Flexible penalized regression for functional data...and other complex data objects

Philip T. Reiss
Flexible penalized regression
for functional data
... and other complex data objects

Philip T. Reiss
New York University and University of Haifa
phil.reiss@nyumc.org http://works.bepress.com/phil_reiss

International Workshop on Advances in Functional Data Analysis
Universidad Carlos III de Madrid, Getafe
November 11, 2015

Research supported in part by U.S. NIH grant 1R01MH095836-01A1
Outline

Penalized smoothing

Functional regression

fMRI

PCoRR

Related work

Signatures
Our story begins with penalized spline regression.

- Data: \((x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2\) assumed to satisfy

\[ y_i = g(x_i) + \varepsilon_i \]

for some smooth function \(g\), where (say) \(\varepsilon_i\) are iid \(N(0, \sigma^2)\) for some \(\sigma > 0\).

- Suppose \(g(x) = \theta^T b(x)\) for some \(\theta \in \mathbb{R}^K\), where \(b(x) = [b_1(x), \ldots, b_K(x)]^T\) denotes a set of basis functions, say B-splines, evaluated at \(x\).
• Estimate coefficient vector $\theta$ in $g(x) = \theta^T b(x)$ by penalized least squares:

$$
\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^K} \left( \|y - B\theta\|^2 + \lambda \theta^T P\theta \right),
$$

where

• $y = (y_1, \ldots, y_n)^T$,

• $B = \begin{bmatrix} b_1(x_1) & \ldots & b_K(x_1) \\ \vdots & \ddots & \vdots \\ b_1(x_n) & \ldots & b_K(x_n) \end{bmatrix}$;

• $\lambda$ is a tuning parameter, and $P$ is a $K \times K$ matrix.
• Roughness penalty $\lambda \theta^T P \theta$ in criterion $\|y - B\theta\|^2 + \lambda \theta^T P \theta$ prevents overfitting by shrinking $\hat{g}(\cdot) = \hat{\theta}^T b(\cdot)$ toward

$$\{\theta^T b(\cdot) : \theta^T P \theta = 0\} = \{\theta^T b(\cdot) : \theta \in \text{null} P\}.$$

• **Smoothing parameter** $\lambda \geq 0$ controls the degree of shrinkage.

• E.g., $P = (\int b_i'' b_j'')_{1 \leq i, j \leq K} \rightarrow \theta^T P \theta = \int g''(x)^2 dx \rightarrow$ shrink toward linear $g$. 

![Graphs showing the effect of different lambda values](image-url)
Connection with linear mixed models

- With a change of basis

\[ B \longrightarrow (X | Z) \text{ where } \text{col}X = \text{null}P \]

(e.g., Wand and Ormerod, 2008), criterion \( \|y - B\theta\|^2 + \lambda \theta^T P\theta \) becomes

\[ \|y - X\beta - Z\uptheta\|^2 + \lambda \uptheta^T \uptheta, \]

which is \( \propto - \log[f(y|\uptheta)f(\uptheta)] \) for the linear mixed model

\[ y = X\beta + Z\uptheta + \varepsilon \]

with \( \uptheta \sim N(0, (\sigma^2/\lambda)I), \varepsilon \sim N(0, \sigma^2I) \).

- Can reduce nonparametric inference to mixed model inference, with smoothing parameter \( \lambda \) treated as a variance parameter (Ruppert et al., 2003).
More specifically, the correspondence between the penalized spline problem

$$\hat{\theta} = \arg\min_{\theta \in \mathcal{R}^K} \left( \| y - B\theta \|^2 + \lambda \theta^T P\theta \right)$$

(1)

and a linear mixed model

$$y = X\beta + Zu + \varepsilon \text{ with } u \sim N(0, (\sigma^2/\lambda)I), \varepsilon \sim N(0, \sigma^2 I),$$

motivates basing inference on the restricted maximum likelihood (REML) criterion

$$\ell_R(\lambda|y) = -\frac{1}{2}(n-p) \log[ y^T \{ V^{-1}_\lambda - V^{-1}_\lambda X(X^T V^{-1}_\lambda X)^{-1} X^T V^{-1}_\lambda \} y ]$$

$$-\frac{1}{2} \log |V_\lambda| - \frac{1}{2} \log |X^T V^{-1}_\lambda X| \quad \text{(with } V_\lambda \equiv I_n + \lambda^{-1} ZZ^T :$$

\rightarrow \text{ Do optimal smoothing by taking } \lambda = \arg\max_{\lambda \geq 0} \ell_R(\lambda|y) \text{ in (1).}

Advantages of REML optimization approach:

1. Easy to incorporate random effects
2. Implement with existing mixed model software (?)
3. REML may be more stable than generalized cross-validation (Reiss and Ogden, 2009; Wood, 2011)
An R implementation

- The \texttt{mgcv} package (Wood, 2006) implements generalized additive models—i.e.,

\[
    h[E(y_i)] = \alpha + \sum_{j=1}^{J} g_j(x_j)
\]

for exponential family link $h$, with optimally chosen penalties for each $g_j$.

- Random effects available in \texttt{mgcv}, or in related package \texttt{gamm4} (Wood and Scheipl, 2014).

- Originally chose smoothing parameters by GCV/AIC, but now offers REML (Wood, 2011):

\[
    \text{gam}(y \sim s(x, \text{ method}="REML"))
\]

- Non-exponential family models recently added.
Outline

Penalized smoothing

Functional regression

fMRI

PCoRR

Related work

Signatures

References
Next let’s apply the penalized regression framework to

1. scalar-on-function regression
2. function-on-scalar regression
Scalar-on-function regression

Consider the functional linear model

\[ y_i = \int_s x_i(s)\beta(s)ds + \varepsilon_i, \]

\(i = 1, \ldots, n\) (intercept omitted for simplicity). Suppose functional predictors observed at grid of points \(s_1, \ldots, s_L\) resulting in data matrix

\[ X = [x_i(s_\ell)]_{1 \leq i \leq n, 1 \leq \ell \leq L}. \]

Taking \(\beta(s) = b(s)^T \theta\) again (Marx and Eilers, 1999) yields matrix form

\[ y = XWBB^\top \theta + \text{error} \]

where

- \(B = [b_k(s_\ell)]_{1 \leq \ell \leq L, 1 \leq k \leq K}\) is the basis evaluation matrix,
- \(W = \text{Diag}\{w_1, \ldots, w_L\}\) where the \(w_\ell\)'s are quadrature weights such that

\[ \sum_{\ell=1}^L x_i(s_\ell)w_\ell b(s_\ell)^T \theta \approx \int_s x_i(s)b(s)^T \theta ds, \]

and we can again proceed by penalized least squares:

\[ \hat{\theta} = \arg\min \left( \|y - XWB\theta\|^2 + \lambda \theta^T P \theta \right). \]

Idea: Why not “hack” penalized smoothing software for optimal choice of \(\lambda\)?
Ramsay and Silverman (2005) and Reiss et al. (2010) consider the model

$$y_i(s) = x_i^T \beta(s) + \epsilon_i(s)$$

Here the penalized sum of squares has a somewhat different form:

$$\| Y - X \Theta B^T \|_F^2 + \text{tr}(\Theta^T \Lambda \Theta P)$$

where $\Lambda = \text{Diag}\{\lambda_1, \ldots, \lambda_p\}$. But a little algebra puts this in the usual form

$$\| y - (B \otimes X)\theta \|^2 + \theta^T (P \otimes \Lambda) \theta$$

where $y = \text{vec}(Y)$, $\theta = \text{vec}(\Theta)$.

$\longrightarrow$ So, can use penalized smoothing software again.
R package refund (Huang et al., 2015) unifies a variety of functional regression problems by:

1. formulating the problem as penalized least squares
2. using mgcv for optimal tuning parameter selection, allowing extensions such as
   - multiple terms (coefficient functions)
   - generalized linear responses (penalized deviance)
   - random effects
Regression functions in the **refund** package

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Scalar</th>
<th>Functional</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Responses</strong></td>
<td><strong>Scalar</strong></td>
<td><strong>pfr</strong>*</td>
</tr>
<tr>
<td><strong>Functional</strong></td>
<td>fosr, pffr</td>
<td>pffr</td>
</tr>
</tbody>
</table>

* Largely supersedes older functions (fpcr, lpfr, peer, lpeer); includes functional principal component regression (Reiss and Ogden, 2007), functional generalized additive models (McLean et al., 2014), variable-domain functional regression (Gellar et al., 2014).

Spinoff packages:

- **refund.wave** (Huo et al., 2014): wavelet-based scalar-on-function regression (Reiss et al., 2015)
- **refund.shiny** (Wrobel and Goldsmith, 2015): interactive plotting for functional data
Outline

Penalized smoothing

Functional regression

fMRI

PCoRR

Related work

Signatures

References
• We consider data from a study that used functional MRI to find a “neurologic signature” of pain (Lindquist, 2012; Wager et al., 2013).

• \( n = 20 \) volunteers had hot (painful) and warm (non-painful) stimuli applied to the left forearm, repeatedly (39–48 trials per person).

• Design for each trial:
  • 0–18 s: Stimulus
  • 18–32 s: Fixation cross
  • 32–46 s: “How painful?”
  • Then, pain rated on scale from 100–550

• We fitted the model

\[
y_{ij} = \alpha_i + \gamma I_{ij}^{\text{hot}} + \int_0^{46} x_{ij}(t) \beta(t) dt + \varepsilon_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, J_i,
\]

in which

• \( y_{ij} \) is the log pain score for the \( i \)th participant’s \( j \)th trial,
• the \( \alpha_i \)'s are iid normally distributed random intercepts,
• \( I_{ij}^{\text{hot}} \) is an indicator for a hot stimulus,
• \( x_{ij}(\cdot) \) is BOLD (fMRI) response in lateral cerebellum,
• \( \varepsilon_{ij} \)'s are iid normally distributed errors with mean zero.

• N.B. Scalar covariate (stimulus type—hot/warm) and repeated measures.
Results for model \( y_{ij} = \alpha_i + \gamma I^{\text{hot}}_{ij} + \int_0^{46} x_{ij}(t) \beta(t) \, dt + \varepsilon_{ij} \):

- Hugely significant hot-vs.-warm effect (\( \hat{\gamma} = 0.599, p < 2 \cdot 10^{-16} \)).
- \( \beta(t) \equiv 0 \) also strongly rejected (Wood, 2013; Swihart et al., 2014).

Estimated coefficient function (center) resembles difference between mean BOLD signal for hot vs. warm trials (left).

\( \rightarrow \) Artifact of confounding?

Right subfigure suggests brain activity mediates the painful effect of heat (cf. Lindquist, 2012).
refund co-authors:

- Lei Huang
- Fabian Scheipl
- Jeff Goldsmith
- Jonathan Gellar
- Jaroslaw Harezlak
- Mathew W. McLean
- Bruce Swihart
- Luo Xiao
- Ciprian Crainiceanu

Thanks to Martin Lindquist for the fMRI data set.
Outline

Penalized smoothing

Functional regression

fMRI

PCoRR

Related work

Signatures
Acknowledgments for second part

This is joint work with

- Dave Miller
- Pei-Shien Wu
- Wen-Yu Hua

with contributions from Lei Huang, Lan Huo and Rong Jiao.
- So far we’ve discussed (generalized) linear models for functional data.

- Ferraty and Vieu (2006) initiated a new nonparametric program for FDA.

- Two advantages of nonparametric scalar-on-function regression:
  1. Sometimes functional (generalized) linear models are inadequate.
  2. Nonparametric models depend on distances (or semidistances) among observations → extend easily from functions to more general data “objects” (images, graphs, trees, etc.) that have a distance defined among them.
Given a set of functional predictors

\[ x_1(\cdot), \ldots, x_n(\cdot) \]

and scalar responses

\[ y_1, \ldots, y_n, \]

when might the functional linear model

\[ y_i = \alpha + \int x_i(t)\beta(t)dt + \varepsilon_i \]

prove inadequate?
Toy example: Suppose $i$th functional predictor $x_i : [0, 1] \rightarrow \mathbb{R}$ is a noisy realization of

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

for some $\tau_i \in [0, 1]$, and the responses arise from the model $y_i = \tau_i + \varepsilon_i$ with IID $\varepsilon_i \sim N(0, .05^2)$.

- $y$ depends strongly on $x(\cdot)$, but
- NOT in a manner that is captured by the linear model $y_i = \alpha + \int x_i(t)\beta(t)dt + \varepsilon_i$.
- → What to do??
A solution (for this example):
Dynamic time warping (DTW) distance

Given discretely observed functions

\[ f(s), s = 1, \ldots, S, \quad g(t), t = 1, \ldots, T, \]

the **DTW distance** between them is the minimized (weighted) average of 
\[ |f(s_k) - g(t_k)| \]
over \((s_1, t_1), \ldots, (s_K, t_K)\), subject to a set of constraints such as

- \(s_1 = t_1 = 1, \quad s_K = S, \quad t_K = T\)
- \(s_{k+1} - s_k, \quad t_{k+1} - t_k \in \{0, 1\}\)
If we add a constraint of the form $|s_k - t_k| \leq W$ (Sakoe-Chiba window),

then

$x_i(\cdot), x_j(\cdot)$ have low DTW distance $\iff \tau_i \approx \tau_j \iff y_i \approx y_j$. 
Performance (GCV) for toy data:
principal component regression (a functional linear model) vs. proposed principal coordinate regression (using DTW distance)
Principal coordinates

• Let $D = (d_{ij})_{1 \leq i,j \leq n}$ be a symmetric matrix of distances with $d_{ii} = 0 \ \forall i$, $d_{ij} \geq 0 \ \forall i, j$.

• Classical multidimensional scaling (Gower, 1966) seeks $n$ points in $\mathbb{R}^q$ (for some $q \leq n$) whose Euclidean distances are “closest” to $D$:
  • Let
    $$K = (I_n - 1_n 1_n^T / n) \left( -\frac{1}{2} d_{ij}^2 \right)_{1 \leq i,j \leq n}$$
  have the eigendecomposition $K = U \Delta U^T$ where $U^T U = I_n$ and $\Delta = \text{Diag}\{\delta_1, \ldots, \delta_n\}$ with $\delta_1 \geq \ldots \geq \delta_n$; assume $\delta_q > 0$.
  • Let $U = (U_q, U_{-q})$ and $\Delta = \begin{pmatrix} \Delta_q & 0 \\ 0 & \Delta_{-q} \end{pmatrix}$.

Then the desired $n$ points in $\mathbb{R}^q$, called the principal coordinates (PCo’s), are given by the rows of the $n \times q$ matrix $U_q \Delta_q^{1/2}$.

• Fundamental “duality” result (Gower, 1966): If $D$ gives Euclidean distances among $n$ points in $\mathbb{R}^m$ ($m \geq q$) then $U_q \Delta_q^{1/2}$ also gives the leading PCs for those data. $\rightarrow$ PCo can be viewed as a generalization of PCA.
Our proposal: Principal coordinate ridge regression

- Given response vector $y$, design matrix $C$ of scalar covariates, and functional predictors characterized by

$$\begin{align*}
\text{distance matrix} & \quad D \\
\text{kernel (?) matrix} & \quad K \\
\text{PCo's} & \quad U_q \Delta_q^{1/2},
\end{align*}$$

our proposal is simply to predict

$$\hat{y} = C\hat{\alpha} + U_q \Delta_q^{1/2} \hat{\gamma}$$

where $(\hat{\alpha}, \hat{\gamma})$ minimizes the ridge-type criterion

$$\|y - C\alpha - U_q \Delta_q^{1/2} \gamma\|^2 + \lambda \gamma^T \gamma$$

for some $\lambda > 0$.

- Generalized additive modeling software (Wood, 2011) can optimally choose $\lambda$ and allows for key extensions: (1) generalized linear responses, (2) multiple functional predictors, (3) random effects.
Example: Signature verification data

• We’ll consider part of the sample data from the First International Signature Verification Competition (SVC 2004), available at http://www.cse.ust.hk/svc2004/ (Geenens, 2011).

• 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 $\approx 150$–$300$ time points.

• Task: Design an algorithm that can distinguish the genuine signatures from the fakes.

• DTW approaches are considered state-of-the-art for signature verification (Kholmatov and Yanikoglu, 2005; Houmani et al., 2012).
Preview:
Logistic regression on DTW-based principal coordinates to distinguish genuine from fake signatures
Outline

Penalized smoothing

Functional regression

fMRI

PCoRR

Related work

Signatures
Relation to previous work, 1

1. **Distance-based regression** (regression on principal coordinates) has been developed extensively (Cuadras and Arenas, 1990; Cuadras et al., 1996), and specifically applied with functional predictors (Boj et al., 2015).

2. **Functional principal component regression**: When the distance among functions is $L^2$ distance, our method reduces to ridge regression on functional principal components (Aguilera et al., 1999).

<table>
<thead>
<tr>
<th>Basis type</th>
<th>Linear (unpenalized)</th>
<th>Ridge (penalized)</th>
<th>Additive</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCo</td>
<td>Cuadras et al. (1996)</td>
<td>PCoRR</td>
<td>—</td>
</tr>
</tbody>
</table>
Recall PCA-PCo duality theorem (Gower, 1966): PCo’s derived from Euclidean distance among \( x_1, \ldots, x_n \in \mathbb{R}^p \) are just PC scores.

More generally, given data \( x_1, \ldots, x_n \) in a Hilbert space \( \mathcal{F} \), define \( T_x : \mathcal{F} \rightarrow \mathbb{R}^n \) by \( T_x f = \left( \langle x_1 - \bar{x}, f \rangle, \ldots, \langle x_n - \bar{x}, f \rangle \right)^T \). Then \( T_x^* T_x \) is \( n - 1 \) times the sample covariance operator on \( \mathcal{F} \) and has eigenexpansion

\[
T_x^* T_x f = \sum_{\ell=1}^{n-1} \delta_\ell \langle \phi_\ell, f \rangle \phi_\ell.
\]

Call \( \phi_1, \phi_2, \ldots, \phi_{n-1} \) a (sample) \( \mathcal{F} \)-principal component basis of \( \mathcal{F} \).

For \( \mathcal{F} \) a space of square-integrable functions, the \( \mathcal{F} \)-PC expansion reduces to the functional PC expansion (Dauxois et al., 1982; Aguilera et al., 1999).

**Theorem.** If \( \delta_1 > \ldots > \delta_q > 0 \) (where \( 1 \leq q \leq n - 1 \)), then for \( \ell = 1, \ldots, q \),

- the vector of \( \ell \)th PCo’s with respect to the distance matrix \( D = (\|x_i - x_j\|_{\mathcal{F}})_{1 \leq i, j \leq n} \) and
- the vector of scores with respect to the \( \ell \)th \( \mathcal{F} \)-PC

are both given by \( T_x \phi_\ell \).
3. **Kernel methods:**

- Whereas we transform distance matrix $D$ to “similarity” matrix $K$, kernel methods (Shawe-Taylor and Cristianini, 2004) *begin* with the (positive semidefinite) matrix $K$.

- Our method is equivalent to minimizing the kernel ridge regression criterion $\|y - C\alpha - K\beta\|^2 + \lambda\beta^T K\beta$ (Hastie et al., 2009) with the restriction $\beta = U_q \Delta_q^{-1/2} \gamma$ for some $\gamma \in \mathbb{R}^q$.

- Preda (2007) proposed a kernel ridge regression approach to the nonparametric regression model

$$y_i = f(x_i) + \varepsilon_i$$

with functional predictors $x_i$.

<table>
<thead>
<tr>
<th>If $x$ is...</th>
<th>Kernel smoothing</th>
<th>Full-rank penalized</th>
<th>Reduced-rank penalized</th>
</tr>
</thead>
</table>
Outline

Penalized smoothing

Functional regression

fMRI

PCoRR

Related work

Signatures
We’ll consider a more difficult set of signatures . . .

. . . and compare logistic ridge regression on leading principal coordinates based on

- $L^2$ distance ($\rightarrow$ functional principal components, i.e. a functional GLM)
- DTW distance (i.e., a nonparametric model).
Penalized smoothing
Functional regression
fMRI
PCoRR
Related work
Signatures
References

DTW: raw data

DTW: 1st differences

DTW: 2nd differences

AIC comparison

First PCo vs. Second PCo for different transformations and AIC comparison.

- DTW: raw data
- DTW: 1st differences
- DTW: 2nd differences
- AIC comparison with various methods (FPCR, FPCR1, FPCR2, DTW, DTW1, DTW2).

The plots show the distribution of genuine and forgery points in the first two principal coordinates (PCo) for each transformation.
A model that includes FPCR and DTW-based PCoRR terms does better than either alone:
Summary

• We have outlined a framework for fitting functional linear models, and a flexible class of extensions, via software for optimally penalized ridge-type regression.

• Ridge regression on principal coordinates (PCoRR) is a new method for nonparametric scalar-on-function regression that takes advantage of the same computational framework.

• Many theoretical and practical issues have yet to be worked out!
Gracias!


References II


References III


