FAStEN: Efficient feature selection and coefficient estimation in functional regression models

Tobia Boschi, **Lorenzo Testa**, Francesca Chiaromonte, Matthew Reimherr

Penn State University (US), Sant'Anna School of Advanced Studies (Italy)

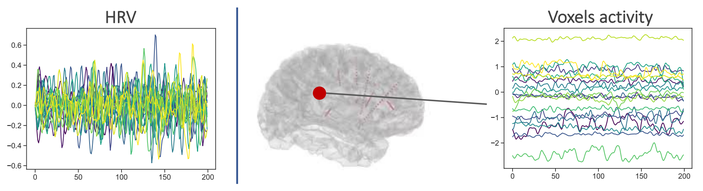
Summary

We propose a new, flexible and ultra-efficient approach to perform **feature selection** and **parameter estimation** simultaneously in a sparse high dimensional **function-on-function regression** framework. We exploit the properties of **Functional Principal Components** and the sparsity inherent to the **Dual Augmented Lagrangian** problem to significantly reduce computational cost, and we introduce an **adaptive scheme** to improve selection accuracy. Through an extensive simulation study, we benchmark our approach to the best existing competitor and demonstrate a massive gain in terms of CPU time and selection performance, without sacrificing the quality of the coefficients' estimation. Finally, we present an application to brain fMRI data from the AOMIC PIOP1 study.

1. Motivating example

AOMIC PIOP1 study: identify, during an anticipation task, which parts of the brain are associated with heart-rate variability (HRV) in 216 university students:

- **Response**: HRV (a convolution of the heart rate trace)
- **Features**: activity of 55551 voxels



Goal: develop a new *efficient* algorithm to solve *feature selection* for a *fully functional* regression setting.
Challenge: current methods cannot deal with such a large number of functional features p in a computationally viable way.

2. Function-on-function regression model

$$\mathcal{Y}_i(t) = \sum_{j=1}^{p} \int_{\mathcal{S}} \mathcal{B}_j(t,s)\mathcal{X}_{ij}(s)ds + \epsilon_i(t) \qquad i = 1, \dots, n$$

where:

- $\mathcal{Y}_i$ are centered functional responses (with 0 mean function) defined over a domain $\mathcal{T}$
- $\mathcal{X}_{ij}$ are p standardized functional features (with 0 mean function and standard deviation equal to 1), defined over domain $\mathcal{S}$
- $\mathcal{B}_j$ are **coefficient surfaces** to be estimated
- ϵ_i are i.i.d. errors, independent of the $\mathcal{X}$'s, with 0 mean function and a common variance operator.

5. Computational efficiency

The challenge is to find the descent direction $D \in \mathbb{R}^{n\times k}$, by solving the linear system $H_\psi(V)\,\mathrm{vec}(D) = -\,\mathrm{vec}\left(\nabla\psi(V)\right)$ efficiently, where:

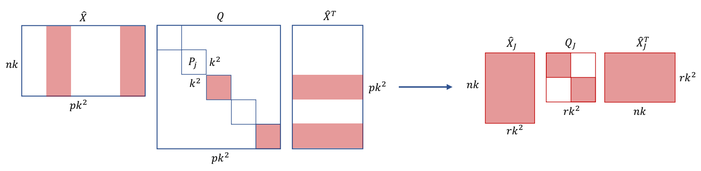
- $\psi(V) = \mathcal{L}_\sigma(V \mid \bar{Z}, B)$
- $H_\psi \in \mathbb{R}^{nk\times nk}$ is the Hessian matrix
- $\nabla\psi \in \mathbb{R}^{n\times k}$ is the gradient matrix
- $\mathrm{vec}\left(\nabla\psi\right) \in \mathbb{R}^{nk}$ is the gradient vector obtained stacking all the columns of $\nabla\psi$

Theorem

Let:

- $\hat{X}_\mathcal{J} = X_\mathcal{J} \otimes_{Kron} I_k$ be the $nk \times rk^2$ Kronecker product between $X_\mathcal{J}$ and the $k \times k$ identity matrix
- $Q_\mathcal{J} \in \mathbb{R}^{rk^2\times rk^2}$ be the block-diagonal matrix formed by the blocks $P_{[j]}$
- $P_{[j]} = \begin{cases} \Lambda & \text{if } \|B_{[j]} - \sigma V^T X_{[j]}\|_2 \geq \sigma\omega_j\lambda_1 \\ 0 & \text{otherwise} \end{cases}$
- $\mathcal{J} = \left\{j \,:\, \|B_{[j]} - \sigma V^T X_{[j]}\|_2 \geq \sigma\omega_j\lambda_1\right\}$ be the active set, with $r = |\mathcal{J}|$.

then $H_\psi(V) = I_{nk} + \sigma\hat{X}_\mathcal{J} Q_\mathcal{J} \hat{X}_\mathcal{J}^T$.



In sparse problems ($r \ll n$), by *Sherman-Morrison-Woodbury* formula, the original total cost (matrix multiplication and *Cholesky* factorization) $\mathcal{O}\left(nk^4(n^2+p^2+np)\right)$ is reduced to $\mathcal{O}\left(rk^4(k^2+r^2+n^2+rnk)\right)$. Hence, **the cost does not depend on** p.

3. FAStEN objective function

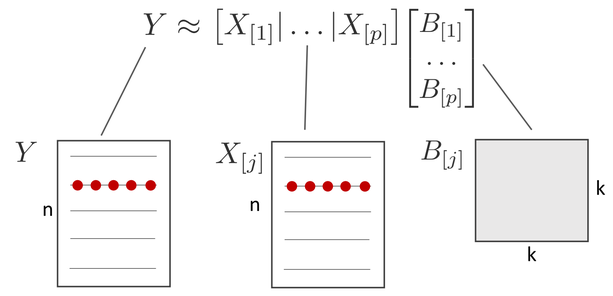
$$\min_{\mathcal{B}_1,\dots,\mathcal{B}_p} \left[\frac{1}{2}\left\| \mathcal{Y} - \sum_{j=1}^{p}\int_{\mathcal{S}} \mathcal{B}_j(t,s)\mathcal{X}_j(s)ds \right\|^2_{\mathbb{L}^2} + \sum_{j=1}^{p}\omega_j\left(\lambda_1\|\mathcal{B}_j\|_{\mathbb{L}^2\times\mathbb{L}^2} + \frac{\lambda_2}{2}\|\mathcal{B}_j\|^2_{\mathbb{L}^2\times\mathbb{L}^2}\right)\right]$$

Matrix representation through Functional Principiapal Components

Idea: use $e_1(t), \dots, e_k(t)$, the first k functional principal components of $\mathcal{Y}$, to represent both response and features. Then:

- $Y \in \mathbb{R}^{n\times k}$, where $Y_{ij} = \langle \mathcal{Y}_i, e_j\rangle_{\mathbb{L}^2}$
- $X \in \mathbb{R}^{n\times pk}$, where for each $X_{[l]} \in \mathbb{R}^{n\times k}$, $l = 1, \dots, p$, it holds $X_{[l](ij)} = \langle \mathcal{X}_{li}, e_j\rangle_{\mathbb{L}^2}$
- $B \in \mathbb{R}^{pk\times k}$, where for each $B_{[l]} \in \mathbb{R}^{k\times k}$, $l = 1, \dots, p$, it holds $B_{[l](ij)} = \langle \mathcal{B}_l, e_i \otimes e_j\rangle_{\mathbb{L}^2\times\mathbb{L}^2} = \int\int \mathcal{B}_l(s,t)e_i(s)e_j(t)\,ds\,dt$.

$$Y \approx \left[X_{[1]}|\dots|X_{[p]}\right]\begin{bmatrix} B_{[1]} \\ \dots \\ B_{[p]} \end{bmatrix}$$

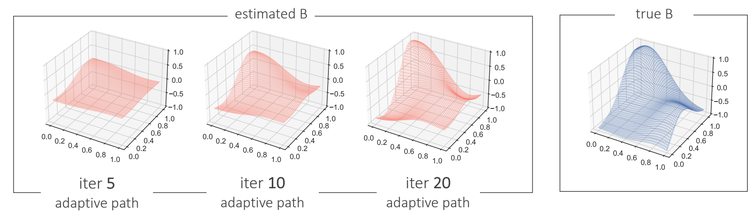


Thus the minimization problem becomes

$$\min_B \frac{1}{2}\|Y - XB\|_2^2 + \sum_{j=1}^{p}\omega_j\left(\lambda_1\|B_{[j]}\|_2 + \frac{\lambda_2}{2}\|B_{[j]}\|_2^2\right) = \min_B h(XB) + \pi(B)$$

where:

- $\|\cdot\|_2$ is the l_2 Frobenius norm
- λ_1 and λ_2 control an Elastic Net-type penalty (creating sparsity)
- ω_j are **adaptive** weights (improving both selection and estimation)
- h and π define the **primal problem**



4. DAL algorithm

A possible **dual formulation** is

$$\min_{V,Z} h^*(V) + \pi^*(Z) \quad s.t. \quad X^TV + Z = 0$$

where $V \in \mathbb{R}^{n\times k}$ and $Z \in \mathbb{R}^{pk\times k}$ are the matrix dual variables and h^* and π^* are the Fenchel-conjugate functions.

Dual Augmented Lagrangian method

Goal: starting at initial values V^0, Z^0, B^0, σ^0, minimize

$$\mathcal{L}_\sigma(V,Z,B) = h^*(V) + \pi^*(Z) - \sum_{j=1}^{p}\langle B_{[j]}, V^TX_{[j]} + Z_{[j]}\rangle + \frac{\sigma}{2}\sum_{j=1}^{p}\|V^TX_{[j]} + Z_{[j]}\|_2^2$$

While not converged:

- Given B^s, find $\left(V^{s+1}, Z^{s+1}\right) \approx \arg\min_{V,Z}\mathcal{L}_\sigma\left(V, Z \mid B^s\right)$

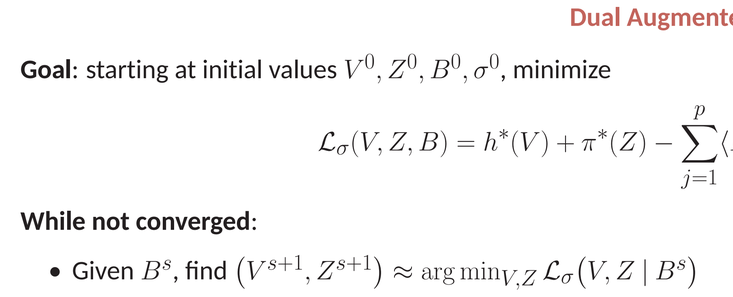
  Inner sub-problem: to find (V^{s+1}, Z^{s+1}) update V, Z *independently*:
  While not converged:
  $V^{m+1} = \arg\min_V \mathcal{L}_\sigma(V \mid Z^m, B^s) \longrightarrow$ **Newton method** (Theorem)
  $Z^{m+1} = \arg\min_Z \mathcal{L}_\sigma(Z, \mid V^{m+1}, B^s) \longrightarrow$ **closed form** (Proposition)

- Update the Lagrangian multiplier B and the parameter σ:

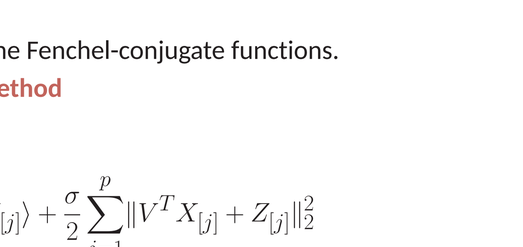
$$B^{s+1} = B^s - \sigma_k\left(X^TV^{s+1} + Z^{s+1}\right)$$
$$\sigma^{s+1} \uparrow \sigma^\infty \leq \infty$$

6. Simulations

FAStEN outperforms the only competitor FRegSimCom



7. AOMIC PIOP 1 study



Orbitofrontal cortex (reward and punishment system) and **mesencephalon** (regulation of respiration and heart rate control)

References

- Boschi, T., Reimherr, M., Chiaromonte, F. (2021), A Highly-Efficient Group Elastic Net Algorithm with an Application to Function-On-Scalar Regression. NeurIPS
- Boschi, T., Testa, L., Chiaromonte, F., Reimherr, M. (2022+), FAStEN: an efficient adaptive method for feature selection and estimation in high-dimensional functional regressions. Submitted

Lorenzo Testa · l.testa@sssup.it · testalorenzo.github.io