

MOX-Report No. 70/2020

## **Smoothly Adaptively Centered Ridge Estimator**

Belli E.

MOX, Dipartimento di Matematica  
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

[mox-dmat@polimi.it](mailto:mox-dmat@polimi.it)

<http://mox.polimi.it>

# Smoothly Adaptively Centered Ridge Estimator

Edoardo Belli  
edoardo.belli@polimi.it

MOX - Modeling and Scientific Computing, Department of  
Mathematics, Politecnico di Milano, Italy

## Abstract

With a focus on linear models with smooth functional covariates, we propose a penalization framework (SACR) based on the nonzero centered ridge, where the center of the penalty is optimally reweighted in a supervised way, starting from the ordinary ridge solution as the initial centerfunction. In particular, we introduce a convex formulation that jointly estimates the model's coefficients and the weight function, with a roughness penalty on the centerfunction and constraints on the weights in order to recover a possibly smooth and/or sparse solution. This allows for a non-iterative and continuous variable selection mechanism, as the weight function can either inflate or deflate the initial center, in order to target the penalty towards a suitable center, with the objective to reduce the unwanted shrinkage on the nonzero coefficients, instead of uniformly shrinking the whole coefficient function. As empirical evidence of the interpretability and predictive power of our method, we provide a simulation study and two real world spectroscopy applications with both classification and regression.

## 1 Introduction

Smooth and highly collinear data generally arises in applications where high frequency acquisition devices are employed, like chemometrics, spectroscopy and electrical engineering. A natural way of modeling these data-generating processes is by means of functional data analysis (FDA) (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006), where each covariate can be seen as a smooth function  $x \in L^2(I)$  that has been evaluated at sequential timesteps, often with missing values and different spacing between the observations. Without loss of generality, we will focus on scalar-valued functions defined on  $I = [0, 1]$ , but this approach can be extended to deal with vector valued functions defined on multidimensional domains. The two main aspects to be considered are the estimation of the underlying functional covariates from the raw data, and the estimation of the predictive model itself. Let  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$  be the training

set with random i.i.d. functions  $x_i \in L^2(I)$  and responses  $y_i \in \mathbb{R}$ , we focus on the scalar on function linear model:

$$y_i = \beta_0 + \int_I x_i(t)\beta(t)dt + \epsilon_i$$

where  $\beta_0 \in \mathbb{R}$  is the intercept,  $\beta$  is the coefficient function and  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$  are random i.i.d. errors. In fact, the penalties that we analyze can also be employed in a generalized linear model (GLM) framework, and we will study classification problems with functional logistic regression as well. Regarding the first aspect, consider the functional covariate  $x$  as a finite expansion  $x(t) = \sum_{j=1}^J \xi_j \psi_j(t)$  with suitable basis functions  $\psi_j$  and coefficients  $\xi_j$ , a common approach is to recover each functional sample  $x_i$  individually, by means of interpolating or smoothing splines, depending on the amount of noise. If the raw data contains a large number of missing observations, this approach fails as some of the functional covariates may have been observed on just a few points over the domain. This issue can be solved by leveraging the information from the whole dataset, estimating the basis and the coefficients of the expansion by means of functional principal components (fPCA) with local smoothing (Yao et al., 2005) or mixed effects (Brumback and Rice, 1998; James et al., 2000). Once the input functions have been recovered, depending on the approach that has been implemented, it is possible to either work directly with the coefficients of the expansion, or to evaluate the estimated functions on the same dense equispaced  $p$ -dimensional grid. Without loss of generality, in this work we opt for the grid approach and we recover the functions individually, but the methods that we propose are not directly tied to this choice, as long as the discretized functional samples have the same dimensionality. Regarding the aspect of estimating the predictive model, the coefficient function  $\beta$  is also expressed as a basis expansion, with some form of regularization as an identifiability constraint, given that the theoretical functional linear model is ill-posed. A parsimonious approach is to restrict the number of basis functions, by either fixing a known suitable basis like a Fourier basis, or by considering only the first  $K$  eigenfunctions of the covariance operator obtained from fPCA (Cai and Hall, 2006), which for any given  $K$  explains most of the variation of the input functions in the  $L^2$  sense. On the opposite side of the spectrum, another approach is instead to employ a rich enough basis while at the same time including some form of penalization, typically an  $L_2$  penalty on  $\beta$  or its derivatives in order to impose smoothness (Cardot et al., 2003, 2007; Crambes et al., 2009; Yuan and Cai, 2010), but  $L_1$ -based penalties have also been used (Matsui and Konishi, 2011). Note that the restricted basis and the penalization approaches are not mutually exclusive, and hybrid techniques have been proposed as well (Marx and Eilers, 1999; James et al., 2009; Lee and Park, 2012). In our setting we choose to adopt the penalization approach, by using the following simple grid basis with  $p$  dense and equispaced knots placed on the  $p$  evaluations corresponding to the evaluation grid of the estimated input functions:

$$\beta(t) = \sum_{j=1}^p \beta_j b_j(t) \quad b_j(t) = \begin{cases} 1 & \text{if } \frac{j-1}{p} < t \leq \frac{j}{p} \\ 0 & \text{otherwise} \end{cases}$$

which is a common solution that enables us to use any multivariate method for the numerical estimation, allowing for a proper comparison between different approaches, as the initial FDA preprocessing is shared between all the tested methods. The main objective of this work is to propose an adaptive penalization approach that is able to fit smooth and sparse coefficient functions (James 2011), ideally being able to recover the regions of the domain in which the covariates have no effect on the response, while at the same time allowing for a smooth behaviour if needed. Given the abundance of  $p \gg N$  applications with different requirements, it is no surprise that the literature on variable selection in linear models has experienced a significant growth in both the statistical and machine learning communities. What appears to be the most successful framework is based on the well known penalized least squares formulation (in the multivariate notation), and in particular the bridge estimator (Frank and Friedman, 1993):

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^N (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|^\gamma$$

where  $\gamma > 0$ , and  $\lambda > 0$  that controls the strength of the penalization (we omit the intercept). It is known that for  $\gamma < 1$  this yields a non-convex optimization problem, where in particular for  $\gamma \rightarrow 0$  the bridge reduces to best subset selection (Fu and Knight, 2000). Besides the computational issues, subset selection methods are also known to be unstable (Breiman, 1996), and for this reason we will focus only on penalty-based approaches, although we are aware of the different stepwise algorithms. Moreover, given that the penalties are not scale-invariant, we will assume that the input data has been standardized. When  $\gamma \geq 1$  the problem is instead convex but we pay the price of unwanted shrinkage of the coefficients, which introduces bias. For  $\gamma = 1$  in particular we obtain the lasso (Tibshirani, 1996), which is a convex relaxation of best subset selection, while  $\gamma = 2$  corresponds to ridge regression (Hoerl and Kennard, 1970). Regarding our specific setting, which deals with high dimensional and highly collinear data, it is not clear which approach to adopt, as the lasso may exclude important variables from the model and produce nonsmooth coefficient functions, the ridge may yield both nonsparse and nonsmooth ones, while the usual FDA roughness penalty may be too smooth and fail to recover any sharp change in the support of the coefficient function  $\beta$ . A possible solution is to impose hybrid penalties, like in the case of the elastic net (Zou and Hastie, 2005) or the smooth lasso (Hebiri and van de Geer, 2011). Our proposed approach is instead exclusively based on the nonzero centered  $L_2$  penalty (Swindel, 1976; Price et al., 2015; van Wieringen, 2019; Bilgrau et al., 2020), and it is also inspired by the adaptive ridge estimator and other reweighted bias reduction techniques, as we will discuss in Section 2. Section 3 describes our method in detail, the applications are

shown in Section 4, with concluding remarks in Section 5.

## 2 Related Work

As previously introduced, a significant issue that follows from the convex formulations of the bridge estimator ( $\gamma \geq 1$ ) is that in order to perform variable selection, we inevitably end up with unwanted shrinkage of the "true" coefficients. This is even worse for  $\gamma > 1$ , where the amount of shrinkage increases with the magnitude of the coefficient being estimated (Fu and Knight, 2000). Moreover, it is known that except when the OLS coefficients are exactly zero, the ridge is not able to yield sparse solutions, although the coefficients can get arbitrarily small for larger values of  $\lambda$ . It follows that when estimating a sparse model with ridge regression, there could be the need to employ some form of manual thresholding, setting to zero the smaller coefficients while at the same time accepting the overshrinkage of the larger ones, with a tradeoff between  $\lambda$  and the threshold. In practice, lasso is usually the preferred choice when sparsity is sought after, as it is able to set coefficients to exactly zero by acting as a soft thresholding operator. However, an issue with the lasso is the fact that the optimal  $\lambda$  with respect to prediction gives inconsistent results from the point of view of variable selection (Meinshausen and Bühlmann, 2006), and for functional data in particular, the irrepresentable condition (Zhao and Yu, 2006) is likely to be violated, given that the curves often have high autocorrelation and the response may depend only on a subset of the domain. The main motivation behind our work is the idea of reducing bias by coefficient-wise adaptive tuning of the penalization. While the ordinary ridge regression corresponds to a sphere centered at the origin, which shrinks all the coefficients uniformly towards zero, Hoerl and Kennard (1970) already introduced a generalized form of ridge regression, which allowed to shrink each coefficient individually, resulting in an ellipsoid. In the usual penalized least squares formulation, the generalized ridge can be expressed as:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^N (y_i - x_i^\top \beta)^2 + \sum_{j=1}^p \lambda_j \beta_j^2$$

where the parameters  $\lambda_j > 0$  control the amount of shrinkage on the corresponding coefficients  $\beta_j$ . This type of penalty is also known as the adaptive ridge estimator, and regardless of the loss function, it has been shown to be equivalent to the lasso, in the sense that they recover the same solution (Grandvalet, 1998) (Grandvalet and Canu, 1999). In principle one would like to optimize with respect to both  $\beta \in \mathbb{R}^p$  and  $\lambda \in \mathbb{R}^p$ , but globally the problem is nonconvex, and the adaptive ridge estimator uses an EM approach that is guaranteed to converge to a local optimum. Instead of alternating optimization, other methods are based on iterative refinements of an initial solution  $\tilde{\beta} \in \mathbb{R}^p$ , and we will refer to such methods as two-stage or multi-stage approaches. One of the first is the non-negative garrote (NNG) (Breiman, 1995), which is closely related to

the EM adaptive ridge and has the following formulation:

$$\begin{aligned} \min_{c \in \mathbb{R}^p} \quad & \sum_{i=1}^N \left( y_i - \sum_{j=1}^p c_j x_{ij} \tilde{\beta}_j \right)^2 + \lambda \sum_{j=1}^p c_j \\ \text{s.t.} \quad & c_j \geq 0 \end{aligned}$$

where  $\lambda > 0$  and the fitted coefficients are recovered as  $\hat{\beta}_j = \hat{c}_j \tilde{\beta}_j$ . The original NNG was initialized with the OLS solution  $\tilde{\beta} = \beta^{ols}$ , but other works experimented with other initial estimators like the ridge, the lasso, and the elastic net for high dimensional scenarios (Yuan and Lin, 2007). On a side note, the NNG was also the inspiration for the original lasso paper (Hastie et al., 2015). A generalization of the NNG (without the sign constraint) is the adaptive lasso (Zou, 2006), which assumes a known weight vector  $w \in \mathbb{R}^p$  and solves:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^N (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^p w_j |\beta_j|$$

with  $w_j = 1/|\beta_j^{ols}|^\gamma$ ,  $\gamma > 0$  and  $\lambda > 0$  selected by cross-validation. This is also a two-stage approach and the final coefficients  $\hat{\beta}_j$  can be computed by setting  $\tilde{x}_{ij} = x_{ij}/w_j$ , solving a lasso problem with  $\tilde{x}_i$  as input data, and finally recovering the coefficients as  $\hat{\beta}_j = \tilde{\beta}_j/w_j$ , with  $\tilde{\beta}$  the solution of the previous lasso problem. As for the NNG, the initial estimator is not restricted to the OLS and the ridge is suggested in case of collinearity. Another reweighted estimator is the broken adaptive ridge (BAR) (Dai et al., 2018), which is a multi-stage approach that starts from a ridge penalized solution  $\hat{\beta}^0$  and at each iteration refines the previous one  $\hat{\beta}^k = g(\hat{\beta}^{k-1})$ , with  $\hat{\beta}^* = \lim_{k \rightarrow \infty} \hat{\beta}^k$  and

$$g(\tilde{\beta}) = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^N (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 / \tilde{\beta}_j^2$$

The subsequent iterations share the same  $\lambda$ , which is fixed starting from  $k = 1$  and not tuned individually at each step. Instead, the initial solution  $\hat{\beta}^0$  is not necessarily obtained with the same  $\lambda$  and could be further tuned, although empirically the BAR estimator was found to be insensitive to the initial value. All the methods that we have discussed share the common idea of using multiplicative weights in order to reduce bias, but in fact this is not the only viable approach. Nonconcave penalties like the SCAD (Fan and Li, 2001) and the MCP (Zhao et al., 2010) are both based on quadratic splines with singularities at the origin, giving rise to nonconvex optimization problems that depend on different parameters and are often regarded as unstable, although more refined optimization algorithms have been proposed (Zou and Li, 2008) (Breheny and Huang, 2011). Our approach is instead based on a convex formulation, but it is worth considering both the (elastic) SCAD and MCP for comparison purposes. Finally, yet another option to reduce bias is the one adopted by the relaxed

lasso (Meinshausen, 2007), which separates the variable selection aspect from the coefficient estimation one by first fitting a standard lasso model, followed by a second lasso but only including the covariates that correspond to the nonzero coefficients, with a relaxation parameter  $\phi \in (0, 1]$  in order to reduce unwanted shrinkage. Both problems share the same fixed  $\lambda$  and therefore this can be done pathwise, unlike from the adaptive lasso where the initial estimator has already been optimized with respect to  $\lambda$ , and then  $\lambda$  is tuned again for the reweighted problem. In particular, let  $\hat{\beta}^\lambda$  be the lasso solution for a fixed  $\lambda$  and let  $\mathcal{A}_\lambda = \{1 \leq j \leq p \mid \hat{\beta}_j^\lambda \neq 0\}$ , the relaxed lasso solution  $\hat{\beta}^{\lambda, \phi}$  is obtained by solving:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^N (y_i - x_i^\top \{\beta \mathbf{1}_{\mathcal{A}_\lambda}\})^2 + \phi \lambda \sum_{j=1}^p |\beta_j|$$

$$\{\beta \mathbf{1}_{\mathcal{A}_\lambda}\}_j = \begin{cases} \beta_j & \text{if } j \in \mathcal{A}_\lambda \\ 0 & \text{otherwise} \end{cases}$$

Our approach in a way is built on a similar relaxation scheme, but instead of performing variable selection and parameter estimation sequentially, we do it jointly and without directly removing any covariate from the initial model, by employing an adaptive weight function that acts on the center of the penalty. Like the ordinary ridge, our penalty is spherical, but is based on the nonzero centered ridge (Swindel, 1976):

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^N (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^p (\beta_j - c_j)^2$$

where the center of the sphere  $c \in \mathbb{R}^p$  is provided by the user and  $\lambda > 0$  is selected by cross-validation. For  $\lambda$  and  $c$  fixed, let  $X \in \mathbb{R}^{N \times p}$  be the design matrix and  $Y \in \mathbb{R}^N$  the response vector, the solution can be computed in closed form as:

$$\hat{\beta}^{\lambda, c} = (X^\top X + \lambda I)^{-1} (X^\top Y + \lambda c)$$

Moreover, the expected value of this estimator is:

$$\mathbb{E}_{Y|X}[\hat{\beta}^{\lambda, c}] = (X^\top X + \lambda I)^{-1} (X^\top X \beta + \lambda c)$$

and therefore it is unbiased for  $c = \beta$ , meaning that the true value of the parameter is used as the center of the penalty. Clearly there would not be the need to fit any model if  $\beta$  was already known, but this suggests that the bias will be low if the fixed  $c$  is a good approximation of the true regression coefficient, and we propose to find  $c$  by adaptively reweighting the ridge solution.

### 3 Smoothly Adaptively Centered Ridge

We already discussed some of the similarities between our approach and other known methods, with specific attention to the generalized/adaptive ridge and the nonzero centered ridge. The main downside of the adaptive ridge is that optimizing with respect to the shrinkage parameters yields a nonconvex problem, while for the nonzero centered ridge we need to specify the center of the penalty. Our focus is on the smooth  $p \gg N$  setting and in particular we will refer to the FDA terminology. We propose a convex formulation that allows for adaptive tuning of the type of shrinkage that is imposed on each region of the domain of the coefficient function  $\beta$ . Instead of employing a variable shrinkage parameter function like in the adaptive ridge, we uniformly shrink  $\beta$  as in the ordinary ridge, while at the same time jointly optimizing the center of the penalty. Let  $\tilde{\beta}^\lambda$  be the solution of the ordinary ridge for a fixed  $\lambda$ , we introduce a smooth weight function  $w : I \rightarrow \mathbb{R}^+$  that acts on  $\tilde{\beta}^\lambda$  and fit our estimator by solving the following convex problem with linear constraints:

$$\begin{aligned} \min_{\beta_0, \beta, w} \sum_{i=1}^N \left[ y_i - \beta_0 - \int_I x_i(t) \beta(t) dt \right]^2 + \lambda \phi \int_I [\beta(t) - w(t) \tilde{\beta}^\lambda(t)]^2 dt \\ + \lambda(1 - \phi) \int_I [D^2 w(t) \tilde{\beta}^\lambda(t)]^2 dt \quad (1) \\ \text{s.t.} \quad \begin{cases} \int_I w(t) dt = |I| \\ w(t) \geq 0 \end{cases} \end{aligned}$$

where  $\beta_0 \in \mathbb{R}$  is the intercept,  $\lambda > 0$  and  $\phi \in (0, 1]$  that controls the balance between the two penalty terms. Both  $\lambda$  and  $\phi$  are selected by cross-validation and the  $\lambda$  used in Problem [1](#) is the same as the one used for computing  $\tilde{\beta}^\lambda$ . The first term of the penalty is a nonzero centered ridge that shrinks  $\beta$  uniformly towards  $w\tilde{\beta}^\lambda$ , while the second term is a roughness penalty on the center of the previous one. The weight function  $w$  can be seen as an adaptive density which either contracts or dilates the initial center  $\tilde{\beta}^\lambda$ , allowing the nonzero centered penalty to selectively shrink  $\beta$  towards zero in the regions of the domain that are not correlated with the response, while reducing the unwanted shrinkage in the informative regions. This sparsity inducing behaviour is motivated by the constraints imposed on  $w$ , which necessarily lead to a tradeoff between inflating and deflating  $\tilde{\beta}^\lambda$ , as proven in Proposition [3.1](#).

**Proposition 3.1.** *Let  $I \subset \mathbb{R}$  be a closed interval and let  $w : I \rightarrow \mathbb{R}^+$  be a smooth function such that  $\int_I w(t) dt = |I|$ . Consider the closed subintervals  $I_i \subset I$  such that  $\mu(I_i) \neq 0$  and  $I_i \cap I_j = \emptyset$  for  $i \neq j$ . Partition  $I$  in disjoint intervals  $I_<$ ,  $I_>$  and  $I_=$  such that  $I = I_> \cup I_< \cup I_=$  where:*

$$\begin{aligned}
I_{>} &= \bigcup_i I_i : w(t) > 1 \quad \forall t \in I_i \\
I_{<} &= \bigcup_i I_i : w(t) < 1 \quad \forall t \in I_i \\
I_{=} &= \bigcup_i I_i : w(t) = 1 \quad \forall t \in I_i
\end{aligned}$$

then  $|I_{>}| \neq 0 \iff |I_{<}| \neq 0$ .

*Proof.* From the first mean value theorem for integrals follows that:

$$\begin{aligned}
\int_{I_{>}} w(t)dt &= w(c_{>})|I_{>}|, \quad c_{>} \in I_{>}^o, \quad w(c_{>}) > 1 \\
\int_{I_{<}} w(t)dt &= w(c_{<})|I_{<}|, \quad c_{<} \in I_{<}^o, \quad w(c_{<}) < 1 \\
\int_{I_{=}} w(t)dt &= w(c_{=})|I_{=}|, \quad c_{=} \in I_{=}^o, \quad w(c_{=}) = 1
\end{aligned}$$

therefore, from  $I = I_{>} \cup I_{<} \cup I_{=}$  results that:

$$\begin{aligned}
\int_I w(t)dt &= \int_{I_{>}} w(t)dt + \int_{I_{<}} w(t)dt + \int_{I_{=}} w(t)dt \\
|I| &= w(c_{>})|I_{>}| + w(c_{<})|I_{<}| + |I_{=}| \\
|I| &= w(c_{>})|I_{>}| + w(c_{<})|I_{<}| + |I| - |I_{>}| - |I_{<}| \\
0 &= |I_{>}|[w(c_{>}) - 1] + |I_{<}|[w(c_{<}) - 1]
\end{aligned}$$

by construction we know that  $[w(c_{>}) - 1] > 1$  and  $[w(c_{<}) - 1] < 1$ , proving that  $I_{>}$  and  $I_{<}$  are either both null sets or both non-null sets.  $\square$

It follows that when the true  $\beta$  is provided as  $\tilde{\beta}^\lambda$ , there is no need to either inflate or deflate the center of the penalty, and therefore the optimal  $w$  is 1 almost everywhere, leading to an unbiased estimator as it is equal to the nonzero centered ridge with  $c = w\tilde{\beta}^\lambda = \beta$ . In practice there is no guarantee that solving Problem [1](#) with the optimal center will lead to uniform unitary weights, as  $w$  is penalized and jointly estimated with  $\beta$  from the data. From the geometrical perspective, the weight function acts as an anisotropic scaling on  $\tilde{\beta}^\lambda$ , which has the effect of adaptively moving the center of the penalty. This has the advantage of nonuniform shrinkage between the coefficients, as in the adaptive ridge with its ellipsoidal penalty, while at the same time keeping the tractable convex formulation of the spherical penalty, since  $\lambda$  is a scalar that is selected by cross-validation. Therefore, the adaptive shrinkage of the coefficient function is the result of an adaptive target, and not of an adaptive shrinkage intensity. Our approach can be described as a continuous way of doing variable selection, which is executed jointly with the estimation of the regression coefficients. As the weight function is not included in the model and is never used for

prediction, we are not adding further parameters to the model itself, although we are doubling the parameters to be estimated. Regarding the two terms of the penalty, it is worth noting that since the roughness penalty is imposed on the center of the first term, the coefficient function is only indirectly penalized with respect to its roughness, by being pushed towards an adaptively scaled center. Imposing the roughness penalty on the center itself, instead of on the weight function only, ensures the option of retrieving a smooth centerfunction, with  $\phi$  controlling the amount of smoothness, and not just a smooth scaling of  $\tilde{\beta}^\lambda$ . The choice of the ridge solution as the initial center is quite natural, as it is stable and almost always not exactly zero. This latter property of the  $L_2$  penalty is often regarded as a problem or at least an inconvenience, but in our case is instead welcomed, as the weight function is multiplicative and would not be able to inflate an initial zero coefficient. It follows that using a sparse coefficient function as the initial center equals to excluding multiple variables from the model, which is not a problem if the initial zero coefficients should indeed be zero, but is also not necessary to produce sparse or at least interpretable solutions, as the variable shrinkage induced by the weight function should already push to zero the coefficients of the unwanted variables. In practice, higher values of  $\lambda$  will correspond to higher shrinkage of the coefficient function towards the centerfunction, where the selected  $\lambda$  depends on how suitable the centerfunction is. Therefore, for a sparse and adequate centerfunction, the selected  $\lambda$  will be high and the fitted coefficient function can get arbitrarily close to zero where needed, without the tradeoff of the ordinary ridge, where we pay the price of unwanted shrinkage on the nonzero coefficients.

Until now we only considered the context of regression, but in fact our approach can be generalized to the GLM framework as follows:

$$\begin{aligned} \min_{\beta_0, \beta, w} \quad & J(\beta_0, \beta, x, y) + Pen_{\lambda\phi}(\beta, w) \\ \text{s.t.} \quad & \begin{cases} \int_I w(t)dt = |I| \\ w(t) \geq 0 \end{cases} \end{aligned}$$

where  $J$  can be any convex loss function, as in the case of functional logistic regression, and it is independent from the weight function.

With respect to the numerical optimization, as the proposed formulation is quadratic with linear equality and inequality constraints, we opted for interior point methods, which are a class of optimization algorithms that are often regarded as state of the art for these types of problems (Gondzio, 2012). In particular, we employ the solver IPOPT (Wächter and Biegler, 2006) which is based on a primal-dual interior point algorithm with filter line search. The worst-case number of iterations is  $O(\sqrt{n})$  with  $n$  the number of variables, although interior point methods usually converge in a few steps. At each iteration the dominating cost is  $O(n^3)$  for applying Newton's method in order to solve a system of equations, and therefore the overall worst-case computational cost is  $O(n^{3.5})$ . In our specific scenario we have  $n = 2p$ , as our formulation doubles the amount of parameters to be estimated.

## 4 Applications

In this section we provide some empirical results of the performance of our method (SACR) in the context of FDA with  $p \gg N$ . In particular, we show a simulation study and two real world applications, one for classification and one for regression. We compare SACR with multiple penalized methods that are known for inducing sparsity and/or smoothness, like the lasso, adaptive lasso, relaxed lasso, NNG, ridge, BAR, elastic net, elastic SCAD, elastic MCP and the roughness penalty. The base implementations for the lasso, ridge and elastic net are the ones from *scikit-learn* (Pedregosa et al., 2011), while for the adaptive lasso, relaxed lasso, NNG and BAR we implemented our own wrappers based on those. The elastic SCAD and elastic MCP are available in the CRAN package *nvcvreg* (Breheny, 2020), while we also use our own python implementation of the roughness penalized functional linear model. We interface with the solver IPOPT by modeling the optimization problems with Pyomo (Hart et al., 2017).

### 4.1 Simulation Study

This is a simulated regression problem where the true coefficient function  $\beta$  is sparse and smooth, in order to show how the adaptive centering is able to jointly shrink towards a mixed target. In particular, we simulate the input curves with the same B-spline model but with two different configurations of dependency between the coefficients, resulting in two separate simulations. The shared base model is a cubic B-spline with inner knots equispaced between  $[-0.5, 1.5]$ , while the spline coefficients are sampled from a multivariate normal for each of the  $N = 50$  observations, with either a diagonal covariance matrix and 35 inner knots as an edge case, or high positive correlation and 50 inner knots in order to simulate a standard FDA setting. The  $N$  input functions  $x_i$  are then evaluated on the same equispaced grid of length  $p = 150$  over  $I = [0, 1]$ , and the responses are computed as  $y_i = \int_I x_i(t)\beta(t)dt + \epsilon_i$  with  $\epsilon_i \sim \mathcal{N}(0, 1)$ . Figure 1 shows the input curves for both simulations and the true coefficient function, while Figures 2 and 4 show the fitted coefficient functions for all methods tested. Regarding the strictly  $L_2$ -based methods, the ridge, the roughness penalty and the SACR provide similar solutions, with varying degrees of smoothness as expected. The SACR in fact reminds of a warped version of the ridge, with a much smoother behaviour in the highly collinear simulation, which is very close to the solution of the roughness penalized model. In Figures 3 and 5 we show the comparison between the initial center, the true coefficient function and the fitted SACR function, together with the corresponding fitted weight function. In particular, while our method is not able to recover exactly zero values of the coefficient function (given the  $L_2$  norm), the overall sparsity pattern is arguably recognizable and is confirmed by the shape of the weight function, which is above one where the initial solution should be inflated, while tapering towards zero in the regions that should be sparse. Note that the fact of having a zero weight does not guarantee an exactly zero coefficient function  $\beta$ , since the weight acts on the center of the penalty and not on the coefficients themselves,

with analogous considerations for very high values of the weight, as shown in the simulation with independent coefficients. The BAR estimator seems to have problems of instability, which could be a numerical issue of our implementation given its asymptotic definition, although the method is competitive in both the subsequent real world case studies. Regarding the pure sparsity inducing penalties like the lasso, the relaxed lasso, the adaptive lasso, and the NNG, there is no clear distinction between the two simulations, and in both cases the methods recover coefficient functions with the typical spikes on some of the variables, failing to recover the exact sparsity pattern and therefore excluding from the model many of the significant predictors. On the other hand, the thresholding effect of the  $L_1$  penalty allows to set to zero the coefficients of the unwanted variables. Finally, the hybrid penalties show a clearly different behaviour in the two simulations, where all three methods visibly leverage the ridge part of the penalty in the highly collinear simulation, including in the model all the correct predictors and many unwanted ones, while in the independent simulation only the elastic net leverages the ridge part, and instead both the elastic SCAD and elastic MCP recover very sparse solutions. We report the regression results in Table 1, obtained by 5-fold cross-validation with 3-fold cross-validation for grid-search hyperparameter selection. It is worth noting that in the simulation with independent coefficients, the adaptive lasso, the elastic SCAD and elastic MCP have lower mean-square error than the SACR, despite the fact that they leave out of the model many of the relevant predictors.

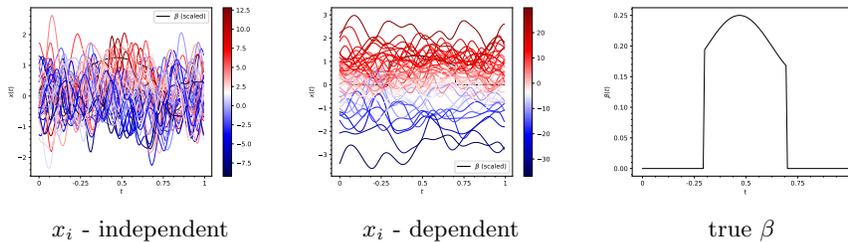


Figure 1  
 Simulated data with independent and highly dependent spline coefficients,  
 true  $\beta$

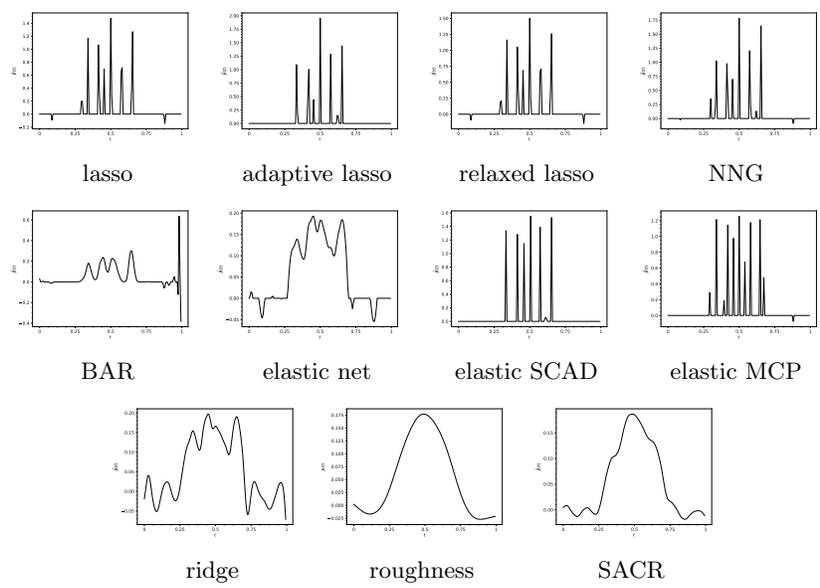


Figure 2  
Independent spline coefficients: fitted coefficient functions  $\hat{\beta}$  by penalty type

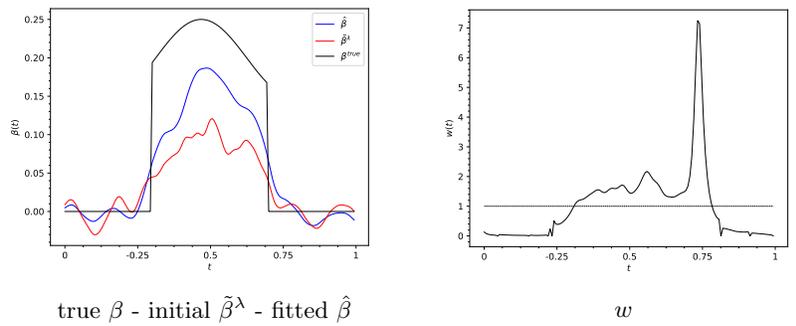


Figure 3  
Independent spline coefficients: comparison between true  $\beta$ , initial centerfunction  $\tilde{\beta}^\lambda$  and fitted SACR estimator  $\hat{\beta}$ , with corresponding weight function  $w$

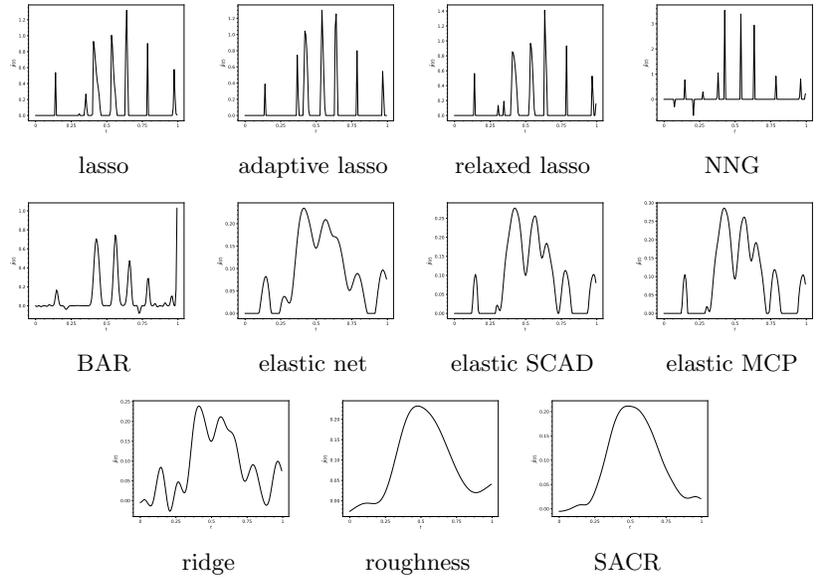


Figure 4  
 Dependent spline coefficients: fitted coefficient functions  $\hat{\beta}$  by penalty type

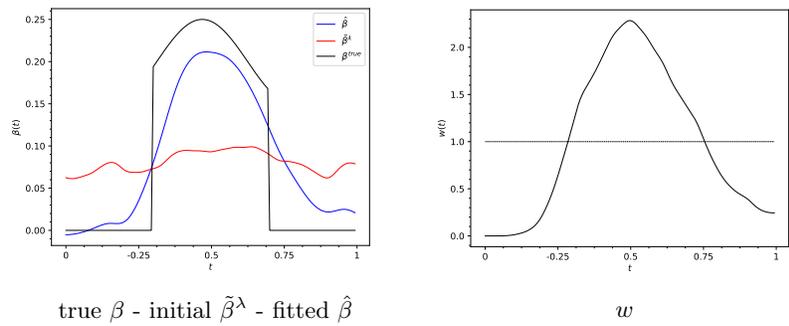


Figure 5  
 Dependent spline coefficients: comparison between true  $\beta$ , initial centerfunction  $\tilde{\beta}^\lambda$  and fitted SACR estimator  $\hat{\beta}$ , with corresponding weight function  $w$

Table 1: Regression results for both simulations: mean-square error

	independent	dependent
lasso	$1.807 \pm 1.03$	$2.011 \pm 1.2$
adaptive lasso	$1.479 \pm .067$	$2.078 \pm 1.4$
relaxed lasso	$1.807 \pm 1.03$	$1.982 \pm 1.2$
NNG	$1.571 \pm 1.01$	$2.171 \pm 1.3$
BAR	$5.669 \pm 2.60$	$2.538 \pm 1.1$
elastic net	$1.708 \pm .548$	$1.949 \pm 1.2$
elastic SCAD	$1.263 \pm .347$	$2.414 \pm 1.2$
elastic MCP	$1.291 \pm .407$	$2.407 \pm 1.1$
ridge	$2.408 \pm .458$	$1.948 \pm 1.2$
roughness	$1.540 \pm .391$	$1.894 \pm 1.3$
SACR	$1.521 \pm .308$	$1.795 \pm 1.1$

## 4.2 IDRC 2018

For regression we present a spectroscopy application that was originally proposed for the on-site competition of the 2018 International Diffuse Reflectance Conference. The data is already smooth and is available at [https://www.cnirs.org/content.aspx?page\\_id=22&club\\_id=409746&module\\_id=276203](https://www.cnirs.org/content.aspx?page_id=22&club_id=409746&module_id=276203), with  $N = 150$  and  $p = 635$ . The response variable has values over the whole dataset of  $27.7 \pm 1$  ( $\mu \pm \sigma$ ), with no information about the nature of the data. Figure 6 shows the input curves with the corresponding (rescaled) coefficient function for each of the methods that we tested, where the dashed line indicates the zero of the coefficient function. The results are reported in Table 2 and are obtained by three random repetitions of 5-fold cross-validation, with 3-fold cross-validation for grid search hyperparameter selection. In this application, the SACR and the ridge both achieve comparable mean-square error (mse) scores, but with a clear difference in the fitted coefficient functions. In fact, while the ridge recovers a noisy solution, the SACR is able to recover a sparse coefficient function which is easier to interpret. Overall, most methods tend to include the same variables in the model, with the lasso variants and the elastic net that show a noisy behaviour similar to the ridge, which may be given by the high collinearity. The roughness penalty instead recovers a smooth but oscillating coefficient function, despite achieving a similar score to the BAR estimator, which is sparse and slightly noisy. Finally, the NNG, the elastic SCAD and elastic MCP all recover very sparse solutions as expected.

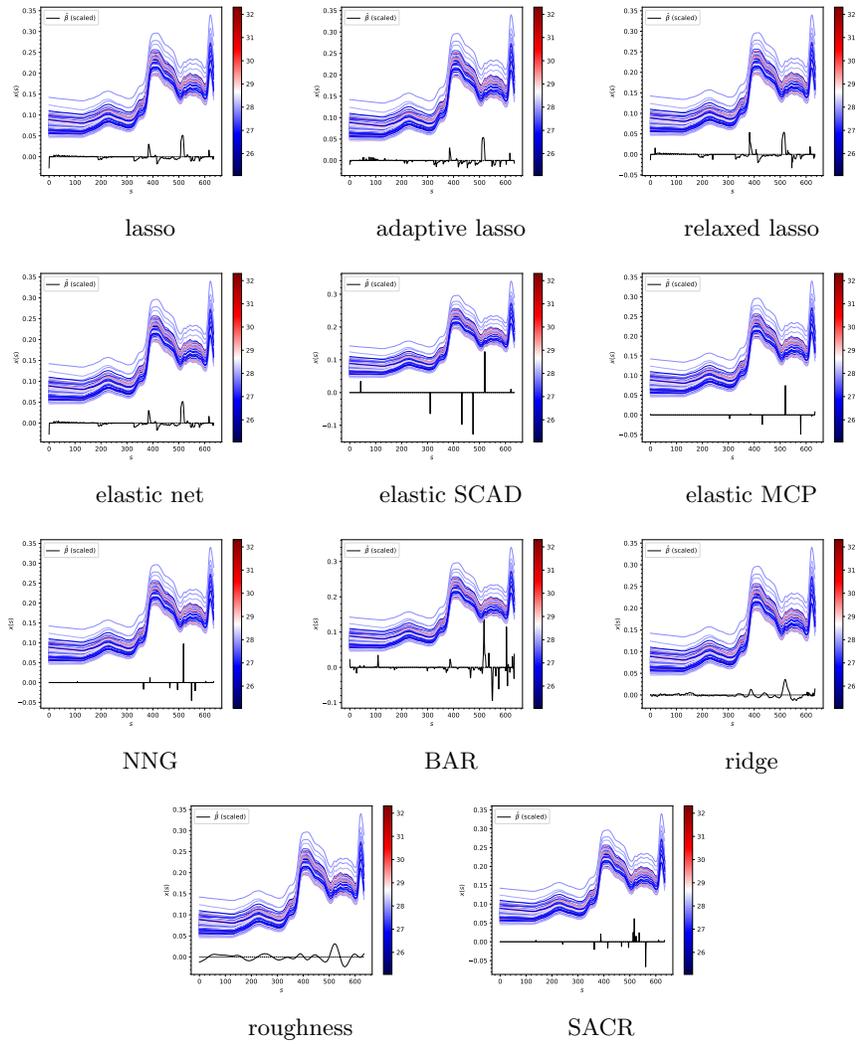


Figure 6  
 IDRC 2018: comparison of the fitted coefficient functions  $\hat{\beta}$  scaled with respect to the spectra, the black dashed line represents the zero level for  $\hat{\beta}$

Table 2: IDRC 2018: regression results, mean-square error

lasso	.1029 $\pm$ .019
adaptive lasso	.0926 $\pm$ .018
relaxed lasso	.0951 $\pm$ .017
NNG	.0750 $\pm$ .015
BAR	.0717 $\pm$ .015
elastic net	.1048 $\pm$ .019
elastic SCAD	.1559 $\pm$ .061
elastic MCP	.1415 $\pm$ .072
ridge	.0695 $\pm$ .013
roughness	.0711 $\pm$ .016
SACR	.0691 $\pm$ .014

### 4.3 Wine

This spectroscopy application is a binary classification problem in which we want to discriminate between two different wine types. The data is available at <http://www.timeseriesclassification.com/description.php?Dataset=Wine> and is already smoothed with  $N = 111$  and  $p = 234$ , although there is no additional information about the acquisition process. The results are reported in Table 3 and are obtained by three random repetitions of 5-fold cross-validation, with additional 3-fold cross-validation for grid search hyperparameter selection. While there is no clear visual distinction between the spectra of the two classes, this problem is not exceptionally hard and it is well suited for a linear model. In particular, the roughness penalty has the second lowest accuracy, suggesting that a very smooth coefficient function is not appropriate. In fact, the SACR provides the highest accuracy without leveraging the smoothing term of the penalty, as the resulting coefficient function is sparse and with multiple spikes, similar to what is usually obtained with  $L_1$ -based methods, as shown in Figure 7. Given the relatively high sample size, the lasso is also able to include many variables in the model, with the adaptive lasso and the relaxed lasso that gradually produce sparser solutions. It is interesting to note that the elastic net instead yields a coefficient function that is almost identical to the one obtained with the ridge, while the elastic SCAD and elastic MCP do not seem to leverage the ridge part of the penalty. Despite that, both approaches include different variables in the model and their resulting accuracy is lower than the other sparse methods, which may be related to the known difficulties of optimizing nonconcave penalties. The NNG has the second highest accuracy and the fitted coefficient function is in fact very similar to the one resulting from the SACR, further suggesting that a sparse solution is indeed adequate for this application, which is also confirmed by the BAR estimator, that for the most part recovers the same variables.

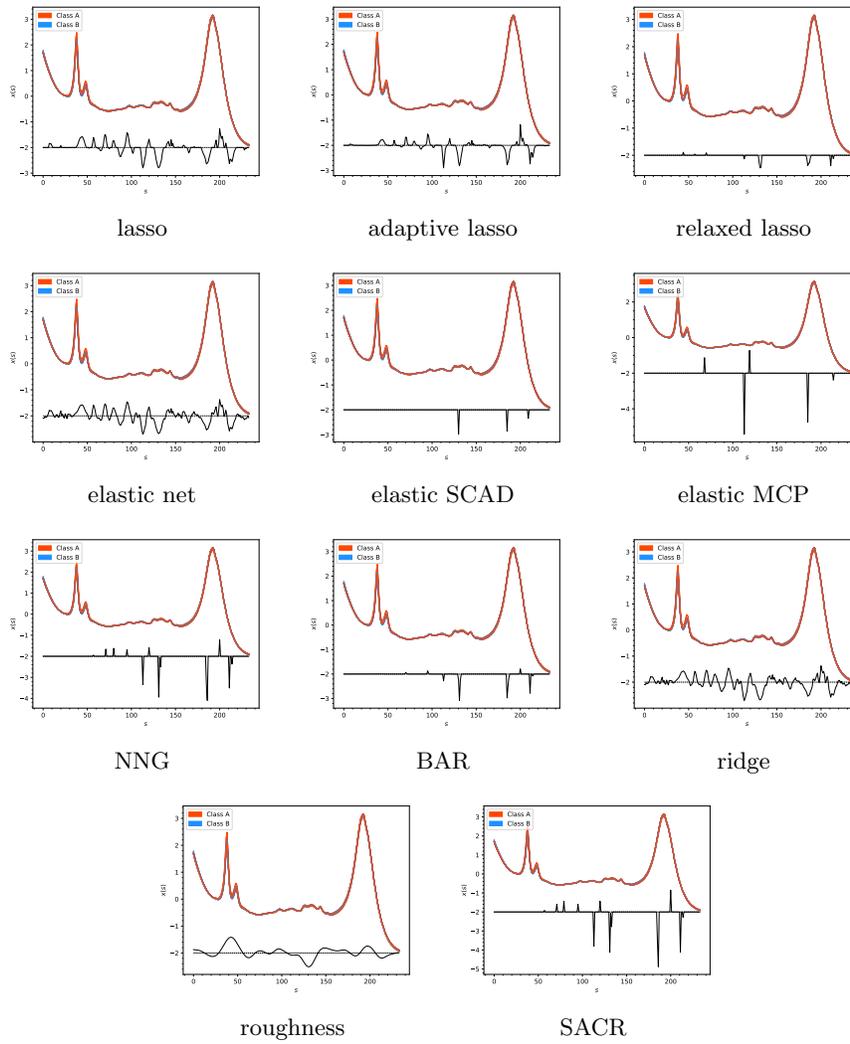


Figure 7  
 Wine: comparison of the fitted coefficient functions  $\hat{\beta}$  scaled with respect to the spectra, the black dashed line represents the zero level for  $\hat{\beta}$

Table 3: Wine: classification results, accuracy (%)

lasso	$94.64 \pm 5.9$
adaptive lasso	$96.77 \pm 4.6$
relaxed lasso	$95.88 \pm 5.0$
NNG	$97.95 \pm 3.1$
BAR	$96.82 \pm 3.3$
elastic net	$94.34 \pm 5.9$
elastic SCAD	$93.35 \pm 3.4$
elastic MCP	$89.49 \pm 6.7$
ridge	$95.25 \pm 5.5$
roughness	$90.28 \pm 8.4$
SACR	$99.44 \pm 1.1$

## 5 Conclusions

In the context of high dimensional linear models, the ordinary ridge penalty is widely known to shrink the coefficients uniformly towards zero, resulting in stable solutions at the price of introducing some bias. In order to reduce unwanted shrinkage on a subset of the coefficients, generalized and adaptive ridge estimators introduce coefficient-wise penalty parameters that allow for a non-uniform regularization effect, with the downside that tuning such parameters is a nonconvex problem. The nonzero centered ridge instead allows for a convex formulation that uniformly shrinks the coefficients towards a specific target, which in turn has to be specified by the user. In this work we have provided a convex formulation that leverages the nonzero centered ridge and allows for variable shrinkage of the coefficient function along its domain, mitigating the downside of uniform shrinkage towards zero, without the need to specify a center for the penalty, as it is learned from the data in a supervised way. In particular, we introduced a constrained weight function that is jointly estimated while fitting the model and acts as a scaling transformation on the initial centerfunction, which is the ordinary ridge solution. We referred to our method as smoothly adaptively centered ridge (SACR), since the centerfunction is adaptively scaled with respect to the loss and is further penalized for its roughness, as it is common in the functional data setting. Regarding the computational aspect, our approach doubles the number of variables to be estimated but not the ones introduced in the model, and for the numerical optimization we resorted to known primal-dual interior point methods with line search. Finally, we provided some empirical evidence with a simulation study, and two real world spectroscopy applications for both classification and regression.

## Acknowledgments

Edoardo Belli was financially supported by the ABB-Politecnico di Milano Joint Research Center through the PhD scholarship *"Development and prototyping of distributed control systems for electric networks based on advanced statistical models for the analysis of complex data"*.

## References

- Bilgrau, A.E., Peeters, C.F., Eriksen, P.S., Boegsted, M., van Wieringen, W.N., 2020. Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *Journal of Machine Learning Research* 26, 1–52.
- Breheny, P., 2020. `ncvreg`: Regularization paths for scad and mcp penalized regression models. <https://CRAN.R-project.org/package=ncvreg>. R package version 3.12.0.
- Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Statistics* 5, 232–253.
- Breiman, L., 1995. Better subset regression using the nonnegative garrote. *Technometrics* 37, 373–384.
- Breiman, L., 1996. Heuristics of instability and stabilization in model selection. *The Annals of Statistics* 24, 2350–2383.
- Brumback, B.A., Rice, J.A., 1998. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 93, 961–976.
- Cai, T.T., Hall, P., 2006. Prediction in functional linear regression. *The Annals of Statistics* 34, 2159–2179.
- Cardot, H., Crambes, C., Kneip, A., Sarda, P., 2007. Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics & Data Analysis* 51, 4832–4848.
- Cardot, H., Ferraty, F., Sarda, P., 2003. Spline estimators for the functional linear model. *Statistica Sinica* 13, 571–591.
- Crambes, C., Kneip, A., Sarda, P., 2009. Smoothing splines estimators for functional linear regression. *The Annals of Statistics* 37, 35–72.
- Dai, L., Chen, K., Sun, Z., Liu, Z., Li, G., 2018. Broken adaptive ridge regression and its asymptotic properties. *Journal of Multivariate Analysis* 168, 334–351.
- Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.
- Ferraty, F., Vieu, P., 2006. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York, NY, USA.
- Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35, 109–135.

- Fu, W., Knight, K., 2000. Asymptotics for lasso-type estimators. *The Annals of Statistics* 28, 1356–1378.
- Gondzio, J., 2012. Interior point methods 25 years later. *European Journal of Operational Research* 218, 587–601.
- Grandvalet, Y., 1998. Least absolute shrinkage is equivalent to quadratic penalization, in: Niklasson, L., Bodén, M., Ziemke, T. (Eds.), *ICANN 98. Perspectives in Neural Computing*. Springer, London, pp. 201–206.
- Grandvalet, Y., Canu, S., 1999. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage, in: Kearns, M.J., Solla, S.A., Cohn, D.A. (Eds.), *Advances in Neural Information Processing Systems 11*. MIT Press, pp. 445–451.
- Hart, W.E., Laird, C.D., Watson, J.P., Woodruff, D.L., Hackebeil, G.A., Nicholson, B.L., Siirola, J.D., 2017. *Pyomo—optimization modeling in python*. Springer Science & Business Media, New York, NY, USA.
- Hastie, T., Tibshirani, R., Wainwright, M., 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press, New York, NY, USA.
- Hebiri, M., van de Geer, S., 2011. The smooth-lasso and other  $\ell_1 + \ell_2$ -penalized methods. *Electronic Journal of Statistics* 5, 1184–1226.
- Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
- James, G., 2011. Sparseness and functional data analysis, in: Ferraty, F., Romain, Y. (Eds.), *The Oxford Handbook of Functional Data Analysis*. Oxford University Press, New York, pp. 298–323.
- James, G.M., Hastie, T.J., Sugar, C.A., 2000. Principal component models for sparse functional data. *Biometrika* 87, 587–602.
- James, G.M., Wang, J., Zhu, J., 2009. Functional linear regression that’s interpretable. *The Annals of Statistics* 37, 2083–2108.
- Lee, E.R., Park, B.U., 2012. Sparse estimation in functional linear regression. *Journal of Multivariate Analysis* 105, 1–17.
- Marx, B.D., Eilers, P.H., 1999. Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics* 41, 1–13.
- Matsui, H., Konishi, S., 2011. Variable selection for functional regression models via the  $L_1$  regularization. *Journal of Multivariate Analysis* 55, 3304–3310.
- Meinshausen, N., 2007. Relaxed lasso. *Computational Statistics & Data Analysis* 52, 374–393.

- Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34, 1436–1462.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12, 2825–2830.
- Price, B.S., Geyer, C.J., Rothman, A.J., 2015. Ridge fusion in statistical learning. *Journal of Computational and Graphical Statistics* 24, 439–454.
- Ramsay, J.O., Silverman, B.W., 2005. *Functional Data Analysis*. Springer, New York, NY, USA.
- Swindel, B.F., 1976. Good ridge estimators based on prior information. *Communications in Statistics - Theory and Methods* 5, 1065–1075.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 58, 267–288.
- Wächter, A., Biegler, L.T., 2006. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* 106, 25–37.
- van Wieringen, W.N., 2019. The generalized ridge estimator of the inverse covariance matrix. *Journal of Computational and Graphical Statistics* 28, 932–942.
- Yao, F., Müller, H., Wang, J., 2005. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100, 577–590.
- Yuan, M., Cai, T.T., 2010. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics* 38, 3412–3444.
- Yuan, M., Lin, Y., 2007. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 69, 143–161.
- Zhao, P., Rocha, G., Yu, B., 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38, 894–942.
- Zhao, P., Yu, B., 2006. On model selection consistency of lasso. *Journal of Machine Learning Research* 7, 2541–2563.
- Zou, H., 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101, 1418–1429.
- Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67, 301–320.
- Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* 36, 1509–1533.

## MOX Technical Reports, last issues

Dipartimento di Matematica  
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

- 67/2020** Caramenti, L.; Menafoglio, A.; Sgobba, S.; Lanzano, G.  
*Multi-Source Geographically Weighted Regression for Regionalized Ground-Motion Models*
- 68/2020** Galvani, M.; Torti, A.; Menafoglio, A.; Vantini S.  
*FunCC: a new bi-clustering algorithm for functional data with misalignment*
- 69/2020** Galvani, M.; Torti, A.; Menafoglio, A.; Vantini S.  
*FunCC: a new bi-clustering algorithm for functional data with misalignment*
- 66/2020** Didkovsky, O.; Ivanov, V.; Papini, M.; Longoni, L.; Menafoglio, A.  
*A comparison between machine learning and functional geostatistics approaches for data-driven analyses of solid transport in a pre-Alpine stream*
- 65/2020** Di Gregorio, S.; Vergara, C.; Montino Pelagi, G.; Baggiano, A.; Zunino, P.; Guglielmo, M.; Fu  
*Prediction of myocardial blood flow under stress conditions by means of a computational model*
- 64/2020** Fiz, F.; Viganò, L.; Gennaro, N.; Costa, G.; La Bella, L.; Boichuk A.; Cavinato, L.; Sollini, M.  
*Radiomics of Liver Metastases: A Systematic Review*
- 63/2020** Tuveri, M.; Milani, E.; Marchegiani, G.; Landoni, L.; Torresani, E.; Capelli, P.; Sperandio, N.;  
*HEMODYNAMICS AND REMODELING OF THE PORTAL CONFLUENCE IN PATIENTS WITH CANCER OF THE PANCREATIC HEAD: A PILOT STUDY*
- 62/2020** Massi, M. C.; Ieva, F.  
*Representation Learning Methods for EEG Cross-Subject Channel Selection and Trial Classification*
- 61/2020** Pozzi, S.; Redaelli, A.; Vergara, C.; Votta, E.; Zunino, P.  
*Mathematical and numerical modeling of atherosclerotic plaque progression based on fluid-structure interaction*
- 60/2020** Lupo Pasini, M; Perotto, S.  
*Hierarchical model reduction driven by a Proper Orthogonal Decomposition for parametrized advection-diffusion-reaction problems*