Variable selection using Adaptive Non-linear Interaction Structures in High dimensions

Gareth James

University of Southern California

15th of January 2011

Joint with Peter Radchenko

Outline

Outline

Outline

Outline

High Dimensional Linear Regression

Recently considerable attention has focussed on fitting the traditional linear regression model,

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad i = 1, \ldots n, \tag{1}$$

where the number of predictors, p, is large relative to the number of observations, n. An important class of variable selection methods utilizes penalized regression. The most well known of these procedures is the Lasso.

Lasso

The Lasso (Tibshirani, 1996) works by fitting a penalized least squares regression of the form

$$\arg\min \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda \sum_{j=1}^{n}|\beta_j| \qquad (2)$$

This has the effect of shrinking the estimated coefficients towards zero and also setting many of the coefficients to exactly zero, hence performing variable selection.

Extending in Two Directions

We wish to extend the linear regression model in two important directions. First, we remove the additive assumption by including interaction terms, using the standard two-way interaction model,

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \sum_{j>k} \beta_{jk} X_{ij} X_{ik} + \epsilon_i, \quad i = 1, \ldots n. \qquad (3)$$

Second, we extend (3) to the more general non-linear domain using,

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{j>k} f_{jk}(X_{ij}, X_{ik}) + \epsilon_i, \quad i = 1, \ldots n. \qquad (4)$$

Difficulties

- While (3) and (4) are well known models, fitting them involves estimating on the order of p^2 terms, most of which, in the case of (4), are two-variate functions.
- Two possible approaches are to
  1. Consider all possible two-way interactions but this will likely result in many false positive interactions swamping the main effects.
  2. Search only for main effects and then attempt to find interactions among the subset of selected variables. However, this approach risks missing important interactions where the main effects are weak.

Difficulties

- While (3) and (4) are well known models, fitting them involves estimating on the order of p^2 terms, most of which, in the case of (4), are two-variate functions.
- Two possible approaches are to
  1. Consider all possible two-way interactions but this will likely result in many false positive interactions swamping the main effects.
  2. Search only for main effects and then attempt to find interactions among the subset of selected variables. However, this approach risks missing important interactions where the main effects are weak.
- We wish to consider a third option which allows all interactions to enter the model but gives preference to those that are associated with variables that have already been selected.

Prior Literature

- **SHIM** (Choi et al., 2010).
  Fits non-additive but still linear model.
- **SpAM** (Ravikumar et al., 2009).
  Fits non-linear but still additive model.
- **Non-linear SIS** (Fan et al., 2010).
  Fits non-linear but still additive model.
- **Non-negative garrote** (Yuan, 2007).
  Discusses non-additive models but concentrates on additive situation.
- **COSSO** (Lin and Zhang, 2006).
  Discusses non-additive models but concentrates on additive situation.

A General Penalization Approach

Our general approach is to minimize the following penalized regression criterion,

$$\frac{1}{2}\left\|\mathbf{Y}-\sum_{j=1}^{p}\mathbf{f}_j-\sum_{j=1}^{p}\sum_{k=j+1}^{p}\mathbf{f}_{jk}\right\|^2+P(f), \tag{5}$$

where $\mathbf{f}_j=(f_j(X_{1j}),...,f_j(X_{nj}))^T$,
$\mathbf{f}_{jk}=(f_{jk}(X_{1j},X_{1k}),...,f_{jk}(X_{nj},X_{nk}))^T$, $\mathbf{Y}$ and ϵ are n-dimensional vectors respectively corresponding to the response and error terms, and $P(f)$ is a penalty function on the $\mathbf{f}_j$ and $\mathbf{f}_{jk}$ terms.

SpIn: An Intuitive Penalty

A simple approach to fit (4) would be to use a penalty function of the form,

$$P(f) = \lambda \left(\sum_{j=1}^{p} \|\mathbf{f}_j\|_+ \sum_{j=1}^{p} \sum_{k=j+1}^{p} \|\mathbf{f}_{jk}\| \right). \tag{6}$$

- This penalty is analogous to fitting a group Lasso.
- It treats main effects and interactions equally which has two problems:
  1. There are many interaction terms so we will end up with a high number of false positives.
  2. Interaction terms are less interpretable so, all other things equal, we would prefer to include main effects.

Variable selection using Adaptive Non-linear Interaction Structures in High dimensions

We use the following penalty function,

$$P(f) = \lambda_1 \sum_{j=1}^{p} \left(\|\mathbf{f}_j\|^2 + \sum_{k:\, k \neq j}^{p} \|\mathbf{f}_{jk}\|^2 \right)^{1/2} + \lambda_2 \sum_{j=1}^{p} \sum_{k=j+1}^{p} \|\mathbf{f}_{jk}\|. \quad (7)$$

λ_1 can be interpreted as the weight of the penalty for each additional predictor included in the model, and λ_2 corresponds to an additional penalty on the interaction terms for the reduction in interpretability of a non-additive model.

VANISH Algorithm: Main Effects

0. Initialize $\widehat{\mathbf{f}}_l = \mathbf{0}, \widehat{\mathbf{f}}_{lk} = \mathbf{0}$ for all $j, k \in \{1, ..., p\}$.

For each $j \in \{1, ..., p\}$,

1. Compute the residual: $\mathbf{R}_j = \mathbf{Y} - \sum_{l:l \neq j} \widehat{\mathbf{f}}_l - \sum_{k>l} \widehat{\mathbf{f}}_{lk}$.
2. Compute $\widehat{\mathbf{P}}_j = S_j \mathbf{R}_j$, where S_j is a linear smoother. This gives the unshrunk estimate of $\mathbf{f}_j$.
3. Set $\widehat{\mathbf{f}}_j = \alpha_j \widehat{\mathbf{P}}_j$ where $0 \leq \alpha_j \leq 1$ is the shrinkage parameter defined below.

- When all $\|\widehat{\mathbf{f}}_{jk}\| = 0$ then the shrinkage parameter can be computed in closed form using $\alpha_j = \left(1 - \lambda_1 / \|\widehat{\mathbf{P}}_j\|\right)_+$.

VANISH Algorithm: Interaction Terms

For each (j, k) with $1 \leq j < k \leq p$,

4. Compute residual: $\mathbf{R}_{jk} = \mathbf{Y} - \sum_{l=1}^{p} \widehat{\mathbf{f}}_l - \sum_{m>l,\,(l,m)\neq(j,k)} \widehat{\mathbf{f}}_{lm}$
5. Set $\widehat{\mathbf{P}}_{jk} = S_{jk}\mathbf{R}_{jk}$. $\widehat{\mathbf{P}}_{jk}$ is the projection of the residuals, $\mathbf{R}_{jk}$, and corresponds to the unshrunk estimate of $\mathbf{f}_{jk}$.
6. Let $\widehat{\mathbf{f}}_{jk} = \alpha_{jk}\hat{\mathbf{P}}_{jk}$ where $0 \leq \alpha_{jk} \leq 1$ is a shrinkage parameter.

- When the main effects associated with variables j and k are missing from the model then $\alpha_{jk} = \left(1 - (2\lambda_1 + \lambda_2)/\|\widehat{\mathbf{P}}_{jk}\|\right)_+$.

Effect of the VANISH Algorithm

The VANISH algorithm will add variabiles to the model iff the norms of their unshrunk estimates are above a given threshold. For the main effect $\mathbf{f}_j$ the threshold is given by

$$\text{Threshold for } \mathbf{f}_j \text{ to enter} = \begin{cases} \lambda_1 & , \|\mathbf{f}_{jk}\| = 0 \text{ for all } k \\ 0 & , \text{otherwise.} \end{cases}$$

The threshold for adding the interaction term $\mathbf{f}_{jk}$ is as follows,

$$\text{Threshold for } \mathbf{f}_{jk} \text{ to enter} \quad = \quad \lambda_2 + \begin{cases} 2\lambda_1 & , \|\mathbf{f}_j\| = \|\mathbf{f}_k\| = 0 \\ \lambda_1 & , \text{either } \|\mathbf{f}_j\| \neq 0 \text{ or } \|\mathbf{f}_k\| \neq 0 \\ 0 & , \|\mathbf{f}_j\| \neq 0 \text{ and } \|\mathbf{f}_k\| \neq 0, \end{cases}$$

Methods

- VANISH Our approach
- SpAM (Ravikumar et al., 2009)
- SpAM$_{LS}$: least squares fits based on each SpAM model
- SpIn: SpAM with interactions, using penalty (6)
- SpIn$_{LS}$: least squares fits based on each SpIn model
- Oracle: least squares fit on the correct model

Note: we use an independently generated validation set to select the tuning parameters for each method

Linear Simulation Study

- Generated 100 training data sets, each with $n = 75$ observations, and $p = 100$ predictors.
- The $p = 100$ case corresponded to $100 \times 101/2 = 5,050$ possible main effects and interactions.
- $s_m = 5$ of the regression coefficients were randomly set to either ± 0.5, or to ± 1.
- Each generated model contained $s_{int} = 0$, $s_{int} = 2$ or $s_{int} = 6$ interaction terms produced by multiplying together two main effects with non-zero coefficients.
- The main effects as well as the error terms came from an uncorrelated standard normal distribution.

Linear Simulation Results

Simulation	Statistic	VANISH	SpIn_{LS}	SpIn	SpAM_{LS}	SpAM	Oracle
	False-Pos Main	0.81	0.16	0.7	0.74	13.38	–
$\beta = \pm 1$	False-Neg Main	0	0.17	0.08	0	0	–
	False-Pos Inter	0.67	5.8	31.42	–	–	–
$S_{int} = 0$	False-Neg Inter	0	0	0	–	–	–
	L2-sq	**0.125**	0.739	1.54	**0.11**	0.376	0.068
	False-Pos Main	2.41	0.2	0.53	1.71	12.88	–
$\beta = \pm 1$	False-Neg Main	0.01	1.24	0.88	0.25	0.07	–
	False-Pos Inter	2.03	8.52	27.62	–	–	–
$S_{int} = 2$	False-Neg Inter	0.06	0.63	0.49	–	–	–
	L2-sq	**0.408**	3.048	3.846	2.79	3.188	0.118
	False-Pos Main	5.81	0.19	0.34	3.03	11.79	–
$\beta = \pm 1$	False-Neg Main	0.22	2.99	2.52	1.25	0.58	–
	False-Pos Inter	6.62	9.99	25.42	–	–	–
$S_{int} = 6$	False-Neg Inter	1.18	3.85	3.46	–	–	–
	L2-sq	**2.758**	14.674	12.313	8.613	8.253	0.221
	False-Pos Main	4.67	0.11	0.27	2.75	9.93	–
$\beta = \pm 0.5$	False-Neg Main	1.07	4.08	3.62	2.09	1.24	–
	False-Pos Inter	5.06	5.77	16.94	–	–	–
$S_{int} = 6$	False-Neg Inter	2.86	5.19	4.68	–	–	–
	L2-sq	**1.671**	4.345	2.996	2.804	2.321	0.199

Non-Linear Simulation Study

- Generated 100 training data sets, with either $n = 300/p_m = 50$ or $n = 75/p_m = 100$ observations/main effects, each independently sampled from a Uniform distribution on the $[0, 1]$ interval.
- The responses were produced, either using the non-linear basis function model,

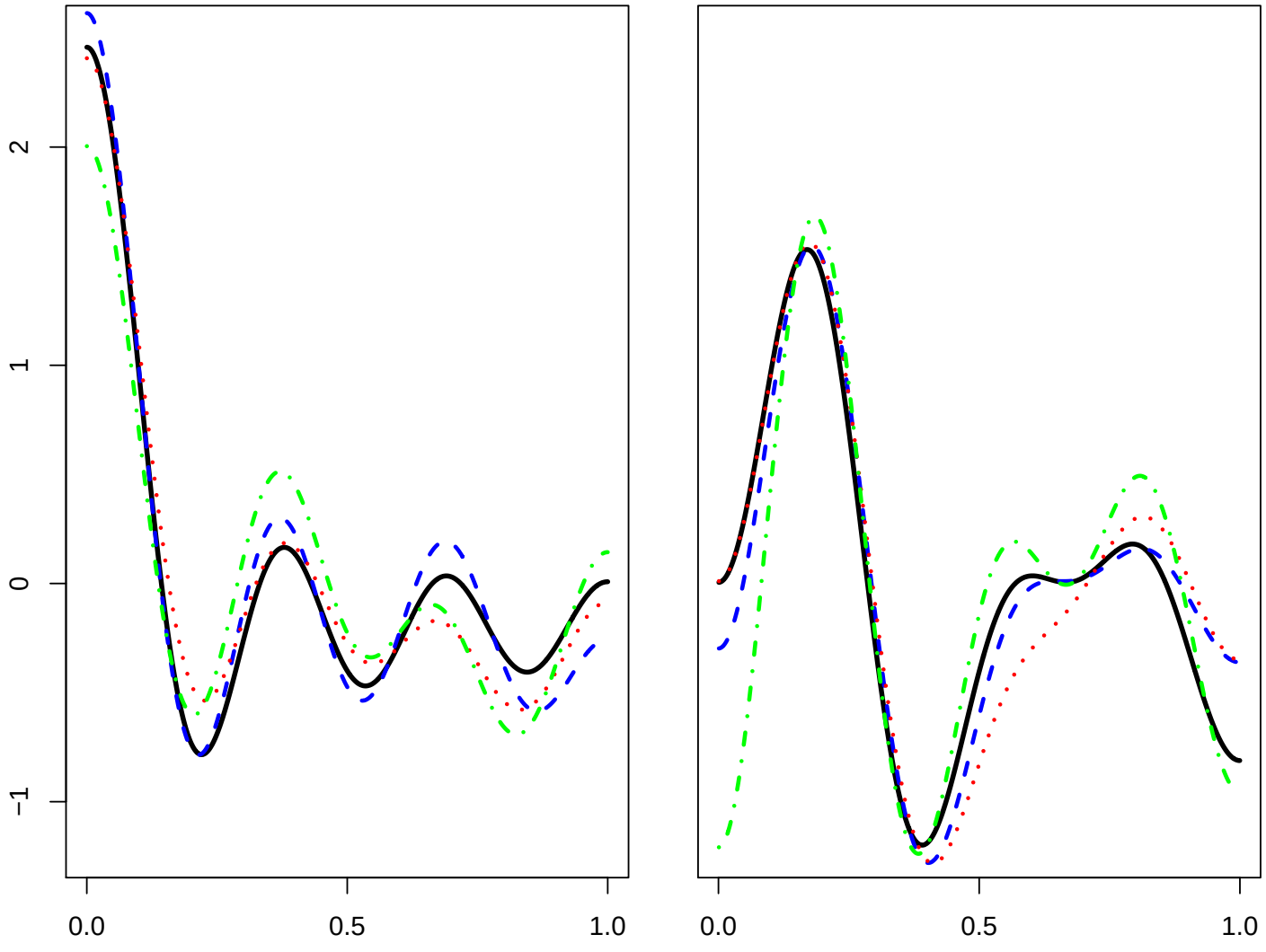
$$Y = f_1(x_1)+f_2(x_2)+f_3(x_3)+f_4(x_4)+f_5(x_5)+f_{12}(x_1, x_2)+f_{13}(x_1, x_3)+\epsilon,$$

  or else the same model without the interaction terms.
- For some simulations f_j and f_{jk} were generated from the same Fourier basis used by the candidate methods.
- For other simulations the functions had a different functional form than the Fourier basis used by the candidate methods.

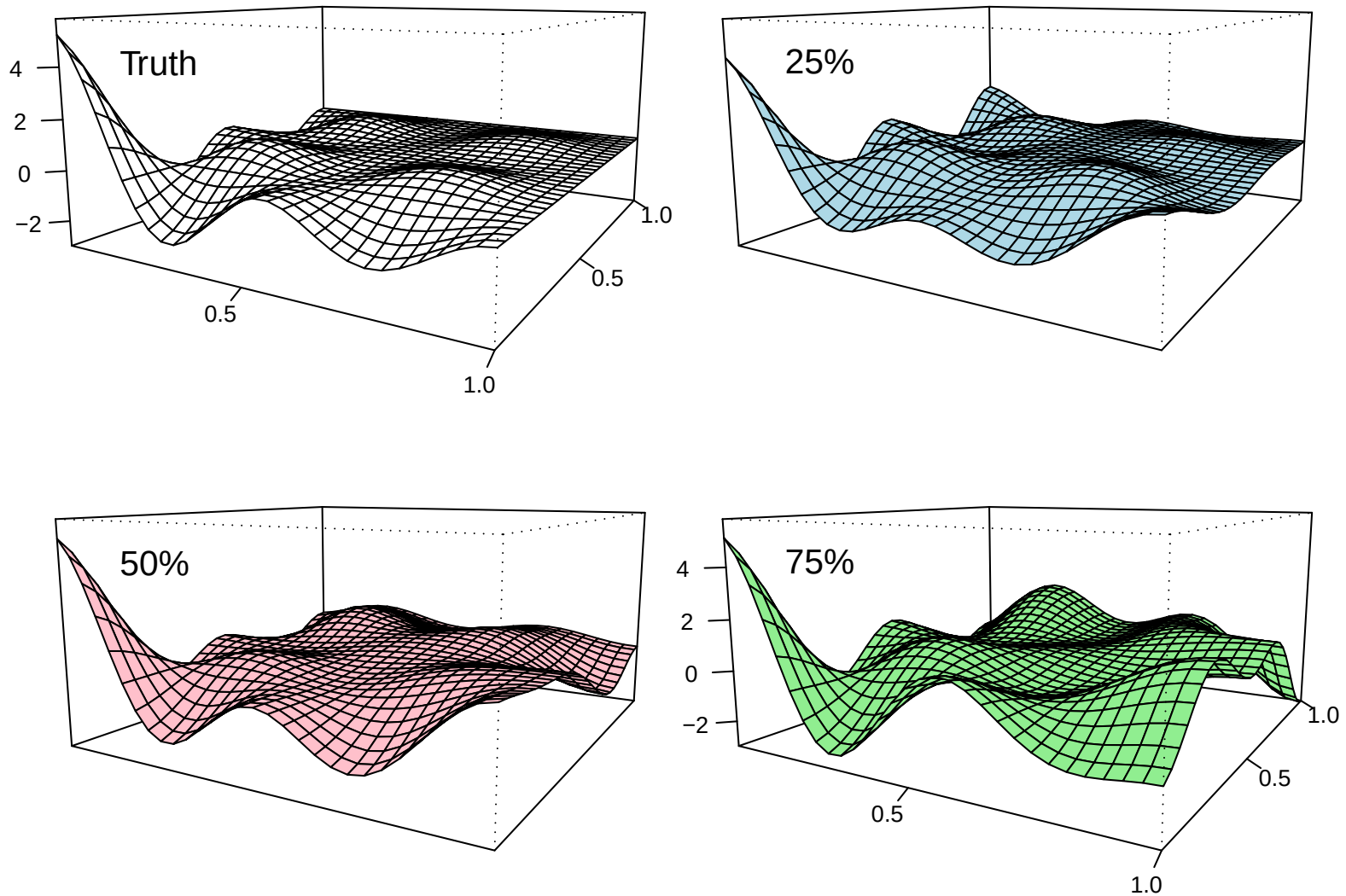
Non-Linear Simulation Results

Simulation	Statistic	VANISH	Spln$_{LS}$	Spln	SpAM$_{LS}$	SpAM	Oracle
	False-Pos Main	0.04	0	0	0.1	15.12	–
Basis	False-Neg Main	0.23	1.42	0.9	0.23	0	–
model	False-Pos Inter	0.51	1.14	9.69	–	–	–
$S_{int} = 2$	False-Neg Inter	0.06	0.21	0.06	–	–	–
	L2-sq	**0.38**	0.879	1.848	1.276	1.437	0.267
	False-Pos Main	0	0	0	0.08	15.48	–
Non Basis	False-Neg Main	0	0.65	0.33	0	0	–
model	False-Pos Inter	0.48	1.96	9.72	–	–	–
$S_{int} = 2$	False-Neg Inter	0	0.2	0.07	–	–	–
	L2-sq	**0.333**	1.023	2.100	1.217	1.405	0.277
	False-Pos Main	0	0	0	0.03	16.58	–
Non Basis	False-Neg Main	0	0.05	0.01	0	0	–
model	False-Pos Inter	0.06	0.72	11.37	–	–	–
$S_{int} = 0$	False-Neg Inter	0	0	0	–	–	–
	L2-sq	**0.116**	0.223	1.064	**0.114**	0.232	0.112
	False-Pos Main	0.05	0	0	0.22	7.47	–
Basis model	False-Neg Main	0.17	1.94	1.85	0.13	0.01	–
$p = 100$	False-Pos Inter	0.04	0.03	1.69	–	–	–
$n = 75$	False-Neg Inter	0.27	0.94	0.85	–	–	–
	L2-sq	**0.452**	2.221	1.793	0.771	0.882	0.217

True and Estimated Main Effects



True and Estimated Interaction Effects



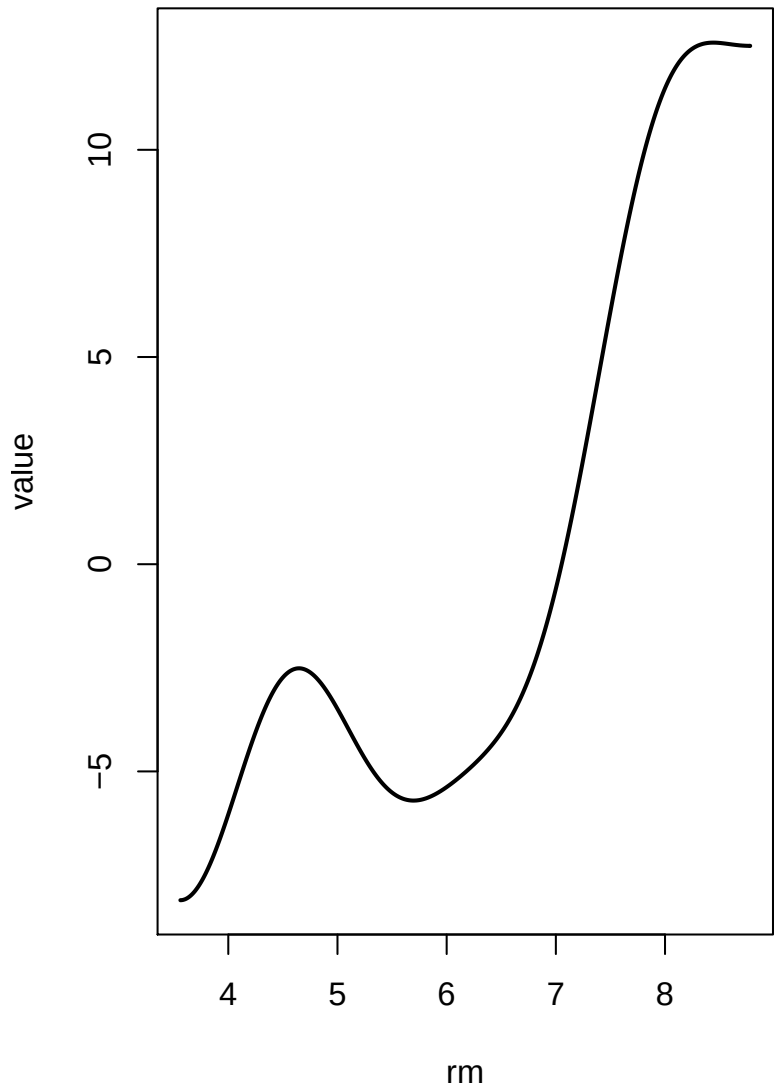
Application on Boston Housing Data

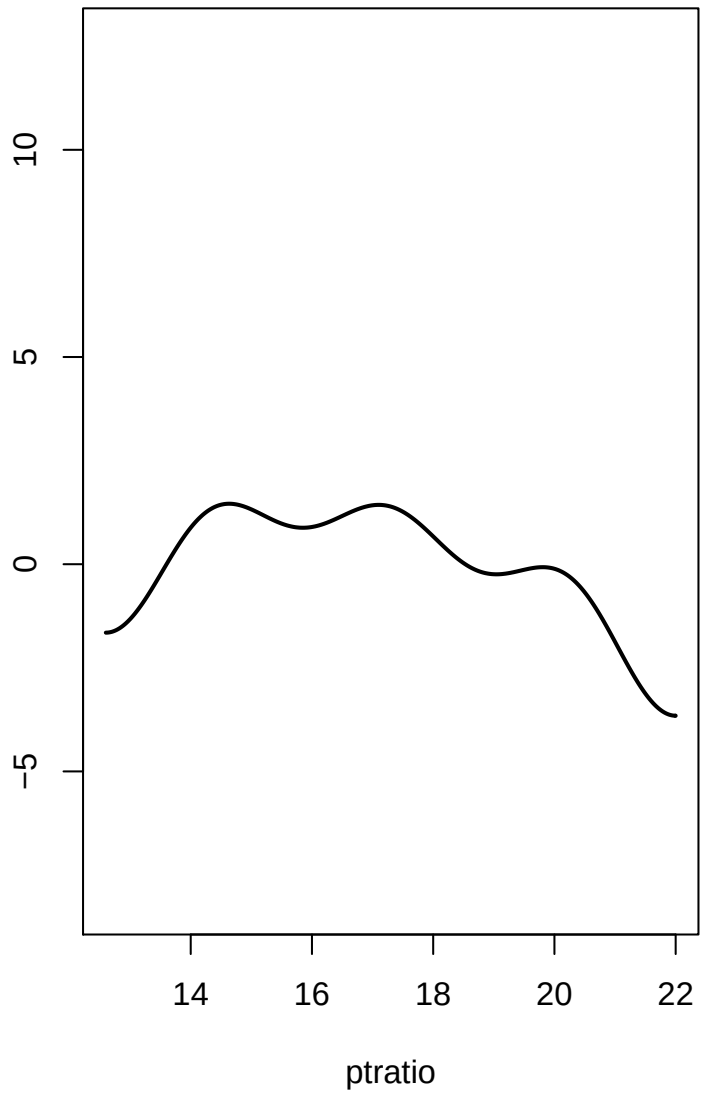
- There are 506 observations and 10 predictors, with the response corresponding to the median house value in each neighborhood.
- We add 30 noise variables, 20 drawn from a Uniform$(0, 1)$ distribution and the remainder generated by permuting the rows of the design matrix.
- Hence the data contained a total of 820 potential main effects and interactions of which 765, or 93%, corresponded to noise terms.
- We used ten-fold cross-validation to select the tuning parameters.
- We first randomly divided the data into a training set of 400 observations and used the remainder as a test set. We then fitted both VANISH and SpIn to the training data, using ten-fold cross-validation to select the tuning parameters.

Boston Housing Results

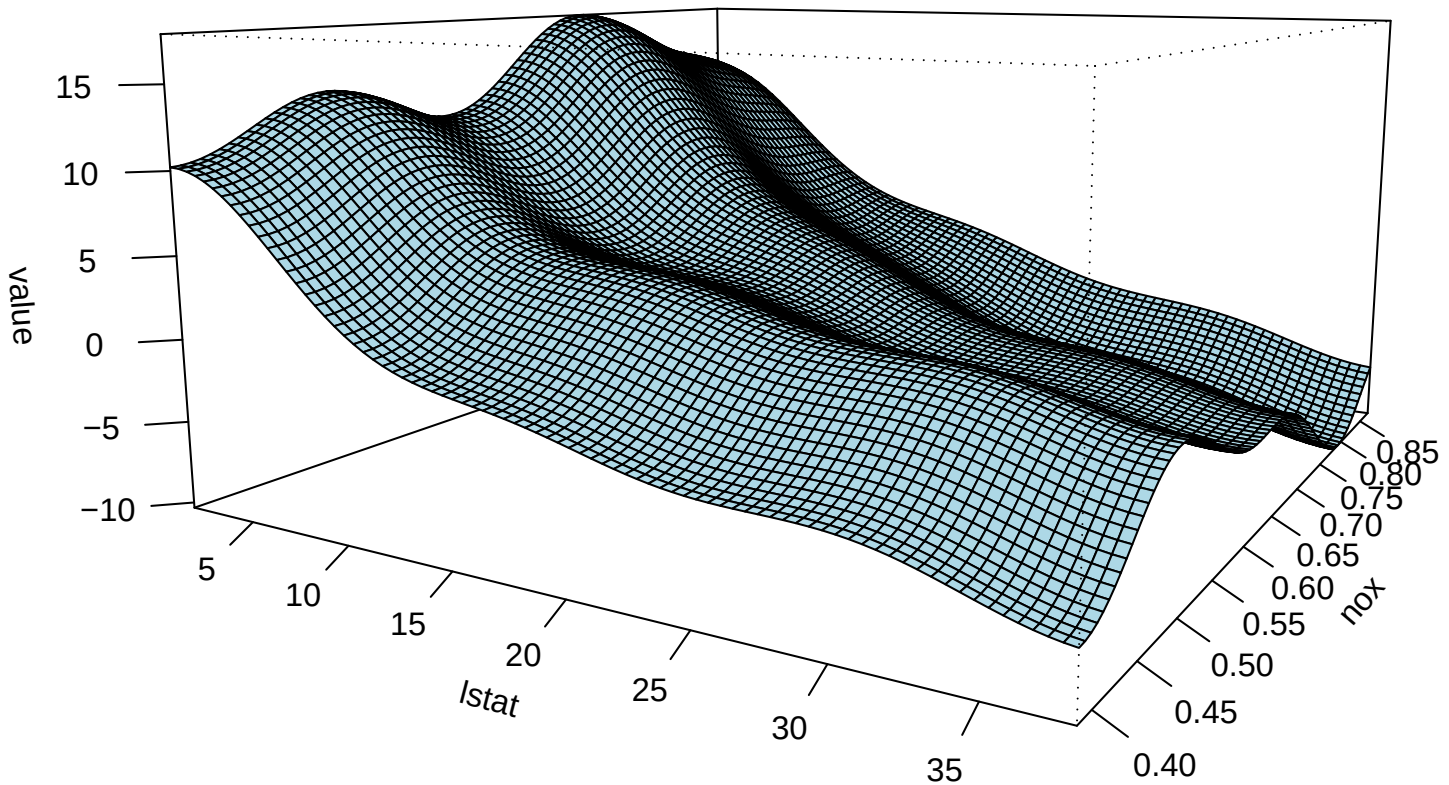
- VANISH correctly excluded the 765 noise terms and selected a model containing four main effects and one interaction term.
- The main effects corresponded to percentage of lower status of the population (lstat), the average number of rooms per dwelling (rm), pupil-teacher ratio by town (ptratio), and nitric oxides concentration in parts per 10 million (nox).
- The interaction term corresponded to lstat and nox.
- By comparison SpIn$_{LS}$ selected only the lstat variable, while the shrunk version of SpIn selected a large 27 variable model including 17 noise variables.
- The five variable VANISH model was superior to the one variable SpIn$_{LS}$ model on 99 of 100 random partitions of the data (MSE of 16.22 vs 29.15).

Main Effects





Interaction Term



Summary

- In order to make this problem feasible we model a sparse response surface in terms of the main effects and the interactions.

Summary

- In order to make this problem feasible we model a sparse response surface in terms of the main effects and the interactions.
- Simulations show that in practice it can produce significant improvements in performance over simpler alternatives.

Summary

- In order to make this problem feasible we model a sparse response surface in terms of the main effects and the interactions.
- Simulations show that in practice it can produce significant improvements in performance over simpler alternatives.
- Even when the true model is additive VANISH is competitive relative to the purely additive SpAM approach.

Summary

- In order to make this problem feasible we model a sparse response surface in terms of the main effects and the interactions.
- Simulations show that in practice it can produce significant improvements in performance over simpler alternatives.
- Even when the true model is additive VANISH is competitive relative to the purely additive SpAM approach.
- The VANISH fitting algorithm is very efficient, allowing it to search through thousands of non-linear two-dimensional surfaces.

Summary

- In order to make this problem feasible we model a sparse response surface in terms of the main effects and the interactions.
- Simulations show that in practice it can produce significant improvements in performance over simpler alternatives.
- Even when the true model is additive VANISH is competitive relative to the purely additive SpAM approach.
- The VANISH fitting algorithm is very efficient, allowing it to search through thousands of non-linear two-dimensional surfaces.
- Finally, VANISH is sparsistent (asymptotically selects the correct model), provided the signal and noise variables are not too highly correlated. The exact condition on the design matrix is similar to that for the Lasso and SpAM methods.