but introduce noise. Therefore a desirable strategy is to identify 'important' components out of a sufficiently large number of candidates, whereas to shrink those 'unimportant' ones to 0.

With these considerations, we seek an entirely new regularization and estimation framework for identifying the sparse structure of the FAM. Model selection that encourages sparse structure has received substantial attention in the last decade mostly due to the rapidly emerging high dimensional data. In the context of linear regression, the seminal works include the lasso (Tibshirani, 1996), the adaptive lasso (Zou, 2006), the smoothly clipped absolute deviation estimator (Fan and Li, 2001) and the references therein. Traditional additive models were considered by Lin and Zhang (2006), Meier *et al.* (2009) and Ravikumar *et al.* (2009); and extensions to generalized additive models were studied by Wood (2006) and Marra and Wood (2011). In comparison with these works, sparse estimation in functional regression has been much less explored. To our knowledge, most existing works are for functional linear models with sparse penalty (James *et al.*, 2009; Zhu *et al.*, 2010) or L^2-type penalty (Goldsmith *et al.*, 2011). Relevant research for additive structures is scant in the literature. In this paper, we consider selection and estimation of the additive components in FAMs that encourage a sparse structure, in the framework of a reproducing kernel Hilbert space (RKHS). Unlike in standard additive models, the FPC scores are not directly observed in FAMs. They need to be firstly estimated from the functional covariates and then plugged into the additive model. The estimated scores are random variables, which creates a major challenge to the theoretical exploration. It is necessary to take into account the influence of the unobservable FPC scores on the resulting estimator properly. Furthermore, the functional curve X is not fully observed either. We typically collect repeated and irregularly spaced sample points, which are subject to measurement errors. Measurement error in data adds extra difficulty for model implementation and inference. All of these issues are tackled in this paper. We propose a two-step estimation procedure to achieve the desired sparse structure estimation in FAMs. For the regularization, we adopt the COSSO (Lin and Zhang, 2006) penalty because of its direct shrinkage effect on functions in the RKHS. On the practical side, the method proposed is easy to implement, by taking advantage of existing algorithms of FPCA.

The rest of the paper is organized as follows. In Section 2, we present the proposed approach and algorithm, as well as the theoretical properties of the resulting estimator. Simulation results in comparison with existing methods are included in Section 3. We apply the proposed method to the Tecator data in Section 4, studying the regression of protein content on the absorbance spectrum. Concluding remarks are provided in Section 5, whereas details of the estimation procedure and technical proofs are deferred to the appendices.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Structured functional additive model regression

Let Y be a scalar response associated with a functional predictor $X(t), t \in \mathcal{T}$, and let $\{y_i, x_i(\cdot)\}_{i=1}^n$ be independent, identically distributed (IID) realizations of the pair $\{Y, X(\cdot)\}$. The trajectories $\{x_i(t) : t \in \mathcal{T}\}$ are observed intermittently on possibly irregular grids $\mathbf{t}_i = (t_{i1}, \ldots, t_{iN_i})^{\mathrm{T}}$. Denote the discretized $x_i(t)$ in vector form by $\mathbf{x}_i = (x_{i1}, \ldots, x_{iN_i})^{\mathrm{T}}$. We also assume that the trajectories are subject to IID measurement error, i.e. $x_{ij} = x_i(t_{ij}) + e_{ij}$ with $E(e_{ij}) = 0$ and $\mathrm{var}(e_{ij}) = \nu^2$. Following the FPCA of Yao *et al.* (2005) and Yao (2007), denote by $\boldsymbol{\xi}_{i,\infty} = (\xi_{i1}, \xi_{i2}, \ldots)^{\mathrm{T}}$ the sequence of FPC scores of x_i, which is associated with eigenvalues $\{\lambda_1, \lambda_2, \ldots\}$ with $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant 0$.

2.1. Proposed methodology

As discussed in Section 1, the theory of FPCA enables isomorphic transformation of random functions to their FPC scores, which brings tremendous convenience to model fitting and theoretical development in functional linear regression. To establish a framework for non-linear and non-parametric regression, we consider regressing the scalar responses $\{y_i\}$ directly on the sequences of FPC scores $\{\boldsymbol{\xi}_{i,\infty}\}$ of $\{x_i\}$. For the convenience of model regularization, we would like to restrict the predictor variables (i.e. FPC scores) to taking values in a closed and bounded subset of the real line, e.g. $[0, 1]$ without loss of generality. This is easy to achieve by taking a transformation of the FPC scores through a monotonic function $\Psi : \Re \to [0, 1]$, for all $\{\xi_{ik}\}$. In fact the choice of Ψ is quite flexible. A wide range of cumulative distribution functions (CDFs) can be used; see assumption 2 in Section 2.2 for the regularity condition. Additionally one may choose Ψ so that the transformed variables have similar or the same variations. This can be achieved by allowing $\Psi(\cdot)$ to depend on the eigenvalues $\{\lambda_k\}$, where $\{\lambda_k\}$ serve as scaling variables. For simplicity, in what follows we use a suitable CDF (e.g. normal), denoted by $\Psi(\cdot, \lambda_k)$, from a location–scale family with zero mean and variance λ_k. It is obvious that, if ξ_{ik}s are normally distributed, the normal CDF leads to uniformly distributed transformed variables on $[0, 1]$.

Denoting the transformed variable of ξ_{ik} by ζ_{ik}, i.e. $\zeta_{ik} = \Psi(\xi_{ik}, \lambda_k)$, and denoting $\boldsymbol{\zeta}_{i,\infty} = (\zeta_{i1}, \zeta_{i2}, \ldots)^{\mathrm{T}}$, we propose an additive model as follows:

$$y_i = b_0 + \sum_{k=1}^{\infty} f_{0k}(\zeta_{ik}) + \varepsilon_i, \tag{3}$$

where $\{\varepsilon_i\}$ are independent errors with zero mean and variance σ^2, and $f_0(\boldsymbol{\zeta}_{i,\infty}) = b_0 + \sum_{k=1}^{\infty} f_{0k}(\zeta_{ik})$ is a smooth function. For each k, let H^k be the lth-order Sobolev Hilbert space on $[0, 1]$, defined by

$$H^k([0,1]) = \{g | g^{(\nu)} \text{ is absolutely continuous for } \nu = 0, 1, \ldots, l-1; g^{(l)} \in L^2\}.$$

One can show that H^k is an RKHS equipped with the norm

$$\|g\|^2 = \sum_{\nu=0}^{l-1} \left\{ \int_0^1 g^{(\nu)}(t)\,\mathrm{d}t \right\}^2 + \int_0^1 g^{(l)}(t)^2\,\mathrm{d}t.$$

See Wahba (1990) and Lin and Zhang (2006) for more details. Note that H^k has the orthogonal decomposition $H^k = \{1\} \oplus \bar{H}^k$. Then the additive function f_0 corresponds to $\mathcal{F}$ which is a direct sum of subspaces, i.e. $\mathcal{F} = \{1\} \oplus \sum_{k=1}^{\infty} \bar{H}^k$ with $f_{0k} \in \bar{H}^k$, for all k. It is easy to check that, for any $f = b + \sum_k f_k \in \mathcal{F}$, we have $\|f\|^2 = b^2 + \sum_{k=1}^{\infty} \|f_k\|^2$. In this paper, we take $l = 2$ but the results can be extended to other cases straightforwardly. To distinguish the Sobolev norm from the L^2-norm, we write $\|\cdot\|$ for the former and $\|\cdot\|_{L^2}$ for the latter.

As motivated in Section 1, it is desirable to impose some type of regularization condition on model (3) to select important components. An important assumption that is commonly made in high dimensional linear regression is the sparse structure of the underlying true model. This assumption is also critical in the context of functional data analysis, which enables us to develop a more systematic strategy than the heuristic truncation that retains the leading FPCs. Although widely adopted, retaining the leading FPCs is a strategy that is guided solely by the covariance operator of the predictor X, and therefore it fails to take into account the response Y. To be more flexible, we assume that the number of important functional additive components that contribute to the response is finite, but not necessarily restricted to the leading terms. In particular, we denote $\mathcal{I}$ the index set of the important components and assume that $|\mathcal{I}| < \infty$, where $|\cdot|$ denotes the cardinality of a set. In other words, there is a sufficiently large s such that $\mathcal{I} \subseteq \{1, \ldots, s\}$, which implies that $f_k \equiv 0$ as long as $k > s$. The FAM is thus equivalent to

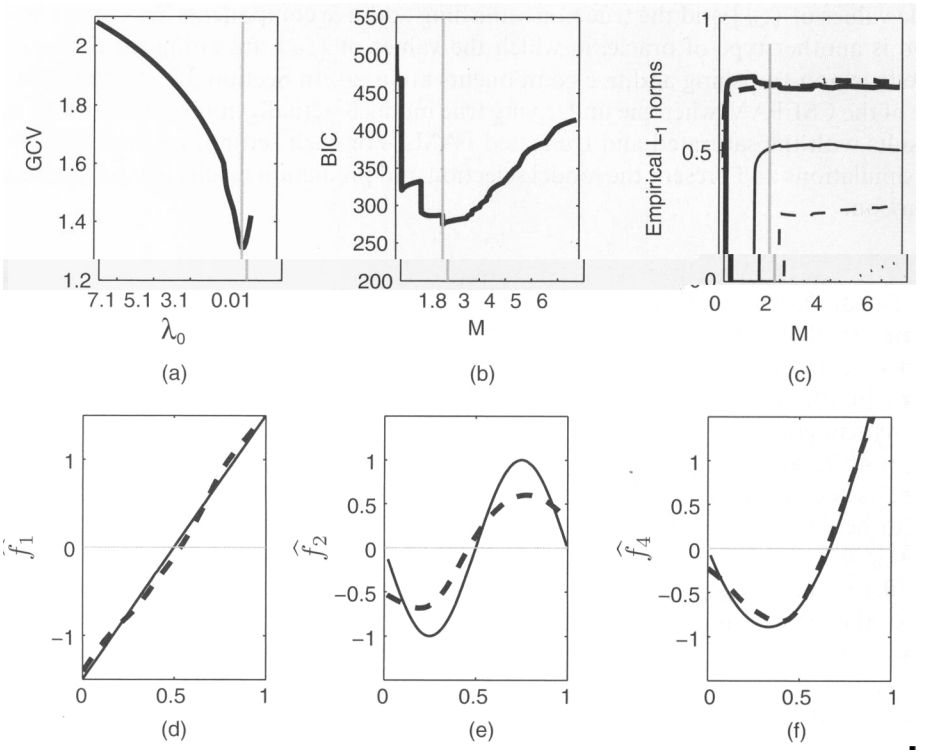

Fig. 1. Plots of component selection and estimation from one simulation: (a) generalized cross-validation *versus* λ_0; (b) BIC *versus* M; (c) empirical L_1-norms at various M-values (———, f_1; ———, f_2; – – –, f_3; ·–·–·, f_4; · · · · · ·, f_5) (⋮, tuning parameter chosen in (a)–(c)); (d)–(f) estimated f_ks (– – –) *versus* true f_ks (———) for $k = 1, 2, 4$

It is noted that the subjective truncation based on the explained variation in X is suboptimal for regression purpose (for conciseness the results are not reported). Therefore, in Table 1, we report (under the 'counts for the following model sizes' columns) the counts of selected numbers of non-vanishing additive components in the CSEFAM, and the counts of the number of significantly non-zero additive components in FAM, FAM_{O1} and FAM_{O2}. For convenience of display, only the counts for model size up to 8 are reported. The 'selection frequencies for the following components' columns of Table 1 record the number of times that each additive component is estimated to be non-zero for the first eight components. For the MARS method, if the jth component $\hat{f}_j$ is selected in one or more basis functions, we counted it as 1 and 0 otherwise. Regarding the prediction error (PE), we use the population estimate from the training set (e.g. the mean, covariance and eigenbasis) to obtain the FPC scores for both training and test set; then we apply the $\{\hat{f}_k\}$ estimated from the training set to obtain predictions for $\{y_i\}$ in the test set. The PEs are calculated by $n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. From the top panel of Table 1, we see that, under the dense design, the CSEFAM chooses the correct models (with model size equal to 3) 61% of the time whereas the FAM_{S} method always overselects ($\alpha = 0.05$ is used to retain significant additive components). The PE of CSEFAM is the smallest among the three non-oracle models. Compared with the oracle methods, the CSEFAM has less prediction power than FAM_{O2} (slightly) and FAM_{O1}, which can be regarded as the price paid by both estimating the ζ and selecting the additive components.

Table 3. Additional simulation for cases with non-sparse additive components†

Type	*Model*	*Counts for the following model sizes:*								*Selection frequencies for the following components:*								*PE*	*AISE of f*
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_4$	$\hat{f}_5$	$\hat{f}_6$	$\hat{f}_7$	$\hat{f}_8$		
I	CSEFAM	0	3	20	34	26	9	6	2	100	88	12	100	17	12	10	10	1.19 (0.08)	0.17
	FAM_S	0	4	16	18	29	17	7	8	100	92	16	99	20	17	19	16	1.33 (0.12)	0.33
	FAM_T	0	4	39	35	15	6	0	1	100	91	15	100	19	15	9	9	1.22 (0.08)	0.18
II	CSEFAM	1	2	4	12	13	20	26	13	46	42	33	42	36	42	38	44	1.25 (0.07)	0.12
	FAM_S	1	6	8	25	14	13	13	10	42	45	29	37	29	38	36	36	1.38 (0.11)	0.42
	FAM_T	13	30	22	20	6	6	2	0	34	35	20	31	25	38	34	30	1.32 (0.08)	0.20

†I, the true model contains both 'larger' and 'smaller' additive components; II, the true model contains only small additive components.

rest are 'smaller' additive components, each randomly selected from $\{f_{01}, f_{02}, f_{04}\}$ with equal probability and rescaled by a smaller constant uniformly chosen from $[1/17, 1/14]$. The data generated have a lower (more challenging) SNR around 0.60, among which 8.7% are from the 'smaller' components. The results are listed in the top panel of Table 3, which shows that the CSEFAM tends to favour smaller model size than FAM_S. We also observe that the model size of FAM_T tends to be smaller than for the CSEFAM since FAM_T adopts more truncation with the 99% threshold. It is important to note that the CSEFAM in fact yields PE and AISE that are substantially smaller than the FAM_S method, and the results of the CSEFAM are comparable with that of FAM_T. In study II, we replace the three larger components by the smaller ones; therefore all additive components have roughly equal small contributions. We select the scaling constant uniformly from $[\frac{1}{8}, \frac{1}{6}]$ so that the total SNR is 0.30 on average. The results listed in the bottom panel of Table 3 suggest that the CSEFAM now tends to select more components (i.e. to produce non-sparse fits) and again yields smaller PE and AISE than both the FAM_S and the FAM_T methods. Overall, this simulation suggests that the proposed CSEFAM is still a reasonable option even if the underlying true model is non-sparse. It is also worth mentioning that the gain of the CSEFAM is more apparent in the challenging settings with low SNR.

4. Real data application

We demonstrate the performance of the proposed method through the regression of protein

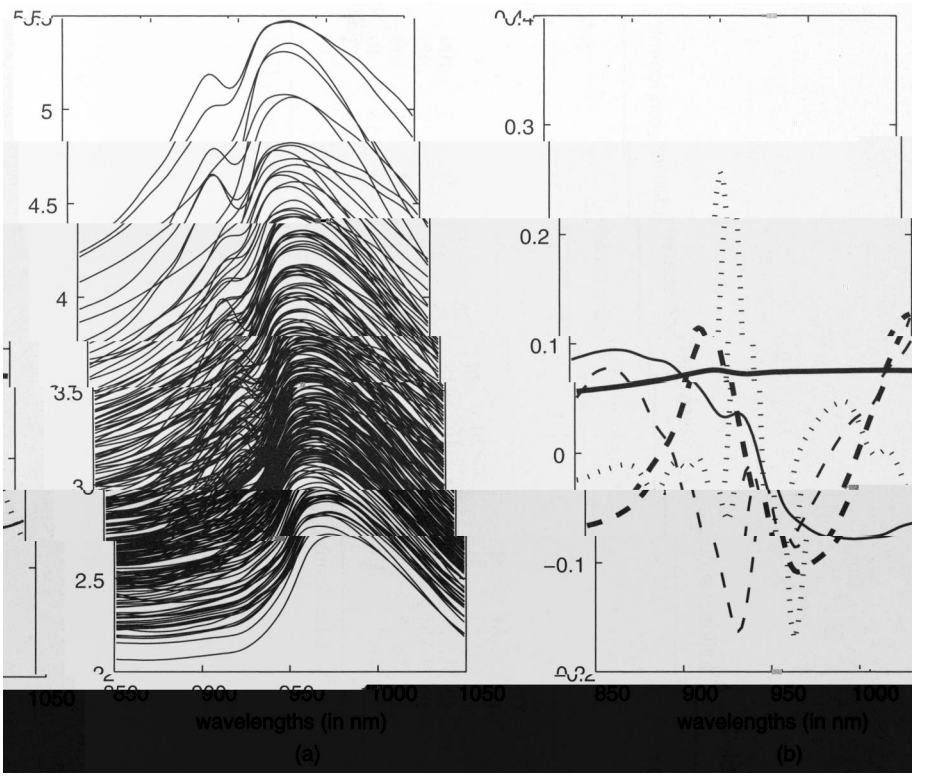

Fig. 2. (a) Near infrared absorbance spectral curves and (b) the first five estimated eigenfunctions (———, $\phi_1(t)$; ———, $\phi_2(t)$; — — —, $\phi_3(t)$; – – –, $\phi_4(t)$; · · · · · ·, $\phi_5(t)$)

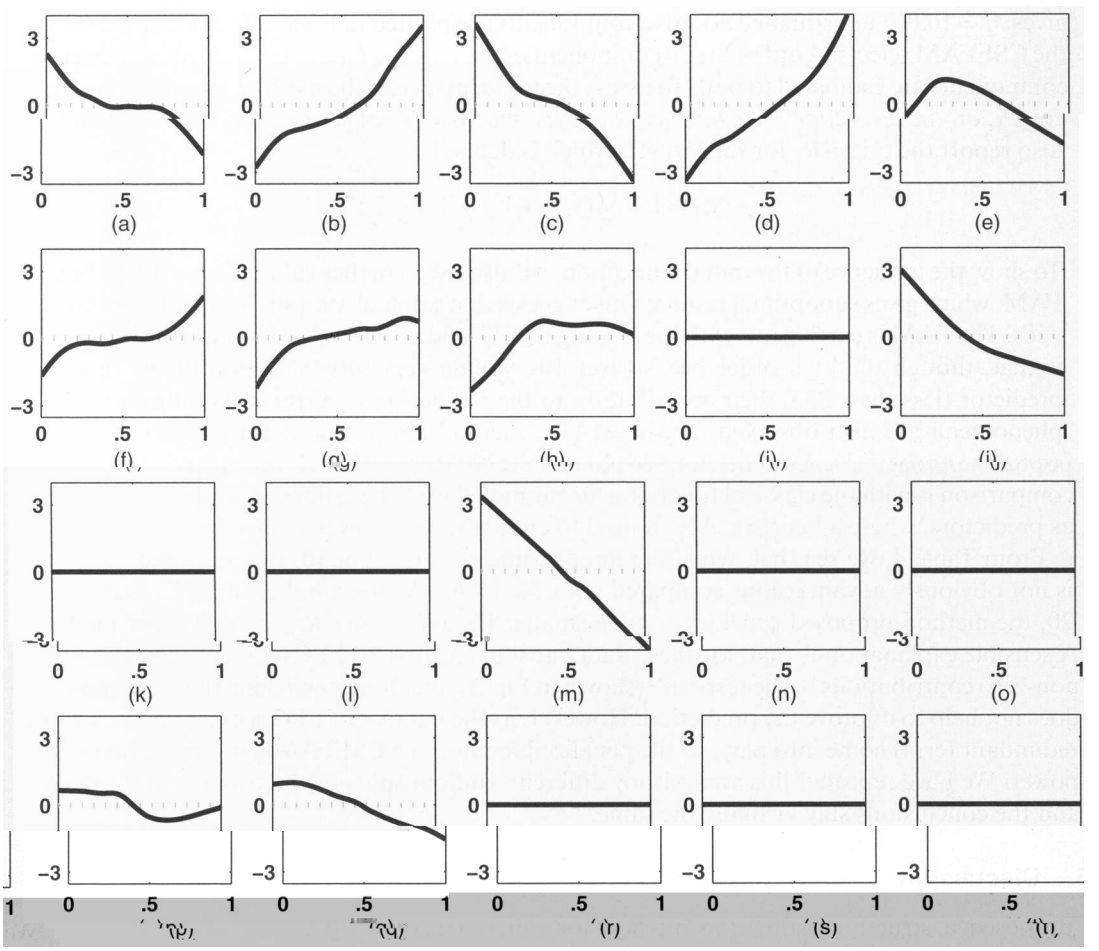

Table 4. Prediction results on the test set compared with several other methods†

	Results for the following methods:							
	CSEFAM		*FAM*			*MARS*	*Partial least squares,*	*Functional linear model,*
	$s=10$	$s=20$	*PC5*	*PC10*	*PC20*	*PC20*	*PLD20*	*AIC7*
PE	2.22	0.72	3.98	2.13	0.84	0.77	1.02	1.50
R^2_Q	0.82	0.94	0.68	0.83	0.93	0.93	0.92	0.88

†PC10 indicates that 10 FPC scores are used. PLD20 indicates that the number of partial least squares directions used is 20. AIC7 indicates that seven FPC scores are used based on the Akaike information criterion.

determination of the tuning parameters in the COSSO step is guided by the generalized cross-validation criterion for λ_0, which gives $\lambda_0 = 0.0013$, and by tenfold cross-validation for M, which gives $M = 10.0$. The estimated additive components are plotted in Fig. 3, from which we see that the CSEFAM selects 12 out of the 20 components, $\{\hat{f}_1, \ldots, \hat{f}_8, \hat{f}_{10}, \hat{f}_{13}, \hat{f}_{16}, \hat{f}_{17}\}$, and the other components are estimated to be 0. To assess the performance of the method proposed, we report the PE on the test set in Table 4, where the PE is calculated in the same way as in Section 3. We also report the quasi-R^2 for the test set, which is defined as

$$R^2_Q = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y}_i)^2.$$

To show the influence of the initial truncation, we also use a smaller value of s, $s = 10$ in the CSEFAM, which gives suboptimal results. This suggests that we shall use a sufficiently large s to begin with. The FAM is carried out with the leading five, 10 and 20 FPCs. An interesting phenomenon is that, though the high order FPCs (over 10) explain very little variation of the functional predictor (less than 1%), their contribution to the prediction is surprisingly substantial. Such phenomena are also observed for the MARS method and partial least squares (which is a popular approach in chemometrics; see Xu *et al.* (2007) and the references therein). One more comparison is with the classical functional linear model with the estimated leading FPCs served as predictors, where a heuristic AIC is used to choose the first seven components.

From Table 4, we see that, when the initial truncation is set at 10, the proposed CSEFAM is not obviously advantageous compared with the FAM. As the number of FPCs increases to 20, the method proposed provides a much smaller PE and higher R^2_Q than all other methods. A sensible explanation is that, for these data, most of the first 10 FPCs (except the ninth) have non-zero contributions to the response (shown in Fig. 3); therefore penalizing these components does not help to improve the prediction. However, as the number of FPC scores increases, more redundant terms come into play, so the penalized method the CSEFAM gains more prediction power. We have repeated this analysis for different random splits of the training and test sets, and the conclusions stay virtually the same.

5. Discussion

We proposed a structure estimation method for functional data regression where a scalar response is regressed on a functional predictor. The model is constructed in the framework of FAMs, where the additive components are functions of the scaled FPC scores. The selection

and estimation of the additive components are performed through penalized least squares using the COSSO penalty in the context of RKHS. The method proposed allows for more general non-parametric relationships between the response and predictors and therefore serves as an important extension of functional linear regression. Through the adoption of the additive structure, it avoids the curse of dimensionality that is caused by the infinite dimensional predictor process. The method proposed provides a way to select the important features of the predictor process and to shrink the unimportant ones to 0 simultaneously. This selection scenario takes into account not only the explained variation of the predictor process, but also its contribution to the response. The theoretical result shows that, under the dense design, the non-parametric rate from component selection and estimation will dominate the discrepancy due to the unobservable FPC scores.

A concern raised is whether the sparsity is necessary in the FAM framework. The sparseness assumption in general helps to balance the trade-off between variance and bias, which may lead to improved model performance. This can be particularly useful when part of the predictor has negligible contribution to the regression. Even if the underlying model is in fact non-sparse and we care only about estimation and prediction, the proposed CSEFAM is still a reasonable option, as illustrated by the simulation in Section 3.3. We also point out that, when all non-zero additive components are linear, the COSSO penalty reduces to the adaptive lasso penalty. An additional simulation (which for conciseness is not reported) has shown that the method proposed produces estimation and prediction results that are comparable with those of the adaptive lasso. Moreover, the COSSO penalty requires that $s < n$, which does not conflict with the requirement that the initial truncation s is chosen sufficiently large to include all important features. In practice the number of FPCs accounting for nearly 100% predictor variation is often far less than the sample size n owing to the fast decay of the eigenvalues. Finally both simulated and real examples indicate that the model performance is not sensitive to s as long as it is chosen to be sufficiently large.

On the computation side, our algorithm takes advantage of both FPCA and COSSO. On a desktop with Intel(R) Core(TM) i5-2400 central processor unit with a 3.10-GHz processor and 8 Gbytes random-access memory each Monte Carlo sample in Section 3.1 takes 30 s and the real data analysis takes about 10 s. As far as the dimensionality is concerned, the capacity and speed depend on the particular FPCA algorithm used. We have used the principal component analysis by conditional expectation algorithm PACE which can deal with fairly large data (`http://anson.ucdavis.edu/~ntyang/PACE/`). For dense functional data with 5000 or more dimensions, pre-binning is suggested to accelerate the computation. An FPCA algorithm geared towards extremely large dimensions (with an identical time grid for all subjects) is also available; for instance, Zipunnikov *et al.* (2011) considered functional magnetic resonance imaging data with dimension of the order of $O(10^7)$ through partitioning the original data matrix to blocks and performing singular value decomposition using blockwise operation.

Although we have focused on the FPC-based analysis in this work, the CSEFAM framework is generally applicable to other basis structures, e.g. splines and wavelets, where the additive components are functions of the corresponding basis coefficients of the predictor process. It may also work for non-parametric penalties other than COSSO, such as the sparsity smoothness penalty that was proposed in Meier *et al.* (2009). The method proposed may be further extended to accommodate categorical responses, where an appropriate link function can be chosen to associate the mean response with the additive structure. Another possible extension is regression with multiple functional predictors, where component selection can be performed for selecting functional predictors. In this case the additive components that are associated with each functional predictor need to be selected in a group manner.

Acknowledgements

This work was conducted through the 'Analysis of object data' programme at the Statistical and Applied Mathematical Sciences Institute, USA. Fang Yao's research was partially supported by an individual discovery grant and discovery accelerator supplement from the Natural Sciences and Engineering Research Council, Canada. Hao Helen Zhang was supported by US National Institutes of Health grant R01 CA-085848 and National Science Foundation grant DMS-0645293.

Appendix A: The estimation procedure

To estimate ζ_i, we assume that the functional predictors are observed with measurement error on a grid of $\mathcal{T}$. We adopt two different procedures for functional data that are either densely or sparsely observed.

(a) *Obtain $\hat{\zeta}_i$ in the dense design.* If $\{x_i(t)\}$ are observed on a sufficiently dense grid for each subject, we apply local linear smoothing to the data $\{t_{ij}, x_{ij}\}_{j=1,\ldots,N_i}$ individually, which gives the smooth approximation $\hat{x}_i(t)$. The mean and covariance function are obtained by $\hat{\mu}(t) = (1/n)\sum_{i=1}^n \hat{x}(t)$ and

$$\hat{G}(s,t) = (1/n)\sum_{i=1}^{n}\{\hat{x}_i(s) - \hat{\mu}(s)\}\{\hat{x}_i(t) - \hat{\mu}(t)\}$$

respectively. The eigenvalues and eigenfunctions are estimated by solving the equation

$$\int_{\mathcal{T}} \hat{G}(s,t)\,\phi_k(s)\,\mathrm{d}s = \lambda_k\,\phi_k(t)$$

for λ_k and $\phi_k(\cdot)$, subject to $\int_{\mathcal{T}} \phi_k^2(t)\,\mathrm{d}t = 1$ and $\int_{\mathcal{T}} \phi_m(t)\,\phi_k(t)\,\mathrm{d}t = 0$ for $m \neq k$, $k, m = 1, \ldots, s$. The FPC scores are obtained by $\hat{\xi}_{ik} = \int_{\mathcal{T}} \{\hat{x}_i(t) - \hat{\mu}(t)\}\,\hat{\phi}_k(t)\,\mathrm{d}t$. Finally CDF transformation yields $\hat{\zeta}_{ik} = \Psi(\hat{\xi}_{ik}; 0, \hat{\lambda}_k)$.

(b) *Obtain ζ_i in the sparse design.* We adopt the principal component analysis through the PACE algorithm that was proposed by Yao *et al.* (2005), where the mean estimate $\hat{\mu}(t)$ is obtained by using local linear smoothers based on the pooled data of all individuals. In particular,

$$\hat{\mu}(t) = \sum_{i=1}^{n}\sum_{j=1}^{N_i} K\{(t_{ij} - t)/b\}\{x_{ij} - \beta_0 - \beta_1(t - t_{ij})\}^2$$

with $K(\cdot)$ a kernel function and b a bandwidth. For the covariance estimation, denote $G_{ijl} = \{x_{ij} - \hat{\mu}(t_{ij})\}\{x_{il} - \hat{\mu}(t_{il})\}$ and let $K_h^*(\cdot,\cdot)$ be a bivariate kernel function with a bandwidth h. One minimizes

$$\sum_{i=1}^{n}\sum_{j\neq l} K^*\{(t_{ij} - s)/h, (t_{il} - t)/h\}\{G_{ijl} - \beta_{00} - \beta_{11}(s - t_{ij}) - \beta_{12}(t - t_{il})\}^2.$$

One may estimate the noise variance ν^2 by taking the difference between the diagonal of the surface estimate $\hat{G}(t,t)$ and the local polynomial estimate obtained from the raw variances $\{(t_{ij}, G_{ijj}) : j = 1, \ldots, N_i; i = 1, \ldots, n\}$. The eigenvalues or eigenfunctions are obtained as in the dense case. To estimate the FPC scores, denote $\mathbf{x}_i = (x_{i1}, \ldots, x_{iN_i})^{\mathrm{T}}$, the PACE estimate is given by $\hat{\xi}_{ik} = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^{\mathrm{T}} \hat{\Sigma}_{\mathbf{x}_i}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i)$, which leads to $\hat{\zeta}_{ik} = \Psi(\hat{\xi}_{ik}; 0, \hat{\lambda}_k)$, $k = 1, \ldots, s$. Here $\boldsymbol{\phi}_{ik} = (\phi_k(t_{i1}), \ldots, \phi_k(t_{iN_i}))^{\mathrm{T}}$, $\boldsymbol{\mu}_i = (\mu(t_{i1}), \ldots, \mu(t_{iN_i}))^{\mathrm{T}}$, and the (j,l)th element $(\boldsymbol{\Sigma}_{\mathbf{x}_i})_{j,l} = G(t_{ij}, t_{il}) + \nu^2\delta_{jl}$ with $\delta_{jl} = 1$ if $j = l$ and $\delta_{jl} = 0$ otherwise, and '$\hat{\ }$' is generic notation for the estimated parameters.

We next estimate $f_0 \in \mathcal{F}^s$ by minimizing expression (6), following the COSSO procedure conditional on the estimated values $\hat{\zeta}_i$. It is important to note that the target function (6) is equivalent to

$$(1/n)\sum_{i=1}^{n}\{y_i - f(\hat{\zeta}_i)\}^2 + \lambda_0\sum_{k=1}^{s}\theta_k^{-1}\|P^k f\|^2 + \lambda\sum_{k=1}^{s}\theta_k,$$

subject to $\theta_k \geqslant 0$ (Lin and Zhang, 2006), which enables a two-step iterative algorithm. Specifically, one first finds $\mathbf{c} \in \mathscr{R}^n$ and $b \in \mathscr{R}$ by minimizing

$$(\mathbf{y} - \mathbf{R}_\theta\mathbf{c} - b\mathbf{1}_n)^{\mathrm{T}}(\mathbf{y} - \mathbf{R}_\theta\mathbf{c} - b\mathbf{1}_n) + n\lambda_0\mathbf{c}^{\mathrm{T}}\mathbf{R}_\theta\mathbf{c}, \tag{9}$$

with fixed $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^{\mathrm{T}}$, where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, λ_0 is the smoothing parameter, $\mathbf{1}_n$ is the $n \times 1$ vector of

The following lemma characterizes the discrepancy between the underlying and estimated transformed variables ζ_{ik}, as well as the boundedness of the derivative of the resulting estimate $\hat{f}$.

Lemma 2. Under assumption 2 in Section 2.2 and condition 1–3, we have

$$|\hat{\zeta}_{ik} - \zeta_{ik}| = O_p[\lambda_k^{\gamma}\{\|\hat{X}_i - X_i\|_{L^2} + (\delta_k^{-1}\|X_i\|_{L^2} + |\xi_{ik}|)\|\hat{G} - G\|_S\}], \tag{14}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{s}|\hat{\zeta}_{ik} - \zeta_{ik}|\right)^2 = O_p(n^{-1}). \tag{15}$$

Additionally, if assumption 1 holds, let $\hat{f}$ be the estimate of f_0 obtained by minimizing expression (6). Then there is a constant $\rho > 0$, such that

$$\left|\frac{\partial \hat{f}(\boldsymbol{\zeta}_i)}{\partial \zeta_{ik}}\right| \leqslant \rho, \tag{16}$$

uniformly over $1 \leqslant k \leqslant s$ and $1 \leqslant i \leqslant n$.

B.1. Proof of lemma 2

From lemma 1 and assumptions 2, we have

$$\begin{aligned}
|\hat{\zeta}_{ik} - \zeta_{ik}| &= \left|(\hat{\xi}_{ik} - \xi_{ik})\frac{\partial}{\partial \xi_{ik}}\Psi(\xi_{ik}, \lambda_k) + (\hat{\lambda}_k - \lambda_k)\frac{\partial}{\partial \lambda_k}\Psi(\xi_{ik}, \lambda_k) + o_p(|\hat{\xi}_{ik} - \xi_{ik}| + |\hat{\lambda}_k - \lambda_k|)\right| \\
&\leqslant |\hat{\xi}_{ik} - \xi_{ik}||\frac{\partial}{\partial \xi_{ik}}\Psi(\xi_{ik}, \lambda_k)| + |\hat{\lambda}_k - \lambda_k||\frac{\partial}{\partial \lambda_k}\Psi(\xi_{ik}, \lambda_k)| + o_p(|\hat{\xi}_{ik} - \xi_{ik}| + (|\hat{\lambda}_k - \lambda_k|) \\
&= O_p[\lambda_k^{\gamma}\{\|\hat{X}_i - X_i\|_{L^2} + (\delta_k^{-1}\|X_i\|_{L^2} + |\xi_{ik}|)\|\hat{G} - G\|_S\}].
\end{aligned}$$

Abbreviate $\Sigma_{i=1}^{n}$ to Σ_i, $\Sigma_{k=1}^{s}$ to Σ_k and $O_p(\cdot)$ to '$\sim$'. Since $E\|\hat{X}_i - X_i\|_{L^2} \leqslant E(\|\hat{X}_i - X_i\|_{L^2}^2)^{1/2} = O(n^{-1/2})$, it is easy to see that $E(n^{-1}\Sigma_i\|\hat{X}_i - X_i\|_{L^2}) = E\|\hat{X}_i - X_i\|_{L^2} = O(n^{-1/2})$, To show result (15) for any fixed s, note that

$$n^{-1}\sum_i\left(\sum_{k=1}^{s}|\hat{\zeta}_{ik} - \zeta_{ik}|\right)^2 \leqslant sn^{-1}\sum_i\sum_{k=1}^{s}|\hat{\zeta}_{ik} - \zeta_{ik}|^2.$$

Then

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{s}(\hat{\zeta}_{ik} - \zeta_{ik})^2 &\sim \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{s}\lambda_k^{2\gamma}\{\|\hat{X}_i - X_i\|_{L^2} + (\delta_k^{-1}\|X_i\|_{L^2} + |\xi_{ik}|)\|\hat{G} - G\|_S\}^2 \\
&\sim \frac{1}{n}\sum_i\sum_k\lambda_k^{2\gamma}\|\hat{X}_i - X_i\|_{L^2}^2 + \frac{1}{n}\sum_i\sum_k\lambda_k^{2\gamma}\delta_k^{-2}\|X_i\|_{L^2}^2\|\hat{G} - G\|_S^2 \\
&+ \frac{1}{n}\sum_i\sum_k\lambda_k^{2\gamma}|\xi_{ik}|^2\|\hat{G} - G\|_S^2 + \frac{1}{n}\sum_i\sum_k\lambda_k^{2\gamma}\|\hat{X}_i - X_i\|_{L^2}\delta_k^{-1}\|X_i\|_{L^2}\|\hat{G} - G\|_S \\
&+ \frac{1}{n}\sum_i\sum_k\lambda_k^{2\gamma}\|\hat{X}_i - X_i\|_{L^2}|\xi_{ik}|\|\hat{G} - G\|_S + \frac{1}{n}\sum_i\sum_k\lambda_k^{2\gamma}\delta_k^{-1}|\xi_{ik}|\|X_i\|_{L^2}\|\hat{G} - G\|_S^2.
\end{aligned}$$

Denoting the additive terms in this formula E_1–E_6, we have

$$\begin{aligned}
E_1 &= \left(\sum_k\lambda_k^{2\gamma}\right)\left(n^{-1}\sum_i\|\hat{X}_i - X_i\|_{L^2}^2\right) = O_p(n^{-1}), \\
E_2 &= \|\hat{G} - G\|_S^2\left(\sum_k\lambda_k^{2\gamma}\delta_k^{-2}\right)\left(n^{-1}\sum_i\|X_i\|_{L^2}^2\right) = O_p(n^{-1}), \\
E_3 &= \|\hat{G} - G\|_S^2\left\{(1/n)\sum_i\sum_k\lambda_k^{2\gamma}|\xi_{ik}|^2\right\} = O_p(n^{-1}),
\end{aligned}$$

as

$$E\left(n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{s}\lambda_k^{2\gamma}|\xi_{ik}|^2\right) = \sum_k\lambda_k^{2\gamma+1} = O(1).$$

For E_4, applying the Cauchy–Schwarz ine ualit n m

$$\leqslant 2C\|\hat{G}-G\|_S\left(\sum_{k=1}^{s}\lambda_k^{2\gamma}\delta_k^{-1}\right)\sqrt{\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\|\hat{X}_i-X_i\|_{L^2}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|_{L^2}^2\right)\right\}}$$
$$= O_p(n^{-1/2})\,O(1)\,O_p(n^{-1/2})\,O_p(1)=O_p(n^{-1}).$$

Similarly, we have $E_5=O_p(n^{-1})$ and $E_6=O_p(n^{-1})$, using the facts that $E\{(\Sigma_{k=1}^{s}\lambda_k^{2\gamma}|\xi_{ik}|)^2\}\leqslant s\Sigma_{k=1}^{s}\lambda_k^{4\gamma+1}=O(1)$ and $E(\Sigma_{k=1}^{s}\lambda_k^{2\gamma}\delta_k^{-1}|\xi_{ik}|)^2\leqslant s\Sigma_{k=1}^{s}\lambda_k^{4\gamma+1}\delta_k^{-2}=O(1)$. This proves result (15).

We now turn to inequality (16). For any $f\in\mathcal{F}^s$, we have

$$f(\zeta_i)=\langle f(\cdot),R(\zeta_i,\cdot)\rangle_{\mathcal{F}^s}\leqslant\|f\|\langle R(\zeta_i,\cdot),R(\zeta_i,\cdot)\rangle_{\mathcal{F}^s}^{1/2}=\|f\|\,R^{1/2}(\zeta_i,\zeta_i),$$

where $R(\cdot,\cdot)$ is the reproducing kernel of space $\mathcal{F}^s$ and $\langle\cdot,\cdot\rangle_{\mathcal{F}^s}$ is the corresponding inner product. Therefore,

$$\frac{\partial f(\zeta_i)}{\partial\zeta_{ik}}=\left\langle f(\cdot),\frac{\partial R(\zeta_i,\cdot)}{\partial\zeta_{ik}}\right\rangle_{\mathcal{F}^s}\leqslant\|f\|\left\langle\frac{\partial R(\zeta_i,\cdot)}{\partial\zeta_{ik}},\frac{\partial R(\zeta_i,\cdot)}{\partial\zeta_{ik}}\right\rangle_{\mathcal{F}^s}^{1/2}.$$

Since $J(f)$ is a convex functional and a pseudonorm, we have

$$\sum_{k=1}^{s}\|P^kf\|^2\leqslant J^2(f)\leqslant s\sum_{k=1}^{s}\|P^kf\|^2. \tag{17}$$

We first claim that $\|f\|\leqslant J(f)$, because $\|f\|^2=b^2+\Sigma_{k=1}^{s}\|P^kf\|^2$. If $b=0$, inequality (17) implies that $\|f\|\leqslant J(f)$. If $b\neq0$, we can write $\tilde{J}(f)=b+J(f)=b+\Sigma_{k=1}^{s}\|P^kf\|$. For minimizing expression (5), it is equivalent to substitute $J(f)$ with $\tilde{J}(f)$, and inequality (17) implies that $\|f\|^2=b^2+\Sigma_{k=1}^{s}\|P^kf\|^2\leqslant b^2+J^2(f)\leqslant\tilde{J}^2(f)$. Therefore we have $\|f\|\leqslant J(f)$ in general. Secondly, owing to the orthogonality of $\{\bar{H}^k\}$, we can write $R(\mathbf{u},\mathbf{v})=R_1(u_1,v_1)+R_2(u_2,v_2)+\ldots+R_s(u_s,v_s)$ by theorem 5 in Berlinet and Thomas-agnan (2004), where $R_k(\cdot,\cdot)$ is the reproducing kernel of the subspace $\bar{H}^k$. For $\bar{H}^k$ being a second-order Sobolev Hilbert space, we have $R_k(s,t)=h_1(s)h_1(t)+h_2(s)h_2(t)-h_4(|s-t|)$, with $h_1(t)=t-\frac{1}{2}$, $h_2(t)=\{h_1^2(t)-1/12\}/2$ and $h_4(t)=\{h_1^4(t)-h_1^2(t)/2+7/240\}/24$. Therefore $R_k(s,t)$ is continuous and differentiable over $[0,1]^2$ and we can find constants a_k and b_k such that

$$\langle R_k(u,\cdot),R_k(u,\cdot)\rangle_{\mathcal{F}^s}<a_k,$$
$$\left\langle\frac{\partial R_k(u,\cdot)}{\partial u},\frac{\partial R_k(u,\cdot)}{\partial u}\right\rangle_{\mathcal{F}^s}\leqslant b_k,$$

for $k=1,\ldots,s$. One can find a uniform bound c with $\langle\partial R(\zeta_i,\cdot)/\partial\zeta_{ik},\partial R(\zeta_i,\cdot)/\partial\zeta_{ik}\rangle_{\mathcal{F}^s}\leqslant c$. However, an $\hat{f}$ minimizing expression (6) is equivalent to minimizing $n^{-1}\Sigma_i\{y_i-f(\zeta_i)\}^2$ under the constraint that $J(f)\leqslant\tilde{c}$ for some $\tilde{c}>0$. Therefore let $\rho=c^{1/2}\tilde{c}$; we have

$$\left|\frac{\partial\hat{f}(\zeta_i)}{\partial\zeta_{ik}}\right|\leqslant\|\hat{f}\|\left\langle\frac{\partial R(\zeta_i,\cdot)}{\partial\zeta_{ik}},\frac{\partial R(\zeta_i,\cdot)}{\partial\zeta_{ik}}\right\rangle_{\mathcal{F}^s}^{1/2}\leqslant J(\hat{f})c^{1/2}\leqslant\tilde{c}c^{1/2}=\rho. \qquad \square$$

Before stating lemma 3, we define the entropy of $\mathcal{F}^s$ with respect to the $\|\cdot\|_n$ metric. For each $\omega>0$, one can find a collection of functions $\{g_1,g_2,\ldots,g_N\}$ in $\mathcal{F}^s$ such that, for each $g\in\mathcal{F}^s$, there is a $j=j(g)\in\{1,2,\ldots,N\}$ satisfying $\|g-g_j\|_n\leqslant\omega$. Let $\mathbb{N}(\omega,\mathcal{F}^s,\|\cdot\|_n)$ be the smallest value of N for which such a cover of balls with radius ω and centres $g_1,g_2,\ldots,g_N$ exists. Then $H(\omega,\mathcal{F}^s,\|\cdot\|_n)=\log\{\mathbb{N}(\omega,\mathcal{F}^s,\|\cdot\|_n)\}$ is called the ω-entropy of $\mathcal{F}^s$.

Lemma 3. Assume that $\mathcal{F}^s=\{1\}\oplus\Sigma_{k=1}^{s}\bar{H}^k$, where $\bar{H}^k$ is the second-order Sobolev space. Denote the ω-entropy of $\{f\in\mathcal{F}^s:J(f)\leqslant1\}$ by $H(\omega,\{f\in\mathcal{F}^s:J(f)\leqslant1\},\|\cdot\|_n)$. Then

$$H(\omega,\{f\in\mathcal{F}^s:J(f)\leqslant1\},\|\cdot\|_n)\leqslant A\omega^{-1/2}, \tag{18}$$

for all $\omega>0$, $n\geqslant1$, and for some constants $A>0$. Furthermore, for $\{\varepsilon_i\}_{i=1}^n$ independent with finite variance and $J(f_0)>0$,

$$\sup_{f\in\mathcal{F}^s}\frac{|(\varepsilon,f-f_0)_n|}{\|f-f_0\|_n^{3/4}\{J(f)+J(f_0)\}^{1/4}}=O_p(n^{-1/2}). \tag{19}$$

Inequality (18) is implied by lemma A.1 of Lin and Zhang (2006). As the $\{\varepsilon_i\}$ satisfy the sub-Gaussian error assumption, the same argument as in Van de Geer (2000) (page 168) leads to result (19). We are now ready to present the proof of the main theorem.