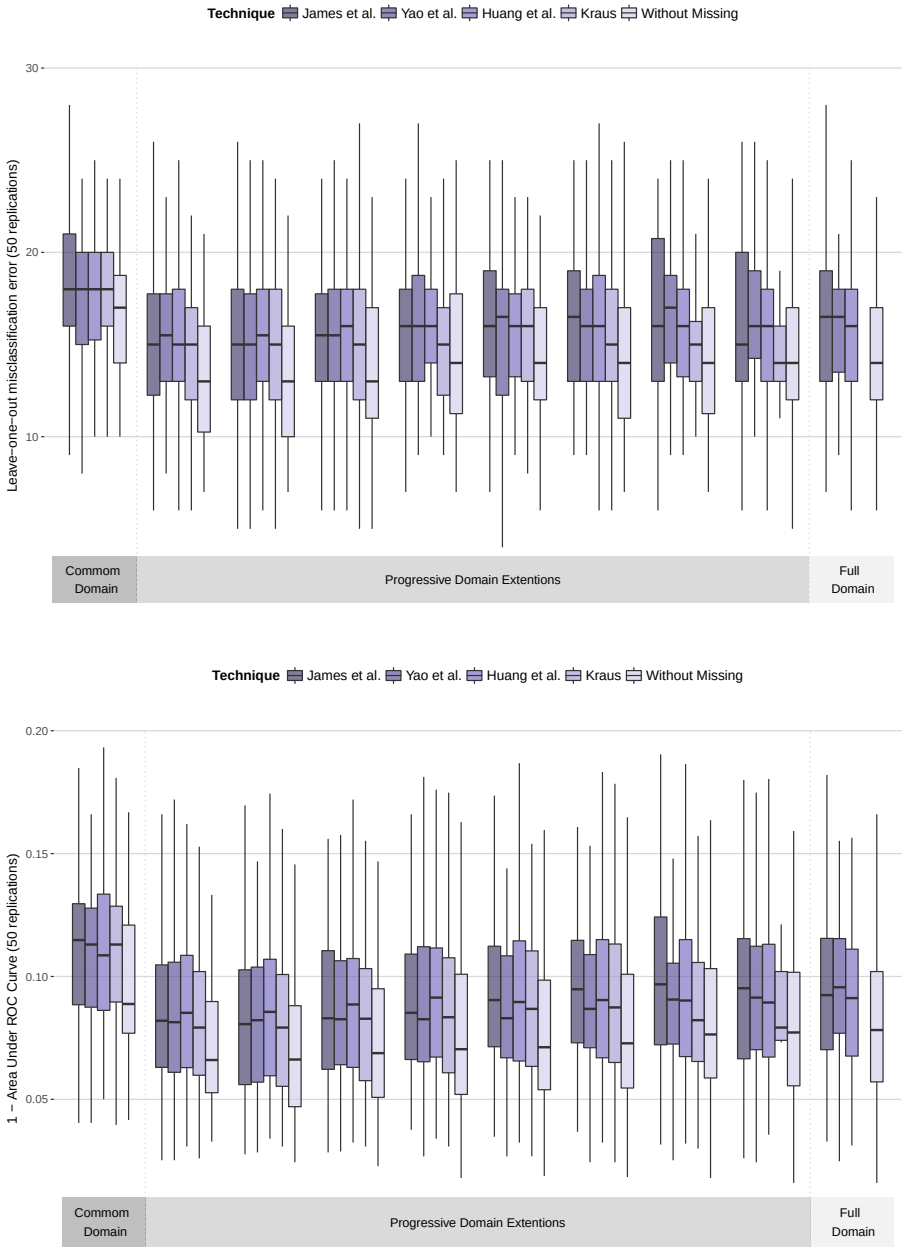

Figure 8: *Simulation study 2. Top: Leave–one–out misclassification error, over the* 50 *simulation replicates, for various domain extensions. Bottom:* 1− *Area Under ROC Curve, over the* 50 *simulation replicates, for various domain extensions.*

Figure 9: *Simulation study 3. Data generated in the first replicate of this simulation. The portion of the domain where the data are not censored is highlighted in light–gray. Two different colors are used for the data in the two groups. Dashed lines represent missing part of the functional data. The small vertical markers along the x-axis indicate the position of the spline knots used for the data generation.*

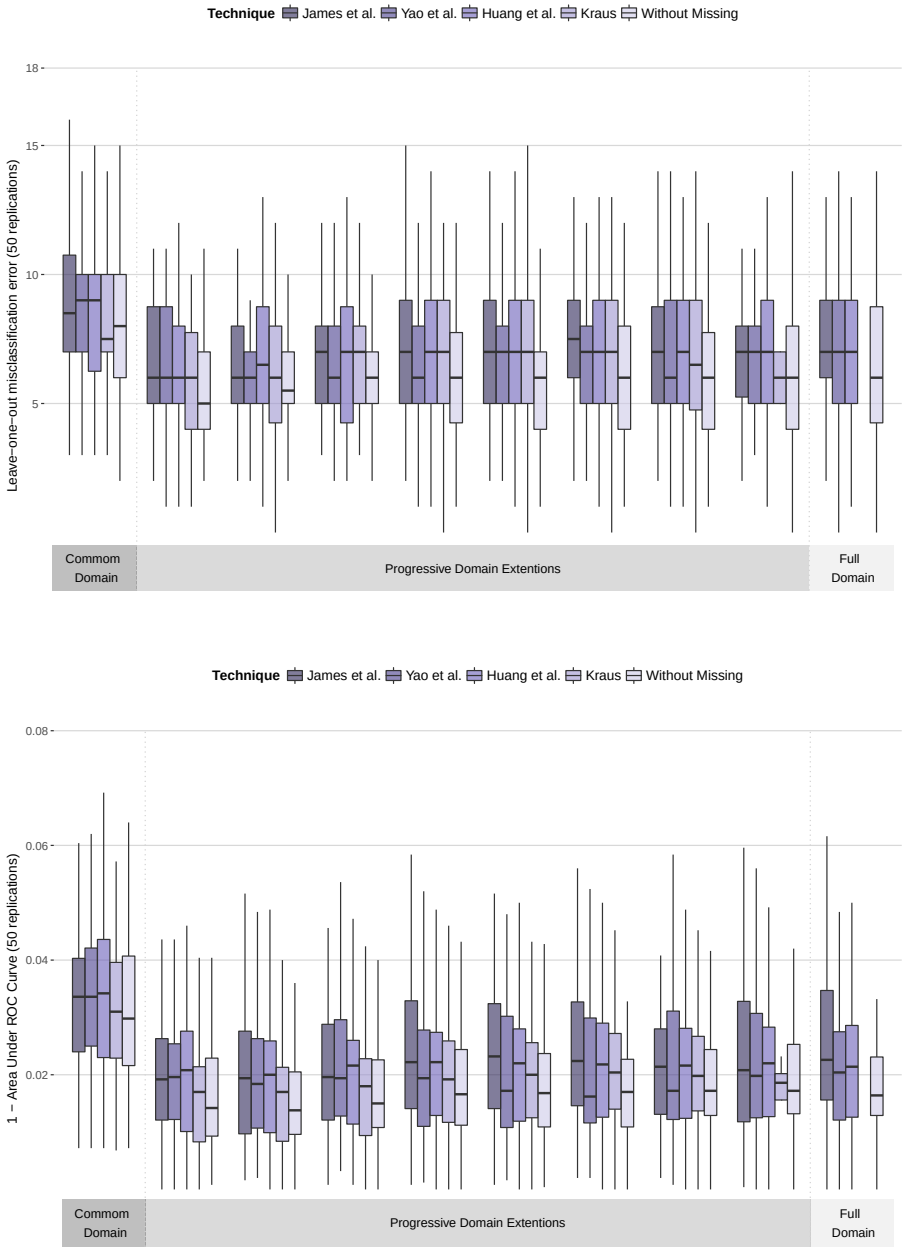

Figure 10: *Simulation study 3. Top: Leave–one–out misclassification error, over the* 50 *simulation replicates, for various domain extensions. Bottom:* 1− *Area Under ROC Curve, over the* 50 *simulation replicates, for various domain extensions.*

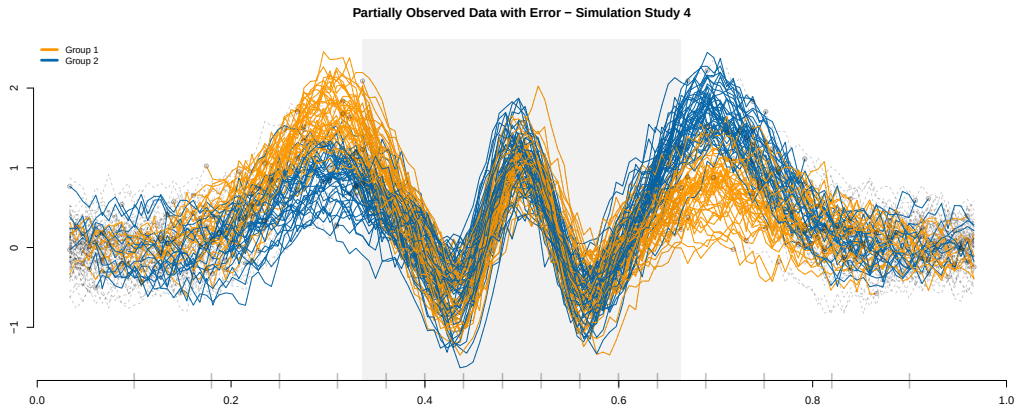

Figure 11: *Simulation study 4. Data generated in the first replicate of this simulation. The portion of the domain where the data are not censored is highlighted in light–gray. Two different colors are used for the data in the two groups. Dashed lines represent missing part of the functional data. The small vertical markers along the x-axis indicate the position of the spline knots used for the data generation.*

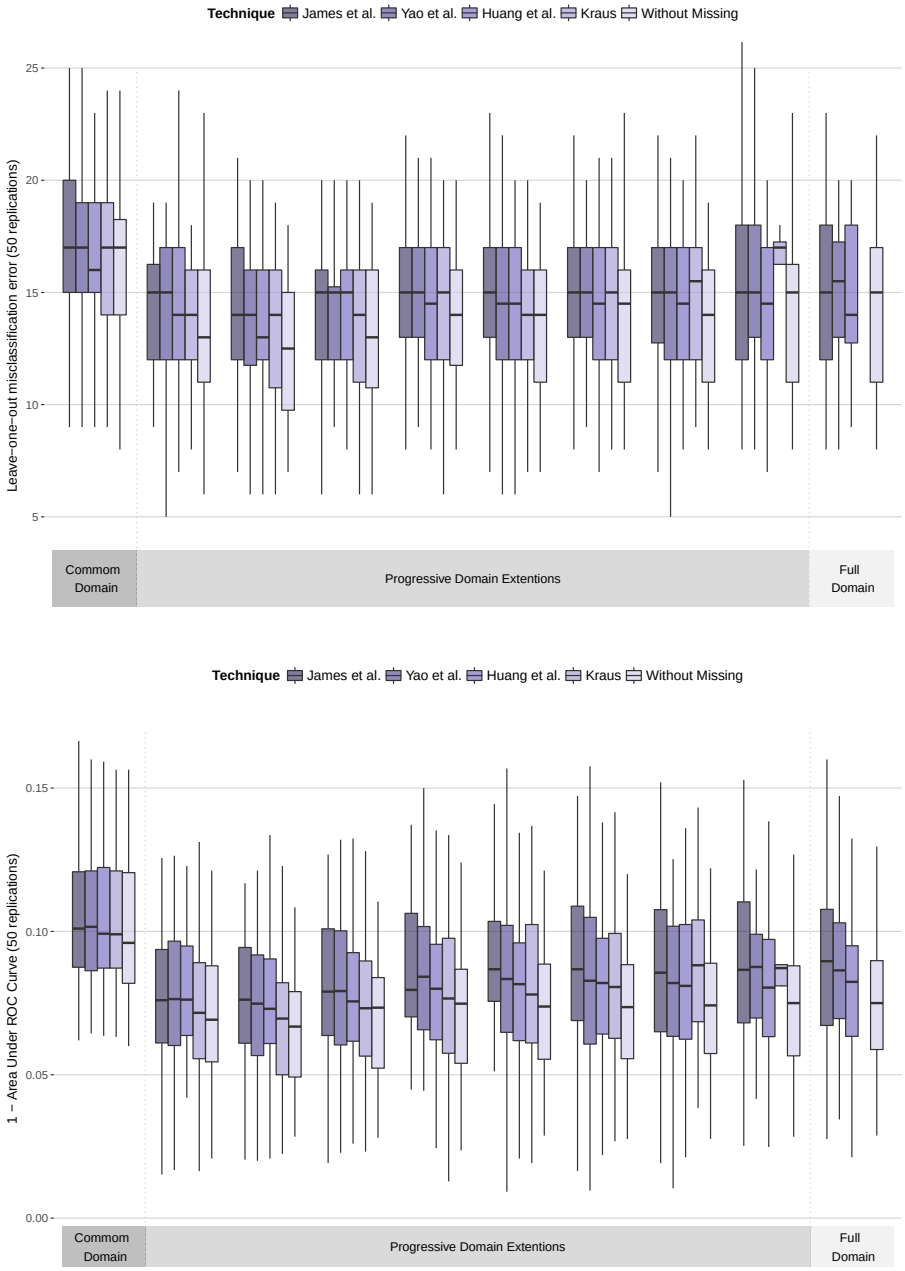

Figure 12: *Simulation study 4. Top: Leave–one–out misclassification error, over the* 50 *simulation replicates, for various domain extensions. Bottom:* 1− *Area Under ROC Curve, over the* 50 *simulation replicates, for various domain extensions.*

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

46/2018 Riccobelli, D.; Ciarletta, P.
Morpho-elastic model of the tortuous tumour vessels

44/2018 Bernardi, M.S.; Sangalli, L.M.
Modelling spatially dependent functional data by spatial regression with differential regularization

45/2018 Bernardi, M.S.; Carey, M.; Ramsay, J.O.; Sangalli, L.M.
Modeling spatial anisotropy via regression with partial di fferential regularization

42/2018 Antonietti, P.F.; Melas, L.
Algebraic multigrid schemes for high-order discontinuous Galerkin methods

43/2018 Fontana, L.; Masci, C.; Ieva, F.; Paganoni, A.M.
Performing Learning Analytics via Generalized Mixed-Effects Trees

41/2018 Mazzieri, I.; Melas, L.; Smerzini, C.; Stupazzini, M.
The role of near-field ground motion on seismic risk assessment in large urban areas

39/2018 Ferro, N.; Micheletti, S.; Perotto, S.
Density-based inverse homogenization with anisotropically adapted elements

40/2018 Chiappa, A.S.; Micheletti, S.; Peli, R.; Perotto, S.
Mesh adaptation-aided image segmentation

38/2018 Domanin, M.; Gallo, D.; Vergara, C.; Biondetti, P.; Forzenigo, L.V.; Morbiducci, U.
Prediction of long term restenosis risk after surgery in the carotid bifurcation by hemodynamic and geometric analysis

37/2018 Bonaventura, L.; Della Rocca A.;
Convergence analysis of a cell centered finite volume diffusion operator on non-orthogonal polyhedral meshes