Appendix to "Penalized Functional Regression", Published in the *Journal of Computational and Graphical Statistics*

Jeff Goldsmith, Jennifer Bobb, Ciprian M. Crainiceanu,
Brian Caffo and Daniel Reich

August 23, 2010

A Additional Simulations

In this online appendix, we carry out simulations to explore the viability of our method in the multivariate, multilevel, and sparsely observed settings. Where applicable, we compare our method to competing approaches. All code used to conduct the simulations is also available online.

A.1 Multivariate Simulations

In this section, we pursue a simulation exercise to study the model presented in section 3.1.

We generate samples from the model

$$\begin{aligned}
Y_i &= \int_0^{10} X_{i1}(t)\beta_1(t)dt + \int_0^{10} X_{i2}(t)\beta_2(t)dt + \epsilon_i, \quad i = 1, \ldots, 200 \\
W_{i1}(t) &= X_{i1}(t) + \delta_{i1}(t);\ W_{i2}(t) = X_{i2}(t) + \delta_{i2}(t) \\
X_{i1}(t) &= u_{i11} + u_{i12}t + \sum_{k=1}^{10} \left\{ v_{ik11} \sin\left(\frac{2\pi k}{10}t\right) + v_{ik12} \cos\left(\frac{2\pi k}{10}t\right) \right\} \\
X_{i2}(t) &= u_{i21} + u_{i22}t + \sum_{k=1}^{10} \left\{ v_{ik21} \sin\left(\frac{2\pi k}{10}t\right) + v_{ik22} \cos\left(\frac{2\pi k}{10}t\right) \right\}
\end{aligned}$$

where $\epsilon_i \sim \text{N}[0, \sigma^2_\epsilon]$, $\delta_{li}(t_g) \sim \text{N}[0, \sigma^2_X]$, $u_{ij1} \sim \text{N}[0, 25]$, $u_{ij2} \sim \text{N}[0, 0.04]$, and $v_{ikj1}, v_{ikj2} \sim \text{N}[0, 1/k^2]$; thus $X_{i1}(t), X_{i2}(t)$ are uncorrelated functions sampled in the manner described in the univariate simulations. We assume $I = 200$ subjects, and select $\beta_1(t) = 4\sin(\pi t/5)$ and $\beta_2(t) = (t/5)^2$; note that the magnitude of the coefficient functions have been altered so that the contributions of the two predictors on the outcome are similar.

We simulate 1000 such datasets, and use the method given in section 3.1 with $K_x = K_b = 20$ to fit the multivariate functional regression. For comparison, we implement a straightforward extension of the PCR-PVE method used in the univariate case. That is, we

regress on the principal component loadings for each of the regressor functions (decomposed separately), choosing the number of loadings for each regressor function based on the percent of the variance explained by the eigenfunctions. Note that, similarly to section 4.1, the $\beta_1(t)$ is exactly a principal component of the $X_{i1}(t)$, which favors PCR-PVE, while $\beta_2(t)$ is an arbitrary smooth function.

Method	True $\beta_1(\cdot)$		True $\beta_2(\cdot)$	
	$\sigma_\epsilon^2 = .5$	$\sigma_\epsilon^2 = 1$	$\sigma_\epsilon^2 = .5$	$\sigma_\epsilon^2 = 1$
PFR				
σ_X^2 = 0	0.0100	0.0176	0.0011	0.0016
σ_X^2 = 1	0.0625	0.0639	0.0095	0.0100
PCR-PVE				
σ_X^2 = 0	0.0390	0.0396	0.0100	0.0106
σ_X^2 = 1	0.0536	0.0541	0.0241	0.0246

Table A.1: Mean MSE over the 1000 simulated multivariate functional regression models for the method presented in this manuscript and an adapted PCR-PVE method.

Table A.1 shows the results of the simulation study of the multivariate functional regression model. While a direct comparison with the univariate simulation is difficult because of the changes in the magnitude of the coefficient functions, we tend to see larger AMSEs in the multivariate setting. This is not surprising: without adding subjects, we have increased the complexity of the model. However, we again see that the PFR approach outperforms the modified PCR-PVE method for $\beta_2(t)$ and is comparable for $\beta_1(t)$. It is interesting to note, however, that PFR outperforms PCR-PVE for $\beta_1(t)$ in the absence of measurement error, which is not the case in the univariate setting.

A.2 Multilevel Simulations

Next we pursue a brief simulation exercise to examine the performance of our proposed method in the multilevel setting (Section 3.2).

We generate samples from the model

$$\begin{aligned} Y_i &= \int_0^{10} X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, \ldots, 200 \\ W_{ij}(t) &= X_i(t) + U_{ij}(t) + \delta_{ij}(t), \quad j = 1, \ldots, 3 \\ X_{i1}(t) &= u_{i1} + u_{i2}t + \sum_{k=1}^{2} \left\{ v_{ik1} \sin\left(\frac{2\pi k}{10}t\right) + v_{ik2} \cos\left(\frac{2\pi k}{10}t\right) \right\} \\ U_{ij}(t) &= a \cdot f_1(t) + b \cdot f_2(t) + c \cdot f_3(t) \end{aligned}$$

where $\epsilon_i \sim \mathrm{N}[0, \sigma_\epsilon^2]$, $\delta_{li}(t_g) \sim \mathrm{N}[0, \sigma_X^2]$, $u_{i1} \sim \mathrm{N}[0, 25]$, $u_{i2} \sim \mathrm{N}[0, 0.04]$, $v_{ik1}, v_{ik2} \sim \mathrm{N}[0, 1/k^2]$, $a \sim N[0, \sigma^2 = 2]$, $b \sim N[0, \sigma^2 = 1]$, and $c \sim N[0, \sigma^2 = .5]$. Further, the components used in the construction of the $U_{ij}(t)$ are given by

$$f_1(t) = \frac{1}{\sqrt{10}};\ f_2(t) = \sqrt{\frac{3}{10}}\left(\frac{t}{5} - 1\right);\ f_3(t) = \sqrt{\frac{5}{10}}\left\{6\left(\frac{t}{10}\right)^2 - 6\left(\frac{t}{10}\right) + 1\right\}.$$

Note that the unobserved subject-specific mean functions have been altered to include only four functions from the Fourier basis. We choose $\beta_1(t) = 2\sin(t)$ and $\beta_2(t) = (t/2.5)^2$.

We generate 100 such data sets and fit the resulting models using the extension detailed in Section 3.2; note that fewer datasets were generated than in other simulations due to the computational expense involved in estimating the $X_i(t)$. Here (as in Di et al. (2009)), we estimate the subject-specific PC loadings in model (5) using Markov chain Monte Carlo because $\boldsymbol{\psi}_j^{(1)}$ and $\boldsymbol{\psi}_j^{(2)}$ are not mutually orthogonal. We compare the penalized functional regression presented here to the functional regression method described in Di et al. (2009), which is an extension of PCR-PVE.

Method	True $\beta_1(\cdot)$		True $\beta_2(\cdot)$	
	$\sigma_\epsilon^2 = .5$	$\sigma_\epsilon^2 = 1$	$\sigma_\epsilon^2 = .5$	$\sigma_\epsilon^2 = 1$
PFR				
σ_X^2 = 0	.0184	.1560	.0200	.1558
σ_X^2 = 0	.0187	.1778	.0193	.1778
PCR-PVE				
σ_X^2 = 1	.0140	.8049	.0151	.8060
σ_X^2 = 1	.0170	1.207	.0181	1.206

Table A.2: Mean MSE over the 100 simulated multivariate functional regression models for the method presented in this manuscript and an adapted PCR-PVE method.

Table A.2 shows the results of this simulation. As in the univariate setting, when the coefficient function is an arbitrary smooth function the PFR method performs several times better than the PCR-PVE method. Also as before, the PCR-PVE method slightly outperforms and PFR when the the coefficient function is taken to be an early principal component, though it is much more sensitive to the presence of measurement error.

A.3 Sparse Data Simulations

Our final simulations test the extension of our method to the case where the functional regressor is sparsely observed at the subject level but densely observed over subjects, as described in Section 3.3.

We generate samples from the model

$$\begin{aligned} Y_i &= \int_0^{10} X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, \ldots, 200 \\ W_i(t) &= X_i(t) + \delta_i(t) \\ X_{i1}(t) &= u_{i1} + u_{i2}t + \sum_{k=1}^{10} \left\{ v_{ik1} \sin\left(\frac{2\pi k}{10}t\right) + v_{ik2} \cos\left(\frac{2\pi k}{10}t\right) \right\} \end{aligned}$$

where $\epsilon_i \sim \mathrm{N}[0, \sigma_\epsilon^2]$, $\delta_{li}(t_g) \sim \mathrm{N}[0, \sigma_X^2]$, $u_{i1} \sim \mathrm{N}[0, 25]$, $u_{i2} \sim \mathrm{N}[0, 0.04]$, and $v_{ik1}, v_{ik2} \sim \mathrm{N}[0, 1/k^2]$. Sparseness at the subject level is introduced by uniformly sampling 10 points in T independently for each subject, so that for each subject we observe $[Y_i, \{W_i(t_{ij}) : t_{ij} \in [0, 10]\}]$, $i = 1, \ldots, 200$, $j = 1, \ldots, 10$; note however that the sampling takes place after the outcome is generated.

We simulate 1000 data sets in this way, and use the extension described in Section 3.3 to fit the functional regression model for sparsely observed subject-level data. For comparison, we include a modified PCR-PVE method for sparsely observed functional data; as elsewhere, we regress on the first few principal component loadings. For our simulations, we estimate the covariance operator $\Sigma^W(s,t)$ on the full grid T rather than a subset thereof, but we note that our code allows one to use a smaller grid to under-smooth the covariance operator. Figure A.1 shows a sample of sparsely observed functions, as well as the estimated function based on the PC decomposition of $\Sigma^W(s,t)$.

Table A.3 gives the results of the sparsely observed functional regression simulation. Because of the presence of a few very large MSEs, we include both the average and the median MSE; also, we indicate whether σ_ϵ^2, the measurement error variance, is known or unknown.

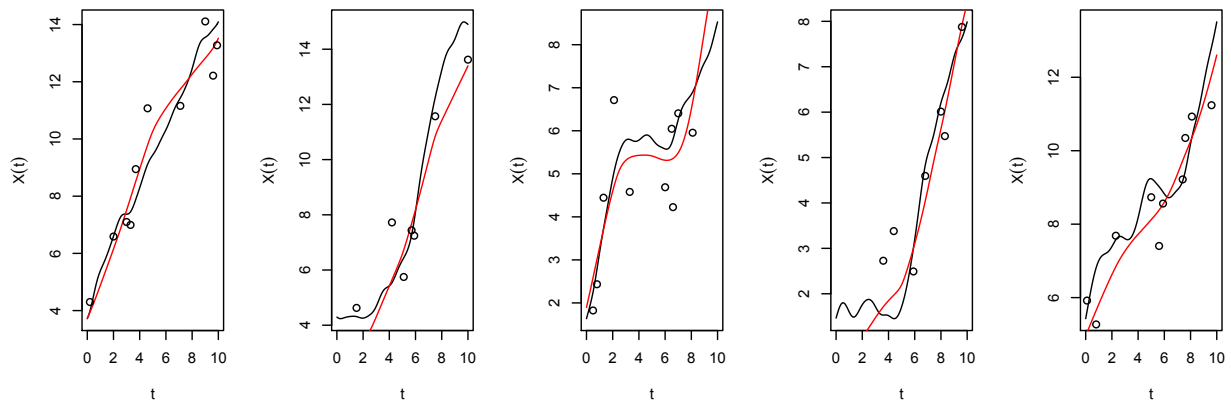

Figure A.1: A sample of sparsely observed functions, measured with error. For each panel, the black curve represents the true $X_i(t)$, the black points are observed points (the $X_i(t_{ij})$), and the red curve is the estimated function.

The results of this simulation are surprising. Focusing for the moment only on the PFR approach, we note that when the measurement error variance is unknown, the presence of error seems to greatly improve the estimation of the coefficient function. This seemingly paradoxical result stems from the estimation of the $X_i(t)$: in the case of no or little measurement error, the functions are systematically over- or underestimated at various regions of $[0, 10]$, leading to similar errors across subjects. In rare cases, this can cause PFR to wildly overestimate the coefficient function and leads to very high AMSEs. We point out, however, that the median MSE indicates that PFR typically performs well when the measurement error variance is unknown. Further, we note that if the measurement error variance is known the chance and severity of this issue is greatly diminished.

The comparison between the PFR and modified PCR-PVE approaches is similar to other settings. For $\beta_2(t)$, PFR outperforms PCR-PVE in AMSE and median MSE with the notable exception of AMSE when $\sigma^2_\epsilon = 0$ and is unknown. For $\beta_1(t)$, the methods have similar performances with, again, the exception of AMSE when the measurement error variance is unknown.

While the issues that arise when the measurement error variance is small and unknown are admittedly troubling, we again iterate that the large AMSEs stem from issues in the estimation of the $X_i(t)$ and are rare. Observing a larger sample size or observing functions on a denser grid would likely alleviate much of the issue. However, we are typically able estimate $\beta(t)$ fairly accurately with comparatively little information for each subject by borrowing information across subjects to estimate the $X_i(t)$.

Method	PFR				PCR-PVE			
	True $\beta_1(\cdot)$		True $\beta_2(\cdot)$		True $\beta_1(\cdot)$		True $\beta_2(\cdot)$	
	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$
σ_X^2 unknown								
σ_X^2 = 0	675900	236.5	166800	165500	.4008	.4009	33.73	33.73
	(.3122)	(.2981)	(.6528)	(.6897)	(.2539)	(.2545)	(23.68)	(23.71)
σ_X^2 = 1	6.112	6.231	7.659	9.208	.2919	.2927	38.87	38.85
	(.3582)	(.3675)	(.4399)	(.4552)	(.2508)	(.2523)	(26.66)	(26.70)
σ_X^2 known								
σ_X^2 = 0	.3980	1.221	8.877	8.254	.2396	.2398	32.67	32.67
	(.0619)	(.0601)	(.6911)	(.6724)	(.2055)	(.2047)	(22.70)	(22.60)
σ_X^2 = 1	6.983	4.919	5.755	6.920	.2719	.2727	36.44	36.43
	(.2698)	(.2736)	(.3656)	(.3795)	(.2368)	(.2358)	(25.23)	(25.21)

Table A.3: Mean MSE over the 1000 repetitions for each combination of the true coefficient function $\beta(t)$, the measurement error variance σ_X^2 and the outcome variance σ_ϵ^2. Median MSE is given in parentheses.