FuSSO: Functional Shrinkage and Selection Operator

Junier B. Oliva[†] **Barnabás Póczos**[†] **Timothy Verstynen**[†] **Aarti Singh**[†]
Jeff Schneider[†] **Fang-Cheng Yeh**[†] **Wen-Yih Tseng**[*]
[†]Carnegie Mellon University [*]National Taiwan University

Abstract

We present the FuSSO, a functional analogue to the LASSO, that efficiently finds a sparse set of functional input covariates to regress a real-valued response against. The FuSSO does so in a semi-parametric fashion, making no parametric assumptions about the nature of input functional covariates and assuming a linear form to the mapping of functional covariates to the response. We provide a statistical backing for use of the FuSSO via proof of asymptotic sparsistency under various conditions. Furthermore, we observe good results on both synthetic and real-world data.

1 Introduction

Modern data collection has allowed us to collect not just more data, but more complex data. In particular, complex objects like sets, distributions, and functions are becoming prevalent in many domains. It would be beneficial to perform machine learning tasks using these complex objects. However, many existing techniques can not handle complex, possibly infinite dimensional, objects; hence one often resorts to the ad-hoc technique of representing these complex object by arbitrary summary statistics.

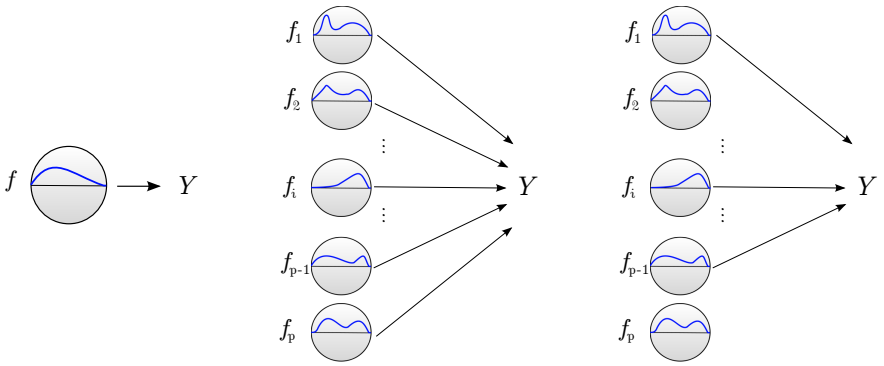
In this paper, we look to perform a regression task when dealing with functional data. Specifically, we look to regress a mapping that takes in many functional input covariates and outputs a real value. Moreover, since we are considering many functional covariates (possibly many more than the number of instances of one's data), we look to find an estimator that performs feature selection by only regressing on a subset of all possible input functional covariates. To this end we present the Functional Shrinkage and Selection Operator (FuSSO), for performing sparse functional regression in a principled, semi-parametric manner.

Indeed, there are a multitude of applications and domains where the study of a mapping that takes in a functional input and outputs a real-value is of interest. That is, if $\mathcal{I}$ is some class of input functions with domain $\Psi \subseteq \mathbb{R}$ and range $\mathbb{R}$, then one may be interested in a mapping $h : \mathcal{I} \mapsto \mathbb{R}$: $h(f) = Y$ (Figure 1(a)). Examples include: a mapping that takes in the time-series of a commodity's price in the past (f is a function with the domain of time and range of price) and outputs the expected price of the commodity in the nearby future; also, a mapping that takes a patient's cardiac monitor's time-series and outputs a health index. Recently, work by [8] has explored this type of regression problem when the input function is a distribution. Furthermore, the general case of an arbitrary functional input is related to functional analysis [2].

However, often it is expected that the response one is interested in regressing is dependent on not just one, but many functions. That is, it may be fruitful to consider a mapping $h : \mathcal{I}_1 \times \ldots \times \mathcal{I}_p \mapsto \mathbb{R}$: $h(f_1, \ldots, f_p) = Y$ (Figure 1(b)). For instance, this is likely the case in regressing the price of a commodity in the future, since the commodity's future price is not only dependent on the history of it own price, but also the history of other commodities' prices as well. A response's dependence on multiple functional covariates is especially common in neurological data, where thousands of voxels in the brain may each contain a corresponding function. In fact, in such domains it is not uncommon to have a number of input functional covariates that far exceeds the number of training instances one has in a data-set. Thus, it would be beneficial to have an estimator that is sparse in the number of functional covariates used to regress the response against. That is, find an estimate, $\hat{h}_s$, that depends on a small subset $\{i_1, \ldots, i_S\} \subset \{1, \ldots, p\}$, such that $\hat{h}(f_1, \ldots, f_p) = \hat{h}_s(f_{i_1}, \ldots, f_{i_S})$ (Figure 1(c)).

Here we present a semi-parametric estimator to perform sparse regression with multiple input functional

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.



(a) Single Functional Covariate (b) Multiple Functional Covariates (c) Sparse Model

Figure 1: (a) Model where mapping takes in a function f and produces a real Y. (b) Model where response Y is dependent on multiple input functions $f_1, \ldots, f_p$. (c) Sparse model where response Y is dependent on a sparse subset of input functions $f_1, \ldots, f_p$.

covariates and a real-valued response, the FuSSO: Functional Shrinkage and Selection Operator. No parametric assumptions are made on the nature of input functions. We shall assume that the response is the result of a sparse set of linear combinations of input functions and other non-paramteric functions $\{g_i\}$: $Y = \sum_j \langle f_j, g_j \rangle$. The resulting method is a LASSO-like [10] estimator that effectively zeros out entire functions from consideration in regressing the response.

Our contributions are as follows. We introduce the FuSSO, an estimator for performing regression with many functional covariates and a real-valued response. Furthermore, we provide a theoretical backing of the FuSSO estimator via proof of asymptotic sparsistency under certain conditions. We also illustrate the estimator with applications on synthetic data as well as in regressing the age of a subject when given orientation distribution function (dODF) [14] data for the subject's white matter.

2 Related Work

As previously mentioned, recently [8] explored regression with a mapping that takes in a probability density function and outputs a real value. Furthermore, [7] studies the case when both the input and outputs are distributions. In addition, functional analysis relates to the study for functional data [2]. In all these works, the mappings studied take in only one functional covariate. Based on them, it is not immediately evident how to expand on these ideas to develop an estimator that simultaneously performs regression and feature selection with multiple function covariates.

To the best of our knowledge, there has been no prior work in studying sparse mappings that take multiple functional inputs and produce a real-valued output. LASSO-like regression estimators that work with functional data include the following. In [6], one has a functional output and several real-valued covariates. Here, the estimator finds a sparse set of functions to scale by the real valued covariates to produce the functional response. Also, [17, 3] study the case when one has one functional covariate f and one real valued response that is linearly dependent on f and some function g: $Y = \langle f, g \rangle = \int fg$. In [17] the estimator searches for sparsity across wavelet basis projection coefficients. In [3], sparsity is in achieved in the time (input) domain of the d^{th} derivative of g; i.e. $[D^d g](t) = 0$ for many values of t where D^d is the differential operator. Hence, roughly speaking, [17, 3] look for sparsity across frequency and time domains respectively, for the regressing function g. However, these methods do not consider the case where one has many input functional covariates $\{f_1, \ldots, f_p\}$, and needs to choose among them. That is, [17, 3] do not provide a method to select among function covariates in an analogous fashion to how the LASSO selects among real-valued covariates.

Lastly, it is worth noting that in our estimator we will have an additive linear model, $\sum_j \langle f_j, g_j \rangle$ where we search for $\{g_i\}$ in a broad, non-parametric family such that many g_j are the zero function. Such a task is similar in nature to the SpAM estimator [9], in which one also has an additive model $\sum_j g_j(X_j)$ (in the dimensions of a real vector X) and searches for $\{g_i\}$ in a broad, non-parametric family such that many g_j are the zero function. Note though, that in the SpAM model, the $\{g_i\}$ functions are applied to real covariates via a function evaluation. In the FuSSO model, $\{g_i\}$ are applied to functional covariates via an inner product; that is, FuSSO works over functional, not real-valued covariates, unlike SpAM.

3 Model

To better understand FuSSO's model we draw several analogies to real-valued linear regression and Group-LASSO [16]. Note that although for simplicity we focus on functions working over a one dimensional domain, it is straightforward to extend the estimator and results to the multidimensional case. Consider a model for typical real-valued linear regression with a data-set of input-output pairs $\{(X_i, Y_i)\}_{i=1}^N$:

$$Y_i = \langle X_i, w\rangle + \epsilon_i,$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^d$, $w \in \mathbb{R}^d$, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, and $\langle X_i, w\rangle = \sum_{j=1}^d X_{ij} w_j$. If instead one were working with functional data $\{(f^{(i)}, Y_i)\}_{i=1}^N$, where $f^{(i)} : [0,1] \mapsto \mathbb{R}$ and $f^{(i)} \in L_2[0,1]$, one may similarly consider a linear model:

$$Y_i = \langle f^{(i)}, g\rangle + \epsilon_i,$$

where, $g : [0,1] \mapsto \mathbb{R}$, and $\langle f^{(i)}, g\rangle = \int_0^1 f^{(i)}(t) g(t) \mathrm{d}t$. If $\Phi = \{\varphi_m\}_{m=1}^\infty$ is an orthonormal basis for $L_2[0,1]$ [11] then we have that

$$f^{(i)}(x) = \sum_{m=1}^\infty \alpha_m^{(i)} \varphi_m(x), \tag{1}$$

where, $\alpha_m^{(i)} = \int_0^1 f^{(i)}(t)\varphi_m(t)\mathrm{d}t$. Similarly, $g(x) = \sum_{m=1}^\infty \beta_m \varphi_m(x)$ where $\beta_m = \int_0^1 g(t)\varphi_m(t)\mathrm{d}t$. Thus,

$$\begin{aligned} Y_i &= \langle f^{(i)}, g\rangle + \epsilon_i \\ &= \langle \sum_{m=1}^\infty \alpha_m^{(i)} \varphi_m(x), \sum_{k=1}^\infty \beta_k \varphi_k(x)\rangle + \epsilon_i \\ &= \sum_{m=1}^\infty \sum_{k=1}^\infty \alpha_m^{(i)} \beta_k \langle \varphi_m(x), \varphi_k(x)\rangle + \epsilon_i \\ &= \sum_{m=1}^\infty \alpha_m^{(i)} \beta_m + \epsilon_i, \end{aligned}$$

where the last step follows from orthonormality of Φ.

Going back to the real-valued covariate case, if instead of having one feature vector per data instance: $X_i \in \mathbb{R}^d$, one had p feature vectors associated to each data instance: $\{X_{ij} \mid 1 \le j \le p,\ X_{ij} \in \mathbb{R}^d\}$, an additive linear model may be used for regression:

$$Y_i = \sum_{j=1}^p \langle X_{ij}, w_j\rangle + \epsilon_i, \text{where } w_1, \ldots, w_p \in \mathbb{R}^d.$$

Similarly, in the functional case one may have p functions associated with data instance i: $\{f_j^{(i)} \mid 1 \le j \le p,\ f_j^{(i)} \in L_2[0,1]\}$. Then, an additive linear model would be:

$$Y_i = \sum_{j=1}^p \langle f_j^{(i)}, g_j\rangle + \epsilon_i = \sum_{j=1}^p \sum_{m=1}^\infty \alpha_{jm}^{(i)} \beta_{jm} + \epsilon_i, \tag{2}$$

where $g_1, \ldots, g_p \in L_2[0,1]$, and $\alpha_{jm}^{(i)}$ and β_{jm} are projection coefficients for $f_j^{(i)}$ and g_j respectively.

Suppose that one has few observations relative to the number of features ($N \ll p$). In the real-valued case, in order to effectively find a solution for $w = (w_1^T, \ldots, w_p^T)^T$ one may search for a group sparse solution where many $w_j = 0$. To do so, one may consider the following Group-LASSO regression:

$$w^\star = \underset{w}{\operatorname{argmin}} \frac{1}{2N} \|Y - \sum_{j=1}^p X_j w_j\|^2 + \lambda_{\mathrm{N}} \sum_{j=1}^p \|w_j\|, \tag{3}$$

where X_j is the $N \times d$ matrix $X_j = [X_{1j} \ldots X_{Nj}]^T$, $Y = (Y_1, \ldots, Y_N)^T$, and $\|\cdot\|$ is the Euclidean norm.

If in the functional case (2) one also has that $N \ll p$, one may set up a similar optimization to (3), whose direct analogue is:

$$g^\star = \underset{g}{\operatorname{argmin}} \frac{1}{2N} \sum_{i=1}^N \left(Y_i - \sum_{j=1}^p \langle f_j^{(i)}, g_j\rangle \right)^2 \tag{4}$$

$$+ \lambda_{\mathrm{N}} \sum_{j=1}^p \|g_j\|; \tag{5}$$

equivalently,

$$\beta^\star = \underset{\beta}{\operatorname{argmin}} \frac{1}{2N} \sum_{i=1}^N \left(Y_i - \sum_{j=1}^p \sum_{m=1}^\infty \alpha_{jm}^{(i)} \beta_{jm} \right)^2 \tag{6}$$

$$+ \lambda_{\mathrm{N}} \sum_{j=1}^p \sqrt{\sum_{m=1}^\infty \beta_{jm}^2}, \tag{7}$$

where $g = \{g_i\}_{i=1}^p = \{\sum_{m=1}^\infty \beta_{im}\varphi_m, \}_{i=1}^p$.

However, it is unfeasible to directly observe functional inputs $\{f_j^{(i)} \mid 1 \le i \le N, 1 \le j \le p\}$. Thus, we shall instead assume that one observes $\{\vec{y}_j^{(i)} \mid 1 \le i \le N, 1 \le j \le p\}$ where

$$\vec{y}_j^{(i)} = \vec{f}_j^{(i)} + \xi_j^{(i)}, \tag{8}$$

$$\vec{f}_j^{(i)} = \left(f_j^{(i)}(1/n),\ f_j^{(i)}(2/n),\ \ldots,\ f_j^{(i)}(1) \right)^T, \tag{9}$$

$$\xi_j^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\xi^2 I_n). \tag{10}$$

That is, we observe a grid of n noisy values for each functional input. Then, one may estimate $\alpha_{jm}^{(i)}$ as:

$$\tilde{\alpha}_{jm}^{(i)} = \frac{1}{n}\vec{\varphi}_m^T \vec{y}_j^{(i)} = \frac{1}{n}\vec{\varphi}_m^T(\vec{f}_j^{(i)} + \xi_j^{(i)}) = \bar{\alpha}_{jm}^{(i)} + \eta_{jm}^{(i)} \tag{11}$$

where $\vec{\varphi}_m = (\varphi_m(1/n),\ \varphi_m(2/n),\ \ldots,\ \varphi_m(1))^T$. Furthermore, we may truncate the number of basis functions used to express $f_j^{(i)}$ to M_n, estimating it as:

$$\tilde{f}_j^{(i)}(x) = \sum_{m=1}^{M_n} \tilde{\alpha}_{jm}^{(i)} \varphi_m(x). \tag{12}$$

Using the truncated estimate (12), one has:

$$\langle \tilde{f}_j^{(i)}(x), g_j \rangle = \sum_{m=1}^{M_n} \tilde{\alpha}_{jm}^{(i)} \beta_{jm}, \text{and}$$

$$\|\tilde{f}_j^{(i)}(x)\| = \sqrt{\sum_{m=1}^{M_n} (\tilde{\alpha}_{jm}^{(i)})^2}.$$

Hence, using the approximations (12), (7) becomes:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2N} \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{p} \sum_{m=1}^{M_n} \tilde{\alpha}_{jm}^{(i)} \beta_{jm} \right)^2 \tag{13}$$

$$+ \lambda_{\mathrm{N}} \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M_n} \beta_{jm}^2} \tag{14}$$

$$= \underset{\beta}{\operatorname{argmin}} \frac{1}{2N} \|Y - \sum_{j=1}^{p} \tilde{A}_j \beta_j\|^2 + \lambda_{\mathrm{N}} \sum_{j=1}^{p} \|\beta_j\|, \tag{15}$$

where $\tilde{A}_j$ is the $N \times M_n$ matrix with values $\tilde{A}_j(i,m) = \tilde{\alpha}_{jm}^{(i)}$ and $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_n})^T$. Note that one need not consider projection coefficients β_{jm} for $m > M_n$ since such projection coefficients will not decrease the MSE term in (13) (because $\tilde{\alpha}_{jm}^{(i)} = 0$ for $m > M_n$), and $\beta_{jm} \neq 0$ for $m > M_n$ increases the norm penalty term in (14). Hence we see that our sparse functional estimates are a Group-LASSO problem on the projection coefficients.

Extensions It is useful to note that there are several straightforward extensions to the FuSSO as presented. First, we would like to note that it may be possible to estimate the inner product of a function $f_j^{(i)}$, and g_j as $\int f_j^{(i)} g_j \approx \langle \vec{y}_j^{(i)}, \frac{1}{n}\vec{g}_j \rangle$, where $\vec{g}_j = (g_j(1/n), \ldots, g_j(1))^T$. This effectively allows one to use a naive approach of simply using Group-LASSO on the $\vec{y}_j^{(i)}$ feature vectors directly (we'll refer to this method as Y-GL). It is important to note, however, that Y-GL will be less robust to noise, and adaptive (and efficient) to smoothness than the FuSSO. Furthermore, we note that it is not necessary to have observations for input functions that are on a grid for the FuSSO, since one may estimate projection coefficients in the case of an irregular design [11]. Moreover, we may also estimate projection coeffincients for density functions with samples drawn from the pdf. Note that the Y-GL would fail to estimate our model in the irregular design case, and would not be possible in the case were functions are pdf. Also, a two-stage estimator as described in [5], where one first uses the regularization penalty with a large λ to find the support, then solves the optimization problem with a smaller λ on just the estimated support to estimate the response, may be more efficient at estimating the response. Furthermore, an analogous problem as (15) may be framed to perform logistic regression and classification.

4 Theory

Next, we show that the FuSSO is able to recover the correct sparsity pattern asymptotically; i.e., that the FuSSO estimate is sparsistent. In order to do so, we shall show that with high probability there is an optimal solution to our optimization problem (15) with the correct sparsity pattern. We follow a similar argument to [12, 9]. We shall use a "witness" technique to show that there is a coefficient/subgradient pair $(\hat{\beta}, \hat{u})$ such that $\operatorname{supp}(\hat{\beta}) = \operatorname{supp}(\beta^*)$, for true response generating β^*. Let $\Omega(\beta) = \sum_{j=1}^{p} \|\beta_j\|_2$, be our penalty term (14). Let S denote the true set of non-zero functions; i.e. $S = \{j \mid \beta_j^* \neq 0\}$, with $s = |S|$. First, we fix $\hat{\beta}_{S^c} = 0$, and set $\hat{u}_S = \partial\Omega(\cdot)(\beta^*)_S$. Note that for a vector β', $\partial\Omega(\cdot)(\beta') = \{u\}$ where: $u_j = \beta_j'/\|\beta'\|_2$, if $\beta_j' \neq 0$; $u_j = \|u_j\|_2 \leq 1$ if $\beta_j' = 0$. We shall show that with high probability, $\forall j \in S,\ \hat{\beta}_j \neq 0$ and $\forall j \in S^c,\ \|u_j\|_2 < 1$, thus showing that there is an optimal solution to our optimization problem (15) that has the true sparsity pattern with high probability.

First, we elaborate on our assumptions.

4.1 Assumptions

Let Φ be the trigonometric basis, $\varphi_1(x) \equiv 1$, $k \geq 2$:

$$\varphi_{2k}(x) \equiv \sqrt{2}\cos(2\pi k x),\ \varphi_{2k+1}(x) \equiv \sqrt{2}\sin(2\pi k x).$$

Let $\mathcal{D} = \{(\{\vec{y}_j^{(i)}\}_{j=1}^p, Y_i)\}_{i=1}^N$, where $\vec{y}_j^{(i)}$ is as (8), and $Y_i = \sum_{j=1}^p \sum_{m=1}^\infty \alpha_{jm}^{(i)} \beta_{jm}^* + \epsilon_i$ as in (2). Assume that $\forall\ 1 \leq i \leq N,\ 1 \leq j \leq p$: $\alpha_j^{(i)} \in \Theta(\gamma, Q)$, where:

$$\Theta(\gamma, Q) = \{\theta\ :\ \sum_{k=1}^{\infty} c_k^2 \theta_k^2 \leq Q\},$$

$$c_k = k^\gamma \text{ if } k \text{ even or one, } (k-1)^\gamma \text{ otherwise,}$$

$$\alpha_j^{(i)} = \{\alpha_{jm}^{(i)} \in \mathbb{R} \mid \alpha_{jm}^{(i)} = \int_0^1 f_j^{(i)} \varphi_m,\ m \in \mathbb{N}^+\}$$

for $0 < Q < \infty$ and $\frac{1}{2} < \gamma < \infty$. Furthermore, assume that that for the true β generating the observed responses Y_i, β^*, $\forall\ 1 \le j \le p$: $\beta_j^* \in \Theta(\gamma, Q)$.

Let A_j be the $N \times M_n$ matrix with entries $A_j(i,m) = \alpha_{jm}^{(i)}$. Let A_S denote the matrix made up from horizontally concatenating the A_j matrices with $j \in S$; i.e. $A_S = [A_{j_1} \dots A_{j_s}]$, where $\{j_1, \dots, j_s\} = S$ and $j_i < j_k$ for $i < k$. Suppose the following:

$$\Lambda_{\max}\left(\tfrac{1}{N} A_S^T A_S\right) \le C_{\max} < \infty \tag{16}$$

$$\Lambda_{\min}\left(\tfrac{1}{N} A_S^T A_S\right) \ge C_{\min} > 0. \tag{17}$$

Also, suppose $\exists \delta \in (0,1]$ s.t. $\forall j \in S^c$

$$\Lambda_{\max}\left(\tfrac{1}{N} A_j^T A_j\right) \le C_{\max} < \infty \tag{18}$$

$$\left\|(\tfrac{1}{N} A_j^T A_S)(\tfrac{1}{N} A_S^T A_S)^{-1}\right\|_2 \le 1 - \delta/\sqrt{s} \tag{19}$$

Let $\bar{A}_j$ be the $N \times M_n$ matrix with entries $\bar{A}_j(i,m) = \bar{\alpha}_{jm}^{(i)} = \frac{1}{n} \vec{\varphi}_m^T \vec{f}_j^{(i)}$. Let H_j be the $N \times M_n$ matrix with entries $H_j(i,m) = \eta_{jm}^{(i)} = \frac{1}{n} \vec{\varphi}_m^T \xi_j^{(i)}$. Thus, $\tilde{A}_j = \bar{A}_j + H_j$. Furthermore, let $E_j = \bar{A}_j - A_j$. Then, $\tilde{A}_j = A_j + E_j + H_j$.

In addition to the aforementioned assumptions, we shall further assume the following:

$$\exists a < 1/2 \quad \text{s.t.} \quad p M_n n^{a-1/2} e^{-n^{1-2a}} \to 0 \tag{20}$$

$$\rho_N^* \equiv \min_{j \in S} \|\beta_j^*\|_\infty > 0 \tag{21}$$

$$\sqrt{s M_n}\left(n^{-\gamma+1/2} + n^{-a}\right) \to 0 \tag{22}$$

$$\frac{1}{\rho_N^*}\left(s^{3/2} M_n^{1/2-2\gamma} + \sqrt{\log(sM_n)/N}\right) \to 0 \tag{23}$$

$$\lambda_{\text{N}} \sqrt{s M_n} / \rho_N^* \to 0 \tag{24}$$

$$\frac{1}{\lambda_{\text{N}}}\left(s\sqrt{M_n} n^{-\gamma+1/2} + \sqrt{\tfrac{s \log(N)}{n}}\right) \to 0 \tag{25}$$

$$\frac{1}{\lambda_{\text{N}}}\left(\frac{s M_n}{n^{\gamma+a-1/2}} + \frac{\sqrt{s M_n \log(N)}}{n^{a+1/2}}\right) \to 0 \tag{26}$$

$$\frac{1}{\lambda_{\text{N}}} \sqrt{M_n \log((p-s)M_n)/N} \to 0 \tag{27}$$

$$s/(\lambda_{\text{N}} N M_n^{2\gamma-1/2}) \to 0, \tag{28}$$

and we assume $\gamma \ge 1$ for the sake of simplification. We may further simplify our assumptions if we take $n = N^{1/2}$ and choose M_n optimally for function estimation: $M_n \asymp n^{1/(2\gamma+1)} = N^{1/(4\gamma+2)}$. Furthermore, take $s = O(1)$, $\rho_N^* \asymp 1$, and $\gamma = 2$. Under these conditions, our assumptions reduce to $\frac{1}{10} < a$ and taking the follow to go to zero:

$$p N^{\frac{10a-3}{20}} e^{-N^{\frac{1}{2}-a}}, \lambda_{\text{N}} N^{1/20}, N^{-\frac{7}{10}}/\lambda_{\text{N}},$$
$$\tfrac{1}{\lambda_{\text{N}}^2} N^{\frac{1}{2}} \log(N), \tfrac{1}{\lambda_{\text{N}}^2} N^{-9/10} \log(pN).$$

4.2 Sparsistency

Theorem 1: $\mathbb{P}\left(\hat{S}_N = S\right) \to 1$.

First, we state some lemmas, whose proofs may be found in the supplementary materials.

4.2.1 Lemmata

Lemma 1 Let X be a non-negative r.v. and $\mathcal{C}$ be an measurable event, then $\mathbb{E}[X|\mathcal{C}]\,\mathbb{P}(\mathcal{C}) \le \mathbb{E}[X]$.

Lemma 2 $\frac{1}{n}\sum_{k=1}^n \varphi_m(k/n)\varphi_l(k/n) = \mathbb{I}\{l = m\}$, for $1 \le l, m \le n-1$.

Lemma 3 Let $H_j^{(i)}$ be the rows of H_j, then $H_j^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \frac{\sigma_\xi^2}{n} I)$, and $H_S^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \frac{\sigma_\xi^2}{n} I)$.

Lemma 4 $\mathbb{P}\left(\|H\|_{\max} \ge n^a\right) \le 2\sigma_\xi p M_n n^{a-1/2} e^{-\frac{n^{1-2a}}{2\sigma_\xi^2}}$

Lemma 5 $\|E_j\|_{\max} \le C_Q n^{-\gamma+1/2}$, where $C_Q \in (0,\infty)$ is a constant depending on Q.

Lemma 6 $\|\beta_S^*\|_2^2 \le Qs$.

Lemma 7 $\exists N_0, n_0, \tilde{C}_{\min}, \tilde{C}_{\max}$, $0 < \tilde{C}_{\min} \le \tilde{C}_{\max} < \infty$, $0 < \tilde{\delta} \le 1$ s.t. if $\|H\|_{\max} < n^{-a}$, and $N > N_0$, $n > n_0$ then

$$\Lambda_{\max}\left(\tfrac{1}{N}\tilde{A}_S^T \tilde{A}_S\right) \le \tilde{C}_{\max} < \infty \tag{29}$$

$$\Lambda_{\min}\left(\tfrac{1}{N}\tilde{A}_S^T \tilde{A}_S\right) \ge \tilde{C}_{\min} > 0 \tag{30}$$

$$\forall j \in S^c,\ \left\|(\tfrac{1}{N}\tilde{A}_j^T \tilde{A}_S)(\tfrac{1}{N}\tilde{A}_S^T \tilde{A}_S)^{-1}\right\|_2 \le \frac{1-\tilde{\delta}}{\sqrt{s}} \tag{31}$$

4.2.2 Proof of Theorem 1

Proposition 1 $\mathbb{P}\left(\forall j \in S\ \hat{\beta}_j \ne 0\right) \to 1$.

Proof. Recall that by (21), $\rho_N^* = \min_{j \in S}\|\beta_j^*\|_\infty > 0$. Thus to prove that $\forall j \in S\ \hat{\beta}_j \ne 0$, it suffices to show that : $\|\hat{\beta}_S - \beta_S^*\|_\infty \le \frac{\rho_N^*}{2}$. To do so we show $\mathbb{P}\left(\|\hat{\beta}_S - \beta_S^*\|_\infty > \frac{\rho_N^*}{2}\right) \to 0$. Let $\mathcal{B}$ be the event that $\|H\|_{\max} < n^{-a}$. Note that:

$$\mathbb{P}\left(\|\hat{\beta}_S - \beta_S^*\|_\infty > \frac{\rho_N^*}{2}\right)$$
$$\le \mathbb{P}\left(\|\hat{\beta}_S - \beta_S^*\|_\infty > \frac{\rho_N^*}{2}\middle|\mathcal{B}\right)\mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{B}^c).$$

Furthermore,

$$\mathbb{P}\left(\|\hat{\beta}_S - \beta_S^*\|_\infty > \frac{\rho_N^*}{2}\middle|\mathcal{B}\right) \le \tfrac{2}{\rho_N^*}\mathbb{E}\left[\|\hat{\beta}_S - \beta_S^*\|_\infty\middle|\mathcal{B}\right].$$

Then, looking at the stationarity condition for the support S:

$$\tfrac{1}{N}\tilde{A}_S^T\left(\tilde{A}_S\hat{\beta}_S - Y\right) + \lambda_{\text{N}}\hat{u}_S = 0. \tag{32}$$

Let V be the $N \times 1$ vector with entries $V_i = \sum_{j\in S}\sum_{m=M_n+1}^{\infty} \alpha_{jm}^{(i)}\beta_{jm}^*$; i.e. the error from truncation. Then, using (32) $Y = A_S\beta_S^* + V + \epsilon \implies$

$$\tfrac{1}{N}\tilde{A}_S^T\left(\tilde{A}_S\hat{\beta}_S - A_S\beta_S^* - V - \epsilon\right) + \lambda_{\mathrm{N}}\hat{u}_S = 0 \implies$$
$$\tfrac{\tilde{A}_S^T}{N}\left(\tilde{A}_S(\hat{\beta}_S - \beta_S^*) - (A_S - \tilde{A}_S)\beta_S^* - V - \epsilon\right) = -\lambda_{\mathrm{N}}\hat{u}_S$$

Thus,

$$\tfrac{1}{N}\tilde{A}_S^T\tilde{A}_S(\hat{\beta}_S - \beta_S^*) = -\tfrac{1}{N}\tilde{A}_S^T(E_S + H_S)\beta_S^* + \tfrac{1}{N}\tilde{A}_S^T V + \tfrac{1}{N}\tilde{A}_S^T\epsilon - \lambda_{\mathrm{N}}\hat{u}_S. \quad (33)$$

Let $\tilde{\Sigma}_{SS}^{-1} = (\frac{1}{N}\tilde{A}_S^T\tilde{A}_S)^{-1}$; we see that,

$$\|\hat{\beta}_S - \beta_S^*\|_\infty \le \|\tilde{\Sigma}_{SS}^{-1}(\tfrac{1}{N}\tilde{A}_S^T)(E_S + H_S)\beta_S^*\|_\infty + \|\tilde{\Sigma}_{SS}^{-1}(\tfrac{1}{N}\tilde{A}_S^T)V\|_\infty + \|\tilde{\Sigma}_{SS}^{-1}(\tfrac{1}{N}\tilde{A}_S^T)\epsilon\|_\infty + \|\tilde{\Sigma}_{SS}^{-1}\lambda_{\mathrm{N}}\hat{u}_S\|_\infty.$$

Thus, we proceed to bound each term on the LHS in expectation. First, note that $\|\tilde{\Sigma}_{SS}^{-1}(\frac{1}{N}\tilde{A}_S^T)(E_S + H_S)\beta_S^*\|_\infty \le \|\tilde{\Sigma}_{SS}^{-1}\|_\infty\|(\frac{1}{N}\tilde{A}_S^T)(E_S + H_S)\beta_S^*\|_\infty = \|\tilde{\Sigma}_{SS}^{-1}\|_\infty\|\frac{1}{N}(A_S^T + E_S^T + H_S^T)(E_S + H_S)\beta_S^*\|_\infty \le \frac{\|\tilde{\Sigma}_{SS}^{-1}\|_\infty}{N}\Big(\|A_S^T(E_S + H_S)\beta_S^*\|_\infty + \|(E_S + H_S)^T(E_S + H_S)\beta_S^*\|_\infty\Big)$. Moreover, given that $\mathcal{B}$ occurs:

$$\|\tilde{\Sigma}_{SS}^{-1}\|_\infty \le \sqrt{sM_n}\|\tilde{\Sigma}_{SS}^{-1}\|_2 \le \frac{\sqrt{sM_n}}{\tilde{C}_{\min}}.$$

Thus, $\mathbb{E}\left[\frac{\|\tilde{\Sigma}_{SS}^{-1}\|_\infty}{N}\|A_S^T(E_S + H_S)\beta_S^*\|_\infty\Big|\mathcal{B}\right]\mathbb{P}(\mathcal{B})$

$$\le \frac{\sqrt{sM_n}}{\tilde{C}_{\min}N}\|A_S^T\|_\infty\mathbb{E}\left[\|(E_S + H_S)\beta_S^*\|_\infty\big|\mathcal{B}\right]\mathbb{P}(\mathcal{B})$$
$$\le \frac{Q\sqrt{sM_n}}{\tilde{C}_{\min}}\left(\|E_S\beta_S^*\|_\infty + \mathbb{E}\left[\|H_S\beta_S^*\|_\infty\big|\mathcal{B}\right]\mathbb{P}(\mathcal{B})\right),$$

noting that $\|A_S^T\|_\infty \le NQ$. Moreover, by Lemma 1:

$$\mathbb{E}\left[\|H_S\beta_S^*\|_\infty\big|\mathcal{B}\right]\mathbb{P}(\mathcal{B}) \le \mathbb{E}\left[\|H_S\beta_S^*\|_\infty\right].$$

Also, $H_S\beta_S^*$ is normally distributed and $\mathrm{Var}[H_S^{(i)T}\beta_S^*]$

$$= \sum_{j=1}^{sM_n}\mathrm{Var}[H_{Sj}^{(i)}\beta_{Sj}^*] = \frac{\sigma_\xi^2}{n}\|\beta_S^*\|_2^2 \le \frac{\sigma_\xi^2 Qs}{n}.$$

Hence, by a Gaussian inequality (e.g. [13]) we have:

$$\mathbb{E}\left[\|H_S\beta_S^*\|_\infty\right] \le \sqrt{2\sigma_\xi^2 Qs\log(N)/n}.$$

Unless otherwise specified, let $X^{(i)}$ be the i^{th} row of matrix X and X_j be the j^{th} column. Also,

$$\|E_S\beta_S^*\|_\infty = \max_{1\le i\le N}|E_S^{(i)T}\beta_S^*| \le \|\beta_S^*\|_2\max_{1\le i\le N}\|E_S^{(i)}\|_2$$
$$\le \sqrt{Qs}\left(C_Q\sqrt{sM_n}n^{-\gamma+1/2}\right)$$
$$= \sqrt{Q}C_Q s\sqrt{M_n}n^{-\gamma+1/2}$$

Thus, $\mathbb{E}\left[\frac{\|\Sigma_{SS}^{-1}\|_\infty}{N}\|A_S^T(E_S + H_S)\beta_S^*\|_\infty\big|\mathcal{B}\right]$

$$= O\left(\sqrt{sM_n}\left(s\sqrt{M_n}n^{-\gamma+1/2} + \sqrt{\frac{s\log(N)}{n}}\right)\right).$$

Furthermore, $\mathbb{E}\left[\|(E_S + H_S)^T(E_S + H_S)\beta_S^*\|_\infty\big|\mathcal{B}\right]\mathbb{P}(\mathcal{B})$

$$= \mathbb{E}\left[\max_{j\le sM_n}|(E_{Sj} + H_{Sj})^T((E_S + H_S)\beta_S^*)|\big|\mathcal{B}\right]\mathbb{P}(\mathcal{B})$$
$$\le \mathbb{E}\left[\max_{j\le sM_n}\|E_{Sj} + H_{Sj}\|_1\|(E_S + H_S)\beta_S^*\|_\infty\big|\mathcal{B}\right]\mathbb{P}(\mathcal{B})$$
$$= \mathbb{E}\left[\|E_S + H_S\|_1\|(E_S + H_S)\beta_S^*\|_\infty\big|\mathcal{B}\right]\mathbb{P}(\mathcal{B}).$$

Then, given that $\mathcal{B}$ occurs $\|E_S + H_S\|_1$

$$\le \|E_S\|_1 + \|H_{Sj}\|_1 \le N(C_Q n^{-\gamma+1/2} + n^{-a}),$$

and, as before: $\mathbb{E}\left[\|(E_S + H_S)\beta_S^*\|_\infty|\mathcal{B}\right]$

$$\le C_2 s\sqrt{M_n}n^{-\gamma+1/2} + C_3\sqrt{\frac{s\log(N)}{n}}.$$

Hence, $\mathbb{E}\left[\|\tilde{\Sigma}_{SS}^{-1}(\frac{1}{N}\tilde{A}_S^T)(E_S + H_S)\beta_S^*\|_\infty\Big|\mathcal{B}\right]\mathbb{P}(\mathcal{B})$

$$= O\Big(\sqrt{sM_n}\left(s\sqrt{M_n}n^{-\gamma+1/2} + \sqrt{s\log(N)/n}\right)\left(1 + n^{-\gamma+1/2} + n^{-a}\right)\Big)$$
$$= O\left(\sqrt{sM_n}\left(s\sqrt{M_n}n^{-\gamma+1/2} + \sqrt{s\log(N)/n}\right)\right)$$

The next terms are bounded as follows[1].

Lemma 8 $\mathbb{E}\left[\|\tilde{\Sigma}_{SS}^{-1}(\frac{1}{N}\tilde{A}_S^T)V\|_\infty\Big|\mathcal{B}\right]\mathbb{P}(\mathcal{B}) = O\left(\frac{s^{3/2}}{M_n^{2\gamma-1/2}}\right)$.

Lemma 9 $\mathbb{E}\left[\|\tilde{\Sigma}_{SS}^{-1}(\frac{1}{N}\tilde{A}_S^T)\epsilon\|_\infty\Big|\mathcal{B}\right]\mathbb{P}(\mathcal{B}) = O\left(\sqrt{\log(sM_n)/N}\right)$.

Lastly,

$$\|\hat{u}_S\|_\infty = \max_{j\in S}\|\hat{u}_j\|_\infty \le \max_{j\in S}\|\hat{u}_j\|_2 \le 1 \implies$$
$$\|\tilde{\Sigma}_{SS}^{-1}\lambda_{\mathrm{N}}\hat{u}_S\|_\infty \le \lambda_{\mathrm{N}}\|\tilde{\Sigma}_{SS}^{-1}\|_\infty\|\hat{u}_S\|_\infty \le \frac{\lambda_{\mathrm{N}}\sqrt{sM_n}}{\tilde{C}_{\min}}.$$

Keeping only leading terms, $\mathbb{E}\left[\|\hat{\beta}_S - \beta_S^*\|_\infty\Big|\mathcal{B}\right]\mathbb{P}(\mathcal{B})$

$$= O\left(s^{3/2}M_n n^{-\gamma+1/2} + s\sqrt{M_n\log(N)/n}\right)$$

[1] See Supplemental Materials for proof.

$$+ O\left(s^{3/2}/M_n^{2\gamma-1/2} + \sqrt{\log(sM_n)/N} + \lambda_{\mathrm{N}}\sqrt{sM_n}\right).$$

Hence, by assumptions (20)-(24) we have $\mathbb{P}\left(\|\hat{\beta}_S - \beta_S^*\|_\infty > \frac{\rho_N^*}{2}\right) \to 0$ □

One may similarly look at the stationarity for $j \in S^c$ to analyze $\hat{u}_j$: $0 = \frac{1}{N}\tilde{A}_j^T\left(\tilde{A}_S\beta_S - Y\right) + \lambda_{\mathrm{N}}\hat{u}_j$

$$= \frac{\tilde{A}_j^T}{N}\left(\tilde{A}_S(\hat{\beta}_S - \beta_S^*) - (A_S - \tilde{A}_S)\beta_S^* - V - \epsilon\right) + \lambda_{\mathrm{N}}\hat{u}_j.$$

Thus,

$$\begin{aligned}\hat{u}_j =& \tfrac{1}{\lambda_{\mathrm{N}}N}\tilde{A}_j^T\tilde{A}_S(\beta_S^* - \hat{\beta}_S) + \tfrac{1}{\lambda_{\mathrm{N}}N}\tilde{A}_j^T(A_S - \tilde{A}_S)\beta_S^* \\ &+ \tfrac{1}{\lambda_{\mathrm{N}}N}\tilde{A}_j^T(V + \epsilon) \\ =& \tfrac{1}{\lambda_{\mathrm{N}}}\tilde{\Sigma}_{jS}\tilde{\Sigma}_{SS}^{-1}\Big(\tfrac{1}{N}\tilde{A}_S^T(E_S + H_S)\beta_S^* - \tfrac{1}{N}\tilde{A}_S^T V \\ &- \tfrac{1}{N}\tilde{A}_S^T\epsilon + \lambda_{\mathrm{N}}\hat{u}_S\Big) - \tfrac{1}{\lambda_{\mathrm{N}}N}\tilde{A}_j^T(E_S + H_S)\beta_S^* \\ &+ \tfrac{1}{\lambda_{\mathrm{N}}N}\tilde{A}_j^T(V + \epsilon),\end{aligned}$$

where $\tilde{\Sigma}_{jS} = \frac{1}{N}\tilde{A}_j^T\tilde{A}_S$ and using (33). We wish to show that $\forall j \in S^c$ $\hat{u}_j$ satisfies the KKT conditions, that is:

Proposition 2 $\mathbb{P}\left(\max_{j\in S^c}\|\hat{u}_j\|_2 < 1\right) \to 1$.

Proof. Let $\mu_j^H \equiv \mathbb{E}\left[\hat{u}_j \middle| H\right]$. We proceed as follows:

$$\begin{aligned}&\mathbb{P}\left(\max_{j\in S^c}\|u_j\|_2 < 1\right) \\ &\geq \mathbb{P}\left(\max_{j\in S^c}\|\mu_j^H\|_2 + \|u_j - \mu_j^H\|_2 < 1\right) \\ &\geq \mathbb{P}\left(\max_{j\in S^c}\|\mu_j^H\|_2 + \sqrt{M_n}\|u_j - \mu_j^H\|_\infty < 1\right) \\ &\geq \mathbb{P}\left(\max_{j\in S^c}\|\mu_j^H\|_2 < 1 - \tfrac{\tilde{\delta}}{2}, \max_{j\in S^c}\|u_j - \mu_j^H\|_\infty < \tfrac{\tilde{\delta}}{2\sqrt{M_n}}\right) \\ &\geq 1 - \mathbb{P}\left(\max_{j\in S^c}\|\mu_j^H\|_2 \geq 1 - \tilde{\delta}\right) \\ &- \mathbb{P}\left(\max_{j\in S^c}\|u_j - \mu_j^H\|_\infty \geq \frac{\tilde{\delta}}{2\sqrt{M_n}}\right).\end{aligned}$$

We obtain the following results:

Lemma 10 $\mathbb{P}\left(\max_{j\in S^c}\|\mu_j^H\|_2 \geq 1 - \frac{\tilde{\delta}}{2}\right) \to 0$

Lemma 11 $\mathbb{P}\left(\max_{j\in S^c}\|\hat{u}_j - \mu_j^H\|_\infty \geq \frac{\tilde{\delta}}{2\sqrt{M_n}}\right) \to 0$

Hence, we have that $\mathbb{P}\left(\max_{j\in S^c}\|\hat{u}_j\|_2 < 1\right) \to 1$. □

5 Experiments

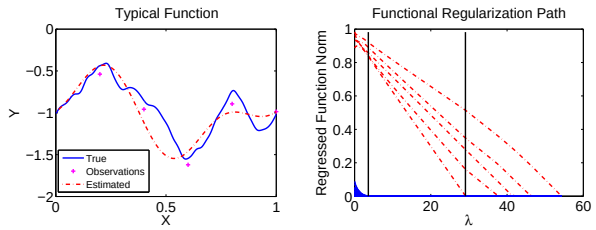
5.1 Synthetic Data

We tested the FuSSO on synthetic data-sets of $\mathcal{D} = \{(\{\vec{y}_j^{(i)}\}_{j=1}^p, Y_i)\}_{i=1}^N$ (where $\vec{y}_j^{(i)}$ as in (8)). The experiments performed were as follows. First, we fix N, n, p, and s. For $i = 1, \ldots, N$, $j = 1, \ldots, p$ we create random functions using a maximum of M projection coefficients as follows: 1) Set $a_{jm} \overset{iid}{\sim}$ Unif$[-1, 1]$ for $m = 1, \ldots, M$; 2) set $a_{jm} = a_{ji}/c_m^2$, where $c_m = m$ if $m = 1$ or is even, $c_m = m - 1$ if m is odd; 3) set $a_{jm} = a_{jm}/\|a_j\|$; 4) set $\alpha_j^{(i)} = a_j$. (See Figures 2(b),2(c) for typical functions.) Similarly, we generate β_j^* for $j = 1, \ldots, s$; for $j = s+1, \ldots, p$, we set $\beta_j^* = 0$. Then, we generate Y_i as $Y_i = \sum_{j=1}^p \langle\beta_j^*, \alpha_j^{(i)}\rangle + \epsilon_i = \sum_{j=1}^s \langle\beta_j^*, \alpha_j^{(i)}\rangle + \epsilon_i$, where $\epsilon \overset{iid}{\sim} \mathcal{N}(0, .1)$. Also, a grid of n noisy function evaluations were generated to make $\vec{y}_j^{(i)}$ as in (8), with $\sigma_\xi = .1$. These were then used to compute $\tilde{\alpha}_{jm}^{(i)}$ for $m = 1, \ldots, M_n$ as in (11), M_n was chosen by cross validation. (See Figures 2(b), 2(c) for typical noisy observations and function estimates for $n = 5$ and $n = 25$ respectively.)

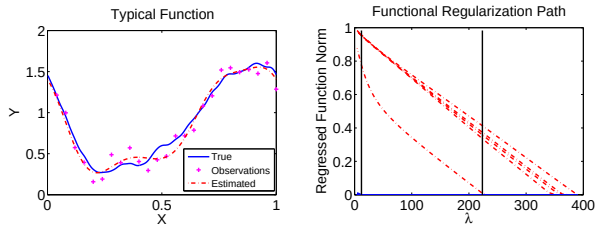
We fixed $s = 5$ and chose the following configurations for the other parameters: $(p, N, n) \in \{(100, 50, 5), (1000, 500, 25), (20000, 500, 25)\}$. For each tuple of (p, N, n) configurations, 100 random trails were performed. We recorded, r, the fraction of the trails that a λ value was able recover the correct sparsity pattern (i.e. that only the first 5 functions are in the support). We also recorded the mean length of the range of λ, Δ_λ, that were able to recover the correct support; i.e. $\Delta_\lambda = \frac{1}{t}\sum_{t=1}^{100}\Delta_\lambda^{(t)}$, where $\Delta_\lambda^{(t)} = (\lambda_f^{(t)} - \lambda_l^{(t)})/\lambda_{\max}^{(t)}$, $\lambda_f^{(t)}$ is the largest λ value found to recover the correct support in the t^{th} trails, $\lambda_l^{(t)}$ the smallest such λ, and $\lambda_{\max}^{(t)}$ is the smallest λ to produce $\hat{\beta} = 0$ ($\Delta_\lambda^{(t)}$ is taken to be zero if no λ recovered the correct support). The results were as follows:

(p, N, n)	r	Δ
(100,50,5)	.68	.2125
(1000,500,25)	1	.4771
(20000,500,25)	1	.4729

Hence we see that even when the number of observations per function is small (5 or 25) and the number of total number of input functional covariates is large (we were able to test up to 20000), the FuSSO can recover the correct support. Also, to illustrate this point that running Group-LASSO on the $\vec{y}_j^{(i)}$ features (Y-GL) is less robust to noise and adaptive to smoothness, we ran noisier trails using the configuration of $(p, N, n) = (1000, 500, 25)$. We increased the standard deviation of the noise on grid function observation and on the response to be 5 and 1 respectively. Under these conditions the FuSSO was able to recover the support in 49% of the trails were as Y-GL recovered the support in 32% of the trails. Furthermore the FuSSO had



(a) Function at $n = 5$ (b) Reg. Path at $p = 100$, $n = 50$, $n = 5$ of $\|\hat{\beta}_j\|$



(c) Function at $n = 25$ (d) Reg. Path at $p = 1000$, $n = 500$, $n = 25$ of $\|\hat{\beta}_j\|$

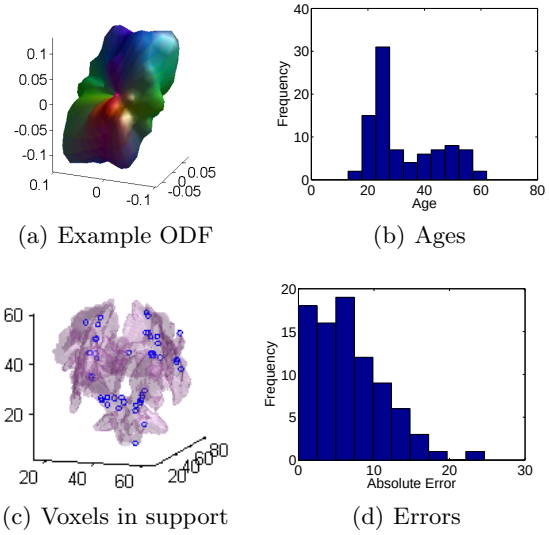
Figure 2: (a)(c) Two typical functions, noisy observations, and estimates. (b)(d) Regularization paths showing the norms of $\hat{\beta}_j$ (in red for j in support, blue otherwise) for a range of λ; rightmost vertical line indicates largest λ able to recover the support, leftmost line for smallest such λ

a $\Delta_\lambda^{(t)} = .0743$ compared to $\Delta_\lambda^{(t)} = .0254$ for Y-GL.

5.2 Neurological Data

We also tested the FuSSO estimator with a neurological data-set, using a total of 89 subjects [14]. Subjects ranged in age from 18 to 60 years old (Figure 3(b)). Our goal was to learn a regression that maps the dODFs at each white matter voxel for each subject to the subject's age. The dODF is a function represents the amount of water molecules, or spins, undergoing diffusion in different orientations over the S^2 sphere[15]. I.e., each dODF is a function with a $2d$ domain (of azimuth, elevation spherical coordinates) and a range of reals representing the strength of water diffusion at the given orientations (see Figure 3(a)). Data was provided for each subject in a template space for white-matter voxels; a total of over 25 thousand voxels' dODFs were regressed on (i.e. $p \approx 25000$). We also compared regression using the FuSSO and functional covariates to using the LASSO and real valued covariates. We used the non-functional collection of quantitative anisotropy (QA) values for the same white matter voxels as with dODF functions. QA values are the estimated amount of spins that undergo diffusion in the direction of the principle fiber orientation, i.e., the peak of the dODF; QAs have been used as a measure of white matter integrity in the underlying voxel hence making for a descriptive and effective summary statistic of an dODF function for age regression [15].

The projection coefficients for the dODFs at each voxel were estimated using the cosine basis. The FuSSO estimator gave a cross-validated MSE of 70.855, where the variance for age was 156.4265; selected voxels in the support may be seen in Figure 3(c). The LASSO estimate using QA values gave a cross-validated MSE of 77.1302. Thus, one may see that considering the entire functional data gave us better results for age regression. We note that we were unable to use the naive approach of Y-GL in this case because of memory constraints and the fact that function evaluation points did not lie on a $2d$ square grid.



(a) Example ODF (b) Ages

(c) Voxels in support (d) Errors

Figure 3: (a) An example ODF for a voxel. (b) Histogram of ages for subjects. (c) Voxels in the support of model shown in blue. (d) Histogram of held out error magnitudes.

6 Conclusion

In conclusion, this paper presents the FuSSO, a functional analogue to the LASSO. The FuSSO allows one to efficiently find a sparse set of functional input covariates to regress a real-valued response against. The FuSSO makes no parametric assumptions about the nature of input functional covariates and assumes a linear form to the mapping of functional covariates to the response. We provide a statistical backing for use of the FuSSO via proof of asymptotic sparsistency.

Acknowledgements

This work is supported in part by NSF grants IIS1247658 and IIS1250350.

References

[1] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer, 1997.

[2] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer, 2006.

[3] Gareth M James, Jing Wang, and Ji Zhu. Functional linear regression that's interpretable. *The Annals of Statistics*, pages 2083–2108, 2009.

[4] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

[5] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[6] Nicola Mingotti, Rosa E Lillo, and Juan Romo. Lasso variable selection in functional regression. 2013.

[7] Junier B Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression. In *International Conference on Machine Learning (ICML)*, page 10491057. ICML, 2013.

[8] B. Poczos, A. Rinaldo, A. Singh, and L Wasserman. Distribution-Free Distribution Regression. *AISTATS*, 2013.

[9] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[10] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[11] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008.

[12] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. *arXiv preprint math/0605740*, 2006.

[13] Larry Wasserman. Probability inequalities. `http://www.stat.cmu.edu/~larry/=stat705/Lecture2.pdf`, August 2012.

[14] Fang-Cheng Yeh and Wen-Yih Isaac Tseng. Ntu-90: a high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction. *NeuroImage*, 58(1):91–99, 2011.

[15] Fang-Cheng Yeh, Van Jay Wedeen, WY Tseng, et al. Generalized q-sampling imaging. *IEEE transactions on medical imaging*, 29(9):1626, 2010.

[16] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[17] Yihong Zhao, R Todd Ogden, and Philip T Reiss. Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617, 2012.