

# MARS via LASSO

Dohyeong Ki

Department of Statistics, UC Berkeley

Jun 1, 2023

**Ki, D.**, Fang, B., and Guntuboyina, A. (2021+) MARS via LASSO.  
Available at <https://arxiv.org/abs/2111.11694>.



R package: <https://github.com/DohyeongKi/regmdc>

# MARS (Multivariate Adaptive Regression Splines)

MARS ([Friedman \[1991\]](#)) is a regression technique that can fit models with **non-linearity** and **interaction** between variables.

Given data  $(x^{(1)}, y_1), \dots, (x^{(n)}, y_n)$  where  $x^{(i)} \in \mathbb{R}^d$  and  $y_i \in \mathbb{R}$ , MARS fits a linear combination of **products** of the hinge (ReLU) functions

$$(x_j - t_j)_+ := \max\{x_j - t_j, 0\} \quad \text{and} \quad (t_j - x_j)_+.$$

Example:  $5.3 + 2.3(x_1 - 2)_+ - 1.4(-2 - x_3)_+ + 4.7(x_1 + 3)_+(1 - x_2)_+$

# The Usual Algorithm for MARS

Model building strategy:

Greedy algorithm (like stepwise regression)

- forward selection
- backward elimination

→ Difficult to guarantee optimality and study theoretical properties

# Our Method

We propose and study a LASSO variant of the MARS method.

Data:  $(x^{(1)}, y_1), \dots, (x^{(n)}, y_n)$  where  $x^{(i)} \in [0, 1]^d$  and  $y_i \in \mathbb{R}$

Two simplifications:

- We only consider  $(x_j - t_j)_+$ .  
 $(\because (t_j - x_j)_+ = (x_j - t_j)_+ - (x_j - 0)_+ + t_j)$   
 $(x_j - t_j)_+$  is linear if  $t_j = 0$ .
- We assume  $t_j \in [0, 1]$ .

We use LASSO to fit a sparse linear combination of basis functions of the form:

$$\prod_{j \in S} (x_j - t_j)_+ \quad \text{where } S \subseteq \{1, \dots, d\} \text{ and } t_j \in [0, 1].$$

We restrict  $S$  to have no more than  $s$  (pre-specified) elements.

We need an infinite-dimensional version of LASSO (Rosset et al. [2007], Bredies and Pikkainen [2013], Condat [2020], ...).

- Parametrize infinite linear combinations with (signed) measures
- Measure complexity in terms of the (total) variation of the involved signed measures

# Our Function Class

$\mathcal{F}_{\infty-\text{mars}}^{d,s}$  is the collection of all the functions of the form

$$f(x_1, \dots, x_d) = c + \sum_{\substack{\emptyset \neq S \subseteq \{1, \dots, d\} \\ |S| \leq s}} \int_{[0,1)^{|S|}} \prod_{j \in S} (x_j - t_j)_+ d\nu_S(t_j, j \in S)$$

$\nu_S$  is a signed measure on  $[0, 1)^{|S|}$  for each  $\emptyset \neq S \subseteq \{1, \dots, d\}$  with  $|S| \leq s$

Examples) (1)  $d = s = 1$

$$f(x_1) = c + \int_{[0,1)} (x_1 - t_1)_+ d\nu_1(t_1)$$

(2)  $d = s = 2$

$$\begin{aligned} f(x_1, x_2) &= c + \int_{[0,1)} (x_1 - t_1)_+ d\nu_1(t_1) + \int_{[0,1)} (x_2 - t_2)_+ d\nu_2(t_2) \\ &\quad + \int_{[0,1]^2} (x_1 - t_1)_+ (x_2 - t_2)_+ d\nu_{1,2}(t_1, t_2) \end{aligned}$$

- The usual MARS functions are special cases.

If  $\nu_S$  is supported on a **finite** set  $\{(t_{\ell j}^S, j \in S) : \ell = 1, \dots, k_S\}$  with

$$\nu_S(\{(t_{\ell j}^S, j \in S)\}) = b_\ell^S \quad \text{for } \ell = 1, \dots, k_S,$$

then the function becomes

$$f(x_1, \dots, x_d) = c + \sum_{\substack{\emptyset \neq S \subseteq \{1, \dots, d\} \\ |S| \leq s}} \sum_{\ell=1}^{k_S} b_\ell^S \cdot \prod_{j \in S} (x_j - t_{\ell j}^S)_+$$

# Complexity Measure

Complexity measure for  $f \in \mathcal{F}_{\infty-\text{mars}}^{d,s}$ :

$$V_{\text{mars}}(f) := \sum_{\substack{\emptyset \neq S \subseteq \{1, \dots, d\} \\ |S| \leq s}} |\nu_S|([0, 1]^{|S|} \setminus \{(0, \dots, 0)\}).$$

- The sum of the variation of the involved signed measures
- $(0, \dots, 0)$  is excluded;  
the products of linear functions are not penalized

If  $\nu_S$  is supported on a finite set  $\{(t_{\ell j}^S, j \in S) : \ell = 1, \dots, k_S\}$  with

$$\nu_S(\{(t_{\ell j}^S, j \in S)\}) = b_\ell^S \quad \text{for } \ell = 1, \dots, k_S,$$

then

$$V_{\text{mars}}(f) = \sum_{\substack{\emptyset \neq S \subseteq \{1, \dots, d\} \\ |S| \leq s}} \sum_{\ell=1}^{k_S} |b_\ell^S| \cdot \mathbf{1}\{(t_{\ell j}^S, j \in S) \neq (0, \dots, 0)\},$$

which is the sum of the absolute values of the coefficients  
(the coefficients of the product of linear functions are excluded).

# Our Estimator

Our infinite-dimensional LASSO estimator for MARS fitting:

$$\hat{f}_{n,V}^{d,s} \in \operatorname{argmin}_f \left\{ \sum_{i=1}^n (y_i - f(x^{(i)}))^2 : f \in \mathcal{F}_{\infty-\text{mars}}^{d,s} \text{ and } V_{\text{mars}}(f) \leq V \right\}$$

$V > 0$  is a single tuning parameter

# Existence and Computation

$\hat{f}_{n,V}^{d,s}$  exists and can be computed by applying finite-dimensional LASSO algorithms to the finite basis of functions

$$\left\{ \prod_{j \in S} (x_j - t_j)_+ : S \subseteq \{1, \dots, d\} \text{ with } |S| \leq s \right. \\ \left. \text{and } t_j \in \{0\} \cup \{x_j^{(1)}, \dots, x_j^{(n)}\} \right\}$$

$x^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)})$  for the  $i^{th}$  design point  $x^{(i)}$

- We can find  $\hat{f}_{n,V}^{d,s}$  that is a sparse linear combination of the basis functions.
- The usual MARS algorithm also works with the same finite basis although no theoretical justification is provided for this reduction.

# Approximation

The number of basis functions in the worst case:  $O(n^s)$   
(ignoring a multiplicative factor in  $d$ )

We also consider the approximate version  $\tilde{f}_{n,V}^{d,s}$  that is obtained by restricting the knots  $t_j$  as

$$t_j \in \left\{ 0, \frac{1}{N_j}, \frac{2}{N_j}, \dots, 1 \right\}$$

for some pre-specified integers  $N_1, \dots, N_d$ .

# Rate of Convergence

Under the assumptions:

- data  $(x^{(1)}, y_1), \dots, (x^{(n)}, y_n)$  are generated according to the model

$$y_i = f^*(x^{(i)}) + \xi_i \quad \text{where } \xi_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

- $x^{(i)}$  are i.i.d. RVs on  $[0, 1]^d$  with pdf  $p_0$  bounded by  $B$ ,
- $f^* \in \mathcal{F}_{\infty-\text{mars}}^{d,s}$  with  $V_{\text{mars}}(f^*) \leq V$  and  $\|f^*\|_\infty \leq M$ ,
- the loss function is

$$\mathcal{L}(\hat{f}_{n,V,M}^{d,s}, f^*) := \int (\hat{f}_{n,V,M}^{d,s}(x) - f^*(x))^2 p_0(x) dx,$$

we prove that

$$\mathbb{E}\mathcal{L}(\hat{f}_{n,V,\textcolor{red}{M}}^{d,s}, f^*) = O_{d,\sigma,V,B,M}\left(n^{-\frac{4}{5}}(\log n)^{\frac{8(s-1)}{5}}\right).$$

Remark)  $d = 1$

It was proved the rate is  $n^{-\frac{4}{5}}$  (see, e.g., Mammen and van de Geer [1997]).

→ Similar results can be proved for the approximate version  $\tilde{f}_{n,V,\textcolor{red}{M}}^{d,s}$

# Minimax Lower Bound

Under the same assumption, we prove that the minimax rate under the loss function  $\mathcal{L}$  over the class

$$\{f \in \mathcal{F}_{\infty-\text{mars}}^{d,s} : V_{\text{mars}}(f) \leq V \text{ and } \|f^*\|_\infty \leq M\}$$

is bounded from below by

$$n^{-\frac{4}{5}} (\log n)^{\frac{8(s-1)}{5}}.$$

# Connection to Smoothness Constrained Estimation

$$d = s = 1$$

The **total variation** of a function  $g : [0, 1] \rightarrow \mathbb{R}$  is defined by

$$V(g) := \sup_{0 = u_0 < u_1 < \dots < u_k = 1} \sum_{i=0}^{k-1} |g(u_{i+1}) - g(u_i)|$$

where the supremum is over all  $k \geq 1$  and partitions  $0 = u_0 < u_1 < \dots < u_k = 1$  of  $[0, 1]$ .

Then, somewhat loosely, we can describe the estimator  $\hat{f}_{n,V}^{1,1}$  as

$$\hat{f}_{n,V}^{1,1} \in \operatorname{argmin}_f \left\{ \sum_{i=1}^n (y_i - f(x^{(i)}))^2 : V(f') \leq V \right\}.$$

Corresponding penalized version:

$$\operatorname{argmin}_f \left\{ \sum_{i=1}^n (y_i - f(x^{(i)}))^2 + \lambda V(f') \right\}$$

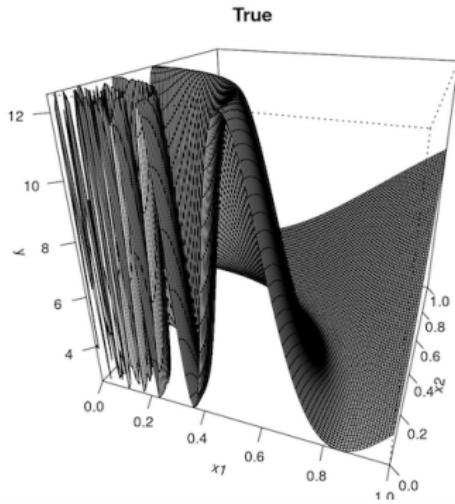
The piecewise linear **locally adaptive regression spline** (LARS) estimator of [Mammen and van de Geer \[1997\]](#)

# Example

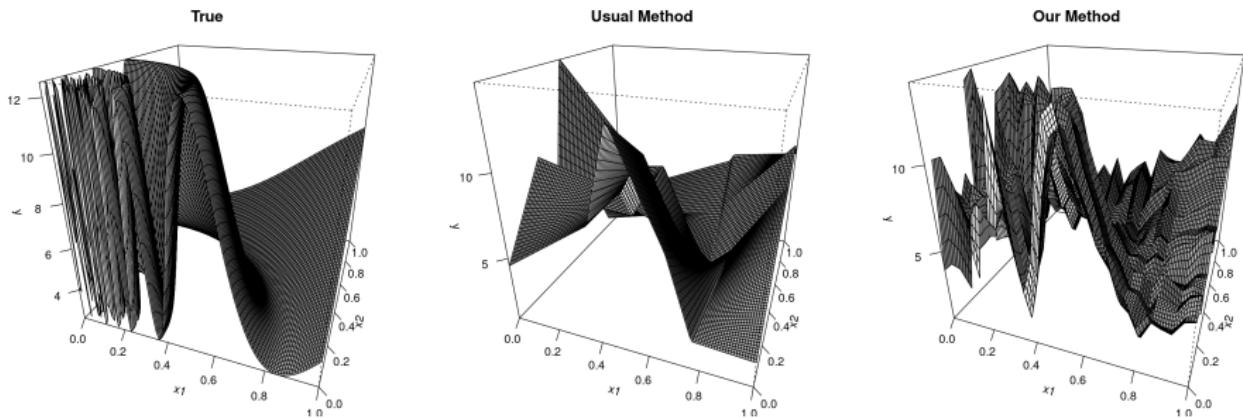
A function with locally varying smoothness (Doppler function):

$$y_i = 5 \cdot \sin \left( 4 / \left( \sqrt{(x_1^{(i)})^2 + (x_2^{(i)})^2} + 0.001 \right) \right) + 7.5 + \xi_i$$

for  $i = 1, \dots, n$ , where  $\xi_i \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ .



$n = 800$ ,  $s = 2$ ,  $N_j = 25$ ,  $V$  is chosen by 10-fold cross validation



Average loss over 25 repetitions

	Usual Method	Our Method
Average loss (Standard error)	3.28 (0.07)	1.51 (0.06)

More examples (simulated data and real data) are in  
<https://github.com/DohyeongKi/mars-lasso-paper>

# Conclusion

- We propose and study an infinite-dimensional LASSO estimator for MARS.
- Our estimator can be computed with finite dimensional LASSO algorithms.
- Our estimator achieves the rate  $n^{-\frac{4}{5}}(\log n)^{\frac{8(s-1)}{5}}$  under the standard nonparametric regression setting.

- The dependence on the dimension of the exponent of the log factor is inevitable.
- It can be considered as a multivariate generalization of the piecewise linear locally adaptive regression spline estimator of Mammen and van de Geer [1997].

- K. Bredies and H. K. Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.
- L. Condat. Atomic norm minimization for decomposition into complex exponentials and optimal transport in Fourier domain. *Journal of Approximation Theory*, 258:105456, 2020.
- J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.
- E. Mammen and S. van de Geer. Locally adaptive regression splines. *Annals of Statistics*, 25(1):387–413, 1997.
- S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu.  $\ell_1$  regularization in infinite dimensional feature spaces. In *International Conference on Computational Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 544–558. Springer, Berlin, 2007.