Sparse Functional Regression

Junier B. Oliva[*], **Barnabás Póczos**[*], **Aarti Singh**[*], **Jeff Schneider**[†], **Timothy Verstynen**[‡]
[*]Machine Learning Department
[†]Robotics Institute
[‡]Psychology Department
Carnegie Mellon University
Pittsburgh, PA 15213

1 Introduction

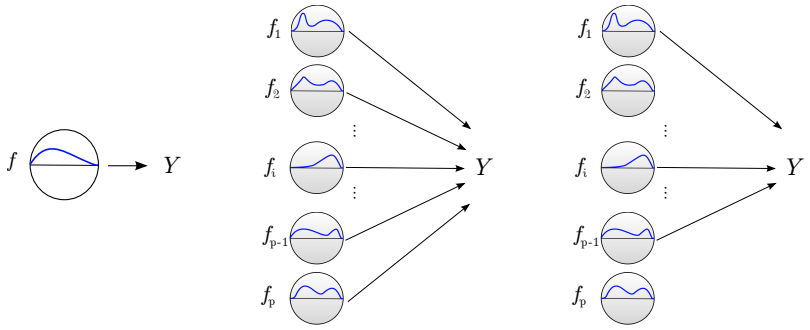
There are a multitude of applications and domains where the study of a mapping that takes in a functional input and outputs a real-value is of interest. That is, if $\mathcal{I}$ is some class of input functions with domain $\Psi \subseteq \mathbb{R}$ and range $\mathbb{R}$, then one may be interested 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 [5] 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 [1].

However, it is often 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, $\hat{h}$, 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 covariates and a real-valued response, 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 $\{g_i\}$: $Y = \sum_j \langle f_j, g_j \rangle + \epsilon$. The resulting method is a LASSO-like [7] estimator that effectively zeros out entire functions from consideration in regressing the response. The estimator was found to be effective in regressing the age of a subject when given orientation distribution function (ODF) data for the subject's white matter.

2 Related Work

As previously mentioned, recently [5] explored regression with a mapping that takes in a probability density function and outputs a real value. Furthermore, [4] studies the case when both the input and outputs are distributions. In addition, functional analysis relates to the study of functional data [1]. In all of these works, the mappings studied take in only one functional covariate. However, 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.



(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 $f_1, \ldots, f_p$.

To 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 [3], 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 a functional response. Also, [10, 2] 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: $Y = \langle f, g \rangle = \int fg$. First, in [10] the estimator searches for sparsity across wavelet basis projection coefficients. In [2], sparsity is 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, [10, 2] look for sparsity across frequency and time domains respectively, for the regessing 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 amongst them. That is, [10, 2] 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 [6], 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

In order to better understand FuSSO's model we draw several analogies to real-valued linear regression and Group-LASSO [9]. First, consider a model for typical real-valued linear regression with a data-set of input-output pairs $\{(X_i, Y_i)\}_{i=1}^N$:

$$Y_i = \langle X_i, w \rangle + \epsilon_i, \text{where } Y_i \in \mathbb{R},\ X_i \in \mathbb{R}^d, w \in \mathbb{R}^d,\ \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \text{ and } \langle X_i, w \rangle = \sum_{j=1}^{d} X_{ij} w_j.$$

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 might similarly consider a linear model:

$$Y_i = \langle f^{(i)}, g \rangle + \epsilon_i, \text{where } g : [0,1] \mapsto \mathbb{R}, \text{and } \langle f^{(i)}, g \rangle = \int_0^1 f^{(i)}(t) g(t) \mathrm{d}t.$$

If $\Phi = \{\varphi_m\}_{m=1}^{\infty}$ is an orthonormal basis for $L_2[0,1]$ [8] then we have that

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

Similarly, $g(x) = \sum_{m=1}^{\infty} \beta_m^* \varphi_m(x)$. Thus,

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

$$= \sum_{m=1}^{\infty} \alpha_m^{(i)}\beta_m^* + \epsilon_i,$$

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 with each data instance: $\{X_{ij} \mid 1 \le j \le p,\ X_{ij} \in \mathbb{R}^d\}$, an additive linear model could be used for regression:

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

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.

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:

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

equivalently,

$$\beta^\star = \underset{\beta}{\operatorname{argmin}} \frac{1}{2N} \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{p}\sum_{m=1}^{\infty} \alpha_{jm}^{(i)}\beta_{jm} \right)^2 + \lambda_{\mathrm{N}} \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{\infty} \beta_{jm}^2}, \tag{5}$$

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

However, it is intractable to assume that one is able 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

$$\vec{y}_j^{(i)} = \vec{f}_j^{(i)} + \xi_j^{(i)}, \quad \vec{f}_j^{(i)} = \left(f_j^{(i)}(1/n),\ f_j^{(i)}(2/n),\ \ldots,\ f_j^{(i)}(1) \right)^T, \quad \xi_j^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\xi^2 I). \tag{6}$$

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

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

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:

$$\tilde{f}_j^{(i)}(x) = \sum_{m=1}^{M_n} \tilde{\alpha}_{jm}^{(i)} \varphi_m(x). \tag{7}$$

Using the truncated estimate (7), 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 (7), (5) becomes:

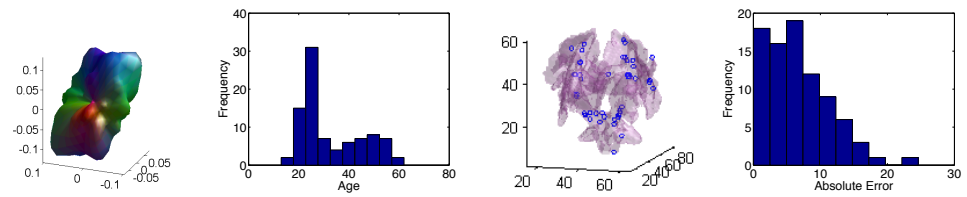
$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{2N} \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{p} \sum_{m=1}^{M_n} \tilde{\alpha}_{jm}^{(i)} \beta_{jm} \right)^2 + \lambda_{\text{N}} \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M_n} \beta_{jm}^2} \tag{8}$$

$$= \underset{\beta}{\text{argmin}} \frac{1}{2N} \|Y - \sum_{j=1}^{p} \tilde{A}_j \beta_j\|^2 + \lambda_{\text{N}} \sum_{j=1}^{p} \|\beta_j\|, \tag{9}$$

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 (8) (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 (8). Hence, we see that our sparse functional estimates are a Group-LASSO problem on the projection coefficients. In a future publication, we shall show that if $\{f_j^{(i)}\}$, and $\{g_j\}$ are in a Sobolev function class and some other mild assumptions hold, then our estimator is asymptotically sparsistent.

4 Experiments

We tested the FuSSO estimator with neurological data. It consisted of 89 total subjects. Orientation distribution function (ODF) (Figure 2(a)) data was provided for each subject in a template space for white-matter voxels; a total of over 25 thousand voxel's ODFs were regressed on. We looked to regress a subject's age given his/her respective ODF data. The projection coefficients for the ODFs at each voxel were estimated using the cosine basis. The FuSSO estimator gave a held out MSE of 70.855, where the variance for age was 156.4265.



(a) Example ODF (b) Ages (c) Active Voxels (d) Errors

Figure 2: (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.

References

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

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

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

[4] Junier B Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression.

[5] B. Poczos, A. Rinaldo, A. Singh, and L Wasserman. Distribution-Free Distribution Regression. *arXiv preprint arXiv:1302.0082*, 2012.

[6] 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.

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

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

[9] 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.

[10] 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.