



Sparse hierarchical regression with polynomials

Dimitris Bertsimas¹ · Bart Van Parys¹

Received: 29 May 2017 / Revised: 9 September 2019 / Accepted: 3 January 2020 /

Published online: 24 January 2020

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2020

Abstract

We present a novel method for sparse polynomial regression. We are interested in that degree r polynomial which depends on at most k inputs, counting at most ℓ monomial terms, and minimizes the sum of the squares of its prediction errors. Such highly structured sparse regression was denoted by Bach (Advances in neural information processing systems, pp 105–112, 2009) as *sparse hierarchical regression* in the context of kernel learning. Hierarchical sparse specification aligns well with modern big data settings where many inputs are not relevant for prediction purposes and the functional complexity of the regressor needs to be controlled as to avoid overfitting. We propose an efficient two-step approach to this hierarchical sparse regression problem. First, we discard irrelevant inputs using an extremely fast input ranking heuristic. Secondly, we take advantage of modern cutting plane methods for integer optimization to solve the remaining reduced hierarchical (k, ℓ) -sparse problem exactly. The ability of our method to identify all k relevant inputs and all ℓ monomial terms is shown empirically to experience a phase transition. Crucially, the same transition also presents itself in our ability to reject all irrelevant features and monomials as well. In the regime where our method is statistically powerful, its computational complexity is interestingly on par with `LASSO` based heuristics. Hierarchical sparsity can retain the flexibility of general nonparametric methods such as nearest neighbors or regression trees (`CART`), without sacrificing much statistical power. The presented work hence fills a void in terms of a lack of powerful disciplined nonlinear sparse regression methods in high-dimensional settings. Our method is shown empirically to scale to regression problems with $n \approx 10,000$ observations for input dimension $p \approx 1000$.

Keywords Nonlinear regression · Sparse regression · Integer optimization · Polynomial learning

Editor: Tong Zhang.

Bart Van Parys is generously supported by the Early Postdoc. Mobility fellowship P2EZP2 165226 of the Swiss National Science Foundation.

✉ Dimitris Bertsimas
dbertsim@mit.edu

Bart Van Parys
vanparys@mit.edu

¹ Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

1 Introduction

We consider the problem of high-dimensional nonlinear regression. Given input $X = (x_1, \dots, x_n) \in \mathbb{R}^{n \times p}$ and response data $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$, we set out to find an unknown underlying nonlinear relationship

$$y_t = g(x_t) + e_t, \quad \forall t \in [n],$$

where $E := (e_1, \dots, e_n)$ in \mathbb{R}^n is the error term. Regression is a problem at the core of machine learning, statistics and signal processing. It is clear that if we are to carry any hope of success, some structure on the nature of the unknown nonlinear relationship g between input and response data must be assumed. Classical statistical learning theory (Vapnik 2013) indeed requires the complexity of the set of considered functions to be bounded in some way. We focus in this work on polynomial regression function primarily for two reasons. Polynomial regression, first, has a long history (Smith 1918) and is mentioned in almost any standard work on machine learning. We will consider in this paper all nonlinear relationships in the form of polynomials in p variables of total degree at most r . We denote with \mathcal{P} the set of polynomials of total degree r . Typically the polynomial which best explains data is defined as the minimizer to the abstract optimization problem

$$\min_{g \in \mathcal{P}} \frac{1}{2} \sum_{t \in [n]} \|y_t - g(x_t)\|^2 + \frac{1}{2\gamma} \|g\|^2, \quad (1)$$

over polynomial functions g in \mathcal{P} . The squared norm $\|g\|^2$ of a polynomial g is taken here to mean the sum of squares of its coefficients. The best polynomial in formulation (1) minimizes a weighted combination of the sum of its squared prediction errors and its coefficient vector. This latter Ridge regularization (Tikhonov 1943; Hoerl and Kennard 1970) term stabilizes its solution and helps reduce overfitting even further. An alternative interpretation of the regularization term as a precaution against errors in the input data matrix X has been given for instance in Bertsimas and Copenhaver (2018). Nevertheless, the value of the hyperparameter γ must in practice be estimated based on historical data using for instance cross validation.

By allowing for nonlinear dependence between input and response data, second, polynomial regression can discover far more complex relationships than standard linear regression. For a sufficiently large degree r in fact, any continuous functional dependence can be discovered up to arbitrary precision (Stone 1948). The previous observation leads to the fact that polynomial regression is a hybrid between parametric and nonparametric regression. Depending on the degree r of the polynomials considered, it falls between textbook parametric regression ($r = 1$) which assumes the functional dependence g between input and response data to be linear and completely nonparametric regression ($r \rightarrow \infty$) where nothing beyond continuity of the dependence g between input and response data is assumed. Although nonparametric approaches are very general and can unveil potentially any continuous relationship between input and response data, they nonetheless seriously lack in statistical power. Indeed, nonparametric methods such as kernel density estimation and nearest neighbors regression (Altman 1992) or decision trees (Breiman 2017) need a large number of samples in order to return statistically meaningful predictions. The amount of observations n needed scales unfavorably with the input dimension p which is commonly in modern data sets.

The polynomial regressor g in the regression problem (1) is a sum of at most $f := \binom{p+r}{r}$ monomial features. A seminal result due to Vapnik (1998) states that the

high-dimensionality of the unconstrained regression problem (1) does not pose an obstacle to its numerical solution. Indeed, the feature dimensionality f can be avoided in its entirety using the now classical kernel representation of polynomials put forward by Mercer (1909). Regression formulations amendable to such a kernel reformulation are typically referred to a kernel learning methods. It is thanks to both the flexibility of the regression formulation (1) and its computational tractability that polynomial and even more general kernel methods have experienced a lot of interest in the learning community (Suykens and Vandewalle 1999). Polynomial regression using kernel learning has indeed been used with success in many applications such as character recognition, speech analysis, image analysis, clinical diagnostics, person identification, machine diagnostics, and industrial process supervision. Today, efficient and mature software implementations of these so called polynomial kernel regressors are widely available, see c.f. Schölkopf and Smola (2002) and Pelckmans et al. (2002). Unfortunately in high-dimensional settings ($f \gg n$), the previously discussed curse of dimensionality and overfitting phenomena do pose a formidable obstacle to the recovery of the correct nonlinear relationship between input and response data. That is, in settings where we have many more monomial input features f than observations n , it becomes unlikely that we recover a statistically meaningful regressor by solving (1).

Here we will work to address the previous issue by providing a sparse counterpart to the polynomial regression problem (1). Sparse regression has recently been identified in the works of Tibshirani (1996), Hastie et al. (2015) and Candès et al. (2006) as an excellent antidote to the malignant phenomena of both dimensionality and overfitting. Interestingly, very few mature machine learning methods seem to have been developed which can deal reliably with sparse nonlinear regressors in a high-dimensional settings despite the obvious relevance of this problem class. One notable exception of direct relevance here is the SPORE algorithm by Huang et al. (2010) which uses an approach based on ℓ_1 -regularization. We subsequently describe a hierarchical sparse regression problem which controls both the dependence and functional complexity of the regression polynomials considered.

1.1 Hierarchical (k, ℓ) -sparsity

The popularity and effectiveness of sparse regression can from a practical perspective be explained by the following two observations. In the digital age obtaining and processing vast amounts of input data is increasingly less of a burden. Nevertheless, we expect only a small number k of all p recorded inputs to be meaningfully correlated to the response data Y . The trouble is that we can not tell the relevant features from the obfuscating bulk of data ahead of time. Sparsity hence firstly describes the limited functional dependence between input and response data. As only a finite amount of data can be recorded, one must avoid overfitting by limiting the complexity of the considered functional relationships. Of the potential f monomials making up the polynomial regressor g , the considered relationships should only depend on a select few of them. We use sparsity to enforce simplicity among the considered relationships. As both these described sparsity considerations are quite different in nature, we believe it is paramount not to conflate them.

We will say that the function $g \in \mathcal{P}_{k,\ell}$ is so called (k, ℓ) -sparse if it is the sum of ℓ monomials in at most k inputs. For instance the regressor $g(x) = x_1^2 + x_2x_3$ would be $(3, 2)$ sparse as it depends on the three inputs x_1, x_2 and x_3 , and is made up of two monomials x_1^2 and x_2x_3 . The resulting problem of hierarchical sparse regression can be cast as the regression problem

$$\min_{g \in \mathcal{P}_{k,\ell}} \frac{1}{2} \sum_{t \in [n]} \|y_t - g(x_t)\|^2 + \frac{1}{2\gamma} \|g\|^2. \quad (2)$$

The previous regression formulation is a structured sparsity constrained version of (1). Using this novel notion of hierarchical sparse regressors, we hope to keep the statistical power of parametric regression while simultaneously allowing highly nonlinear relationships between input and response data as well. Although structured hierarchical sparsity patterns were studied already by Zhao et al. (2009), they were never considered in our polynomial regression context directly. A related hierarchical kernel learning approach to a convex proxy of problem (2) is studied in Bach (2009) and from which we retained the word hierarchical to describe the structured sparsity of $\mathcal{P}_{k,\ell}$. The regression problem (2) carries the additional benefit of automatically yielding highly interpretable regressors with only a few nonlinear terms and input dependencies. By explicitly controlling both the dependence complexity k of used inputs as well as the functional complexity ℓ of the regression polynomials, the hierarchical sparse regression problem (2) promises to deliver nonlinear regressors with significant statistical power even in high-dimensional settings.

Unfortunately, solving the hierarchical sparse regression problem (2) can be challenging. Bringing to bear the power of modern integer optimization algorithms combined with smart heuristics, we will nevertheless show that many hierarchical sparse regression problems can nevertheless be dealt with.

1.2 Exact algorithms

The problem of sparse linear regression has been studied extensively in the literature. Despite being provably hard in the sense of complexity theory's NP hardness, in practice many successful algorithms are available. Historically, the first heuristic methods for sparse approximation seem to have arisen in the signal processing community (c.f. the work of Miller 2002; Mallat and Zhang 1993 and references therein) and typically are of an iterative thresholding type. More recently, one popular class of sparse regression heuristics solve the convex surrogate

$$\min_{\|g\|_1 \leq \lambda} \frac{1}{2} \sum_{t \in [n]} \|y_t - g(x_t)\|^2 + \frac{1}{2\gamma} \|g\|^2. \quad (3)$$

to the sparse regression formulation (2) where λ is a sparsity inducing hyperparameter. Here the norm $\|g\|_1$ of the polynomial g is meant to denote the sum of the absolute values of its coefficients. The convex proxy reformulation (3) is a direct adaptation of the seminal Lasso method of Tibshirani (1996) to the polynomial regression problem (1). The discussed SPORE algorithm by Huang et al. (2010) provides an implementation of this idea specific to the polynomial regression context considered here. Where the convex heuristic (3) does not incorporate the hierarchical sparsity structure of our exact formulation (2), more refined methods such as Group Lasso (Zhao et al. 2009; Bach 2008) could in principle do so by considering structured norm constraints. There is an elegant theory for convex proxy schemes promising large improvements over the more myopic iterative thresholding methods. Indeed, a truly impressive amount of high-quality work (Bühlmann and van de Geer 2011; Hastie et al. 2015; Wainwright 2009) has been written on characterizing when exact solutions can be recovered, albeit through making strong probabilistic assumptions on the data.

The problem of exact sparse nonlinear regression however seems, despite its importance, not to have been studied extensively. Although they are well studied separately, combining nonlinear and sparse regression never received much attention. Our recent work (Bertsimas and Van Parys 2017, and earlier in Bertsimas et al. 2016) has revealed that despite complexity results, exact sparse linear regression is not outside the realm of the possible even for very high-dimensional problems with a number of features f and samples n in the 100,000 s. Contrary to traditional complexity theory which suggests that the difficulty of a problem increases with size, the sparse regression problems seem to have the property that for a small number of samples n , exact regression is not easy to accomplish, but most importantly its solution does not recover the truth. However, for a large number of samples n , exact sparse regression can be done fast and perfectly separates the true features in the data from the obfuscating bulk. These results warrant also the possibility of nonlinear feature discovery in regression tasks.

1.3 Triage heuristic

Unfortunately, hierarchical (k, ℓ) -sparse regression problems quickly becomes problematic for all but small problem instances. The number of monomial features f is indeed combinatorial and hence scales quite badly in both the number of regressors p as well of the degree r of the considered polynomials. In order to provide a scalable algorithm to the problem of hierarchical kernel regression it is clear that the dimensionality of the problem needs to be reduced.

Our key insight in this paper will be to use polynomial kernel regression to rank the potential inputs. This heuristic method is very helpful in rejecting many irrelevant candidate inputs without missing out on the actual underlying nonlinear relationship between input and response data. Exact sparse hierarchical kernel regression described before will then be used to identify the relevant nonlinearities from among the most promising candidate inputs; see Fig. 1. We remark that the input triage step in our two step approach to the sparse hierarchical regression (2) renders the overall method potentially inexact. Indeed, if any relevant input is eliminated erroneously it can not be recovered later by the subsequent exact sparse hierarchical regression step. In this paper we nevertheless set out to show that this combination of smart heuristic and exact sparse regression goes a long way to solve hierarchical sparse regression problems (2) of practical size.

1.4 Contributions

In this paper, we first and foremost want to promote a notion of hierarchical (k, ℓ) -sparsity that rhymes well with the challenges of the big data era. Capturing limited functional dependence and complexity in big data problems is crucial to allow statistically

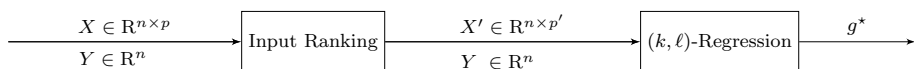


Fig. 1 Our two-step approach to (k, ℓ) -sparse regression for observations Y in \mathbb{R}^n and regressors X in $\mathbb{R}^{n \times p}$. In a first step we select the $p' \ll p$ most promising inputs out of the p candidates. The second step then performs exact (k, ℓ) -sparse regression on these most promising candidates. We set out to show that this combination of a smart ranking heuristics and exact sparse regression goes a long way to solve hierarchical sparse regression problems (2) of practical size

meaningful regression in high-dimensional and nonlinear settings. Hierarchical (k, ℓ) -sparse regression definition is in this regard a first step in the direction of lifting nonlinear regression into high-dimensional settings as well. In particular, we hope that the method presented here will show a more disciplined approach to nonlinear discovery than some more black box methods such as artificial neural networks.

Secondly, we also offer scalable algorithms able to address these hierarchical (k, ℓ) -sparse regression problems using modern optimization techniques. In accordance with previous results (Bertsimas and Van Parys 2017), we show that sparse regression is indeed tractable in practice even for very high-dimensional problems with a number of features f and samples n in the 100,000 s. We will indeed show that we can reliably discover nonlinear relationships using a combination of smart heuristics and exact sparse regression via a cutting plane approach for convex integer optimization.

In order to judge the quality of a proposed regressor g^* , we will measure on the one hand to what extent all the relevant monomial features are discovered. In order to do so, while at the same time avoiding notational clutter, we need to establish a way to refer to each of the monomials of degree r in p variables in an efficient fashion. Let $m_j : \mathbb{R}^p \rightarrow \mathbb{R}$ for each j in $[f]$ denote a distinct monomial in p inputs of degree at most r . We define the accuracy of a regressor g^* as

$$A\% := \frac{|\text{supp}(g^*) \cap \text{supp}(g_{\text{true}})|}{|\text{supp}(g_{\text{true}})|}, \quad (4)$$

where $\text{supp}(g)$ represents all j such that monomial m_j contributes to the polynomial g . This accuracy measure $A\%$ thus represents the proportion of true underlying monomial features discovered by the proposed polynomial regressor g^* . On the other hand, we can use

$$F\% := \frac{|\text{supp}(g^*) \setminus \text{supp}(g_{\text{true}})|}{|\text{supp}(g^*)|} \quad (5)$$

to quantify how many irrelevant features were wrongly included in the process. Perfect recovery occurs when the method gives the whole truth ($A\% = 100$) and nothing but the truth ($F\% = 0$). In practice, however, machine learning methods must inevitable make a choice between both desirables.

We intend to illustrate that exact sparse regression methods have an inherent edge over proxy based sparse heuristics. Proxy based methods such as `Lasso` do indeed have several well documented shortcomings. First and foremost, as argued by Bertsimas et al. (2016) they do not recover very well the sparsity pattern. Furthermore, the `Lasso` leads to biased regression regressors, since the ℓ_1 -norm penalizes both large and small coefficients uniformly. The ability of our method to identify all relevant features is shown empirically to experience a phase transition. There exists a critical number of data samples n_0 such that when presented sufficient data $n > n_0$ our method recovers the ground truth ($A\% \approx 100$) completely, whereas otherwise its accuracy $A\%$ tends to zero. Crucially, the same number of samples n_0 also enables our method to reject most irrelevant features ($F\% \approx 0$) as well. We thus show that we significantly outperform `Lasso` in terms of offering regressors with a larger number of relevant features (bigger $A\%$) for far fewer nonzero coefficients (smaller $F\%$) enjoying at the same time a marginally better prediction performance. In the regime $n > n_0$ where our method is statistically powerful ($A\% \approx 100$, $F\% \approx 0$), its computational complexity is furthermore on par with sparse

heuristics such as `LASSO`. This last observation takes away the main propelling justification for most heuristic based sparsity approaches.

Finally, we illustrate that sparse hierarchical regression its predictive performance on real data sets is on par with more established nonparametric methods such as nearest neighbors or `CART`, while offering interpretable models. We hence argue that hierarchical sparse regression is a competitive nonlinear regression approach in particular when faced with many modern high-dimensional regression problems.

1.5 Notation

The knapsack set S_k^p denotes here the binary set $S_k^p := \{s \in \{0, 1\}^p : \sum_{j \in [p]} s_j \leq k\}$, which contains all binary vectors s selecting k components out of p possibilities. Assume that (y_1, \dots, y_p) is a collection of elements and suppose that s is an element of S_k^p , then $y_s \in \mathbb{R}^{|s|}$ denotes the sub-collection of y_j where $s_j = 1$. Similarly, we use $\text{supp}(x) = \{s \in \{0, 1\}^p : s_j = 1 \iff x_j \neq 0\}$ to denote those indices of a vector x which are nonzero. We denote by S_+^n (S_{++}^n) the cone of $n \times n$ positive semidefinite (definite) matrices. Given a matrix $K \in \mathbb{R}^{n \times n}$, we denote its element-wise r th power or Hadamard power as K^{or} , i.e., we have that

$$K^{or} := \begin{pmatrix} K_{11}^r & K_{12}^r & \dots & K_{1n}^r \\ K_{21}^r & K_{22}^r & \dots & K_{2n}^r \\ \vdots & \vdots & \ddots & \vdots \\ K_{n1}^r & K_{n2}^r & \dots & K_{nn}^r \end{pmatrix}.$$

2 Hierarchical (k, ℓ) -sparse polynomial regression

In this section, we discuss two formulations of the hierarchical sparse regression problem (2) through a standard integer optimization lens. In the last twenty plus years the computational power of integer optimization solves has increased at a dramatic speed (Bertsimas et al. 2016). Where 20 years ago integer optimization for statistics was branded impossible, recent work (Bertsimas and Van Parys 2017; Bertsimas et al. 2016) has shown convincingly that this position needs to be revisited. The position that exact sparse regression is an unattainable goal only to be striven for via admittedly elegant convex heuristics such as `LASSO` should not be held any longer.

We considered in the sparse regression problem (2) as features the set of all monomials of degree at most r . It is clear that this sparse hierarchical regression problem over polynomials can equivalently stated as an optimization problem over their coefficients in the monomial basis m_j for j in $[f]$. We will avoid the need to explicitly order each these monomials as follows. Define the dependents $D(i)$ of any data input i as the set of indices j such that the monomial m_j depends on input i . Similarly, we define the ancestors $A(j)$ of any j as the multiset of inputs making up the monomial m_j . Instead of using the classical monomial basis, we consider a scaled variant in which we take $m_j(\mathbb{1}_p)$ to coincide with the square root of the number of distinct ways to order the multiset $A(j)$. This rescaling of the monomials comes in very handy when discussing the solution of the regression problem (1) in Sect. 3. In fact, the same scaling is implicitly made by the Mercer kernel based approach as well.

To make the discussion more concrete, we consider a data problem with $p = 3$ inputs and associated data $(y_t, x_{t,1}, x_{2,t}, x_{3,t})$ for $t \in [n]$. We consider all monomials on the three

inputs of degree at most two, i.e., we consider the monomials and their corresponding indices as given below.

Monomial m_j	1	$\sqrt{2}x_1$	$\sqrt{2}x_2$	$\sqrt{2}x_3$	x_1^2	$\sqrt{2}x_1x_2$	$\sqrt{2}x_1x_3$	x_2^2	$\sqrt{2}x_2x_3$	x_3^3
Index j	1	2	3	4	5	6	7	8	9	10

The set $D(i)$ corresponds to the set of indices of the monomials in which input x_i participates. In our example above we have $D(1) = \{2, 5, 6, 7\}$ corresponding to the monomials $\{\sqrt{2}x_1, x_1^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3\}$. The set $A(j)$ corresponds to all inputs that are involved in the monomial with index j . Again as an illustration, in our example $A(7) = \{1, 3\}$ corresponding to the inputs x_1 and x_3 . The fact that $A(7)$ has two distinct permutations causes the scaling $\sqrt{2}$, as opposed to $A(5) = \{1, 1\}$ which only has one distinct permutation and hence is not scaled.

2.1 Mixed integer formulation

With the help of the previous definitions, we can now cast the hierarchical (k, ℓ) -sparse regression problem (2) as a standard optimization problem. The problem of hierarchical sparse regression can indeed be cast as the following (MIO) problem

$$\begin{aligned}
 \min & \frac{1}{2} \sum_{i \in [n]} \|y_i - \sum_{j \in [f]} w_j \cdot m_j(x_i)\|^2 + \frac{1}{2\gamma} \|w\|^2 \\
 \text{s.t.} & w \in \mathbb{R}^f, s \in \mathbb{S}_\ell^f, h \in \mathbb{S}_k^p, \\
 & s_j \leq h_i \quad \forall i \in A(j), \forall j \in [f], \\
 & -\mathcal{M}s_j \leq w_j \leq \mathcal{M}s_j \quad \forall j \in [f],
 \end{aligned} \tag{6}$$

using a big- \mathcal{M} formulation. Its optimal solution w^* gives the coefficients of the polynomial $g^*(x) = \sum_{j \in [f]} w_j^* m_j(x)$ in (2) best describing the relationship between input and observations in our monomial basis $\{m_j\}$. The coefficient w_j of any monomial m_j is only then nonzero when $s_j = 1$ as per the first constraint in (6). The constant \mathcal{M} needs to be chosen sufficiently large for the given data as to make none of the ultimate constraints in (6) binding. Although nontrivial, this can be done using the strategy found in Bertsimas et al. (2016). Mixed integer formulations employing the previously outlined strategy are commonly denoted as Big- \mathcal{M} formulations. In any case, the binary variable $s \in \mathbb{S}_\ell^f$ represents the sparsity pattern in the monomials, i.e., which ℓ monomials are used out of f potential candidates. The first constraint of formulation (6) encodes the hierarchical nature of our (k, ℓ) -sparsity requirement. Only those monomials m_j such that the input $h_i = 1$ is selected for all its ancestors $i \in A(j)$ are considered as potential regressors. In all, the binary constraint

$$(s, h) \in \mathbb{S}_{\ell, k}^{f, p} := \{s \in \mathbb{S}_\ell^f, h \in \mathbb{S}_k^p : s_j \leq h_i, \forall i \in A(j), \forall j \in [f]\}$$

hence represents the postulated hierarchical (k, ℓ) -sparsity pattern.

To use the example discussed before in which we have three inputs and monomials of order two, the monomial $m_7(x) = \sqrt{2}x_1x_3$ can only be included as a regressor if the variable $s_7 = 1$. The variable s_7 can only then be nonzero if both inputs x_1 and x_2 are selected which requires that the variables $h_1 = 1$ and $h_3 = 1$. The resulting optimal

regressor polynomial $h^*(x) = \sum_{j \in [f]} w_j^* \cdot m_j(x)$ thus counts at most ℓ monomials depending on at most k regressor inputs.

Although the direct formulation (6) of the hierarchical (k, ℓ) -sparse regression problem results in a well posed MIO problem, the constant \mathcal{M} needs to be chosen with extreme care as not to impede its numerical solution. The choice of this data dependent constant \mathcal{M} indeed affects the strength of the MIO formulation (6) and is critical for obtaining solutions quickly in practice. Consequently, solving the Big- \mathcal{M} formulation directly is only been practical for problems of dimension $n \approx 1000$ and $f \approx 100$ (Bertsimas et al. 2016) far smaller than the problems we hope to address. Furthermore, as the regression dimension p grows, explicitly constructing the MIO problem (6), let alone solving it, becomes burdensome. In order to develop an exact scalable method a different perspective on sparse regression is needed. We will employ a perspective found in our previous work (Bertsimas and Van Parys 2018) albeit applied now instead to the nonlinear sparse regression problem (2). In the subsequent section we develop an alternative exact formulation which avoids using a big- \mathcal{M} formulation while at the same time will prove to be more amenable to a scale algorithms.

2.2 Convex integer formulation

We next establish that the sparse regression problem (6) can in fact be represented as a pure binary optimization problem. By doing so we will eliminate the dependence of our formulation on the data dependent constant \mathcal{M} . We will need the help of the following supporting lemma regarding linear regression.

Lemma 1 (The regression loss function c) *The least-squares regression cost $c(ZZ^T) := \min_w \frac{1}{2} \|Y - Zw\|^2 + \frac{1}{2\gamma} \|w\|^2$ admits the following explicit characterization*

$$c(ZZ^T) = \frac{1}{2} Y^T (\mathbb{I}_n + \gamma ZZ^T)^{-1} Y. \tag{7}$$

Proof As the regression problem over w in \mathbb{R}^p is an unconstrained (QO) problem, the optimal value w^* satisfies the linear relationship $(\mathbb{I}_p/\gamma + Z^T Z)w^* = Z^T Y$. Substituting the expression for the optimal linear regressor w^* back into optimization problem, we arrive at $c(ZZ^T) = 1/2 Y^T Y - 1/2 Y^T Z(\mathbb{I}_p/\gamma + Z^T Z)^{-1} Z^T Y$. The final characterization can be derived from the previous result with the help of the matrix inversion lemma found stating the identity $(\mathbb{I}_n + \gamma ZZ^T)^{-1} = \mathbb{I}_n - Z(\mathbb{I}_p/\gamma + Z^T Z)^{-1} Z^T$. \square

Lemma 1 will enable us to eliminate the continuous variable w out of the MIO sparse regression formulation (6). The proof of the next Theorem is inspired by a similar but distinct result from Bertsimas and Van Parys (2017). The following result provides a different pure integer approach to hierarchical sparse regression. It will form the basis to our attempts to solve hierarchical regression problems exactly.

Theorem 1 (Hierarchical (k, ℓ) -sparse regression) *The hierarchical (k, ℓ) -sparse regression problem (2) can be reformulated as the pure (CIO) problem*

$$\begin{aligned}
 & \min \frac{1}{2} Y^\top \left(\mathbb{I}_n + \gamma \sum_{j \in [f]} s_j K_j \right)^{-1} Y \\
 & \text{s.t. } s \in S_\rho^f, h \in S_k^p, \\
 & \quad s_j \leq h_i \quad \forall i \in A(j), \forall j \in [f],
 \end{aligned} \tag{8}$$

where the micro kernel matrices K_j in S_+^n are defined as the dyadic outer products $K_j := m_j(X) \cdot m_j(X)^\top$.

Proof We start the proof by separating the optimization variable w in the sparse regression problem (6) into its support $s := \text{supp } w$ and the corresponding non-negative entries w_s . Evidently, we can now write the sparse regression problem (6) as the bilevel minimization problem

$$\min_{(s,h) \in S_{k,\ell}^{p,r}} \left[\min_{w \in \mathbb{R}^k} \frac{1}{2\gamma} \|w\|^2 + \frac{1}{2} \sum_{t \in [n]} \|y_t - \sum_{\{j \in [n] : s_j=1\}} w_j \cdot m_j(x_t)\|^2 \right]. \tag{9}$$

It now remains to be shown that the inner minimum can be found explicitly as the objective function of the optimization problem (8). Using Lemma 1, the minimization problem can be reduced to the minimization problem $\min\{c(m_s(X) \cdot m_s(X)^\top) : (s, h) \in S_{k,\ell}^{p,r}\}$. We finally remark that the outer product can be decomposed as the sum $m_s(X) \cdot m_s(X)^\top = \sum_{j \in [p]} s_j \cdot m_j(X) \cdot m_j(X)^\top$, thereby completing the proof. \square

For all but small problems, exact hierarchical sparse regression nevertheless quickly becomes problematic. The effective number of regression features f which determines the dimension of our formulation (8) is indeed combinatorial in the number of inputs p and degree r of the considered polynomials. Our key insight is to triage the inputs first heuristically using an efficient input ranking method described in the subsequent section. Later in Sect. 5 we will show that this two-step procedure outlined in Fig. 1 goes a long way to solve practical hierarchical sparse regression problems.

3 Polynomial kernel input ranking

Input selection is an important and well known technique for high-dimensional data reduction. Most commonly this is done by ranking the inputs based on some performance criteria. We point to Fan and Lv (2010) for a recent review on the topic. For high-dimensional data problems surely independence screening (SIS, Fan and Lv 2008) and Lasso (Tibshirani 1996) are the most popular methods. Both these methods implicitly assume a linear regression model. Indeed, neither of these feature-wise methods takes into account nonlinear feature interactions. A large body of work, c.f. Kong et al. (2017), Hall and Xue (2014), Hao and Zhang (2014) and further references therein, tries to consider feature interactions effects in the input selection method as well. We do remark though that these methods are generally limited to finding interactions only between at most two distinct variables at once. Translated to our polynomial regression setting this means its applicability is limited to the case of quadratic ($r = 2$) polynomials.

The objective in this section is hence to present an efficient method which can address the high-dimensional nature of sparse hierarchical regression by ignoring irrelevant regression inputs and working with promising candidates only while considering

potential interaction effects between at most r variables. The reader might wonder at this point whether any such attempt would not entail the solution of the original hierarchical sparse regression problem. It should indeed be clear though that any input triage heuristic renders our overall two step procedure (see Fig. 1) potentially inexact. Here, however, we do not claim to triage the inputs optimally, but rather aim for a simple approximate yet fast method. In practice however, we will observe that this potential loss of optimality often comes with only minor consequences. We will attempt to triage inputs efficiently by leveraging the fact that polynomial regression problem without sparse constraints can be solved optimally.

A seminal result due to Vapnik (1998) states that the feature dimensionality f of the unconstrained regression problem (1) surprisingly does not play any role in its numerical solution. Indeed, the feature dimensionality f can be done away with in its entirety using the now classical Mercer (1909) kernel representation theorem. We can state the polynomial regression problem (1) as an optimization problem in terms of coefficients in the monomial basis

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{t \in [n]} \|y_t - \sum_{j \in [f]} w_j \cdot m_j(x_t)\|^2 + \frac{1}{2\gamma} \|w\|^2 \\ \text{s.t.} \quad & w \in \mathbb{R}^f. \end{aligned} \quad (10)$$

We state the Mercer kernel representation in Theorem 2 for the sake of completeness regarding the dual of the regression problem (10). It should be noted that surprisingly the dimension f does not play a role but instead the number of samples n is of importance. This previous observation is what has propelled kernel learning algorithms as viable nonlinear regression methods (Schölkopf and Smola 2002).

Theorem 2 (Mercer Kernel Representation (Vapnik 1998)) *The polynomial regression problem (10) can equivalently be formulated as the unconstrained maximization problem*

$$\begin{aligned} c(K) = \max \quad & -\frac{\gamma}{2} \alpha^\top K \alpha - \frac{1}{2} \alpha^\top \alpha + Y^\top \alpha \\ \text{s.t.} \quad & \alpha \in \mathbb{R}^n, \end{aligned} \quad (11)$$

where the positive semidefinite kernel matrix $K := m(X) \cdot m(X)^\top$ allows for an efficient characterization as the Hadamard power $K = (XX^\top + \mathbb{1}_{n \times n})^{or}$.

Theorem 2 uses the Mercer kernel representation which establishes that the outer product $m(X) \cdot m(X)^\top$ can be characterized as the element-wise Hadamard power $(XX^\top + \mathbb{1}_{n \times n})^{or}$ for our specific polynomial bases. Indeed, for any t and t' in $[n]$ we have

$$\begin{aligned} K(t, t') &:= m(x_t)^\top \cdot m(x_{t'}) \\ &= [\sqrt{\|A(1)\|} \cdot m_1(x_t), \dots, \sqrt{\|A(f)\|} \cdot m_f(x_t)]^\top \cdot [\sqrt{\|A(1)\|} \cdot m_1(x_{t'}), \dots, \sqrt{\|A(f)\|} \cdot m_f(x_{t'})] \\ &= \sum_{j \in [f]} \|A(j)\| m_j(x_t) \cdot m_j(x_{t'}) \\ &= \sum_{j \in [f]} \|A(j)\| m_j([x_{t,1} \cdot x_{t',1}, \dots, x_{t,p} \cdot x_{t',p}]) \\ &= (1 + \sum_{i \in [p]} x_{t,i} \cdot x_{t',i})^r \\ &= (1 + x_t^\top x_{t'})^r \end{aligned}$$

where $\|A(j)\|$ denotes here the number of distinct ways to order the multiset $A(j)$. The penultimate equality is recognized as the binomial expansion theorem. Note that for the Mercer kernel representation to hold, the monomial basis had indeed to be properly normalized

using $\sqrt{\|A(j)\|}$ as explained in the beginning of Sect. 2. This well known but crucial observation seems to have been made first by Poggio (1975).

The size of the kernelized regression problem (11) scales only with the number of data points n rather than the feature dimension f . It could be remarked that as the kernelized regression problem is unconstrained it admits a closed form solution in the form of the linear system $(\mathbb{1}_n + \gamma K)\alpha^* = Y$ which can be solved requiring $\mathcal{O}(n^3)$ computation. Indeed, the matrix K is a full matrix without any obvious sparsity pattern to exploit. The optimal regression coefficients w^* in formulation (10) are linearly related to the optimal dual variable α^* in formulation (11) via the complementarity conditions which here read

$$w_j^* = \gamma \cdot m_j(X)^\top \alpha^* \quad \forall j \in [f]. \tag{12}$$

Although this last relationship is linear, computing the coefficients in the monomial basis might still prove a daunting task merely because of the sheer number of them. In the following, we show that we can nevertheless compute the Euclidean norm of the optimal coefficients w_j^* in front of the monomials depending on a certain input i efficiently.

Proposition 1 *The Euclidean norm of the coefficients w_j^* in front of all monomials m_j depending on input i is related to the dual optimal variable α^* in (11) as*

$$\|w_{D(i)}^*\|^2 = \sum_{j \in D(i)} \|w_j^*\|^2 = \gamma^2 \cdot \alpha^{*\top} K_i \alpha^*, \tag{13}$$

where the kernel matrix K_i can be characterized explicitly as $K - (XX^\top - X_i X_i^\top + \mathbb{1}_{n \times n})^{or}$.

Proof From the linear relationship (12) between the optimal coefficients w^* and dual variable α it follows immediately that $\|w_j^*\|^2 = \gamma^2 \cdot \alpha^{*\top} (\sum_{j \in D(i)} K_j) \alpha^*$. Through simple expansion it is quite easy to see that $(XX^\top - X_i X_i^\top + \mathbb{1}_{n \times n})^{or}$ coincides exactly with $\sum_{j \notin D(i)} K_j$. Hence, the kernel matrix $K_i := \sum_{j \in D(i)} K_j$ is found as its complement $K - (XX^\top - X_i X_i^\top + \mathbb{1}_{n \times n})^{or}$. \square

Hence, despite the fact that the size of the optimal coefficients w_j in front of the monomials depending on a certain input i consists of the sum of squares of as many as $\binom{p+r-1}{r-1}$ components, it can nevertheless be computed without much effort. Unfortunately, the optimal regressors coefficients w^* in (10) are not expected to be sparse. Nevertheless, the optimal regressors coefficients can be used to provide a heuristic ranking of the importance of the p data inputs. The Euclidean norm of the coefficients of the monomials which depend on input i can indeed be used as a proxy for the relevance of the input of interest. Fortunately, the quantities (13) are very efficient to compute once the optimal dual variable α^* has been found. Indeed, each quantity can be computed in $\mathcal{O}(n^2)$ time independent of the dimension f of the polynomials considered.

The complete computation is given in Algorithm 1. Though not exact, it gives a good indication of the significance of each of the p inputs. In fact it is very closely related the backward elimination wrapper methods discussed in Guyon and Elisseeff (2003).

Algorithm 1: Input Ranking

input : $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ and $r \in \mathbb{N}$
output: $r \in \mathbb{R}^p$
 $K \leftarrow (XX^\top + \mathbb{1}_{n \times n})$
 $\alpha^* = (\mathbb{0}_n + \gamma K^{or})^{-1}Y$
for i **in** $[p]$ **do**
 $K_i \leftarrow K^{or} - (K - X_i X_i^\top)^{or}$
 $r_i \leftarrow \gamma^2 \cdot \alpha^{*\top} K_i \alpha^*$

An alternative way of looking at our input ranking algorithm is through the subgradients of the convex regression loss function c defined in the dual problem (11). We note that the subgradient of the function c can be computed explicitly using its dual characterization given in (11) as well.

Proposition 2 (Derivatives of the optimal regression loss function) *We have that the subgradient of the regression loss function c as a function of the kernel K can be stated as*

$$\nabla c = -\frac{\gamma}{2} \cdot \alpha^{*\top} \nabla K \alpha^*,$$

where α^* maximizes (11).

Proof From the dual definition of the regression loss function c it is clear that we have the inequality

$$c(\bar{K}) \geq -\frac{\gamma}{2} \cdot \alpha^{*\top} \bar{K} \alpha^* - \frac{1}{2} \alpha^{*\top} \alpha^* + Y^\top \alpha^* \tag{14}$$

for all \bar{K} . From the very definition of α^* as the maximizer of (11) it follows that the previous inequality becomes tight for $\bar{K} = K$. This proves that the left hand side of (14) is a subgradient to the regression loss function c at the point K . □

When comparing the previous proposition with the result in Theorem 1, it should be noted that the derivatives of c at K agree up to the constant -2γ with the sum of squares of the coefficients w^* of the optimal polynomial. In essence thus, our Algorithm 1 ranks the inputs according to the linearized loss of regression performance ∇c , caused by ignoring the inputs using the polynomial regression (1). The quantity r_i characterizes up to first-order the loss in predictive power when not using the input i . The bigger this caused loss, the more importance is assigned to including input i as a regressor.

Before we can move on to the next section, the computational efficiency of our input ranking heuristic needs to be addressed first. The dominant cost in computing the input ranking r is clearly finding the optimal dual maximizer α^* which requires $\mathcal{O}(n^3)$ computation. This computational cost becomes a practical issue for problems counting much more than a few thousand observations. To speed up computations, in practice we split such very large data sets in smaller chunks of size approximately $n = 2000$ and compute for each of those chunks a distinct ranking r using Algorithm 1. We then take as overall approximate ranking the average of the rankings obtained on each of the distinct data chunks. As was noted in the beginning of the section, the input ranking method presented in Algorithm 1 does not aspire to find all k relevant inputs exactly. Rather, it is only meant to eliminate

the most unpromising inputs and keep p' high potential candidates as illustrated in Fig. 1. Hence, although it would be better not separate the data into chunks at all in terms of ranking performance, it is very beneficial for the main reason of using the input heuristic; speed. Among those inputs which are deemed promising we will then solve the hierarchical (k, ℓ) sparse regression problem (2) exactly as explained in the subsequent section.

4 A cutting plane algorithm for hierarchical sparse regression

We point out again that our pure integer formulation (8) of the hierarchical (k, ℓ) -sparse regression problem (2) circumvents the introduction of a big- \mathcal{M} constant which is simultaneously hard to estimate and crucial for its numerical efficacy. Solving the Big- \mathcal{M} formulation directly is consequently only practical for small problems of dimension $n \approx 1000$ and $f \approx 100$ (Bertsimas et al. 2016). That being said, explicitly constructing the optimization problem (8) results in the integer semidefinite optimization problem

$$\min_{(s,h) \in S_{\ell,k}^{f,p}} c\left(\sum_{j \in [f]} s_j K_j\right)$$

which might prove daunting as well. The regression loss function c is indeed a semidefinite representable function (Nesterov and Nemirovskii 1994) to be optimized over the discrete set $S_{\ell,k}^{f,p}$. Without even taking into account the discrete nature of the optimization problem (8), solving a semidefinite optimization (SDO) of size the number of samples might even prove in a convex case impractical for medium size problems with $n \approx 1000$. In order to solve our CIO formulation (8), we take here an alternative route using the outer approximation approach introduced by Duran and Grossmann (1986) and given explicitly in Algorithm 2. As this is a standard algorithm applied, albeit applied to our novel pure integer formulation (8), we try to keep its discussion high-level.

Algorithm 2: The outer approximation process

input : $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$ and $k \in [1, p]$
output: $s^* \in S_k^p$ and $w^* \in \mathbb{R}^p$
 $s_1 \leftarrow$ warm start
 $\eta_1 \leftarrow 0$
 $a \leftarrow 1$
while $\eta_a < c(s_a)$ **do**
 $s_{a+1}, \eta_{a+1} \leftarrow \arg \min_{s, \eta} \{ \eta \in \mathbb{R}_+ \text{ s.t. } (s, h) \in S_{\ell,k}^{f,p}, \eta \geq c(s^u) + \nabla c(s^u)(s - s^u), \forall u \in [a] \}$
 $a \leftarrow a + 1$
 $s^* \leftarrow s^a$
 $w^* \leftarrow 0, \quad w_{s^*}^* \leftarrow (\mathbb{I}_p / \gamma + X_{s^*}^\top X_{s^*})^{-1} X_{s^*}^\top Y$

Theorem 3 (Exact Sparse Regression (Fletcher and Leyffer 1994)) *Algorithm 2 returns the exact sparse solution w^* of the hierarchical (k, ℓ) -sparse regression problem (6) in finite time.*

Despite the previous encouraging theorem, it nevertheless remains the case that from a theoretical point of view we may need to compute exponentially many cutting planes in the worst-case, thus potentially rendering our approach impractical. Indeed, in the worst-case

Algorithm 2 considers all integer point in $S_{\ell,k}^{f,p}$ forcing us to minimize the function so constructed

$$\bar{c}(s) := \max_{(\bar{s}, \bar{h}) \in S_{\ell,k}^{f,p}} c(\bar{s}) + \nabla c(\bar{s})(s - \bar{s})$$

over the hierarchical binary constraint set $S_{\ell,k}^{f,p}$. As the number of integer points in the constraint set is potentially extremely large, the previous full explicit construction should evidently be avoided. This complexity behavior is however to be expected as exact sparse regression is known to be an NP-hard problem. In practice usually very few cutting planes need to be considered making the outer approximation method an efficient approach.

In general, outer approximation methods such as Algorithm 2 are known as *multi-tree* methods because every time a cutting plane is added, a slightly different integer optimization problem is to be solved anew by constructing a branch-and-bound tree. Over the course of the iterative cutting plane algorithm, a naive implementation would require that multiple branch and bound trees are built in order to solve the successive integer optimization problems. We employed a *single tree* implementation, instead of the iteration algorithm 2 directly, by using dynamic constraint generation (Barnhart et al. 1998). Such single tree implementations save the rework of rebuilding a new branch-and-bound tree every time a new binary solution is found in Algorithm 2.

Bertsimas and Van Parys (2017) provide a closely related algorithm intended to solve sparse linear regression problem up to dimensions f and n in the order of 100, 000s based on a cutting plane formulation for integer optimization. Contrary to traditional complexity theory which suggests that the difficulty of sparse regression problem increases with size, there seems to exist a critical number of observations n_0 such that the sparse regression problems has the property that for a small number of samples $n < n_0$, an exact regressor is not easy to obtain, but most importantly its solution does not recover the truth ($A\% \approx 0$ and $F\% \approx 100$). For a large number of samples $n > n_0$ however, exact sparse regression can be done extremely fast and perfectly separates ($A\% \approx 100$ and $F\% \approx 0$) the true monomial features from the obfuscating bulk. We will show that similar behavior is observed in our nonlinear sparse regression setting as well.

5 Numerical results

To evaluate the effectiveness of hierarchical sparse polynomial regression discussed in this paper, we report its performance first on synthetic sparse data and subsequently on real data from the UCI Machine Learning Repository as well. All algorithms in this document are implemented in Julia and executed on a standard Intel(R) Xeon(R) CPU E5-2690 @ 2.90GHz running CentOS release 6.7. All optimization was done with the help of the commercial mathematical optimization distribution Gurobi version 6.5 interfaced through the JuMP package developed by Lubin and Dunning (2015).

5.1 Benchmarks and data

In the first part we will describe the performance of our cutting plane algorithm for polynomial sparse regression on synthetic data. Working with synthetic data will allow for experiments concerning the efficacy of our hierarchical sparse regression method in terms of its

accuracy $A\%$ and false alarm rate $F\%$ which would otherwise not be impossible on real data. All synthetic data is generated as detailed in the following section.

Synthetic data The synthetic observations Y and input data X satisfy the linear relationship

$$\begin{aligned} Y &= g_{\text{true}}(X) + E \\ &= m(X) \cdot w_{\text{true}} + E. \end{aligned}$$

The unobserved true regressor w_{true} has exactly ℓ nonzero components at indices j selected uniformly at random without replacement from \mathcal{J} . This previous subset \mathcal{J} is itself furthermore constructed as $\{j \in [f] : A(j) \subseteq \mathcal{I}\}$ where each of the k elements in \mathcal{I} are uniformly selected out of $[p]$. The previous discussed construction thus guarantees that the ground truth w_{true} is (k, ℓ) sparse. Additionally, the nonzero coefficients in w_{true} are drawn uniformly at random from the set $\{-1, +1\}$. The observation Y consists of the signal $S := Xw_{\text{true}}$ corrupted by the noise vector E . The noise components e_t for t in $[n]$ are drawn independent identically distributed (i.i.d.) from a normal distribution and scaled such that the signal-to-noise ratio equals

$$\sqrt{\text{SNR}} := \|S\|_2 / \|E\|_2.$$

Evidently as the signal-to-noise ratio SNR increases, recovery of the unobserved true regressor w_{true} from the noisy observations can be done with higher precision. We have yet to specify how the input matrix X is chosen. We assume here that the input data samples $X = (x_1, \dots, x_n)$ are drawn from an i.i.d. source with Gaussian distribution. Although the columns of the data matrix X are left uncorrelated, the features $m_j(X)$ will be correlated. For instance, it is clear that the second and forth power of the first input can only be positively correlated.

We will compare the performance of hierarchical sparse regression with four other benchmark regression approaches. The first two approaches were chosen as to investigate the impact of both sparsity and nonlinearity on the performance of our method. The latter two approaches are of a general nonparametric nature and investigate the effect of our polynomial regressor assumption. We now describe the particularities of these four benchmarks more closely.

Polynomial Kernel Regression As a first benchmark we consider ordinary polynomial regression with regularization. The primary advantage of this formulation stems from the fact that the optimal polynomial regressor in (1) can be found efficiently using

$$g_2^*(x) = \gamma \sum_{t \in [n]} \alpha_t^* (x^\top x_t + 1)^r$$

where $\alpha^* = (\mathbb{1}_n + \gamma K)^{-1} Y$ is the maximizer of (11) with kernel matrix K as given in Theorem 2. Computing the maximizer α^* requires $\mathcal{O}(n^3)$ effort which quickly becomes prohibitive for problems counting much more than a few thousand samples. An approximation of some sort is required for such large scale data sets. Approximation methods such as the Nystrom method (Drineas and Mahoney 2005) would be one possibility here. We however opted for an approximation method in line with the approximation method suggested for the ranking heuristic when faced with large data sets at the end of Sect. 3. To speed up

computations, we split large data sets in smaller chunks of size approximately $n = 2000$ samples and compute for each of these chunks a distinct polynomial regressor. We use as an approximation to g_2^* the average of the regressors obtained on the different data chunks. We find our approximation to be both simple yet effective.

As ordinary polynomial regression does not yield sparse regressors, this benchmark will show us the merit of sparsity in terms of prediction performance. Classical Ridge regression is found as a special case for $r = 1$, enabling us to see what fruits polynomial nonlinearity brings us.

ℓ_1 -Heuristic Regression In order to determine the effect of exact sparse regression in our two step procedure, we will also use a close variant of the `SPORE` algorithm developed by Huang et al. (2010) as a benchmark. The potentially large number of polynomial features $f = \binom{p+r}{p}$ also here necessitates the use of an input ranking heuristic for data sets of size observed in practice. Using the input ranking method discussed in Sect. 3, we determine first the p' most relevant inputs heuristically. Using the remaining inputs $X' \in \mathbb{R}^{n \times p'}$ and response data $Y \in \mathbb{R}^n$ we then consider the maximizer g_1^* of (3) as a heuristic sparse regressor for varying sparsity parameter λ and the standard choice $\gamma = \infty$. The best (p, ℓ) -sparse model is then ultimately taken as the least regularized model (highest λ) still counting at most ℓ nonzero feature coefficients. This two-step regression procedure hence shares the structure outlined in Fig. 1 with our hierarchical exact sparse regression algorithm. As to have a comparable number of hyper parameters as our two-step approach, we finally perform Ridge regression on the thus selected features using a Tikhonov regularization parameter γ selected using cross validation. The final Ridge regression comes with the added benefit that it debiases the sparse regressor obtained via the standard lasso procedure which uses its hyperparameter λ for both regularization and sparsification simultaneously.

Theoretical considerations (Bühlmann and van de Geer 2011; Hastie et al. 2015; Wainwright 2009) and empirical evidence (Donoho and Stodden 2006) suggests that the ability to recover the support of the correct regressor w_{true} from noisy data using the `Lasso` heuristic experiences a phase transition. While it is theoretically understood (Gamarnik and Zadik 2017) that a similar phase transition must occur in case of exact sparse regression, due to a lack of scalable algorithms such a transition was never empirically reported. The scalable cutting plane algorithm developed in Sect. 4 offers us the means to do so however. Our main observation is that exact regression is significantly better than convex heuristics such as `Lasso` in discovering all true relevant features ($A\% \approx 100$), while truly outperforming their ability to reject the obfuscating ones ($F\% \approx 0$).

Nonparametric Regression To be able to investigate the efficacy of our polynomial regressor assumption, we also compare our hierarchical sparse method with two classical nonparametric regression methods. The simplest such method may be nearest neighbors regression (Altman 1992). This method is theoretically very well understood and even promises near optimal performance in the asymptotic regime $n \rightarrow \infty$ as discussed by Kpotufe (2011). A closely related nonparametric method is regression trees which have been empirically observed to perform excellent on many types of data sets (Breiman 2017). We used standard implementations of both these nonparametric regression methods found in the Julia packages `NearestNeighbors.jl` and `DecisionTree.jl`, respectively.

5.2 Phase transitions

In Fig. 2 we depict the performance of the hierarchical sparse regression procedure outlined in Fig. 1 in its ability to discover all relevant regression features ($A\%$) and the running time T in seconds as a function of the sample size n for a regression problem with $p' = p = 25$ inputs expanded with the help of all cubic monomials into $f = 3276$ possible features. As $p' = p$ the input ranking heuristic is irrelevant here and instead all inputs are considered by the exact sparse regression procedure. The reported results are the average of 20 independent sparse synthetic data sets where the error bars visualize one standard deviation of the inter data set variations. For the purpose of this section, we assume that we know that only $\ell = 20$ features are relevant but do not know which. In practice though, the parameter ℓ must be estimated from data. In order to play into the ballpark of the `Lasso` method, no hierarchical structure ($k = p$) is imposed. This synthetic data is furthermore lightly corrupted by Gaussian noise with $\sqrt{SNR} = 20$. Furthermore, if the optimal sparse regressor was not found by the outer approximation Algorithm 2 within 2 min, the best solution found up to that point is considered.

It is clear that our ability to uncover all relevant features ($A\% \approx 100$) experiences a phase transition at around $n_0 \approx 600$ data samples. That is, when given more than n_0 data points, the accuracy of the sparse regression method is perfect. With fewer data points our ability to discover the relevant features quickly diminishes. For comparison, we also give the accuracy performance of the `Lasso` heuristic described in (3). It is clear that exact sparse regression dominates the `Lasso` heuristic and needs fewer samples for the same accuracy $A\%$.

Especially surprising is that the time T it takes Algorithm 2 to solve the corresponding (k, ℓ) -sparse problems exactly experiences the same phase transition as well. That is, when given more than n_0 data points, the accuracy of the sparse regression method is not only perfect but easy to obtain. In fact, in case $n > n_0$ our method is empirically as fast as the `Lasso` based heuristic. This complexity transition can be characterized equivalently in terms of the number of cutting planes necessary for our outer approximation Algorithm 2 to return the optimal hierarchical sparse regressor. While potentially exponentially many ($|S_{\ell,k}^{f,p}|$ in fact) cutting planes might be necessary in the worst-case, Table 1 list the actual average number of cutting planes considered on the twenty instances previously discussed in this section. When $n > n_0$ only a few cutting planes suffice, whereas for $n < n_0$ an exponential number seem to be necessary.

5.3 The whole truth, and nothing but the truth

In the previous section we demonstrated that exact sparse regression is marginally better in discovering all relevant features $A\%$. Nevertheless, the true advantage of exact sparse regression in comparison to heuristics such as `Lasso` is found to lie elsewhere. Indeed, both our method and the `Lasso` heuristic are with a sufficient amount of data eventually able to discover all relevant ($A\% \approx 100$) features. In this section, we will investigate the ability of both methods to reject irrelevant features. Indeed in order for a method to tell the truth, it must not only tell the whole truth ($A\% \approx 100$), but nothing but the truth ($F\% \approx 0$). It is in the latter aspect that exact sparse regression truly shines.

In Fig. 3 we show the performance of the (p, ℓ) -sparse regressors found with our exact method and the `Lasso` heuristic on the same synthetic data discussed in the previous

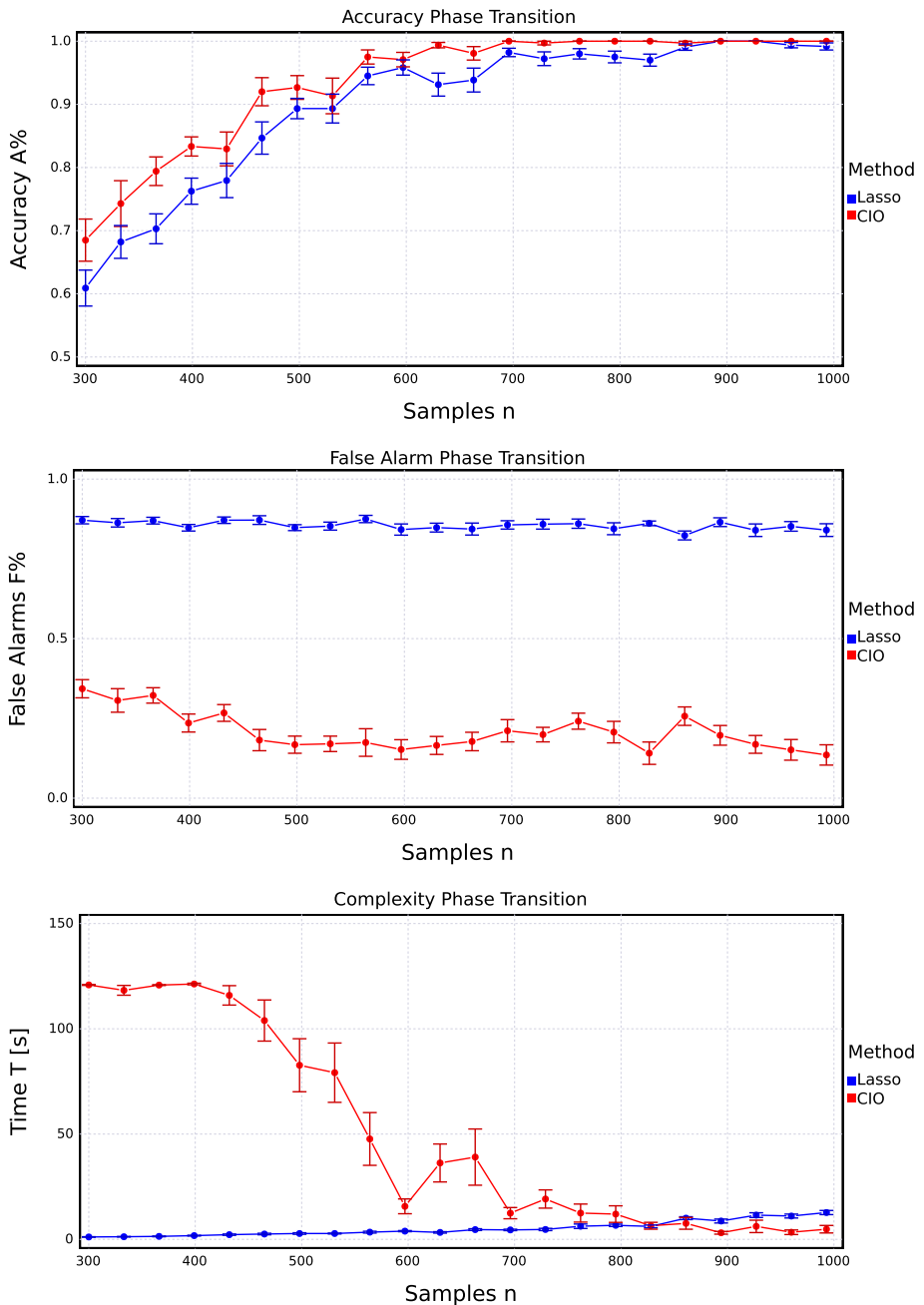


Fig. 2 The performance of exact sparse regression and the Lasso heuristic on synthetic data in terms of accuracy $A\%$, false alarm rate $F\%$ and time T in seconds. The reported results are the average of 20 independent sparse synthetic data sets where the error bars visualize one standard deviation of the inter data set variations

Table 1 Number of cutting planes considered in the outer approximation Algorithm 2 as a function of the sample size n

Samples n	300	400	500	600	700	800	900	1000
Cutting planes	> 300	> 300	298	200	70	59	25	31

For the smallest sample sizes n the optimal solution could not be computed within the allocated maximum solution time

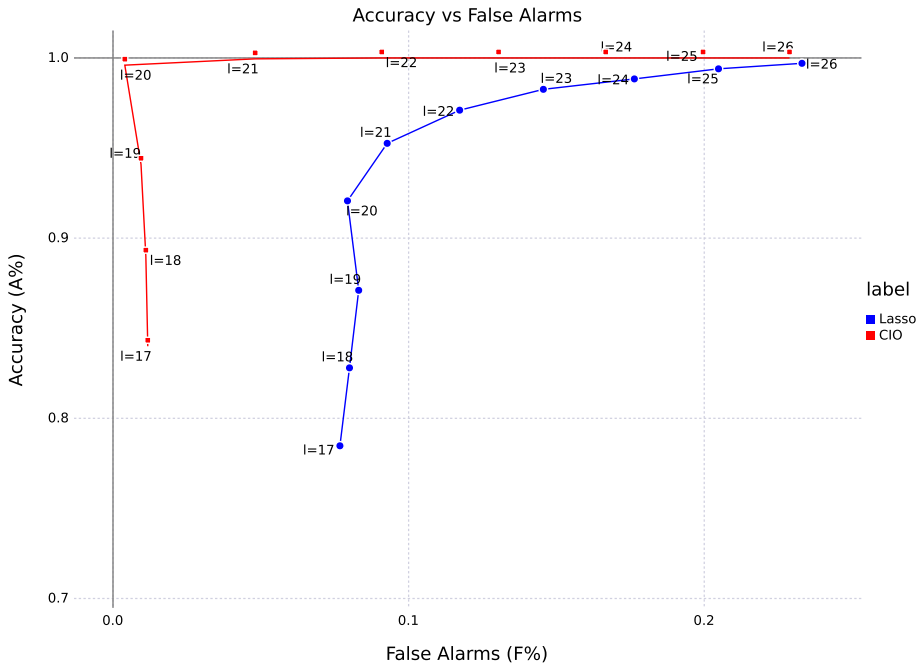


Fig. 3 The performance of exact sparse regression and the Lasso heuristic on synthetic data in terms of accuracy $A\%$ and false alarms $F\%$

section in terms of both their accuracy $A\%$ and false alarm rate $F\%$ in function of ℓ . Again, the reported results are averages over 20 distinct synthetic data sets. Each of these data sets consisted of $n = 560$ observations. Among all potential third degree monomial features, again only 20 were chosen to be relevant for explaining the observed data Y . Whereas in previous section, the true number of features was treated as a given, in practical problems ℓ must also be estimated from data. As we vary ℓ over the regression path $[f]$, we implicitly trade lower false alarm rates for higher accuracy. We have indeed a choice between including too many features in our model resulting in a high false alarm rate but hopefully discovering many relevant features, or limiting the number of features thus keeping the false alarm rate low but at the cost of missing features. One method is better than another when it makes this tradeoff better, i.e., obtains higher accuracy for a given false alarm rate or conversely a lower false alarm rate for the same accuracy. It is clear from the results shown in Fig. 3 that exact sparse regression dominates the Lasso heuristic in terms of

keeping a smaller false alarm rate while at the same time discovering more relevant monomial features.

Hence although both exact sparse regression and the `Lasso` heuristic are eventually capable to find all relevant monomial features, only the exact method finds a truly sparse regressor by rejecting most irrelevant monomials from the obfuscating bulk. In practical situations where interpretability of the resulting regression model is key, the ability to reject irrelevant features can be a game changer. While all the results so far are demonstrated on synthetic data, we shall argue in the next section that also for real data sets similar encouraging conclusions can be drawn.

5.4 Polynomial input ranking

In the preceding discussions, the first step of our approach outlined in Fig. 1 did not come into play as it was assumed that $p' = p$. Hence, no preselection of the inputs using the ranking heuristic took place. Because of the exponential number of regression features f as a function of the input dimension p , such a direct approach may not be tractable when p becomes large. In this part we will argue that the performance of the input ranking heuristic discussed in Sect. 3 is sufficient in identifying the relevant features while ignoring the obfuscating bulk.

To make our case, we consider synthetic data with hierarchical sparsity $(k, \ell) = (20, 40)$ generated in the fashion detailed in Sect. 5.1. That is, only 20 inputs in 40 monomial features of degree $r = 3$ are relevant for the purpose of regression. In Table 2 we report the average reduced input dimension p' necessary for the input ranking heuristic to recover all relevant inputs. That is, all relevant k inputs are among the top p' ranked inputs. For example, for $n = 2000$ and $p = 1000$ the input ranking heuristic needs to include $p' = 286$ to cover all 20 true features, while for $n = 10,000$ and $p = 1000$ the input ranking heuristic needs to include only $p' = 78$. We note that as n increases the input ranking heuristic needs a smaller number of features to identify the relevant ones. Note, however, that the false alarm rate of the input ranking heuristic remains high, that is, among the top p' inputs many inputs were in fact irrelevant ($p' > k$). However, we do not aspire here to find all k relevant inputs exactly, rather we hope to reduce the dimension to p' without missing out any relevant inputs. Reducing the false alarm rate is done by means of exact hierarchical sparse regression in the second step of our overall algorithm.

5.5 Real data sets

In the final part of the paper we report the results of the presented methods on several data sets found in the UCI Machine Learning Repository found under <https://archive.ics.uci.edu/ml/datasets.html>. Each of the datasets was folded ten

Table 2 Average size p' necessary for our input ranking heuristic to identify all relevant features

Dimension p'	$n = 2 \times 10^3$	$n = 4 \times 10^3$	$n = 6 \times 10^3$	$n = 8 \times 10^3$	$n = 10 \times 10^3$
$p = 200$	72	41	36	54	27
$p = 400$	117	109	75	80	62
$p = 600$	222	166	128	68	104
$p = 800$	361	230	96	95	102
$p = 1000$	286	264	142	103	78

times into 80% training data, 10% validation data and 10% test data \mathcal{T} . No preprocessing was performed on the data besides rescaling the inputs to have a unit norm.

We report the prediction performance of each of the mentioned regression methods on the test data sets. The first regression method we consider is ordinary Ridge regression. Ridge regression allows us to find out whether considering nonlinear regressors has merit. The second method we consider is polynomial kernel regression with degree r polynomials as in (1). This nonlinear but non-sparse method on its part will allow us to find out the merits of sparsity for the purpose of out-of-sample prediction performance. The third method is the ℓ_1 -heuristic described before and which will allow us to see whether our sparse formulation bring any benefits. We used the input ranking algorithm described in Sect. 3 to limit the number of potentially relevant inputs to at most $p' = 20$. Furthermore, we limit the time spent on computing our sparse hierarchical regression to at most $T_{\max} = 60$ s. We also report an upper bound on the suboptimality of the best obtained solution \tilde{s} after T_{\max} seconds as the relative accuracy gap $= (c(\sum_{j \in [p']} \tilde{s}_j K_j) - \tilde{c})/\tilde{c}$ where \tilde{c} is a provable lower bound on optimal value of the sparse hierarchical optimization formulation (8) and can be obtained at no additional cost as the quantity η defined in the outer approximation Algorithm 2.

Each of these preceding regression methods assumes the data to follow a polynomial model as discussed in Sect. 1. To be able to judge the impact of the polynomial model assumption, we also report the prediction performance of nonparametric nearest-neighbors and regression tree (CART) methods. Each of these methods is compared to our hierarchical exact regressor in terms of out-of-sample test error

$$\text{TE} := \frac{\sum_{t \in \mathcal{T}} \|y_t - h(x_t)\|_2}{\sqrt{|\mathcal{T}|}}.$$

The hyperparameters of each method were chosen as those best performing on the validation data from among $k \in [p']$, $\ell \in [100]$ and polynomial degree ranging in $r \in [4]$. The average out-of-sample performance on the test data and sparsity of the obtained regressors using the Lasso heuristic and exact hierarchical sparse regression is shown in Table 3.

As one could expect, considering nonlinear monomial features is not beneficial for prediction in all data sets. Indeed, in six data sets ordinary Ridge regression (RR) provides the same out-of-sample performance as its polynomial kernel (SVM) counterpart. In these cases, sticking with $r = 1$ was validated to be best. A similar remark can be made with regards to sparsity. That is, for all but four data sets adding sparsity using either polynomial lasso (PL) or our hierarchical regression method (CIO) does not immediately yield any benefits in terms of prediction power. Also this should not be that surprising. If the underlying data is simply not well explained by a sparse model, considering sparse models heuristically or exactly will not be beneficial to prediction performance. Nevertheless, adding sparsity using our integer optimization approach (CIO) is empirically found to lead to better prediction performance than with Lasso (PL) in all but five instances.

When we compare the performance of any of the four parametric methods (RR, SVM, PL, CIO) to either nonparametric methods (NN, CART) which do not assume a polynomial model, then in all but five instances they have better prediction performance. It should be remarked that our method (CIO) has two potential sources of prediction performance degradation: (1) the inexact nature of the input ranking heuristic when $p > p' = 20$ and (2) the fact that the obtained regressor solves the sparse regression problem (8) only up to relative accuracy gap > 0 . Increasing T_{\max} from 60 to 3600 s had no statistically significant effect on either the test error (TE) or relative accuracy (gap) for the data sets in Table 3 which

Table 3 Out-of-sample performance of regression trees (CART), nearest neighbors (NN), ridge regression (RR), polynomial kernel regression (SVM), polynomial lasso (PL) and exact hierarchical regression (CIO) on several UCI Data sets

Problem	n	p	CART	NN	RR	SVM	PL	CIO	
			TE	TE	TE	TE	TE	TE	Gap
Airfoil self-noise	1502	5	3.07	2.28	4.76	3.09	6.10	3.25	0.15
Concrete compressive strength	1030	8	7.00	7.67	10.23	6.12	7.63	6.84	0.16
Energy efficiency I	768	8	0.62	3.06	2.95	0.72	2.49	2.39	0.13
Energy efficiency II	768	8	2.00	3.50	3.31	1.66	3.32	3.43	0.20
Wine quality red	1599	11	0.74	0.68	0.69	0.69	0.67	0.78	0.18
Wine quality white	4898	11	0.78	0.75	0.76	0.71	0.76	0.77	0.17
UJIIndoorLoc I	199,937	519	11.73	8.57	36.8	36.8	121.08	75.3	0.18
UJIIndoorLoc II	199,937	519	0.06	0.04	0.21	0.09	0.36	0.35	0.18
Monitoring	11,933	16	0.0014	0.0016	0.0022	0.00010	0.010	0.0056	0.03
Music I	1058	68	20.06	16.16	16.81	16.81	21.97	15.1	0.0
Music II	1058	68	53.89	42.21	43.65	43.65	47.73	45.26	0.11
Residential I	372	107	257.51	672.36	157.79	157.79	119.27	120.56	0.15
Residential II	372	107	46.49	88.12	35.90	35.90	29.53	35.34	0.17
Slice localization	53,500	384	2.83	0.204	8.47	5.32	15.06	14.69	0.0
Average rank			3.43	2.64	3.29	4.36	3.79	3.5	

The last row gives the rank of each method as averaged over all data sets

have at most 20 inputs. Hence, we do not know with certainty whether the observed loss in prediction performance stems from the imposed polynomial nature of the regression models or the heuristic nature of our structured sparse regression algorithm. As the polynomial kernel (SVM) regression method has the worst overall average ranking, the former seems however a more likely explanation than the latter.

6 Conclusions

We discussed a scalable hierarchical sparse regression method based on a smart heuristic and modern integer optimization for nonlinear regression. We consider as the best regressor that degree r polynomial of the input data which depends on at most k inputs counting at most ℓ monomial terms which minimizes the sum of squares prediction errors with a Tikhonov loss. This hierarchical sparse specification aligns well with big data settings where many inputs are not relevant for prediction purposes and the functional complexity of the regressor needs to be controlled to avoid overfitting. Using a modern cutting plane algorithm, we can use exact sparse regression on regression problems of practical size. The ability of our method to identify all k relevant inputs as well as all ℓ relevant monomial terms and reject all others was shown empirically to experience a phase transition. In the regime where our method is statistically powerful, the computational complexity of exact hierarchical regression was empirically on par with LASSO based heuristics taking away their main propelling justification. We have empirically shown that we can outperform

heuristic methods in both finding all relevant nonlinearities as well as rejecting obfuscating ones.

References

- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185.
- Bach, F. (2008). Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, *9*(Jun), 1179–1225.
- Bach, F. (2009). Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in neural information processing systems* (pp. 105–112).
- Barnhart, C., Johnson, E., Nemhauser, G., Savelsbergh, M., & Vance, P. (1998). Branch-and-price: Column generation for solving huge integer programs. *Operations Research*, *46*(3), 316–329.
- Bertsimas, D., & Copenhaver, M. (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, *270*, 931–942.
- Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Annals of Statistics*, *44*(2), 813–852.
- Bertsimas, D., & Van Parys, B. (2017). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. Submitted to the Annals of Statistics. <https://arxiv.org/abs/1709.10029>.
- Breiman, L. (2017). *Classification and regression trees*. London: Routledge.
- Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Berlin: Springer.
- Candès, E., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, *52*(2), 489–509.
- Donoho, D., & Stodden, V. (2006). Breakdown point of model selection when the number of variables exceeds the number of observations. In *International joint conference on neural networks* (pp. 1916–1921). IEEE.
- Drineas, P., & Mahoney, M. (2005). On the Nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, *6*(Dec), 2153–2175.
- Duran, M., & Grossmann, I. (1986). An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming*, *36*(3), 307–339.
- Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.
- Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.
- Fletcher, R., & Leyffer, S. (1994). Solving mixed integer nonlinear programs by outer approximation. *Mathematical Programming*, *66*(1), 327–349.
- Gamarnik, D., & Zadik, I. (2017). High-dimensional regression with binary coefficients. Estimating squared error and a phase transition. <https://arxiv.org/abs/1701.04455>.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.
- Hall, P., & Xue, J. H. (2014). On selecting interacting features from high-dimensional data. *Computational Statistics & Data Analysis*, *71*, 694–708.
- Hao, N., & Zhang, H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, *109*(507), 1285–1301.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton: CRC Press.
- Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
- Huang, L., Jia, J., Yu, B., Chun, B. G., Maniatis, P., & Naik, M. (2010). Predicting execution time of computer programs using sparse polynomial regression. In *Advances in neural information processing systems* (pp. 883–891).
- Kong, Y., Li, D., Fan, Y., Lv, J., et al. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics*, *45*(2), 897–922.
- Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. In *Advances in neural information processing systems* (pp. 729–737).
- Lubin, M., & Dunning, I. (2015). Computing in operations research using Julia. *INFORMS Journal on Computing*, *27*(2), 238–248.

- Mallat, S., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12), 3397–3415.
- Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London*, 209, 415–446.
- Miller, A. (2002). *Subset selection in regression*. Boca Raton: Chapman and Hall/CRC.
- Nesterov, Y., & Nemirovskii, A. (1994). *Interior-point polynomial algorithms in convex programming*. Philadelphia: SIAM.
- Pelckmans, K., Suykens, J., Van Gestel, T., De Brabanter, J., Lukas, L., Hamers, B., De Moor, B., & Vandewalle, J. (2002). LS-SVMlab: A Matlab/C toolbox for least squares support vector machines. Technical report, K.U.Leuven
- Poggio, T. (1975). On optimal nonlinear associative recall. *Biological Cybernetics*, 19(4), 201–209.
- Schölkopf, B., & Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge: MIT press.
- Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2), 1–85.
- Stone, M. (1948). The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21(5), 237–254.
- Suykens, J., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3), 293–300.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Tikhonov, A. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39(5), 195–198.
- Vapnik, V. (1998). The support vector method of function estimation. In *Nonlinear modeling* (pp. 55–85), Springer.
- Vapnik, V. (2013). *The nature of statistical learning theory*. Berlin: Springer.
- Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5), 2183–2202.
- Zhao, P., Rocha, G., & Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37, 3468–3497.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.