Published in final edited form as:

J Comput Graph Stat. 2017 ; 26(3): 569–578. doi:10.1080/10618600.2016.1217227.

Penalized nonparametric scalar-on-function regression via principal coordinates

Philip T. Reiss*

Department of Child and Adolescent Psychiatry and Department of Population Health, New York University, USA and Department of Statistics, University of Haifa, Israel

David L. Miller

Integrated Statistics, Woods Hole, Massachusetts, USA and Centre for Research into Ecological and Environmental Modelling and School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland, United Kingdom

Pei-Shien Wu and Wen-Yu Hua

Department of Child and Adolescent Psychiatry, New York University

Abstract

A number of classical approaches to nonparametric regression have recently been extended to the case of functional predictors. This paper introduces a new method of this type, which extends intermediate-rank penalized smoothing to scalar-on-function regression. In the proposed method, which we call *principal coordinate ridge regression*, one regresses the response on leading principal coordinates defined by a relevant distance among the functional predictors, while applying a ridge penalty. Our publicly available implementation, based on generalized additive modeling software, allows for fast optimal tuning parameter selection and for extensions to multiple functional predictors, exponential family-valued responses, and mixed-effects models. In an application to signature verification data, principal coordinate ridge regression, with dynamic time warping distance used to define the principal coordinates, is shown to outperform a functional generalized linear model.

Keywords

dynamic time warping; functional regression; generalized additive model; kernel ridge regression; multidimensional scaling

1 Introduction

A central problem in functional data analysis is to relate scalar responses y_i to functional predictors $x_i(\cdot)$ ($i = 1, \dots, n$) by a regression model. In the terminology of Reiss et al. (2010), models of this type are known as “scalar-on-function” regression, to distinguish

*Philip Reiss, Pei-Shien Wu and Wen-Yu Hua gratefully acknowledge the support of the U.S. National Institute of Mental Health (grant 1R01MH095836-01A1).

SUPPLEMENTARY MATERIAL

R code for analyses: Code to reproduce the analyses of the toy data and the signature verification data. (GNU zipped tar file)

them from models for functional responses (“function-on-scalar” or “function-on-function” regression). The standard approach (Marx and Eilers, 1999; Cardot et al., 1999; Ramsay and Silverman, 2005) is to estimate the intercept α and slope or coefficient function $\beta(\cdot)$ in the functional linear model

$$y_i = \alpha + \int_{\tau} x_i(t) \beta(t) dt + \varepsilon_i \quad (1)$$

where τ is the domain of the functional predictors and $E(\varepsilon_i) = 0$.

But as with ordinary (scalar-predictor) regression, a linear model may not be the best way to describe the relationship of interest. While this point will be illustrated in §4 with an application to signature data, it may be demonstrated more sharply by a toy example. Suppose that $x_i: [0, 1] \rightarrow \mathbb{R}$, the functional predictor for the i th observation, is a noisy realization of

$$x_i^0(t) = \begin{cases} -1, & t - \tau_i \in [-.05, 0); \\ 1, & t - \tau_i \in [0, .05); \\ 0, & \text{otherwise,} \end{cases} \quad \text{for some } \tau_i \in [0, 1], \quad (2)$$

as in the first two panels of Figure 1. Suppose further that we have responses arising as

$$y_i = \tau_i + \varepsilon_i \quad (3)$$

with independent identically distributed mean-zero normal errors ε_i . In this setup, y_i depends strongly on $x_i(\cdot)$, via the parameter τ_i which characterizes $x_i(\cdot)$. But this dependence is not captured by the linear model (1).

On the other hand, if we can define a distance between functional predictors such that $x_i(\cdot)$, $x_j(\cdot)$ are close when τ_i , τ_j are, then nearby $x(\cdot)$'s (according to this distance) will be associated with similar y 's. Intuitively, then, the principle that *nearby $x(\cdot)$'s imply similar y 's* may offer useful guidance for predicting y . Thus a regression procedure based on this principle may succeed where the functional linear model fails.

This core idea underlies the method proposed here, a new approach to scalar-on-function regression that is *nonparametric* in the sense of Ferraty and Vieu (2006). More specifically, given a distance among the functional predictors that is relevant to predicting y in a particular application, we propose to

1. extract the leading principal coordinates (defined below), which can be thought of as n points in Euclidean space whose distances are approximately equal to those among the functional predictors;
2. then, treating these coordinates as a set of predictors, use them to predict y via ridge regression.

We call this approach principal coordinate ridge regression (PCoRR).

As a brief illustration of PCoRR's utility, we used windowed dynamic time warping (DTW; Sakoe and Chiba, 1978; Giorgino, 2009; Faraway, 2012) to define distances among the functional predictors shown in Figure 1 (see §4). For this toy data set, the windowed DTW distance between $x_i(\cdot)$, $x_j(\cdot)$ is low if and only if $\tau_i \approx \tau_j$. This can be seen in the upper left subfigure of Figure 2, which shows the DTW distances among the functional observations $x_i(\cdot)$, arranged in increasing order of τ_i in (2). Hence, in light of (3), DTW distance tends to be low for pairs of x 's with similar corresponding y 's. This property is what makes principal coordinates based on DTW distance efficacious for predicting y , as shown in the lower subfigure of Figure 2: PCoRR, using DTW-based principal coordinates, dramatically outperforms an analogous functional linear model in the sense of attaining much lower generalized cross-validation scores (GCV; see §2.5).

For some insight into how PCoRR parlayes a well-chosen distance into good predictive performance, consider the hat matrix $\mathcal{H} = (h_{ij})_{1 \leq i, j \leq n}$ such that for each i , the i th fitted value $\hat{y}_i$ is given by the weighted mean

$$\hat{y}_i = \sum_{j=1}^n h_{ij} y_j \quad (4)$$

of the responses (see §2.2). This matrix is displayed, for a PCoRR fit with a low GCV score, in the upper right subfigure of Figure 2. Observe that high values in the distance matrix correspond to near-zero values in the hat matrix. Thus for each i , the “equivalent kernel” weights $h_{i1}, \dots, h_{in}$ in (4) (Wasserman, 2006) are such that $\hat{y}_i$ is determined mainly by the most relevant observations: essentially those with nearby $x(\cdot)$, and hence similar y , to the i th observation.

The same basic concept—namely, predicting the response by preferentially weighting observations that are nearby in a relevant sense—underlies previous methodology for non-parametric scalar-on-function regression (see Reiss et al., 2016, for a review). But unlike previous proposals, our approach can be implemented using existing software for generalized additive and related models (e.g., Wood, 2006, 2011). This allows it to be extended readily to the wide range of models available with such software, including models with generalized linear responses, multiple functional predictors, and/or random effects.

In §2, after providing the necessary background on principal coordinates, we present PCoRR more formally and describe our implementation. In §3 we relate PCoRR to existing linear and nonparametric approaches to scalar-on-function regression and to kernel methods. An application to signature verification in §4 shows how PCoRR can outperform existing methodology in a real data setting, and concluding remarks are offered in §5.

2 Methodology

2.1 A quick introduction to principal coordinates

Let $\mathbf{D} = (d_{ij})_{1 \leq i, j \leq n}$ be a symmetric matrix of distances with $d_{ii} = 0$ for all i and $d_{ij} > 0$ for all i, j . Classical multidimensional scaling (Torgerson, 1952; Gower, 1966) seeks n points in $\mathbb{R}^q$

(for some $q < n$) whose Euclidean distances are closest to $\mathbf{D}$. Appendix A explains what we mean by “closest,” and illustrates this optimal approximation for our toy data example.

Define the $n \times n$ centering matrix $\mathbf{H} = \mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n^T / n$ and let

$$\mathbf{K} = \mathbf{H} \left(-\frac{1}{2} d_{ij}^2 \right)_{1 \leq i, j \leq n} \mathbf{H} \quad (5)$$

have leading eigenvalues $\delta_1 \dots \delta_q > 0$, with corresponding eigenvectors $\mathbf{w}_{\cdot 1}, \dots, \mathbf{w}_{\cdot q}$ scaled so that $\|\mathbf{w}_{\cdot \ell}\|^2 = \delta_\ell$ for $\ell = 1, \dots, q$. Then the desired n points are given by the rows of the $n \times q$ matrix $\mathbf{W}_q = (\mathbf{w}_{ij})_{1 \leq i \leq n, 1 \leq j \leq q} = (\mathbf{w}_{\cdot 1} \dots \mathbf{w}_{\cdot q})$. These row vectors, which we may denote by $\mathbf{w}_1, \dots, \mathbf{w}_n$, are called the *principal coordinates* (PCo's) of the data.

To construct the PCo matrix $\mathbf{W}_q$ explicitly, consider the eigendecomposition

$$\mathbf{K} = \mathbf{U} \mathbf{\Delta} \mathbf{U}^T \quad (6)$$

where $\mathbf{U}^T \mathbf{U} = \mathbf{I}_n$ and $\mathbf{\Delta} = \text{Diag}\{\delta_1, \dots, \delta_n\}$ with $\delta_1 \dots \delta_n$. Let $\mathbf{U} = (\mathbf{U}_q \mathbf{U}_{-q})$ and

$\mathbf{\Delta} = \begin{pmatrix} \mathbf{\Delta}_q & 0 \\ 0 & \mathbf{\Delta}_{-q} \end{pmatrix}$ where $\mathbf{U}_q$ is $n \times q$ and $\mathbf{\Delta}_q$ is $q \times q$. From the above definition, we have

$$\mathbf{W}_q = \mathbf{U}_q \mathbf{\Delta}_q^{1/2}. \quad (7)$$

2.2 Ridge regression on principal coordinates

PCoRR was devised for situations, such as our toy example, in which there is a distance among the functional predictors $x_1, \dots, x_n$ such that nearby x 's tend to imply similar responses. In such settings, PCo's with respect to that distance offers a natural space in which to represent the functional predictors as multivariate data. Suppose, then, that we have a response vector $\mathbf{y} \in \mathbb{R}^n$, an $n \times p$ design matrix $\mathbf{M}$ of scalar covariates, and the $n \times q$ matrix $\mathbf{W}_q$ of PCo's based on an appropriate distance, with $p + q < n$. We could simply perform multiple regression with design matrix $(\mathbf{M} \mid \mathbf{W}_q)$. But in view of the possibly high dimension of this matrix, and for reduced sensitivity to the choice of q , it is preferable to perform this regression with a ridge penalty. The fitted values are then

$$\hat{\mathbf{y}} = \mathbf{M} \hat{\boldsymbol{\alpha}} + \mathbf{W}_q \hat{\boldsymbol{\gamma}} \quad (8)$$

where $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ minimizes the criterion

$$\|\mathbf{y} - \mathbf{M} \boldsymbol{\alpha} - \mathbf{W}_q \boldsymbol{\gamma}\|^2 + \lambda \boldsymbol{\gamma}^T \boldsymbol{\gamma} \quad (9)$$

for some $\lambda > 0$. This is the proposed PCoRR solution in its basic form.

Remark 1. When $\mathbf{M}$ is of full rank and orthogonal to the principal coordinates ($\mathbf{M}^T \mathbf{W}_q = \mathbf{0}$), the fitted value matrix has the explicit form

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{M}(\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{y} + \mathbf{U}_q \text{Diag} \left\{ \frac{\delta_\ell}{\delta_\ell + \lambda} \right\}_{1 \leq \ell \leq q} \mathbf{U}_q^T \mathbf{y} \\ &= \mathbf{M}(\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{y} + \sum_{1 \leq \ell \leq q} \frac{\delta_\ell}{\delta_\ell + \lambda} \mathbf{u}_\ell \mathbf{u}_\ell^T \mathbf{y},\end{aligned}\quad (10)$$

where $\mathbf{u}_\ell$ is the ℓ th column of $\mathbf{U}$. In other words, the projection of $\mathbf{y}$ on the ℓ th PCo direction is

multiplied by the “filter factor” $\frac{\delta_\ell}{\delta_\ell + \lambda}$ (Randolph et al., 2012), with preference (larger factors, or less shrinkage) given to directions corresponding to larger eigenvalues, resulting

in a fit with effective degrees of freedom $p + \sum_{\ell=1}^q \frac{\delta_\ell}{\delta_\ell + \lambda}$.

Remark 2. For a fit with an intercept only ($\mathbf{M} = \mathbf{1}_n$), (10) leads to the formula

$$h_{ij} = \frac{1}{n} + \sum_{k=1}^q \frac{w_{ik} w_{jk}}{\delta_k + \lambda} \quad (11)$$

for the hat matrix elements (4). (It is easily shown that $\sum_{j=1}^n h_{ij} = 1$ for each i , justifying the above interpretation of (4) as a weighted mean of the responses.) Expression (11) suggests that when the distance between the i th and j th observations is low, implying similar PCo's ($w_{ik} \approx w_{jk}$ for most or all k 's), h_{ij} will tend to be relatively high—as was the case above in the upper right subfigure of Figure 2.

2.3 Prediction for new observations

Expression (8) tells us only the fitted values for the n given observations (“design points”). Suppose we have n^* new observations of the functional predictor, along with an $n^* \times p$ matrix $\mathbf{M}^*$ of covariate values. To obtain predictions for these non-design points we must first “insert” them into the PCo configuration. In other words, we need an $n^* \times q$ matrix $\mathbf{W}_q^*$ of coordinates with respect to this configuration, which would yield predicted values

$$\hat{\mathbf{y}}^* = \mathbf{M}^* \hat{\boldsymbol{\alpha}} + \mathbf{W}_q^* \hat{\boldsymbol{\gamma}}. \quad (12)$$

A solution is given by applying Gower's interpolation (Gower, 1968) to each point. In Appendix B we provide a novel formulation of Gower's interpolation that is suitable for our regression setting.

2.4 Extending the model

Generalized additive modeling software, specifically the mgcv package (Wood, 2006, 2011) for R (R Core Team, 2016), allows for several important extensions to the PCoRR fit (8):

1. A generalized linear model (GLM) extension is achieved by minimizing the penalized deviance $D(\mathbf{a}, \boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^T \boldsymbol{\gamma}$, as opposed to the penalized sum of squares (9).
2. Since mgcv can perform fast optimal selection of multiple smoothing parameters, one can incorporate multiple functional predictors. Letting $\mathbf{W}_{q_r}^{(r)}$ denote the PCo matrix for the r th of R functional predictors, criterion (9) is then replaced by

$$\| \mathbf{y} - \mathbf{M}\boldsymbol{\alpha} - \sum_{r=1}^R \mathbf{W}_{q_r}^{(r)} \boldsymbol{\gamma}_r \|^2 + \sum_{r=1}^R \lambda_r \boldsymbol{\gamma}_r^T \boldsymbol{\gamma}_r, \quad (13)$$

with sum of squared errors again replaced by deviance for the generalized linear case.

3. Ruppert et al. (2003), Reiss and Ogden (2009) and Wood (2011), among others, propose to choose the penalty parameter λ (or parameters $\lambda_1, \dots, \lambda_R$) by maximum likelihood or restricted maximum likelihood (REML), or approximate versions thereof in the generalized linear case. Since these criteria are also the standard procedures for fitting mixed-effect models, their use allows for straightforward incorporation of random effects in the model.

2.5 Tuning parameter selection

In (13), we must choose both the dimension q_r and the ridge penalty parameter λ_r for $r = 1, \dots, R$. As noted above, we favor maximum likelihood or REML for the choice of λ_r . For q_r , in line with Wood (2006), we use generalized cross-validation (GCV; Craven and Wahba, 1979) for GLMs with unknown scale parameter (e.g., linear regression), and the Akaike information criterion (AIC; Akaike, 1973) for those with known scale parameter (e.g., logistic regression). Typically there is a minimum number of coordinates above which these criteria attain an approximate plateau (Miller and Wood, 2014), much as has been found regarding the number of knots for penalized splines (Ruppert, 2002); the lower subfigure of Figure 2 illustrates this phenomenon for our toy example. Consequently, when $R > 1$ it seems adequate to fix $q_1 = \dots = q_R = q$ for a sufficiently high value of q , perhaps 10 or 20, rather than painstakingly optimizing GCV or AIC over $(q_1, \dots, q_R)$.

2.6 Implementation

We have implemented PCoRR in the R package poridge (Principal co-Ordinate Ridge regression), which is publicly available at <https://github.com/dill/poridge>, with a vignette at <https://github.com/dill/poridge/blob/master/vignettes/PCoRR-vignette.pdf> that walks through the toy example of the Introduction. Essentially the package is an add-on to the mgcv package (Wood, 2006, 2011). It implements a non-standard “smooth constructor”

function that inputs a distance matrix $\mathbf{D}$ and a number of coordinates q , and enables the workhorse function `s()` to add the resulting term $+\mathbf{W}_q\boldsymbol{\gamma}$ to a generalized additive model, along with a ridge penalty. The functions in `poridge` have also been incorporated into the `refund` package for regression with functional data (Goldsmith et al. 2016), which is available on CRAN.

3 Relationships with other work

A number of previous approaches in the functional data literature and elsewhere have connections with PCoRR, some of which may not be immediately apparent. This section explains some of these relationships. Situating our method in relation to others may offer new insights and may help to suggest alternative avenues for future work.

3.1 Functional principal component regression

Relating our method to functional principal component regression requires some background. We first need to recall the notion of PCo-PCA “duality” developed by Gower (1966) for multivariate data. Given an $n \times p$ matrix $\mathbf{X}$ of n observations in $\mathbb{R}^p$, the key result is that the principal coordinates arising from Euclidean distances among the observations are equal to the principal component scores. For our purposes, this result needs to be extended to a more abstract setting that encompasses functional data. Consider, then, observations $x_1, \dots, x_n$ in a separable Hilbert space $\mathcal{F}$, such as $L^2(\boldsymbol{\tau})$ or a subspace thereof, equipped with an inner product $\langle \cdot, \cdot \rangle$ and the associated norm $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle}$. Principal components for such observations, which we shall refer to as $\mathcal{F}$ -PCs, are defined in Appendix C.1, and the following generalized duality result is proved in Appendix C.2.

Theorem 1. *Assume a set of observations $x_1, \dots, x_n$ in a separable Hilbert space $\mathcal{F}$ satisfies $\delta_1 > \dots > \delta_q > 0$ in (A.4) where $q \in \{1, \dots, n-1\}$. Then for $\ell = 1, \dots, q$, the vector of ℓ th principal coordinates with respect to the distance matrix $\mathbf{D} = (\|x_i - x_j\|_{\mathcal{F}})_{i,j=1}^n$ is equal, with a possible sign change, to the vector of scores with respect to the ℓ th $\mathcal{F}$ -PC.*

If $\mathcal{F}$ is a space of square-integrable functions on $\boldsymbol{\tau}$, Theorem 1 says that principal

coordinates with respect to the usual L^2 distance $\|f - g\| = \sqrt{\int_{\boldsymbol{\tau}} [f(t) - g(t)]^2 dt}$ are simply FPC scores. Consequently, regressing on PCo's with respect to an *arbitrary* distance among functions is a direct generalization of functional principal component regression.

In light of this, we can view several FPC-based methods for scalar-on-function regression, referred to in the first row of Table 1, as being generalized by the corresponding PCo-based methods in the second row. The first column of Table 1 refers to unpenalized linear regression on leading FPC scores, a popular approach to scalar-on-function regression (e.g., Cardot et al., 1999; Müller and Stadtmüller, 2005), as a special case of regression on leading PCo's. The latter method has been discussed extensively by Cuadras and Arenas (1990) and Cuadras et al. (1996), who refer to it as “distance-based regression,” and has been recently applied to functional predictors (Boj et al., 2015). In contrast to these unpenalized methods, the methods in the second column of Table 1 incorporate a penalty: the “FPCR_R” method of

Reiss and Ogden (2007) regresses on leading FPCs with a roughness penalty, whereas PCoRR adopts a ridge penalty (and thus is not an exact generalization of FPCR_R). As in the case of scatterplot smoothing, introducing a penalty enables one to use a richer basis, and the simulation results of Reiss and Ogden (2007) suggest that this yields improved performance for scalar-on-function regression (see the discussion of Horváth and Kokoszka, 2012). To be sure, Cuadras et al. (1996) also consider extensions of simple linear regression, including ridge regression; but our PCoRR formulation, which achieves the extensions outlined in §2.4 with automatically optimized tuning parameters, is new. The third column of Table 1 refers to additive models. These have been developed for FPC scores by Müller and Yao (2008), and might be studied for general PCo's in future work.

3.2 Nonparametric functional regression

A more flexible alternative to the functional linear model (1) is the nonparametric model

$$y = m(x) + \varepsilon, \quad (14)$$

where m is some mapping from the function space of interest to $\mathbb{R}$ and $E(\varepsilon|x) = 0$. To estimate m , one may extend nonparametric regression methodology from the case of scalar x to the case of functional x (Geenens, 2011). For example, the original proposal for nonparametric scalar-on-function regression (Ferraty and Vieu, 2006) was a functional version of the kernel smoothing method of Nadaraya (1964) and Watson (1964). Likewise, Baíllo and Grané (2009) extended local linear smoothing to the functional predictor setting.

Somewhat less transparently, roughness penalty approaches to nonparametric regression also have functional-predictor counterparts, as suggested by the two rightmost columns of Table 2. Given observations $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$, the smoothing spline approach to nonparametric regression (Wahba, 1990) posits a model of the form (14), and finds the estimator $\hat{m}_n$ that minimizes a penalized loss over a reproducing kernel Hilbert space $\mathcal{H}$ of functions $m: \mathbb{R} \rightarrow \mathbb{R}$. Considering for simplicity the case of squared error loss and $\|\cdot\|_{\mathcal{H}}$ having trivial null space, the smoothing spline estimator is

$$\hat{m}_n = \arg \min_{m \in \mathcal{H}} \left[\sum_{i=1}^n \{y_i - m(x_i)\}^2 + \lambda \|m\|_{\mathcal{H}}^2 \right] \quad (15)$$

for some $\lambda > 0$. The nonparametric functional regression estimator of Preda (2007) has the same form, but here $x_1, \dots, x_n$ belong to a function space $\mathcal{F}$, and $\mathcal{H}$ is an RKHS of maps $m: \mathcal{F} \rightarrow \mathbb{R}$. Briefly, the RKHS $\mathcal{H}$ considered by Preda (2007) is generated by a kernel $k: \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$ such as $k(f, g) = \exp(-\|f - g\|_{\mathcal{F}}^2 / 2\sigma^2)$ ($\sigma > 0$), in the sense that $\mathcal{H}$ is the completion of

$$\left\{ \sum_{\ell=1}^L a_{\ell} k(\cdot, f_{\ell}) : a_1, \dots, a_L \in \mathbb{R}, f_1, \dots, f_L \in \mathcal{F} \right\}$$

with respect to a specific inner product. By the representer theorem (Schölkopf et al., 2001), the minimizer in (15) is of the form $m(x) = \sum_{i=1}^n \beta_i k(x, x_i)$. For $\hat{m}$ of this form the right side of (15) can be shown to equal

$$\|y - K\beta\|^2 + \lambda \beta^T K \beta \quad (16)$$

where

$$K = [k(x_i, x_j)]_{1 \leq i, j \leq n} \quad (17)$$

and $\beta = (\beta_1, \dots, \beta_n)^T$. Thus $\hat{m}(x) = \sum_{i=1}^n \hat{\beta}_i k(x, x_i)$ where $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_n)^T$ minimizes (16). In the machine learning literature this is referred to as kernel ridge regression (KRR; e.g., Shawe-Taylor and Cristianini, 2004).

As noted in the Introduction, PCoRR can be viewed as a new approach to nonparametric functional regression. To situate it in relation to previous approaches, let us proceed to the fourth column of Table 2. In the top entry of that column we find penalized intermediate-rank splines (e.g., Green and Silverman, 1994; Ruppert et al., 2003; Wood, 2006). In comparison to smoothing splines, these approaches can be viewed as approximate RKHS optimizers offering greater computational efficiency. Similarly PCoRR can be thought of as approximating the RKHS approach of Preda (2007) to functional regression. To see this, suppose that K in (16) arises not as the Gram matrix (17), but from a distance matrix as in (5). (See Appendix D for discussion of the correspondence between distance and kernel matrices.) Suppose we impose the restriction

$$\beta = U_q \Delta_q^{-1/2} \gamma \quad (18)$$

for some $\gamma \in \mathbb{R}^q$. Substituting into (16), and using (6) and (7), yields the criterion $\|y - W_q \gamma\|^2 + \lambda \gamma^T \gamma$, which is simply the PCoRR criterion (9) (where the covariate part $M\alpha$ has been omitted for simplicity). Since restriction (18) confines $K\beta$ to the column space of U_q , an optimal q -dimensional approximation in the sense of Eckart and Young (1936) (as in Reiss and Ogden, 2010), PCoRR can be viewed as optimally truncated KRR. PCoRR, then, seems particularly reminiscent of thin plate regression splines (Wood, 2003), which are essentially thin plate splines with a similar optimal truncation step. At any rate, our aim here, in highlighting the analogy between PCoRR and intermediate-rank spline smoothing, has been to present our method as a natural, and potentially fruitful, next step in the development of nonparametric scalar-on-function regression.

3.3 Other kernel methods

Our interpretation of PCoRR as truncated KRR implies that multiple-term models, such as the optimizer of (13), constitute a form of *multiple kernel learning* (see Gönen and

Alpaydm, 2011, for a review). The standard viewpoint of multiple kernel learning differs from ours in several respects, however: one starts with kernels $k_1(\cdot, \cdot), \dots, k_K(\cdot, \cdot)$ as opposed to distances, the aim is usually classification (as opposed to general exponential family outcomes), and the idea is to combine the R kernels into a single kernel in some optimal manner.

The viewpoint of Corrada Bravo et al. (2009) is in a sense intermediate between ours and that of the kernel literature. Like us, these authors begin not with a kernel, but with a distance function that may not be Euclidean and need not even satisfy the triangle inequality (although they reserve the term “distance” for functions that do satisfy the triangle inequality, and use “dissimilarity” for those that do not). But as in most of the kernel literature, Corrada Bravo et al. (2009) require a positive semidefinite matrix $\mathbf{K}$. The “pseudo-attributes” entered into their smoothing spline ANOVA model have the same form as the PCo’s (7), but arise from a positive semidefinite $\mathbf{K}$, which they obtain by regularized kernel estimation (Lu et al., 2005; Zhang et al., 2016). A simpler way to derive a positive semidefinite $\mathbf{K}$ from a non-Euclidean distance matrix $\mathbf{D}$ is to add a suitable positive constant to the non-diagonal entries of $\mathbf{D}$, and then apply (5). The smallest such constant that yields a positive semidefinite $\mathbf{K}$ was derived by Cailliez (1983). This approach is available as an option in our PCoRR implementation.

4 Application: Signature verification

4.1 The data set

We now consider part of the sample data from the First International Signature Verification Competition (SVC 2004), available at <http://www.cse.ust.hk/svc2004/>. Each individual in the sample contributed 20 genuine signatures, which were accompanied by 20 skilled forgeries. For each signature we have x - and y -coordinates recorded at ≈ 150 – 300 time points. The challenge is to design an algorithm that can distinguish the genuine signatures from the fakes. Geenens (2011) considers a functional NW estimator for these data. For illustration we consider the data for one individual. Figure 3 displays five of the 20 instances of this individual’s Chinese signature, along with five of the corresponding forgeries.

4.2 Dynamic time warping

Noting that dynamic time warping approaches are considered state-of-the-art for signature verification (Kholmatov and Yanikoglu, 2005; Houmani et al., 2012), we apply DTW-based PCoRR to this data set. Briefly, DTW refers to a set of dynamic programming algorithms for optimally aligning discretely observed functions $f(s)$, $s = 1, \dots, S$ and $g(t)$, $t = 1, \dots, T$. The alignment is optimal in the sense that one chooses pairs $(s_1, t_1), \dots, (s_K, t_K)$ to minimize a normalized weighted average of

$$|f(s_k) - g(t_k)|, \quad (19)$$

subject to a set of constraints including (i) $s_1 = t_1 = 1$, $s_K = S$, $t_K = T$ and (ii) $s_{k+1} - s_k, t_{k+1} - t_k \in \{0, 1\}$ for each k . Constraints (i) and (ii) mean that we proceed from the beginning to

the end of both discretized functions by taking one step at a time along one or both. The minimized weighted average, referred to as the DTW distance, need not be a metric. Another possible constraint is $|s_k - t_k| \leq W$ for some $W > 0$ (Sakoe and Chiba, 1978); this was used in the toy example of the introduction to ensure that pairs of curves with nearby values of τ would have low distance.

4.3 Comparison

For this illustration we compare logistic PCoRR based on the DTW distance (a nonparametric approach, in the sense of §3.2) versus logistic functional PCR (a functional GLM) with a ridge penalty. We implemented FPCR as PCoRR based on L^2 distance (see §3.1) to eliminate irrelevant differences in implementation between FPCR and DTW-based PCoRR. Both approaches perform “online” signature verification, meaning that they use pen timing information in addition to the shape of the signature.

To implement FPCR, we transformed each signature’s time scale to the interval $[0, 1]$. Using the R package *fda* (Ramsay et al., 2014), we set up a basis of 40 quintic B-spline functions on $[0, 1]$, with which we performed penalized smoothing of each signature’s x - and y -coordinates with a second derivative penalty. Due to the irregularly spaced time points, automatic selection of the smoothing parameter λ (defined analogously to λ in (9) or (15)) led to undersmoothing for some of the curves, so a fixed value $\lambda = 2 \times 10^7$ was used. We then formed distance matrices between 1) the original curves, 2) the first derivatives and 3) the second derivatives; all three distance matrices were attempted as inputs to PCoRR. The smoothed x - and y -curves and their derivative estimates are displayed in Figure 4.

For DTW-based PCoRR, we calculated DTW distances among the original signatures and first- and second-differenced versions thereof, using the R package *dtw* (Giorgino, 2009). For bivariate time series such as these, Euclidean distance in $\mathbb{R}^2$ is substituted for absolute value in (19) to define the DTW distance. First vs. second PCo’s are plotted for the original data and for the differenced data in Figure 5. The scatterplots suggest that the leading DTW-based PCo’s for both the first- and the second-differenced data do a good job of separating the two groups.

This suggestion appears to be borne out by our comparative results, displayed in the lower subfigures of Figure 5, for two metrics of predictive performance:

1. The new modified AIC implemented in *mgcv*, which adjusts the degrees of freedom to account for smoothing parameter uncertainty (Wood et al., 2016).
2. Cross-validated area under the ROC curve (AUC; LeDell et al., 2015), implemented in the *cvAUC* package for R (LeDell et al., 2014). Here we averaged over 10 repetitions of 5-fold cross-validation.

In terms of AIC, DTW-based logistic PCoRR with differenced data attains better performance than either DTW with the raw data or FPCR with either the function estimates or their derivatives. DTW with differenced data generally attains the highest cross-validated AUC as well, although by this criterion, FPCR (without taking derivatives) performs comparably.

In §2.4 we referred to multiple-term models which, in the linear case, minimize criterion (13). For logistic regression with $R = 2$ terms and no scalar covariates, such a model might be expressed as

$$\text{logit}(\mathbf{p}) = \mathbf{W}_{q_1}^{(1)}\gamma_1 + \mathbf{W}_{q_2}^{(2)}\gamma_2 \quad (20)$$

where $\mathbf{p} = [\Pr(y_1 = 1), \dots, \Pr(y_n = 1)]^T$. To illustrate the potential utility of such models, we fitted logistic PCoRR models, with 2 PCo's, based on (i) L^2 distance among (i.e., FPCR with) the 2nd derivative curves, (ii) DTW distance among the 2nd-differenced data, and (iii) model (20) combining both of the above terms, i.e., $\mathbf{W}_{q_1}^{(1)} = \mathbf{W}_2^{(1)}$ based on L^2 distance and $\mathbf{W}_{q_2}^{(2)} = \mathbf{W}_2^{(2)}$ based on DTW distance. Fitted values (estimated probabilities that the signatures are genuine) are shown in Figure 6, with the first 20 and last 20 points, respectively, representing the genuine and forged signatures. For the two-term model (20), the estimated probability of being genuine is near 1 for all but two true signatures and near 0 for most of the forgeries. Does this near-perfect in-sample prediction reflect superior expected out-of-sample prediction, or mere overfitting? To this question, our two predictive performance criteria give conflicting answers: the two-term model has lower AIC than either of the single-term models, but slightly lower cross-validated AUC than the DTW-only model. Our goal here, however, is not to show unequivocally that this two-term model is best in this particular application, but rather to demonstrate how our framework makes it routine to fit multiple-term models.

5 Discussion

While our toy and real-data examples have involved DTW distance, PCoRR can equally well be applied with other distances among functional predictors, such as ISOMAP distance (Chen and Müller, 2012), the optimally weighted L^2 distance of Chen et al. (2014), or Mahalanobis distance (Galeano et al., 2015). PCoRR applies as well to non-functional data objects among which a distance or kernel is defined, as in the references cited in §3.3. Thus our contribution may be viewed either as a new approach to scalar-on-function regression, or more broadly as an extension of generalized additive models to incorporate data expressed as distances or kernels.

Our only requirements for the distance among data points are those given at the beginning of §2.1. A downside of this flexibility is that for non-Euclidean distances the matrix $\mathbf{K}$ in (5) is not positive semidefinite and thus mainstream kernel learning theory does not apply. It may, however, be possible to derive error results for PCoRR that combine kernel-based error bounds as in Steinwart and Christmann (2008) with bounds on the approximation error due to truncation.

Another limitation of PCoRR is that it does not yield a coefficient function and hence is less interpretable than the functional linear model (1). It would therefore be helpful to have straightforward methodology for testing the null model (1) against the PCoRR alternative. We hope to present such methodology in a forthcoming paper.

Acknowledgments

The authors thank Lan Huo, Lei Huang, Huaihou Chen, Rong Jiao and Fabian Scheipl for their assistance in implementing the methods proposed here, and the Associate Editor and referees for their very helpful feedback. Preliminary versions of this paper were presented at the Banff International Research Station workshop “Frontiers in Functional Data Analysis” in July 2015, and at the International Workshop on Advances in Functional Data Analysis held at Universidad Carlos III de Madrid in November 2015. We thank the participants of both workshops for their helpful feedback.

Appendix A: Optimality of the principal coordinate representation

Here we explain in what sense classical multidimensional scaling offers an optimal approximation to the values in the $n \times n$ distance matrix $\mathbf{D}$, by presenting two optimality results whose proofs can be found in Mardia et al. (1979), Section 14.4.

Suppose first that $\mathbf{D}$ is Euclidean, i.e., there exist points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ for some p such that, for each i, j , d_{ij} equals the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Without loss of generality $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ may be taken to be centered ($\mathbf{X} = \mathbf{H}\mathbf{X}$). For $q < p$, any $p \times q$ matrix $\mathbf{L}$ with orthonormal columns defines a q -dimensional configuration

$$\hat{\mathbf{X}} = (\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n)^T = \mathbf{X}\mathbf{L}, \quad (\text{A.1})$$

which may be thought of as projecting each of the n data points from $\mathbb{R}^p$ onto $\mathbb{R}^q$ via the transformation $\hat{\mathbf{x}}_i = \mathbf{L}\mathbf{x}_i$; hence, letting $\hat{d}_{ij}$ denote the Euclidean distance between $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$, we have $\hat{d}_{ij} \leq d_{ij}$. The principal coordinate matrix $\mathbf{W}_q$ (7) is of the form (A.1), and it can be shown (Gower, 1966) to minimize $\sum_{i,j=1}^n (d_{ij}^2 - \hat{d}_{ij}^2)$ over all configurations of this form.

When $\mathbf{D}$ is not Euclidean we can no longer speak of an optimal projection from $\mathbb{R}^p$ to $\mathbb{R}^q$ as above; but principal coordinates attain an optimal approximation of the matrix $\mathbf{K} = (k_{ij})_{1 \leq i,j \leq n}$ in (5), in the following sense. Any $n \times q$ matrix $\hat{\mathbf{X}}$ gives rise to Euclidean distances $\hat{d}_{ij}$

among its rows and hence to a matrix $\hat{\mathbf{K}} = (\hat{k}_{ij})_{1 \leq i,j \leq n} = \mathbf{H} \left(-\frac{1}{2} \hat{d}_{ij}^2 \right)_{1 \leq i,j \leq n} \mathbf{H}$,

analogous to (5). Mardia (1978) showed that $\sum_{i,j=1}^n (k_{ij} - \hat{k}_{ij})^2$ is minimized by taking $\hat{\mathbf{X}} = \mathbf{W}_q$.

For the toy example of the Introduction, although the DTW distance matrix $\mathbf{D}$ is not Euclidean, Figure A.1 shows how distances in PCo space provide an increasingly good approximation to $\mathbf{D}$ as the dimension q increases. The approximation is quite good for $q = 9$, which is roughly where GCV attains its lowest point in the lower subfigure of Figure 2.

Figure A.1.

Distances among the toy-data functional observations in q -dimensional PCo space, for $q = 1, 3, 5, 9$, compared with the exact windowed DTW distance matrix.

Appendix B: Gower's interpolation

To obtain the required W_q^* in (12), let d_{mi}^* denote the distance from the m th new point to the i th design point, and let k_{ii} be the i th diagonal element of $\mathbf{K}$ in (5). We then take

$$W_q^* = \mathbf{K}^* \mathbf{U}_q \Delta_q^{-1/2}, \quad (\text{A.2})$$

where $\mathbf{K}^*$ is the $n^* \times n$ matrix with (m, i) entry $-\frac{1}{2} (d_{mi}^{*2} - k_{ii})$. Subtracting k_{ii} is a form of centering (Gower, 1968), and we use the notation $\mathbf{K}^*$ to suggest that this matrix is constructed by transforming the distances d_{mi}^* in a roughly similar manner to the transformation of the original distance matrix $\mathbf{D}$ to obtain $\mathbf{K}$.

At first glance, formula (A.2) has the appealing consequence that if we had $n^* = n$ new points such that $\mathbf{K}^* = \mathbf{K}$, the interpolated coordinates would be given by

$W_q^* = \mathbf{K} \mathbf{U}_q \Delta_q^{-1/2} = W_q$. The equality $\mathbf{K}^* = \mathbf{K}$ does not, however, hold in general given n new observations that are identical to the original ones. What *is* true in that case, assuming $\delta_q > 0$, is that

$$\mathbf{K}^* \mathbf{U}_q = \mathbf{K} \mathbf{U}_q. \quad (\text{A.3})$$

To see this, define $a_{rs} = -\frac{1}{2} d_{rs}^2$, $\bar{a}_r = \frac{1}{n} \sum_{s=1}^n a_{rs}$ and

$\bar{a} = \frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n a_{rs}$. Then $\mathbf{K} - \mathbf{K}^*$ has (m, i) entry

$\left(a_{mi} - \bar{a}_m - \bar{a}_i + \bar{a} \right) - \left(a_{mi} - \bar{a}_i + \frac{1}{2} \bar{a} \right) = \frac{1}{2} \bar{a} - \bar{a}_m$. Thus the m th row of $\mathbf{K} - \mathbf{K}^*$ is

$\left(\frac{1}{2} \bar{a} - \bar{a}_m \right) \mathbf{1}^T$, and that of $(\mathbf{K} - \mathbf{K}^*) \mathbf{U}_q$ is

$\left(\frac{1}{2} \bar{a} - \bar{a}_m \right) \mathbf{1}^T \mathbf{U}_q = \left(\frac{1}{2} \bar{a} - \bar{a}_m \right) \mathbf{1}^T \mathbf{K} \mathbf{U}_q \Delta_q^{-1} = 0^T$. Equation (A.3) follows,

implying that $W_q^* = W_q$ and hence, as one would expect in this case, the predicted values $\hat{y}^*$ are the same as the fitted values $\hat{y}$.

Appendix C: PCo-PCA duality in Hilbert spaces

C.1 $\mathcal{F}$ -principal components

The proof of PCo-PCA duality for multivariate data relies on the one-to-one correspondence between eigenvectors of $\mathbf{X}_c \mathbf{X}_c^T$ and those of $\mathbf{X}_c^T \mathbf{X}_c$, where $\mathbf{X}_c = \mathbf{H}\mathbf{X}$. To generalize the result, we need to introduce an abstract counterpart of the matrix $\mathbf{X}_c$, namely the bounded linear transformation $T_x: \mathcal{F} \rightarrow \mathbb{R}^n$ defined by

$$T_x f = \left(\langle x_1 - \bar{x}, f \rangle, \dots, \langle x_n - \bar{x}, f \rangle \right)^T.$$

It is readily checked that the adjoint of T_x is $T_x^*: \mathbb{R}^n \rightarrow \mathcal{F}$ given by $T_x^* v = \sum_{i=1}^n (v_i - \bar{v}) x_i$ where $v = (v_1, \dots, v_n)^T$. We then have, for $f \in \mathcal{F}$,

$$T_x^* T_x f = \sum_{i=1}^n \langle x_i - \bar{x}, f \rangle (x_i - \bar{x}) = \sum_{i=1}^n \langle x_i - \bar{x}, f \rangle x_i,$$

i.e., $T_x^* T_x$ is $n-1$ times the sample covariance operator on $\mathcal{F}$; and for $v \in \mathbb{R}^n$,

$$T_x T_x^* v = \mathbf{H} \mathbf{C} \mathbf{H} v \text{ where } \mathbf{C} = (\langle x_i, x_j \rangle)_{1 \leq i, j \leq n}$$

$T_x^* T_x$ has eigenexpansion

$$T_x^* T_x f = \sum_{\ell=1}^{n-1} \delta_\ell \langle \phi_\ell, f \rangle \phi_\ell \quad (\text{A.4})$$

where $\delta_1, \delta_2, \dots, \delta_{n-1} \geq 0$ and the ϕ_ℓ 's are orthogonal. One way to see that $n-1$ terms suffice in this expansion is to observe that for each eigenvalue-eigenelement pair (δ_ℓ, ϕ_ℓ) of $T_x^* T_x$, $\mathbf{H} \mathbf{C} \mathbf{H}$ has eigenvalue δ_ℓ since

$$\mathbf{H} \mathbf{C} \mathbf{H} (T_x \phi_\ell) = (T_x T_x^*) T_x \phi_\ell = T_x (T_x^* T_x) \phi_\ell = \delta_\ell T_x \phi_\ell; \quad (\text{A.5})$$

but $\mathbf{H} \mathbf{C} \mathbf{H} \mathbf{1}_n = \mathbf{0}$, so $\mathbf{H} \mathbf{C} \mathbf{H}$ has at most $n-1$ positive eigenvalues. For $\mathcal{F}$ a space of square-integrable functions, the $\mathcal{F}$ -PC expansion is essentially the FPC expansion as formulated by, for example, Dauxois et al. (1982) and Aguilera et al. (1999).

C.2 Proof of Theorem 1

Letting $c_{ij} = \langle x_i, x_j \rangle$ and $d_{ij} = \|x_i - x_j\|_{\mathcal{F}}$, we have

$$d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}, \quad (\text{A.6})$$

i.e., $\mathbf{D}$ is obtained from $\mathbf{C}$ by the “standard transformation” of Mardia et al. (1979), p. 402. Hence by Theorem 14.2.2 of Mardia et al. (1979), $\mathbf{HCH}$ equals the matrix $\mathbf{K}$ of (5). By the distinctness of the leading q eigenvalues, it suffices to show (i) that $T_x \phi_\ell$ the vector of scores with respect to the $\mathcal{H}$ -PC, is an eigenvector of $\mathbf{HCH}$ corresponding to that matrix’s $\mathcal{H}$ largest eigenvalue δ_ℓ and (ii) that $T_x \phi_\ell$ has squared norm δ_ℓ . But we have already established (i) in (A.5), and (ii) is clear since $\|T_x \phi_\ell\|^2 = \phi_\ell^T (T_x^* T_x \phi_\ell) = \delta_\ell (\phi_\ell^T \phi_\ell) = \delta_\ell$.

Appendix D: Relating distances and kernels

Expression (5) can be viewed as establishing a rough correspondence between distance matrices $\mathbf{D}$ and kernel matrices $\mathbf{K}$. However, the kernel (or Gram) matrix encountered in the kernel literature (given by $[k(x_i, x_j)]_{i,j=1}^n$ for a kernel function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$) is generally positive semidefinite, whereas $\mathbf{K}$ in (5) need not be. A fundamental result (Theorem 14.2.1 of Mardia et al., 1979) says that $\mathbf{K}$ is positive semidefinite if and only if $\mathbf{D}$ is Euclidean.

Working in the opposite direction, given a positive semidefinite kernel matrix $\mathbf{C}$, the standard transformation (A.6) defines a corresponding distance matrix $\mathbf{D}$. As noted in the proof of Theorem 1, the matrix $\mathbf{K}$ (5) associated with that distance matrix is equal to $\mathbf{HCH}$, i.e., to the centered version of the original kernel matrix $\mathbf{C}$ (Shawe-Taylor and Cristianini, 2004).

In summary, (5) and (A.6) establish a one-to-one correspondence between *Euclidean* distance matrices and *centered* kernel matrices.

It is worth noting that when $\mathbf{K}$ is defined as a kernel matrix—as opposed to being derived from a distance matrix via (5)—the matrix $\mathbf{W}_q$ defined as in §2.2 gives the leading *kernel principal components* of the data (Schölkopf et al., 1998). For further discussion of the connection between distances and kernels, see Schölkopf (2001) and Faraway (2012).

References

- Aguilera AM, Ocaña FA, Valderrama MJ. Forecasting with unequally spaced data by a functional principal component approach. *Test*. 1999; 8(1):233–253.
- Akaike, H. Second International Symposium on Information Theory. Akademiai Kiado; 1973. Information theory and an extension of the maximum likelihood principle; p. 267–281.
- Baíllo A, Grané A. Local linear regression for functional predictor and scalar response. *Journal of Multivariate Analysis*. 2009; 100(1):102–111.
- Boj E, Caballé A, Delicado P, Esteve A, Fortiana J. Global and local distance-based generalized linear models. *TEST*. 2015 in press.
- Cailliez F. The analytical solution of the additive constant problem. *Psychometrika*. 1983; 48(2):305–308.
- Cardot H, Ferraty F, Sarda P. Functional linear model. *Statistics and Probability Letters*. 1999; 45(1): 11–22.
- Chen D, Müller H-G. Nonlinear manifold representations for functional data. *Annals of Statistics*. 2012; 40(1):1–29.
- Chen H, Reiss PT, Tarpey T. Optimally weighted L^2 distance for functional data. *Biometrics*. 2014; 70:516–525. [PubMed: 26228660]

- Corrada Bravo H, Lee KE, Klein BE, Klein R, Iyengar SK, Wahba G. Examining the relative influence of familial, genetic, and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences*. 2009; 106(20):8128–8133.
- Craven P, Wahba G. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*. 1979; 31(4): 317–403.
- Cuadras C, Arenas C. A distance based regression model for prediction with mixed data. *Communications in Statistics—Theory and Methods*. 1990; 19(6):2261–2279.
- Cuadras CM, Arenas C, Fortiana J. Some computational aspects of a distance-based model for prediction. *Communications in Statistics—Simulation and Computation*. 1996; 25(3):593–609.
- Dauxois J, Pousse A, Romain Y. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*. 1982; 12(1):136–154.
- Eckart C, Young G. The approximation of one matrix by another of lower rank. *Psychometrika*. 1936; 1(3):211–218.
- Eilers PHC, Marx BD. Flexible smoothing with B -splines and penalties (with discussion). *Statistical Science*. 1996; 11(2):89–121.
- Fan, J., Gijbels, I. *Local Polynomial Modelling and its Applications*. Chapman & Hall; London: 1996.
- Faraway JJ. Backscoring in principal coordinates analysis. *Journal of Computational and Graphical Statistics*. 2012; 21(2):394–412.
- Ferraty, F., Vieu, P. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer; New York: 2006.
- Galeano P, Joseph E, Lillo RE. The Mahalanobis distance for functional data with applications to classification. *Technometrics*. 2015; 57(2):281–291.
- Geenens G. Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys*. 2011; 5:30–43.
- Giorgino T. Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of Statistical Software*. 2009; 31(7):1–24.
- Gönen M, Alpaydın E. Multiple kernel learning algorithms. *Journal of Machine Learning Research*. 2011; 12:2211–2268.
- Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. 1966; 53(3-4):325–338.
- Gower JC. Adding a point to vector diagrams in multivariate analysis. *Biometrika*. 1968; 55(3):582.
- Green, PJ., Silverman, BW. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall; Boca Raton, FL: 1994.
- Horváth, L., Kokoszka, P. *Inference for Functional Data with Applications*. Springer Science & Business Media; New York: 2012.
- Houmani N, Mayoue A, Garcia-Salicetti S, Dorizzi B, Khalil MI, Moustafa MN, Abbas H, Muramatsu D, Yanikoglu B, Kholmatov A, Martinez-Diaz M, Fierrez J, Ortega-Garcia J, Alcobé R, Fabregas J, Faundez-Zanuy M, Pascual-Gaspar JM, Cardeñoso Payo V, Vivaracho-Pascual C. BioSecure signature evaluation campaign (BSEC'2009): Evaluating online signature algorithms depending on the quality of signatures. *Pattern Recognition*. 2012; 45(3):993–1003.
- Goldsmith J, Scheipl F, Huang L, Wrobel J, Gellar J, Harezlak J, McLean MW, Swihart B, Xiao L, Crainiceanu C, Reiss PT. refund: Regression with Functional Data. 2016 R package version 0.1-16.
- Hyndman RJ, Shang HL. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*. 2010; 19(1):29–45.
- Kholmatov A, Yanikoglu B. Identity authentication using improved online signature verification method. *Pattern Recognition Letters*. 2005; 26(15):2400–2408.
- LeDell E, Petersen M, van der Laan M. cvAUC: Cross-validated area under the ROC curve confidence intervals. 2014 R package version 1.1.0.

- LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic Journal of Statistics*. 2015; 9(1):1583–1607. [PubMed: 26279737]
- Lu F, Kele S, Wright SJ, Wahba G. Framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(35):12332–12337. [PubMed: 16109767]
- Mardia KV. Some properties of classical multi-dimensional scaling. *Communications in Statistics—Theory and Methods*. 1978; 7(13):1233–1241.
- Mardia, KV., Kent, JT., Bibby, JM. *Multivariate Analysis*. Academic Press; New York: 1979.
- Marx BD, Eilers PHC. Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics*. 1999; 41(1):1–13.
- Miller DL, Wood SN. Finite area smoothing with generalized distance splines. *Environmental and Ecological Statistics*. 2014; 21(4):715–731.
- Müller H-G, Stadtmüller U. Generalized functional linear models. *Annals of Statistics*. 2005; 33(2): 774–805.
- Müller H-G, Yao F. Functional additive models. *Journal of the American Statistical Association*. 2008; 103:1534–1544.
- Nadaraya EA. On estimating regression. *Theory of Probability & Its Applications*. 1964; 9(1):141–142.
- Preda C. Regression models for functional data by reproducing kernel hilbert spaces methods. *Journal of Statistical Planning and Inference*. 2007; 137(3):829–840.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2016.
- Ramsay, JO., Silverman, BW. *Functional Data Analysis*. 2nd. Springer; New York: 2005.
- Ramsay JO, Wickham H, Graves S, Hooker G. *fda: Functional Data Analysis*. 2014 R package version 2.4.4.
- Randolph TW, Harezlak J, Feng Z. Structured penalties for functional linear models—partially empirical eigenvectors for regression. *Electronic Journal of Statistics*. 2012; 6:323. [PubMed: 22639702]
- Reiss PT, Goldsmith J, Shang HL, Ogden RT. Methods for scalar-on-function regression. *International Statistical Review*. 2016 to appear.
- Reiss PT, Huang L, Mennes M. Fast function-on-scalar regression with penalized basis expansions. *International Journal of Biostatistics*. 2010; 6(1) article 28.
- Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*. 2007; 102:984–996.
- Reiss PT, Ogden RT. Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B*. 2009; 71(2):505–523.
- Reiss PT, Ogden RT. Functional generalized linear models with images as predictors. *Biometrics*. 2010; 66(1):61–69. [PubMed: 19432766]
- Ruppert D. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*. 2002; 11(4):735–757.
- Ruppert, D., Wand, MP., Carroll, RJ. *Semiparametric Regression*. Cambridge University Press; Cambridge: 2003.
- Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 1978; 26(1):43–49.
- Schölkopf, B. The kernel trick for distances; *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*; MIT Press; 2001. p. 301–307.
- Schölkopf, B., Herbrich, R., Smola, AJ. In: Helmbold, D., Williamson, B., editors. *A generalized representer theorem*; *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT*; Berlin and Heidelberg: Springer; 2001. p. 416–426.
- Schölkopf B, Smola A, Müller K-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*. 1998; 10(5):1299–1319.

- Shawe-Taylor, J., Cristianini, N. Kernel Methods for Pattern Analysis. Cambridge University Press; New York: 2004.
- Steinwart, I., Christmann, A. Support Vector Machines. Springer; New York: 2008.
- Torgerson WS. Multidimensional scaling: I. Theory and method. *Psychometrika*. 1952; 17(4):401–419.
- Wahba, G. Spline Models for Observational Data. Society for Industrial Mathematics; Philadelphia: 1990.
- Wasserman, L. All of Nonparametric Statistics. Springer; New York: 2006.
- Watson GS. Smooth regression analysis. *Sankhya A*. 1964; 26:359–372.
- Wood SN. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B*. 2003; 65(1): 95–114.
- Wood, SN. Generalized Additive Models: An Introduction with R. Chapman & Hall; Boca Raton, FL: 2006.
- Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*. 2011; 73(1):3–36.
- Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*. 2016 in press.
- Zhang L, Wahba G, Yuan M. Distance shrinkage and Euclidean embedding via regularized kernel estimation. *Journal of the Royal Statistical Society: Series B*. 2016 to appear.

Figure 1.

Two instances of the noiseless functional predictors $x_i^0(2)$ in the toy example are shown in the left panel, with corresponding observed functions x_i shown in the middle panel. The right panel shows $n = 30$ such functional predictors, color-coded as in Hyndman and Shang (2010) by the corresponding responses (3).

Figure 2.

Upper left: 30×30 matrix of windowed DTW distances among the functional observations $x_i(\cdot)$, arranged in increasing order of τ_i . Lower: Generalized cross-validation scores for ridge regression on principal component scores, a linear approach to scalar-on-function regression, vs. the proposed PCoRR with DTW-based principal coordinates. Upper right: Hat matrix for DTW-based PCoRR with 10 principal coordinates.

Figure 3.

Five instances of a Chinese signature (above) along with five skilled forgeries thereof (below).

Figure 4.

First row: Smoothed x -coordinate curves and their first and second derivatives (in the online version, the 20 true signatures appear in black and the 20 forgeries in red). Second row: Same, for the y -coordinates. Third row: Plots of first vs. second FPC scores.

Figure 5.

Upper subfigures: First vs. second DTW-based principal coordinates for the original and differenced data. Lower subfigures: Performance comparisons of logistic ridge FPCR (L^2 -based PCoRR) vs. logistic DTW-based PCoRR for different numbers of PCo's. In the legend, “1” and “2” refer to first and second derivatives (for FPCR) or differences (for DTW).

Figure 6.

Fitted values from logistic PCoRR models based on (i) L^2 distance among 2nd-derivative curves (equivalent to logistic FPCR), (ii) DTW distance among 2nd-differenced curves, and (iii) model (20) incorporating both terms. In each case 5 PCo's are used.

Table 1

Selected references for some approaches to scalar-on-function regression via functional principal component bases (first row) and principal coordinate bases (second row).

Basis type	Linear (unpenalized)	Ridge (penalized)	Additive
FPC	Cardot et al. (1999)	Reiss and Ogden (2007)	Müller and Yao (2008)
PCo	Cuadras et al. (1996)	PCoRR	–

Table 2

Selected references for some approaches to regression with scalar predictors (first row) and functional predictors (second row).

If x is . . .	Kernel smoothing	Local polynomial	Full-rank penalized	Reduced-rank penalized
Scalar	Nadaraya (1964); Watson (1964)	Fan and Gijbels (1996)	Wahba (1990)	Eilers and Marx (1996); Ruppert et al. (2003); Wood (2003)
Functional	Ferraty and Vieu (2006)	Baíllo and Grané (2009)	Preda (2007)	PCoRR