Outline

Highly Adaptive Lasso

For a cadlag function $\psi : [0, \tau] \subset \mathbb{R}^d \to \mathbb{R}$ with finite variation norm (and thus generates a signed measure), we have

$$\psi(x) \;=\; \sum_{s \subset \{1,\ldots,d\}} \int I(x_s \geq u_s) d\psi_s(u_s),$$

where $\psi_s(u) = \psi(u_s, 0_{s^c})$ is the section of ψ that sets the coordinates in s equal to zero. Here $x_s = (x(j) : j \in s)$ and the sum is over all subsets of $\{1, \ldots, d\}$. The variation norm of ψ can be defined as:

$$\parallel \psi \parallel_v = \sum_{s \subset \{1,\ldots,d\}} \int \mid d\psi_s(u_s \mid .$$

For discrete measures $d\psi_s$ with support points $\{u_{s,j} : j\}$ one obtains the following linear combination of indicator basis functions:

$$\psi(x) = \sum_{s \subset \{1,\dots,d\}} \sum_j \beta_{s,j} \phi_{u_{s,j}}(x),$$

where $\beta_{s,j} = d\psi_s(u_{s,j})$, and

$$\begin{aligned} \| \psi \|_v &= \sum_{s \subset \{1,\dots,d\}} \sum_j \mid \beta_{s,j} \mid \\ &\equiv \| \beta \|_1 . \end{aligned}$$

Highly Adaptive Lasso

Consider a loss function $L(\psi)$ such as $L(\psi)(X, Y) = (Y - \psi(X))^2$, let $\psi_0 = \arg\min_\psi P_0 L(\psi)$ and let

$$d_0(\psi, \psi_0) = P_0 L(\psi) - P_0 L(\psi_0)$$

be the loss-based dissimilarity. Consider the constrained MLE:

$$\psi_{n,M} = \arg \min_{\psi, \|\psi\|_v < M} P_n L(\psi).$$

Given that this MLE is attained at a discrete measure $d\psi_{n,M}$, this MLE is given by $\psi_{n,M} = \sum_{s \subset \{1,\ldots,d\}} \beta_{n,M,s,j} \phi_{u_{s,j}}$, where

$$\beta_{n,M} = \arg \min_{\beta, \|\beta\|_1 < M} \frac{1}{n} \sum_{i-1}^{n} (Y_i - \sum_{s,j} \beta_{s,j} \phi_{u_{s,j}}(X_i))^2.$$

In other words, $\beta_{n,M}$ is computed with the Lasso.

As in the Lasso, we select M with cross-validation. Let M_n be the cross-validation selector and

$$\psi_n = \psi_{n,M_n}.$$

We refer to ψ_n as the Highly Adaptive Lasso estimator (HAL-E).

Guaranteed rate faster than $n^{-1/4}$

We have

$$d_0(\psi_{n,M}, \psi_{0,M}) = o_P(n^{-(1/2+\alpha(d)/4)}),$$

where $\alpha(d) = 1/(d+1)$. Thus, if we select $M > \parallel \psi_0 \parallel_v$, then

$$d_0(\psi_{n,M}, \psi_0) = o_P(n^{-(1/2+\alpha(d)/4)}).$$

Due to oracle inequality for the cross-validation selector M_n, as long as $\parallel \psi_0 \parallel_v < \infty$, we have

$$d_0(\psi_n = \psi_{n,M_n}, \psi_0) = o_P(n^{-(1/2+\alpha(d)/4)}).$$

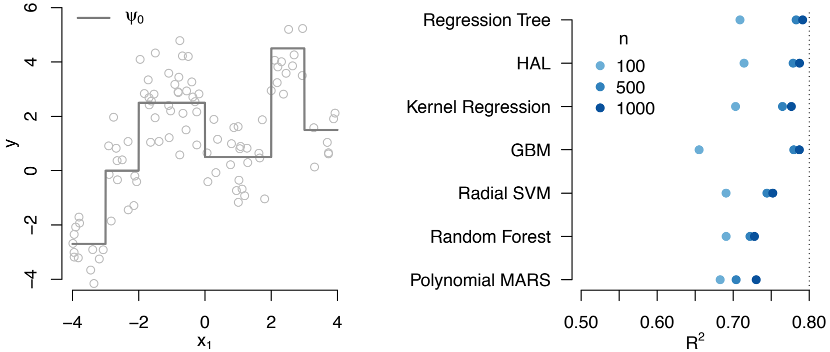
Super Learner should use HAL to guarantee this minimal rate of convergence: HAL-SL

HAL Performance

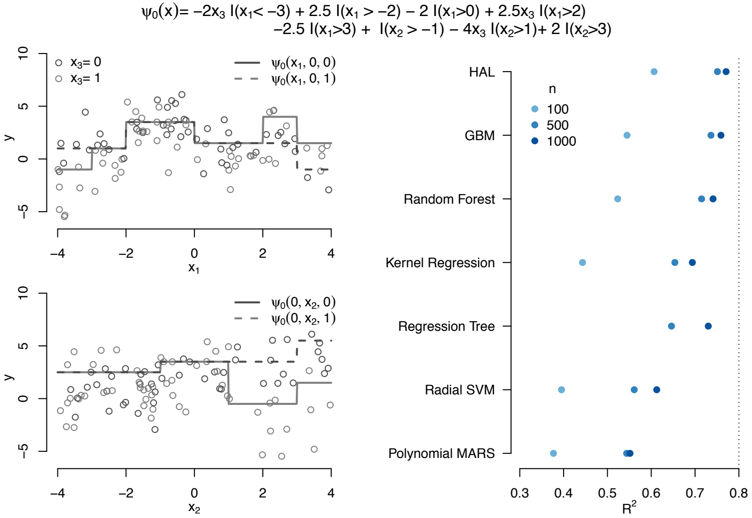
For each simulations 20 data sets of sample size n were drawn from P_0. Each data generating mechanism was chosen such that the optimal $R^2 = 0.8$.

HAL was evaluated against competitor algorithms based on R^2 calculated on an independent evaluation data set of size 10,000 averaged across the 20 data sets.

Jump functions, $d = 1$



Jump functions, $d = 3$



Overall Performance