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

PETER RADCHENKO AND GARETH M. JAMES [*]

Abstract

Numerous penalization based methods have been proposed for fitting a traditional linear regression model in which the number of predictors, p, is large relative to the number of observations, n. Most of these approaches assume sparsity in the underlying coefficients and perform some form of variable selection. Recently, some of this work has been extended to non-linear additive regression models. However, in many contexts one wishes to allow for the possibility of interactions among the predictors. This poses serious statistical and computational difficulties when p is large, as the number of candidate interaction terms is of order p^2. We introduce a new approach, "Variable selection using Adaptive Non-linear Interaction Structures in High dimensions" (VANISH), that is based on a penalized least squares criterion and is designed for high dimensional non-linear problems. Our criterion is convex and enforces the heredity constraint, in other words if an interaction term is added to the model, then the corresponding main effects are automatically included. We provide theoretical conditions under which VANISH will select the correct main effects and interactions. These conditions suggest that VANISH should outperform certain natural competitors when the true interaction structure is sufficiently sparse. Detailed simulation results are also provided, demonstrating that VANISH is computationally efficient and can be applied to non-linear models involving thousands of terms while producing superior predictive performance over other approaches.

Some key words: Non-Linear Regression; Interactions; Heredity structure; Regularization; Variable Selection

1 Introduction

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

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

[*] Marshall School of Business, University of Southern California.

when the number of predictors, p, is large relative to the number of observations, n. In this situation there are many methods that outperform ordinary least squares (Frank and Friedman, 1993). One common approach is to assume that the true number of regression coefficients, i.e. the number of nonzero β_j^*'s, is small, in which case estimation results can be improved by performing some form of variable selection. An important class of variable selection methods utilizes penalized regression. The most well known of these procedures is the Lasso (Tibshirani, 1996) which imposes an L_1 penalty on the coefficients. Numerous alternatives and extensions have been suggested. A few examples include SCAD (Fan and Li, 2001), the Elastic Net (Zou and Hastie, 2005), the Adaptive Lasso (Zou, 2006), the Group Lasso (Yuan and Lin, 2006), the Dantzig selector (Candes and Tao, 2007), the Relaxed Lasso (Meinshausen, 2007), VISA (Radchenko and James, 2008), and the Double Dantzig (James and Radchenko, 2009).

Penalized regression methods have now been extensively studied for (1). This paper extends 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,

$$Y_i = \beta_0^* + \sum_{j=1}^{p} \beta_j^* X_{ij} + \sum_{j>k} \beta_{jk}^* X_{ij} X_{ik} + \epsilon_i, \quad i = 1, \ldots n. \tag{2}$$

Second, we extend (2) to the more general non-linear domain using,

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

While (2) and (3) are well known models, fitting them involves estimating on the order of p^2 terms, most of which, in the case of (3), are two-variate functions. Thus fitting these models presents a considerable computational and statistical challenge for large p.

A relatively small number of papers have been written on sparse high dimensional models involving interactions or non-linearity. Choi *et al.* (2010) propose an approach, SHIM, for fitting (2) and also extend SHIM to generalized linear models. A nice aspect of SHIM is that it enforces a hierarchical structure where main effects are automatically added to a model at the same time as the corresponding interaction term. SHIM also possesses the oracle property of Fan and Li (2001). However, its optimization criterion is non-convex so it is only examined with up to $p = 10$ predictors, corresponding to 45 interaction terms. SHIM is not used to fit the non-linear model, (3). The SpAM approach of Ravikumar *et al.* (2009) fits a sparse additive model by imposing a penalty, $\lambda \sum_{j=1}^{p} \|f_j\|_2$, on the empirical L_2 norms of the main effects. Meier *et al.* (2009) fit the same model but also incorporate a smoothness term in the penalty function, $\lambda \sum_{j=1}^{p} \sqrt{\|f_j\|_2^2 + \int f_j''(x)^2 dx}$, leading to interesting theoetical properties. Fan *et al.* (2010) extend the sure independence screening approach of Fan and Lv (2008) to the ultrahigh dimensional non-linear setting. However, while the approaches of Ravikumar *et al.* (2009), Meier *et al.* (2009), and Fan and Lv (2008), can all work well on additive models, they are not designed to fit non-linear interaction models such as (3). Yuan (2007) use a non-negative garrote to fit a non-linear regression model. However, while they discuss fitting (3), in practice they only implement additive models. Lin and Zhang (2006) propose a method called COSSO that can fit (3) but only appears feasible for relatively low dimensional settings.

A simple approach to fit (3) would be to use a penalty function of the form

$$P(f) = \lambda \left(\sum_{j=1}^{p} \|f_j\|_2 + \sum_{j=1}^{p} \sum_{k=j+1}^{p} \|f_{jk}\|_2 \right). \tag{4}$$

Minimizing the usual sum of squares plus the penalty (4) has the effect of shrinking most of the main effect and interaction terms to zero, in a similar manner to that of the Lasso in the linear setting. This is a natural extension of SpAM so we call it the "SpAM with Interactions" (SpIn) method. However, SpIn has some significant drawbacks. First, it is inefficient, because it treats interactions and main effects similarly, when in fact an entry of an interaction into the model generally adds more predictors than an entry of a main effect, and is also harder to interpret. Second, for sufficiently large p it is computationally prohibitive to naively estimate p^2 different terms.

Instead we introduce a novel convex penalty function that enforces the heredity constraint and also automatically adjusts the degree of shrinkage on the interactions depending on whether the main effects are already present in the model. The penalty function motivates a computationally efficient block coordinate descent algorithm that handles models involving thousands of interaction terms. One consequence of the algorithm is to make it easier to enter the model for interaction terms corresponding to predictors that have already been added. Thus, it reduces the false positive rate among interaction terms.

The paper is set out as follows. In Section 2 we present our approach, called "Variable selection using Adaptive Non-linear Interaction Structures in High dimensions" or VANISH for short. VANISH extends the high dimensional linear model (1) both by incorporating interaction terms and by allowing the main effects and interactions to be non-linear. We also provide an efficient coordinate descent algorithm for fitting VANISH. Our theoretical results are given in Section 3. Here we provide conditions under which VANISH will select the correct model with probability tending to one, as n and p tend to infinity. Further, these conditions suggest that VANISH should outperform SpIn when the true interaction structure is sufficiently sparse. A number of detailed simulation results, both for linear and non-linear models, are surveyed in Section 4. These simulations involve up to $5,000$ interaction terms and demonstrate that VANISH is computationally efficient for moderate scale data sets and has better estimation accuracy than competing methods. We end with an application of VANISH on a real data set in Section 5 and a discussion in Section 6.

2 Methodology

In this section we present the VANISH method. The model and the optimization criterion are detailed in Section 2.1. Then in Section 2.2 we derive a coordinate descent algorithm. Methods for accelerating the fitting algorithm are covered in Section 2.3.

2.1 Optimization Criterion

Our goal is to fit the general non-linear model (3). We assume that p is large, but only a small fraction of the main effects and interaction terms are present in the true model. We can express (3) as

$$\mathbf{Y} = \sum_{j=1}^{p} \mathbf{f}_j^* + \sum_{j=1}^{p} \sum_{k=j+1}^{p} \mathbf{f}_{jk}^* + \boldsymbol{\epsilon}, \tag{5}$$

where $\mathbf{f}_j^* = \left(f_j^*(X_{1j}), ..., f_j^*(X_{nj})\right)^T$, $\mathbf{f}_{jk}^* = \left(f_{jk}^*(X_{1j}, X_{1k}), ..., f_{jk}^*(X_{nj}, X_{nk})\right)^T$, and $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are n-dimensional vectors corresponding to the response and the error terms, respectively. Here it is understood that $f_{jk}^*(a, b) = f_{kj}^*(b, a)$ for all a and b and all $j \neq k$. We don't include an intercept term in this model, because we center $\mathbf{Y}$, the $\mathbf{f}_j^*$'s and the $\mathbf{f}_{jk}^*$'s. The estimate for the intercept, which we do not penalize, can be computed from the fitted model. We assume, for concreteness, that $X_{ij} \in [0, 1]$ for all i and j.

We consider candidate vectors $\{\mathbf{f}_j, \mathbf{f}_{jk}\}$ that are defined analogously to their true counterparts. The corresponding functions are assumed to belong to some prespecified finite dimensional space. Our general approach for fitting (5) is to minimize the following penalized regression criterion,

$$\frac{1}{2} \left\|\mathbf{Y} - \mathbf{f}\right\|^2 + P(\mathbf{f}), \tag{6}$$

where

$$\mathbf{f} = \sum_{j=1}^{p} \mathbf{f}_j + \sum_{j=1}^{p} \sum_{k=j+1}^{p} \mathbf{f}_{jk},$$

and $P(\mathbf{f})$ is a penalty function on $\mathbf{f}$. However, the choice of the penalty function is crucial to the performance of the method.

The SpIn approach, which uses the penalty function given by (4), is an obvious candidate. However, as discussed in the introduction, a significant disadvantage to SpIn is that it treats the main effects and interactions equivalently. We argue that, "all else equal", there are two reasons one would prefer to add main effects to the model ahead of interaction terms. First, adding an interaction when the corresponding main effects are not present results in two new predictors. In terms of model sparsity this is the equivalent of adding two main effects. Second, interaction terms are more difficult to interpret than main effects, thus, given similar predictive ability, one would prefer to add a main effect ahead of an interaction. SpIn does not account for either of these concerns.

Instead we suggest a more refined penalty function,

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

which automatically addresses the above issues. In the discussion section we provide an extension of (7) to higher order interactions. The norm $\|\cdot\|$ that we use is the usual Euclidean vector norm. We show in the next few sections that the VANISH algorithm resulting from (7) has several desirable properties. In particular, it turns out that λ_1 can be interpreted as the weight of the penalty for each additional predictor included in the model, while λ_2 corresponds to an additional penalty on the interaction terms for the reduction in interpretability of a non-additive model. We also show in Sections 3 and 4 that the VANISH estimator has desirable theoretical properties and results in strong performance in comparison to other methods. In addition, penalty function (7) imposes the heredity constraint and, unlike the approach of Choi *et al.* (2010), has the advantage of producing a convex optimization criterion.

In order for (6) to have a non-trivial solution some form of smoothness constraint must be imposed on the f_j's and f_{jk}'s. Two standard approaches are to include a smoothness penalty in the optimization criterion or, alternatively, to restrict the functions to some finite-dimensional class. In this setting either approach could be adopted but we use the latter method. More specifically, we represent the candidate main effect functions using a preselected finite orthonormal basis with respect to the Lebesgue measure on the unit interval, $\{\psi_1(\cdot), ..., \psi_{d_m}(\cdot)\}$, and we represent the interaction functions using a preselected orthonormal basis with respect to the Lebesgue measure on the unit square, $\{\phi_1(\cdot,\cdot), ..., \phi_{d_{in}}(\cdot,\cdot)\}$. The assumption we are making is that f_j^* and f_{jk}^* are well approximated by the ψ's and the ϕ's respectively. The exact statement of this assumption and further details on the construction of the basis are given in the theory section.

Recall that we center all the candidate functions, so the basis functions are centered as well. Let Ψ_j denote the n by d_m matrix with the (i,k)-th entry given by $\psi_k(X_{ij})$, and let Φ_{jk} denote the n by d_{in} matrix with the (i,l)-th entry given by $\phi_l(X_{ij}, X_{ik})$. Hence the main effects and interaction terms can be expressed as $\mathbf{f}_j = \Psi_j \boldsymbol{\beta}_j$ and $\mathbf{f}_{jk} = \Phi_{jk}\boldsymbol{\beta}_{jk}$ where $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_{jk}$ respectively denote the d_m and d_{in}-dimensional vectors of basis coefficients for the j-th main effect, and the jk-th interaction term. We write $\widehat{\boldsymbol{\beta}}_j, \widehat{\boldsymbol{\beta}}_{jk}, \widehat{\mathbf{f}}_j$ and $\widehat{\mathbf{f}}_{jk}$ for the corresponding estimates and assume that $\Phi_{jk} = \Phi_{kj}$ and $\boldsymbol{\beta}_{jk} = \boldsymbol{\beta}_{kj}$. Using this basis function representation, the optimization criterion given in (6) and (7) can be expressed as

$$\frac{1}{2}\left\|\mathbf{Y}-\sum_{j=1}^{p}\Psi_j\boldsymbol{\beta}_j-\sum_{j=1}^{p}\sum_{k=j+1}^{p}\Phi_{jk}\boldsymbol{\beta}_{jk}\right\|^2 + \\ \lambda_1\sum_{j=1}^{p}\left(\left\|\Psi_j\boldsymbol{\beta}_j\right\|^2+\sum_{k:\,k\neq j}^{p}\left\|\Phi_{jk}\boldsymbol{\beta}_{jk}\right\|^2\right)^{1/2}+\lambda_2\sum_{j=1}^{p}\sum_{k=j+1}^{p}\left\|\Phi_{jk}\boldsymbol{\beta}_{jk}\right\| \tag{8}$$

Our general approach is to minimize (8) over $\boldsymbol{\beta}_j \in \mathbb{R}^{d_m}, \boldsymbol{\beta}_{jk} \in \mathbb{R}^{d_{in}}$. Note that the criterion is strictly convex in the parameters as long as the columns of each Ψ and Φ matrix are linearly independent. The last requirement will hold with probability one if d_m and d_{in} are less than n.

2.2 Fitting the Penalized Optimization

We use the block coordinate descent method to fit the VANISH model (Fu, 1998; Friedman *et al.*, 2007; Wu and Lange, 2008). This approach works by cycling through all the terms in the expansion, i.e. all the main effects and all the interaction terms, at each step holding all but one term fixed. The VANISH penalty function is not separable with respect to the different terms, hence the coordinate descent need not optimize the criterion for all values of the tuning parameters. However, for the types of sparse fits we are interested in achieving one can write out sufficient conditions for the algorithm to converge to the correct solution. For example, the algorithm will optimize the criterion when λ_2 is sufficiently large, which falls in line with the sparse interaction setting that VANISH is designed to fit.

Direct calculation shows that the following iterative method provides the coordinate descent algorithm corresponding to (8).

VANISH Algorithm

0. Initialize $\widehat{\boldsymbol{\beta}}_j, \widehat{\boldsymbol{\beta}}_{jk}$ for all $j,k \in \{1,...,p\}$. Let $S_j = \Psi_j(\Psi_j^T\Psi_j)^{-1}\Psi_j^T$ and $S_{jk} = \Phi_{jk}(\Phi_{jk}^T\Phi_{jk})^{-1}\Phi_{jk}^T$ represent the projection matrices for the main effects and interaction terms respectively.

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 $\widehat{\mathbf{P}}_j = S_j\mathbf{R}_j$, the projection of the residual onto the space spanned by the columns of Ψ_j. 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.

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

4. Compute the 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. Compute $\widehat{\mathbf{P}}_{jk} = S_{jk}\mathbf{R}_{jk}$, the projection of the residual onto the space spanned by the columns of Φ_{jk}. This gives the unshrunk estimate of $\mathbf{f}_{jk}$.

6. Set $\widehat{\mathbf{f}}_{jk} = \alpha_{jk}\hat{\mathbf{P}}_{jk}$ where $0 \leq \alpha_{jk} \leq 1$ is the shrinkage parameter defined below.

Iterate the steps 1 through 6 until convergence.

> **Remark.** It is well known that penalized regression methods can over shrink coefficient estimates. A common solution is to use the unshrunk least squares fits based on the currently selected model. See for example the Relaxed Lasso approach of Meinshausen (2007). For the remainder of this paper we take a similar approach, producing, for each λ, final estimates for the functions using the least squares fits based on the current VANISH model.

When the quantity $c_j = \sum_{k:k\neq j} \|\widehat{\mathbf{f}}_{jk}\|^2$ is zero, the shrinkage parameter from step 3 can be computed in closed form using the equation $\alpha_j = \left(1 - \lambda_1/\|\widehat{\mathbf{P}}_j\|\right)_+$, where $(\cdot)_+$ represents the positive part. When c_j is nonzero, the shrinkage parameter is derived by solving the equation

$$\alpha_j \left(1 + \frac{\lambda_1}{\sqrt{\alpha_j^2 \|\widehat{\mathbf{P}}_j\|^2 + c_j}}\right) = 1. \tag{9}$$

Equation (9) can be solved by applying the Newton-Raphson method, but instead of iterating until convergence, we only do one step. We have found that in practice the number of sweeps through all the terms does not increase after this simplification. Generally VANISH is fitted using a grid of tuning parameters. In this case we initialize Newton-Raphson using the corresponding α_j from the previous fit on the grid. Typically, the fitted models contain few interaction terms, so $c_j = 0$ for most j, and α_j can be computed directly. Note that after solving for α_j we set $\widehat{\boldsymbol{\beta}}_j = \alpha_j(\Psi_j^T\Psi_j)^{-1}\Psi_j^T\mathbf{R}_j$ and $\widehat{\mathbf{f}}_j = \Psi_j\widehat{\boldsymbol{\beta}}_j = \alpha_j\widehat{\mathbf{P}}_j$.

The shrinkage parameter for the interaction terms from step 6 can be computed in a similar fashion. Let $c_{jk}^1 = \|\widehat{\mathbf{f}}_j\|^2 + \sum_{l\notin\{j,k\}} \|\widehat{\mathbf{f}}_{jl}\|^2$ and $c_{jk}^2 = \|\widehat{\mathbf{f}}_k\|^2 + \sum_{l\notin\{j,k\}} \|\widehat{\mathbf{f}}_{kl}\|^2$. If $c_{jk}^1 = c_{jk}^2 = 0$ then the shrinkage parameter can be computed in closed form using the equation $\alpha_{jk} = \left(1 - (2\lambda_1 + \lambda_2)/\|\widehat{\mathbf{P}}_{jk}\|\right)_+$. Alternatively, if both c_{jk}^1 and c_{jk}^2 are nonzero, the shrinkage parameter is derived by solving the equation

$$\alpha_{jk}\|\widehat{\mathbf{P}}_{jk}\| \left(1 + \frac{\lambda_1}{\sqrt{\alpha_{jk}^2\|\widehat{\mathbf{P}}_{jk}\|^2 + c_{jk}^1}} + \frac{\lambda_1}{\sqrt{\alpha_{jk}^2\|\widehat{\mathbf{P}}_{jk}\|^2 + c_{jk}^2}}\right) = \left(\|\widehat{\mathbf{P}}_{jk}\| - \lambda_2\right)_+. \tag{10}$$

Finally, if only one of these quantities is nonzero, say c^1_{jk}, then the shrinkage parameter satisfies

$$\alpha_{jk}\|\widehat{\mathbf{P}}_{jk}\|\left(1+\frac{\lambda_1}{\sqrt{\alpha^2_{jk}\|\widehat{\mathbf{P}}_{jk}\|^2+c^1_{jk}}}\right)=\left(\|\widehat{\mathbf{P}}_{jk}\|-\lambda_1-\lambda_2\right)_+. \tag{11}$$

As before, we solve for α_{jk} using one step of the Newton-Raphson method which again proved to be more computationally efficient in practice. Note that after solving for α_{jk} we set $\widehat{\boldsymbol{\beta}}_{jk}=\alpha_{jk}(\Phi^T_{jk}\Phi_{jk})^{-1}\Phi^T_{jk}\mathbf{R}_{jk}$ and $\widehat{\mathbf{f}}_{jk}=\Phi_j\widehat{\boldsymbol{\beta}}_{jk}=\alpha_{jk}\widehat{\mathbf{P}}_{jk}$.

Examination of the above equations for α_j and α_{jk} shows that terms will be added to the model if and only if the norms of their unshrunk estimates, $\|\widehat{\mathbf{P}}\|$, are above a given threshold, with the threshold varying for different terms. For the main effect, $\mathbf{f}_j$, the threshold is given by

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

If none of the interactions associated with $\mathbf{f}_j$ have entered the model, then its threshold equals λ_1, the same as for the SpAM method. However, if any $\mathbf{f}_{jk}$ enters the model, then the threshold for $\mathbf{f}_j$ drops to zero. This is intuitive, because in this case adding the main effect would not introduce any new predictors. This change in threshold can be seen by noting that $c_j>0$ implies, through equation (9), that $\alpha_j>0$. Hence, if an interaction term enters the model, the two corresponding main effects must also enter the model, automatically implementing the standard hierarchical principal. For the remainder of this paper we make the assumption that the heredity structure holds for the true model. This is a common assumption that significantly reduces the complexity of the high dimensional data, making the problem more tractable.

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

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

i.e. λ_2 plus λ_1 multiplied by the number of main effects, $\mathbf{f}_j$ and $\mathbf{f}_k$, that are absent from the model. For example, if both $\mathbf{f}_j$ and $\mathbf{f}_k$ are already in the model, adding $\mathbf{f}_{jk}$ introduces no new predictors, so equation (10) shows that the threshold drops to λ_2, corresponding to the penalty for moving away from an additive model. However, if both $\mathbf{f}_j$ and $\mathbf{f}_k$ are absent from the model, then including $\mathbf{f}_{jk}$ introduces two new predictors, so the threshold for its entry rises to $2\lambda_1+\lambda_2$. Finally, when exactly one of the main effects is present, adding the interaction term introduces one new predictor, so equation (11) shows that the threshold becomes $\lambda_1+\lambda_2$.

One can generalize our algorithm, following the reasoning of Ravikumar *et al.*

(2009), by allowing S_j and S_{jk} to be general linear smoothers. Using this approach the algorithm is identical to the one we outline above, except that the least squares calculations for $\|\widehat{\mathbf{P}}_j\|$ and $\|\widehat{\mathbf{P}}_{jk}\|$ are replaced by a more general operator such as a kernel smoother or a smoothing spline.

So far we have presented the most general version of VANISH involving both λ_1 and λ_2. However, in our experience we have generally obtained good results by setting the tuning parameters equal and selecting a single $\lambda = \lambda_1 = \lambda_2$. This approach amounts to imposing the same degree of penalty on model size (λ_1) as on interpretability of the interaction terms (λ_2). Several criteria, such as CV, GCV, BIC or AIC can be used to select λ. In the simulation studies we used a validation set, and for the real data we used cross-validation; both methods worked well. We constructed the path of VANISH solutions by selecting a fine grid of λ's and iteratively applying the VANISH algorithm to each grid point, using the previous solution as a warm start. From here on when we refer to the VANISH estimator, we will mean a point on the solution path constructed according to this iterative grid approach.

2.3 Accelerating VANISH

A significant difficulty when incorporating interaction terms into a regression context is that we need to fit of order p^2 different terms, requiring multiple sweeps through p^2 variables for each value of the tuning parameter. While coordinate descent algorithms are generally very fast (Friedman *et al.*, 2007; Wu and Lange, 2008), for sufficiently large p this can be extremely costly computationally. It is also inefficient, because in practice VANISH will prevent almost all the interaction terms from entering the model.

Let $\mathcal{A}_\lambda$ represent the active set of variables associated with a tuning parameter λ, i.e. the variables with non-zero coefficients. When constructing the path of solutions as a function of the tuning parameter, one typically chooses a grid of λ's and computes the solution iteratively at each point. For adjacent points λ and λ' it is usually the case that $\mathcal{A}_\lambda$ and $\mathcal{A}_{\lambda'}$ are identical or differ by at most one element. Hence, Peng *et al.* (2010) and others suggest that, given the current active set, $\mathcal{A}_\lambda$, one should first assume $\mathcal{A}_\lambda = \mathcal{A}_{\lambda'}$ and iterate through the small number of active variables to produce the estimated fit. Once convergence is achieved on the candidate active set, one performs a single sweep through all the variables to ensure that no new predictors enter the model. If the active set is unchanged, the correct solution has been found. If the set changes, then the algorithm again iterates through the new $\mathcal{A}_{\lambda'}$. Peng *et al.* (2010) show that this procedure involves many fewer sweeps through all the variables and hence provides large computational savings.

We implement a version of this approach. However, even a single sweep through all p^2 terms is expensive and should be avoided where possible. In practice, given our current fit, most of the interaction terms have almost no chance of entering the model for a small change in λ. Hence we primarily restrict ourselves to examining a small set of candidate variables, $\mathcal{C}_\lambda$. The set $\mathcal{C}_\lambda$ is defined as all main effects plus the K highest ranked interaction terms with respect to the difference between their threshold value for entry and their unshrunk norm, $\|\widehat{\mathbf{P}}_{jk}\|$. There are various strategies for selecting K. One reasonable approach is to adjust K according to the rank of interactions entering the model. For example, if an interaction ranked near the top of $\mathcal{C}_\lambda$ enters, this would suggest K could safely be kept relatively low. However, if interactions ranked near the bottom of $\mathcal{C}_\lambda$ start being selected, one would be concerned that $\mathcal{C}_\lambda$ was too small so K should be enlarged. A simple rule might be to keep K at least twice as large as the largest observed jump. In practice we have found that when fixing K between 10% and 20% of p and using a sufficiently dense grid of λ's it was extremely rare for an interaction term outside $\mathcal{C}_\lambda$ to enter the model.

Using the candidate set $\mathcal{C}_\lambda$ and the current active set $\mathcal{A}_\lambda$ we compute the VANISH solution at a nearby λ' using the following accelerated algorithm.

Accelerated Algorithm

0. Set $\mathcal{A}_{\lambda'} \leftarrow \mathcal{A}_\lambda$ and $\mathcal{C}_{\lambda'} \leftarrow \mathcal{C}_\lambda$.

1. Iterate the VANISH algorithm on the variables in $\mathcal{A}_{\lambda'}$ until convergence.

2. Iterate through the candidate variables in $\mathcal{C}_{\lambda'}$ until convergence. Usually this only involves a single sweep, because the active set is often the same for λ and λ'. If no new variables become active then stop.

3. If the active set changes at Step 2, iterate through all variables until convergence (typically a single sweep) and update $\mathcal{C}_{\lambda'}$ with the K highest ranked interaction terms.

Using this algorithm we often only iterate through the active set $\mathcal{A}_{\lambda'}$, with a single sweep through the small number of elements in $\mathcal{C}_{\lambda'}$. It is only necessary to sweep through all the variables in situations where the active set changes. Usually this only involves one sweep, except for the rare situation where $\mathcal{C}_{\lambda'}$ does not include all variables that enter the model. We have found that restricting to this candidate set significantly accelerates the algorithm without affecting the performance.

3 Theory

As mentioned earlier, we assume the heredity structure for the true model. Let $\mathcal{K}_m$ and $\mathcal{K}_{in}$, respectively, denote the index sets of the true main effects and the true interaction terms. We define the corresponding estimated sets $\widehat{\mathcal{K}}_m$ and $\widehat{\mathcal{K}}_{in}$ by analogy. Further, define $s_m = |\mathcal{K}_m|$, $s_{in} = |\mathcal{K}_{in}|$ and $s = s_m + s_{in}$. In this section we establish conditions for *sparsistency* and *persistency* of VANISH. Sparsistency means

$$P(\widehat{\mathcal{K}}_m = \mathcal{K}_m, \widehat{\mathcal{K}}_{in} = \mathcal{K}_{in}) \to 1 \text{ as } n \text{ goes to infinity.}$$

We have omitted the superscript n for simplicity of the notation, but we treat all the introduced quantities as functions of the sample size. Thus, for example, s_m may tend to infinity along with n. For clarity of exposition we first present our sparsistency results for the linear VANISH setting in Section 3.1 and then for the more general non-linear setting in Section 3.2. The definition of persistency and the corresponding results are provided in Section 3.3. In what follows we assume that the error terms are i.i.d. gaussian with zero mean and finite variance, although this assumption could be relaxed to include subgaussian random variables.

3.1 Linear Results

Let $b = \min\{|\beta_j^*|, |\beta_{jk}^*|, j \in \mathcal{K}_m, jk \in \mathcal{K}_{in}\}$ denote the smallest signal size in the true model. We use $X_{\mathcal{K}}$ to denote the matrix with columns of the form $X_j, j \in \mathcal{K}_m$ and $X_{jk}, jk \in \mathcal{K}_{in}$ and let $\Sigma_{\mathcal{K}} = X_{\mathcal{K}}^T X_{\mathcal{K}}$. Theorem 1 provides conditions under which VANISH is sparsistent in the linear setting. It is a consequence of a more general nonlinear result given in Theorem 2.

Theorem 1 *Set $\lambda_1 = \lambda_2$ and let s and b be bounded above and away from zero, respectively. In the linear setting VANISH is sparsistent for p as large as $\exp(n^{1-\epsilon})$, with arbitrarily small positive ϵ, as long as $\lambda_1 \asymp \sqrt{(\log p)(\log n)}$, and conditions*

$$\frac{\sqrt{s}}{1-\delta}\|X_j^T X_{\mathcal{K}} \Sigma_{\mathcal{K}}^{-1}\| \;\leq\; (1+2\gamma)^{-1/2}, \quad j \notin \mathcal{K}_m \tag{12}$$

$$\frac{\sqrt{s}}{1-\delta}\|X_{jk}^T X_{\mathcal{K}} \Sigma_{\mathcal{K}}^{-1}\| \;\leq\; \begin{cases} \left(\frac{1}{9}+\frac{8}{9}\gamma\right)^{-1/2} & , jk \notin \mathcal{K}_{in}, j \text{ and } k \notin \mathcal{K}_m \\ \left(\frac{1}{4}+2\gamma\right)^{-1/2} & , jk \notin \mathcal{K}_{in}, \text{either } j \text{ or } k \in \mathcal{K}_m \\ (1+8\gamma)^{-1/2} & , jk \notin \mathcal{K}_{in}, j \text{ and } k \in \mathcal{K}_m \end{cases} \tag{13}$$

are satisfied for all j, k where $\gamma = s_{in}/s$ and $\delta > 0$ is an arbitrarily small constant.

The Lasso bounds generally involve setting the tuning parameter proportional to $\sqrt{\log p}$. In fact, a close analysis of our proof shows that one could also set λ_1 proportional to $\sqrt{\log p}$ without affecting the results of Theorem 1 in any way. Note that we set $\lambda_1 = \lambda_2$ in Theorem 1 for notational convenience, but a similar result holds in the general case. As with the analogous Lasso conditions, (12) and (13) cannot be verified in practice but are still useful because they allow us to compare VANISH to the linear version of SpIn, introduced in Section 1. It follows directly from the proofs in the Appendix that for the SpIn method the right hand sides in the four inequalities above would equal 1. Hence, since $0 \leq \gamma < 1$, two of the conditions are weaker for SpIn, i.e. (12) and the third inequality of (13), one is weaker for VANISH, i.e. the first inequality of (13), while the final condition depends on the relative sizes of s_m and s_{in}.

As mentioned in the introduction, our goal for VANISH is to relax the restrictive additive structures associated with most high dimensional methods, which effectively assume $\gamma = 0$. However, introducing two way interactions requires considering of order p^2 terms. Hence, even for moderate p, an assumption of sparsity in the interaction terms is necessary, both for computational and statistical reasons. With this in mind, we designed VANISH for the small γ setting. One way to interpret equations (12) and (13) is that as $\gamma \to 0$, i.e. the fraction of true interaction terms becomes small, the VANISH conditions are strictly weaker than their SpIn counterparts for the first two inequalities in (13), while the other two conditions converge to their SpIn counterparts. Thus, for low values of γ one may expect VANISH to dominate SpIn.

If one wishes to assume a large γ then neither method dominates in terms of these conditions. However, it is worth noting that even in this situation the two conditions that are weaker for SpIn involve only a total of $O(p)$ inequalities while the condition that is weaker for VANISH involves $O(p^2)$ inequalities, using the assumption $s = O(1)$.

3.2 Non-Linear Results

In this section we extend the linear sparsistency results, developed in the previous section, to the non-linear setting. Let $\{\psi_0 \equiv 1, \psi_1, \psi_2, ...\}$ be a uniformly bounded orthonormal basis in $L_2[0,1]$. Consequently, the set $\{\phi_{l_1 l_2}(x,y) = \psi_{l_1}(x)\psi_{l_2}(y),\ l_1, l_2 = 0, 1, ...\}$ of the corresponding tensor products forms an orthonormal basis in $L_2[0,1]^2$. We center functions $\phi_{l_1 l_2}$ and, to simplify the notation, we let them retain their original names. We then do exactly the same for the univariate functions ψ_l. Recall that all the true effects are assumed to be centered, thus we can represent the j'th true main effect as $\sum_{l=1}^{\infty} \beta^*_{lj}\psi_l(x)$ and the jk'th true interaction term as $\sum_{l_1,l_2=1}^{\infty} \beta^*_{l_1 l_2 jk}\phi_{l_1 l_2}(x,y)$. We will follow Ravikumar *et al.* (2009) and require that all true main effects belong to the Sobolev space of order two: $\sum_{l=1}^{\infty}(\beta^*_{lj})^2 l^4 < C^2$ for each j, with C being some constant independent of n. We will use $S^2_C([0,1])$ to denote this space of main effects. We will also impose the same smoothness requirement on the two univariate projections corresponding to each true interaction function: $\sum_{l_1}^{\infty}(\beta^*_{l_1 l_2 jk})^2 l_1^4 < C^2$ and $\sum_{l_2}^{\infty}(\beta^*_{l_1 l_2 jk})^2 l_2^4 < C^2$. We will refer to this space of interaction terms as $S^2_C([0,1]^2)$. Let d be the dimension of the univariate basis used in the VANISH algorithm. We will assume that it is growing to infinity, but slower than n. For notational convenience, we will also assume that the dimension of the two-variate basis used in the algorithm is growing at the order d^2.

We write $\boldsymbol{\beta}^*_j$ and $\boldsymbol{\beta}^*_{jk}$, respectively, for the vectors of basis coefficients for the j-th main effect, and the jk-th interaction term. Let $b = \min\{\|\boldsymbol{\beta}^*_j\|_\infty, \|\boldsymbol{\beta}^*_{jk}\|_\infty, j \in \mathcal{K}_m, jk \in \mathcal{K}_{in}\}$ denote the smallest signal size in the true model relative to the selected basis. We let $\Theta_{\mathcal{K}}$ denote the matrix with columns $\{\Psi_j, j \in \mathcal{K}_m\}$ and $\{\Phi_{jk}, jk \in \mathcal{K}_{in}\}$, and we define $\Sigma_{\mathcal{K}} = \Theta^T_{\mathcal{K}}\Theta_{\mathcal{K}}$. For the remainder of this section we will assume that the eigenvalues of $\Sigma_{\mathcal{K}}/n$ stay bounded above and away from zero as n and p grow. We define $C_{\mathcal{K}}$ and $c_{\mathcal{K}}$ as the largest and the smallest eigenvalues, respectively, of the matrices of the form $\Psi^T_j\Psi_j$ or $\Phi^T_{jk}\Phi_{jk}$ for $j \in \mathcal{K}_m$ and $jk \in \mathcal{K}_{in}$. Theorem 2 is our most general result, providing conditions to guarantee sparsistency in the non-linear setting, for arbitrary p, s, d, b, λ_1 and λ_2. It is proved in the Appendix.

Theorem 2 *Suppose that conditions*

$$\begin{aligned} M\,\|\Psi_j^T\Theta_{\mathcal{K}}\Sigma_{\mathcal{K}}^{-1}\| &\leq [s_m + s_{in}(2+\lambda_2/\lambda_1)]^{-1/2}, \quad j \notin \mathcal{K}_m & (14)\\ M\,\|\Phi_{jk}^T\Theta_{\mathcal{K}}\Sigma_{\mathcal{K}}^{-1}\| &\leq \frac{1_{\{j\notin\mathcal{K}_m\}} + 1_{\{k\notin\mathcal{K}_m\}} + \lambda_2/\lambda_1}{(s_m + s_{in}[2+\lambda_2/\lambda_1]^2)^{1/2}}, \quad jk \notin \mathcal{K}_{in}, & (15)\end{aligned}$$

and

$$\frac{\sqrt{\log(sd)}}{b\sqrt{n}} + \frac{s\sqrt{s}}{b\sqrt{d}} + \frac{(\lambda_1 \vee \lambda_2)\sqrt{s}}{b\sqrt{n}} + \frac{s\sqrt{n}}{(\lambda_1 \wedge \lambda_2)\sqrt{d}} + \frac{d^2\log(dp)}{\lambda_1^2} + \frac{d^2\log(ds_m)}{\lambda_2^2} \to 0 \quad (16)$$

hold for all j, k, where $M = \frac{1}{1-\delta}\sqrt{\frac{C_{\mathcal{K}}}{c_{\mathcal{K}}}}$ and $\delta > 0$ is arbitrarily small. Then VANISH is sparsistent.

Conditions (14) and (15) are generalizations of their linear versions, (12) and (13), from Theorem 1. Condition (16) constrains the relative sizes of s, d, n, s_m, λ_1 and λ_2. To better understand it, suppose that the basis dimension grows at the one dimensional minimax rate $n^{1/5}$, the true model size is bounded, and the smallest signal size is bounded away from zero. Then condition (16) will be satisfied for p as large as $\exp(n^{3/5-\epsilon})$, with arbitrarily small positive ϵ, if we set $\lambda_1 \asymp \lambda_2 \asymp (n^{1/2}/\log n)$.

Corollary 1 reexpresses the conditions from Theorem 2 in the format of Theorem 1, with $\lambda_1 = \lambda_2$.

Corollary 1 *Set $\lambda_1 = \lambda_2$ and let s and b be bounded above and away from zero, respectively. Then VANISH is sparsistent for p as large as $\exp(n^{1-\epsilon}/d^2)$, with arbitrarily small positive ϵ, as long as $\lambda_1 \asymp \sqrt{n/(\log n)}$, $d \gg \log n$, and conditions*

$$\begin{aligned} \sqrt{s}M\|\Psi_j^T\Theta_{\mathcal{K}}\Sigma_{\mathcal{K}}^{-1}\| &\leq (1+2\gamma)^{-1/2}, \quad j \notin \mathcal{K}_m \\ \sqrt{s}M\|\Phi_{jk}^T\Theta_{\mathcal{K}}\Sigma_{\mathcal{K}}^{-1}\| &\leq \begin{cases} \left(\frac{1}{9} + \frac{8}{9}\gamma\right)^{-1/2} & , jk \notin \mathcal{K}_{in}, j \text{ and } k \notin \mathcal{K}_m \\ \left(\frac{1}{4} + 2\gamma\right)^{-1/2} & , jk \notin \mathcal{K}_{in}, \text{either } j \text{ or } k \in \mathcal{K}_m \\ (1+8\gamma)^{-1/2} & , jk \notin \mathcal{K}_{in}, j \text{ and } k \in \mathcal{K}_m \end{cases}\end{aligned}$$

are satisfied for all j, k, where $M = \frac{1}{1-\delta}\sqrt{\frac{C_{\mathcal{K}}}{c_{\mathcal{K}}}}$ and $\delta > 0$ is arbitrarily small.

As in the linear case, the right hand sides in the four inequalities above would equal 1 for the SpIn method. The same reasoning applies as in the linear case, in other words the VANISH conditions will dominate those for SpIn as $\gamma \to 0$.

3.3 Persistence

For a given function $m(\cdot)$, which may depend on the observed data, predictive risk $R(m)$ is defined as $\mathbb{E}\big(Y - m(X_1, ..., X_p)\big)^2$, where the expected value is taken with respect to an independently generated random vector $(Y, X_1, ..., X_p)$. For a given functional class $\mathcal{M}_n$, let m_n^* denote the predictive oracle, i.e. the minimizer of the predictive risk over $\mathcal{M}_n$. The empirical counterpart $\widehat{m}_n$, which minimizes the sample sum of squares over the class $\mathcal{M}_n$, is said to be *persistent* relative to $\mathcal{M}_n$ if $R(\widehat{m}_n) - R(m_n^*) \to 0$ in probability as n and p tend to infinity. In this section we derive conditions for the persistence of the estimator corresponding to the VANISH optimization criterion.

Here we do not assume a particular structure for the true regression function. The only assumption we impose is that the regression function is uniformly bounded. For simplicity of exposition we focus on the case $\lambda_1 = \lambda_2$, although the result that follows is given for the general case. It is now convenient to view $\widehat{m}_n$ as the minimizer of the sum of squares $\sum_{i=1}^n \big(Y_i - m(X_{i1}, ..., X_{ip})\big)^2$ over the class $\mathcal{M}_n$ of functions $m(x) = \sum_{j=1}^p \beta_j g_j(x_j) + \sum_{j=1}^p \sum_{k=j+1}^p \beta_{jk} g_{jk}(x_{jk})$, such that $g_j \in S_C^2[0,1])$ and $g_{jk} \in S_C^2([0,1]^2)$ for all j and k, and

$$\sum_{j=1}^p \big(|\beta_j|^2 + \sum_{k:\, k\neq j}^p |\beta_{jk}|^2\big)^{1/2} + \sum_{j=1}^p \sum_{k=j+1}^p |\beta_{jk}| \ \leq\ L_n. \tag{17}$$

Greenshtein and Ritov (2004) show that the Lasso is persistent for $L_n = o\big([n/\log n]^{1/4}\big)$ when p grows polynomially in n. Ravikumar *et al.* (2009) establish a similar result for the SpAM method, and our Theorem 3 provides the corresponding result for VANISH.

Theorem 3 *If $n = O(p)$, then*

$$R(\widehat{m}_n) = R(m^*) + O_p\left(L_n^2 \sqrt{\frac{\log p}{n}}\right).$$

In particular, if $L_n = o\big([n/\log p]^{1/4}\big)$ then $\widehat{m}_n$ is persistent over the class $\mathcal{M}_n$. If $p = o(n)$, then the above results hold with the $\log p$ *factor replaced by* $\log n$.

4 Simulation Results

In this section we report results from two simulation studies conducted to compare the performance of VANISH with other possible competing methods. In Section 4.1 we test VANISH using linear models, while Section 4.2 covers the more general non-linear situation. Throughout the simulation study we set the two VANISH tuning parameters, λ_1 and λ_2, equal to each other.

4.1 Linear

Our first set of results is from high dimensional data generated using a standard linear regression model. A total of six sets of simulations were performed; corresponding to differences in the magnitudes of the coefficients and number of interaction terms. For each simulation we generated 100 training data sets, each with $n = 75$ observations, and $p = 100$ main effects. This corresponded to $100 \times 101/2 = 5,050$ possible main effects and interactions. Of these main effects $s_m = 5$ of the regression coefficients were randomly set to either ± 0.5, or to ± 1, and the remainder were set to zero. In addition, 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.

In addition to our VANISH approach we fitted five competing methods, "SpAM", "SpAM$_{LS}$", "SpIn", "SpIn$_{LS}$" and "Oracle". SpAM corresponded to the approach of Ravikumar *et al.* (2009), with no interaction terms, except that all fitted functions were restricted to be linear. In this setting SpAM simply amounted to a Lasso fit. SpAM$_{LS}$ was similar to SpAM except that the final estimates for the non-zero coefficients were produced using their least squares estimates rather than their shrunk counterparts. SpIn also used a linear version of SpAM but included interaction terms, created by multiplying together all possible pairs of main effects. In this setup interaction terms and main effects were treated similarly. SpIn$_{LS}$ took the SpIn models and estimated the non-zero coefficients using their least squares fits. Finally, the Oracle approach provided a best case scenario by assuming the correct model was known and using the least squares estimate to compute the corresponding coefficients. This method could not be used in a real life situation but represented a gold standard with which to compare the other methods. The tuning parameters for each method were selected by generating an independent validation data set, with the same sample size and characteristics as the original training data, and choosing the parameters that gave the lowest prediction error on the validation data.

The results from the six simulations are shown in Table 1. For each simulation and method we report five statistics. "False-Pos Main" is the number of noise main effects that each method incorrectly included in the final model, while "False-Neg Main" is the number of true main effects that are incorrectly excluded. "False-Pos Inter" and "False-Neg Inter" are the corresponding counterparts computed for the interactions. Finally, "L2-sq" corresponds to the squared Euclidean distance between the vectors of true and estimated regression coefficients. For this last statistic we placed in bold font the results that correspond to the best method, or else a method that is not statistically worse than the best method. The Oracle results were excluded from this comparison since they are not achievable in practice. False positive and negative values were not reported for the Oracle because these are, by definition, zero. Similarly the interaction false positive and negative values were not reported for SpAM because the method does not fit interaction terms.

VANISH was statistically superior, in terms of L2-sq, to all methods but the Oracle in all the simulations where the true regression function included interaction terms. Alternatively, in the settings with $S_{int} = 0$ SpAM$_{LS}$ was either statistically indistinguishable from VANISH or statistically better. Because SpAM$_{LS}$ fitted a model

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	2.03	0.14	0.56	2.77	14.14	–
$\beta = \pm 0.5$	False-Neg Main	0.53	2.75	2.17	0.41	0.07	–
	False-Pos Inter	1.2	5.96	20.98	–	–	–
$S_{int} = 0$	False-Neg Inter	0	0	0	–	–	–
	L2-sq	0.331	1.066	1.066	**0.309**	0.383	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	2.96	0.08	0.31	2.82	12.46	–
$\beta = \pm 0.5$	False-Neg Main	0.75	3.75	3.11	1.1	0.52	–
	False-Pos Inter	2.4	3.65	16.77	–	–	–
$S_{int} = 2$	False-Neg Inter	0.69	1.45	1.18	–	–	–
	L2-sq	**0.676**	2.023	1.63	1.071	1.033	0.11
	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

Table 1: *Simulation results in the linear setting.*

with no interaction terms, while VANISH could include interactions, this simulation scenario was specifically designed to favor SpAM. The fact that VANISH provided a roughly similar level of performance demonstrates that it need not be significantly handicapped even if the true relationship turns out to be purely additive. The relative improvement in performance for VANISH also grew with the number of true interaction terms; see for example the $s_{int} = 6$ simulation. Relative to VANISH, both SpIn and SpAM tended to have many more false negatives, among both the main effects and the interactions. SpIn also tended to have more false positives among the interactions. In general, the methods using least squares fits outperformed the Lasso type fits with shrunk coefficients. While VANISH could not match the idealized performance of the Oracle, it was considerably closer than the examined competitors.

4.2 Non-Linear

We also tested the more general implementation of VANISH using five non-linear simulation scenarios. For the first simulation we generated 100 data sets, each containing $n = 300$ observations. For each observation $p = 50$ predictors were produced, each independently sampled from a Uniform distribution on the $[0, 1]$ interval. The responses were produced using the following non-linear basis function model,

$$Y = f_1(X_1)+f_2(X_2)+f_3(X_3)+f_4(X_4)+f_5(X_5)+f_{12}(X_1, X_2)+f_{13}(X_1, X_3)+\epsilon, \quad \epsilon \sim N(0, 1)$$

The main effects and interactions were generated as

$$f_j(x) \propto \sum_{l=1}^{6} \beta_{lj} b_l(x) \quad \text{and} \quad f_{jk}(x_j, x_k) \propto \sum_{l=1}^{4}\sum_{m=1}^{4} \beta_{lm,jk} b_l(x_j), b_m(x_k)$$

where each $b_l(x)$ was an element of the Fourier basis, and all the β coefficients were independently sampled from a standard normal distribution. All the main effects and interactions were scaled to ensure that $Var[f_j(X_j)] = Var[f_{jk}(X_j, X_k)] = 0.5$ for $X_j, X_k \sim U(0, 1)$. We then fitted the non-linear version of VANISH and the five competing methods to each data set. For each method we used the same functional basis as the one that generated the data.

The second simulation was identical to the first, except that the responses were generated using no interaction terms. The third and fourth simulations tested the harder situation where the true main effects and interactions were not generated from the basis functions used in VANISH and the competing methods. Instead of using the basis functions, the five true main effects were initially generated as

$$f_1(x) = x, \quad f_2(x) = \frac{1}{(1 + x)}, \quad f_3(x) = \sin(x), \quad f_4(x) = \exp(x), \quad f_5(x) = x^2.$$

Then each f_j was standardized by subtracting $E[f_j(X)]$ and dividing by $SD[f_j(X)]$

with $X \sim U(0,1)$. The interaction functions in simulation three were generated by multiplying together the standardized main effects,

$$f_{jk}(x_j, x_k) = f_j(x_j) \times f_k(x_k).$$

The responses were then produced using the non basis function model

$$Y = \sqrt{0.5}[f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + f_5(x_5) + f_{12}(x_1, x_2) + f_{13}(x_1, x_3)] + \epsilon,$$

where the $\sqrt{0.5}$ scaling factor was used to ensure that each of the seven terms had variance equal to 0.5, the same as in the previous simulations. We used a Cosine basis to fit all methods. No interaction functions were generated in simulation four.

In simulation five we examined the case $p > n$ by using 100 predictors and 75 observations with the noise level set to 0.5. We generated two true main effect functions and one interaction function using the basis function model described earlier, but with fixed values of β coefficients. More specifically, rather than generating the coefficients from $N(0, 1)$, we set all them to one, except the last three coefficients for the second main effect, which we set to minus one.

The results from these five simulations are shown in Table 2. The summary statistics are the same as for the linear simulations except that L2-sq is now calculated using the integrated squared difference between the true and the estimated functions. Qualitatively the results are very similar to those from the linear simulations. VANISH is again statistically superior, in terms of L2-sq, among all methods other than Oracle. The only exception is the simulation settings involving no interaction terms, where SpAM$_{LS}$ performs somewhat better than VANISH. This is not surprising, as SpAM fits no interaction models by design. In comparison to the SpIn and SpAM methods VANISH has lower false negative rates and roughly similar main effect false positive rates. It also has significantly lower interaction false positive rates than SpIn.

Figure 1 plots the true main effects in simulation five together with some representative estimates from VANISH. We ordered the 100 simulation runs by the L_2 distance between the estimated regression function and the truth. We then identified the 25-th, 50-th and 75-th best simulations and plotted their main effect estimates. We can see that the shapes of the true main effects were reasonably well estimated in each of the selected simulations. A similar conclusion can be made for the corresponding estimates of the interaction effect, which are illustrated in Figure 2.

5 Applications

We illustrate VANISH on the Boston housing data (Harrison and Rubinfeld, 1978) because this is one of the data sets used for the SpAM method of Ravikumar *et al.* (2009). The data has also been examined in other papers (Härdle *et al.*, 2004; Lin and Zhang, 2006). There are 506 observations and 10 predictors, with the response corre-

Simulation	Statistic	VANISH	SpIn$_{LS}$	SpIn	SpAM$_{LS}$	SpAM	Oracle
Basis function model $S_{int} = 2$	False-Pos Main	0.04	0	0	0.1	15.12	–
	False-Neg Main	0.23	1.42	0.9	0.23	0	–
	False-Pos Inter	0.51	1.14	9.69	–	–	–
	False-Neg Inter	0.06	0.21	0.06	–	–	–
	L2-sq	**0.38**	0.879	1.848	1.276	1.437	0.267
Basis function model $S_{int} = 0$	False-Pos Main	0	0	0	0.02	16.54	–
	False-Neg Main	0.12	0.93	0.46	0.04	0	–
	False-Pos Inter	0.16	0.36	11.7	–	–	–
	False-Neg Inter	0	0	0	–	–	–
	L2-sq	0.132	0.344	1.012	**0.109**	0.227	0.106
Non Basis function model $S_{int} = 2$	False-Pos Main	0	0	0	0.08	15.48	–
	False-Neg Main	0	0.65	0.33	0	0	–
	False-Pos Inter	0.48	1.96	9.72	–	–	–
	False-Neg Inter	0	0.2	0.07	–	–	–
	L2-sq	**0.333**	1.023	2.100	1.217	1.405	0.277
Non Basis function model $S_{int} = 0$	False-Pos Main	0	0	0	0.03	16.58	–
	False-Neg Main	0	0.05	0.01	0	0	–
	False-Pos Inter	0.06	0.72	11.37	–	–	–
	False-Neg Inter	0	0	0	–	–	–
	L2-sq	**0.116**	0.223	1.064	**0.114**	0.232	0.112
Basis function model $p > n$	False-Pos Main	0.05	0	0	0.22	7.47	–
	False-Neg Main	0.17	1.94	1.85	0.13	0.01	–
	False-Pos Inter	0.04	0.03	1.69	–	–	–
	False-Neg Inter	0.27	0.94	0.85	–	–	–
	L2-sq	**0.452**	2.221	1.793	0.771	0.882	0.217

Table 2: *Simulation results for each method in the nonlinear setting.*

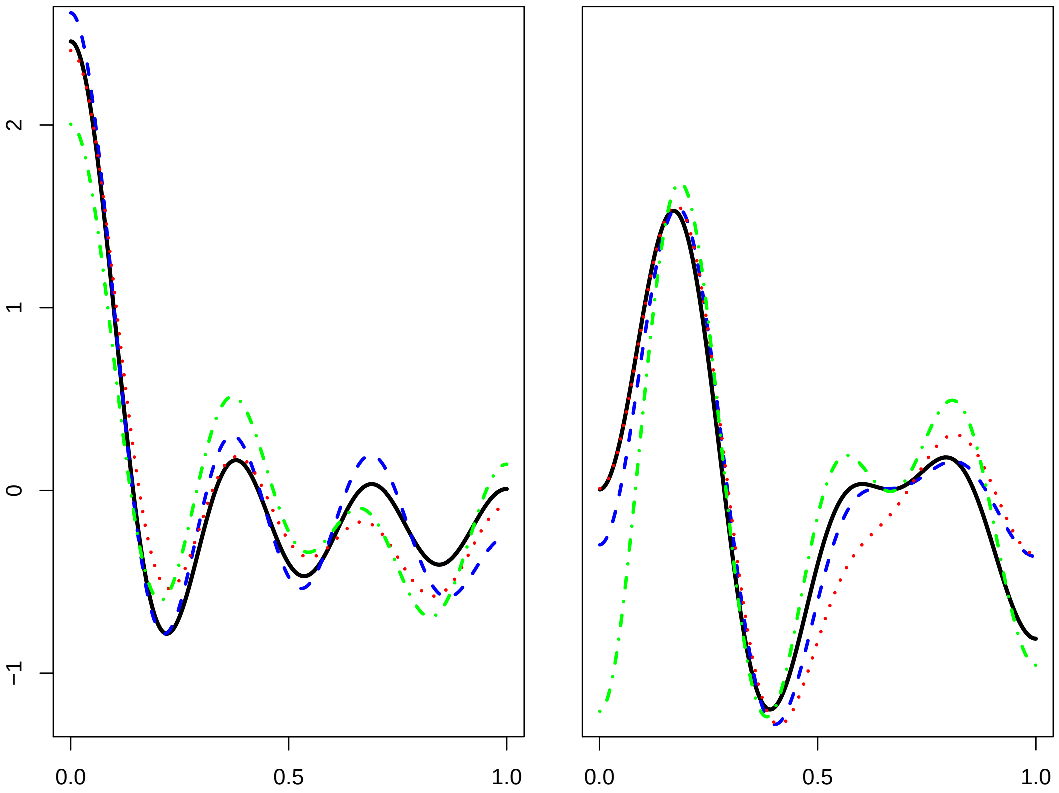

Figure 1: *The two main effects from Simulation 5. Truth (black solid), and VANISH estimates, 25th (blue dash), 50th (red dot), and 75th (green dash-dot) percentiles.*

sponding to the median house value in each neighborhood. Ravikumar *et al.* (2009) added 20 noise variables to the data to test whether SpAM correctly removed these from the model. Likewise we added 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 first randomly split the data into a training set and a test set, so that the training set contained 400 observations. We then fitted both VANISH and SpIn to the training data, using ten-fold cross-validation to select the tuning parameters. We tested two possible values for λ_2 corresponding to $\lambda_2 = \lambda_1$ and $\lambda_2 = 1.25\lambda_1$. The CV statistic favored the latter value so we used this for our analysis, but both settings for λ_2 gave reasonable models that did not include the noise terms. Using the tuning parameters chosen via cross-validation 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 economic 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 the variables lstat and nox. Ravikumar *et al.* (2009) found that SpAM also chose lstat, rm and ptratio plus a crime variable. They found nox to be

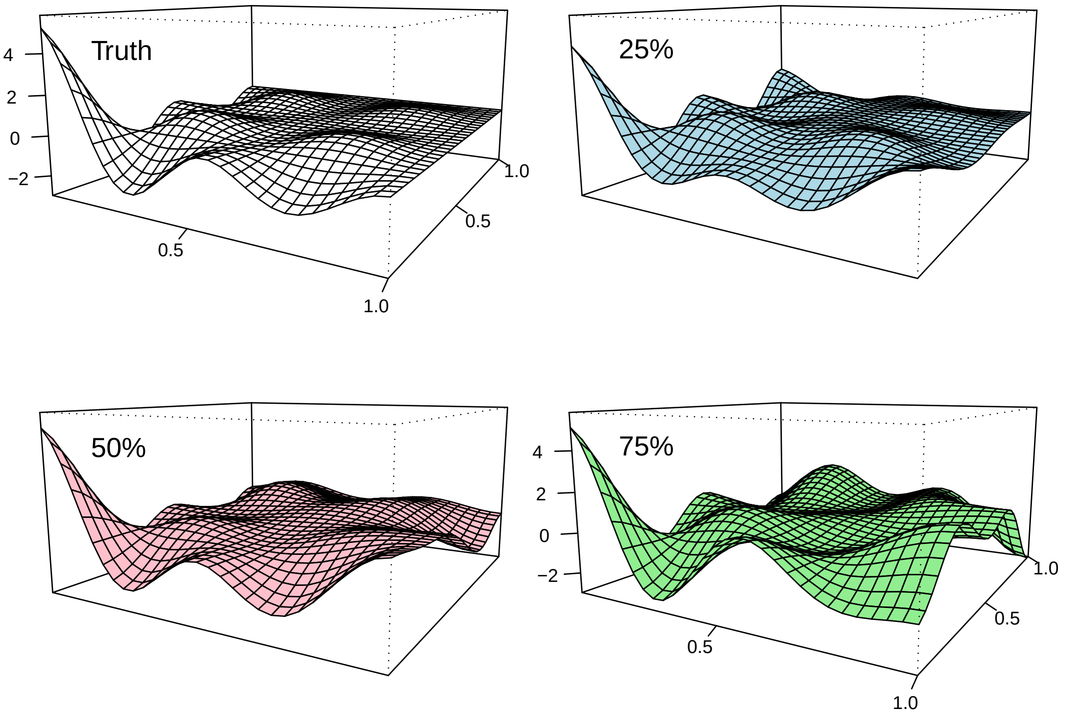

Figure 2: *The true and estimated interaction effects for the last nonlinear simulation.*

borderline. SpAM was not capable of selecting the interaction term.

Figure 3 provides plots of the estimated relationships between rm and house value and between ptratio and house value. The rm variable shows a sharp increase in value moving from 6 to 8 rooms, while ptratio suggests a more gradual decrease in value as the pupil to teacher ratio increases. Figure 4 plots the main effects and interaction combined for the lstat and nox variables. Not surprisingly, there is a significant decrease in housing value for poorer neighborhoods. There is also a decrease in values for more polluted neighborhoods. However, the figure also suggests that the effect of lstat is more pronounced for neighborhoods with higher pollution levels.

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. To test the predictive accuracy of VANISH versus SpIn we fixed the five variable active set for VANISH and the one variable active set for SpIn_{LS}. We then randomly generated 100 partitions of the data into training and test sets. For each training set we used least squares to fit the five variable VANISH and the one variable SpIn_{LS} models. Next we computed the mean squared error (MSE) of each method on the test data. The average MSEs over the 100 data sets were 16.22 for VANISH and 29.15 for SpIn_{LS}, with VANISH superior on 99 of the 100 data sets. Finally, we compared the four variable model including only the main effects to the five variable model including the interaction term. The larger model was superior on approximately 2/3rds of the

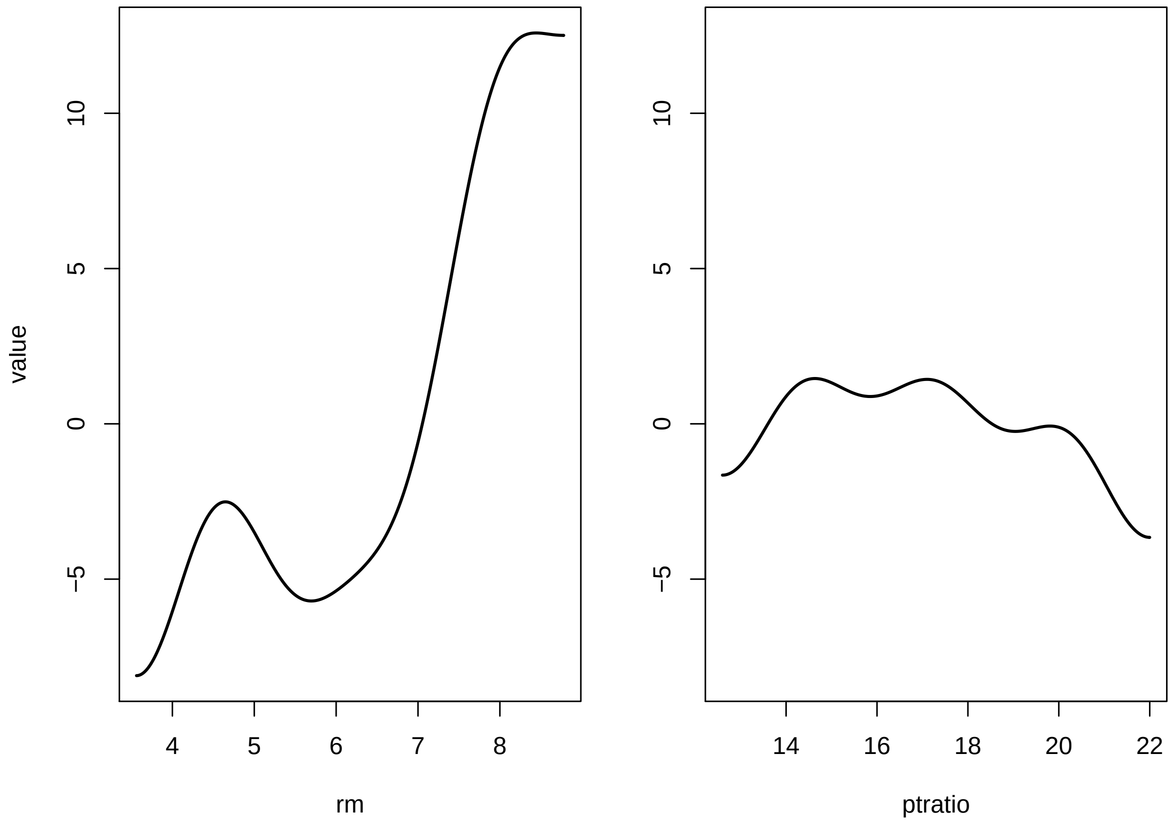

Figure 3: *Estimated main effects from VANISH for the rm and ptratio variables.*

data sets, suggesting that the interaction term was a real effect.

6 Discussion

VANISH is attempting to address the difficult problem of fitting a non-additive, nonlinear model in a high dimensional space. In order to make this problem feasible we model a sparse response surface in terms of the main effects and the interactions. We further assume that all interactions correspond to nonzero main effects. As a simple example, if we let $Y = g_{12}(X_1, X_2) + \epsilon$ then $E_{X_1} g_{12}(X_1, X_2) = f_2(X_2) \neq 0$ and $E_{X_2} g(X_1, X_2) = f_1(X_1) \neq 0$. This assumption seems reasonable and makes the problem far more tractable, because it concentrates the search on a much smaller subset of interactions. Note that VANISH does not prevent interactions entering the model when the corresponding main effects are not currently present, but it does raise the threshold for such terms, significantly lowering the false positive rate.

A simple alternative would be to adapt the SpIn method by imposing different penalties on the main effects and interactions. However, such an approach does not differentiate between "more likely" and "less likely" interactions, based on whether the main effects are present in the model. Hence, in addition to introducing another

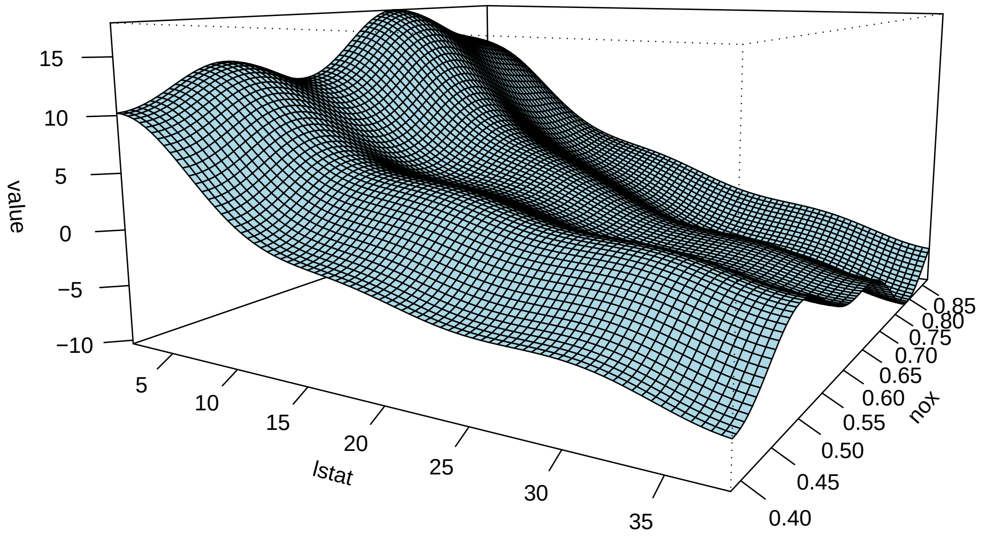

Figure 4: *Estimated two-dimensional response surface, including the main effects and interaction term, for the lstat and nox variables.*

tuning parameter, such an approach would likely either miss true interactions or include noise ones.

VANISH could be extended to higher order interaction terms using a similar penalty function to (7). For example, one could implement VANISH with third order interactions, $\mathbf{f}_{jkl}$, using the following penalty function,

$$\lambda_1 \sum_j \left(\|\mathbf{f}_j\|^2 + \sum_{k:\, k\neq j}^{p} \|\mathbf{f}_{jk}\|^2 \right)^{1/2} + \lambda_2 \sum_{j<k} \left(\|\mathbf{f}_{jk}\|^2 + \sum_{l:\, l\neq j,k}^{p} \|\mathbf{f}_{jkl}\|^2 \right)^{1/2} + \lambda_3 \sum_{j<k<l} \|\mathbf{f}_{jkl}\|.$$

Fitting the corresponding optimization criterion would use a similar algorithm except now we would fit three dimensional functions. The main practical limitation is that one would need to fit of order p^3 terms which may not be possible for large p. More generally, we recently became aware of a large class of "CAP" penalty functions (Zhao *et al.*, 2009). CAP only covers the linear setting but it turns out that the VANISH penalty can be considered as a non-linear generalization of one of the CAP penalties. This connection between CAP and VANISH suggests many other potential non-linear penalty functions for these types of models.

Our theoretical results imply that VANISH should perform best when the model is non-additive but has few interaction terms. Simulation studies 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. Finally, the VANISH fitting algorithm is very efficient, allowing it to search through thousands of non-linear two-dimensional surfaces.

Acknowledgments

We would like to thank the Associate Editor, two referees and Kyle Matoba for their useful comments and suggestions. This work was partially supported by NSF Grants DMS-0705312 and DMS-0906784.

A Proof of Theorem 2

The general approach we follow is similar to the one in the SpAM method paper by Ravikumar, Lafferty, Liu and Wasserman, to which we refer as RLLW from here on. A similar argument was earlier used in the linear case by Wainwright (2009). Our job is complicated by the fact that we consider models with interactions and use a more complex penalty function.

For concreteness we will take the dimension of the two-variate basis exactly equal to d^2 and let D stand for $dp + d^2p(p-1)/2$. The VANISH estimator, $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^D$, is constructed along a path as described in Section 2.3. The ratio of the two tuning parameters is kept fixed throughout the construction, and one of the tuning parameters is decreased along a grid, from a large value that corresponds to an empty model down towards zero. We will assume that in between two grid points there is no more than one change in the active set (here we count an entry of an interaction that brings in the corresponding main effects as only one change). For convenience we introduce a "modified penalty function" for a given index set $\mathcal{K}'$,

$$P_{\mathcal{K}'}(\boldsymbol{\beta}) = \lambda_1' \sum_j \sqrt{\left\|\Psi_j\boldsymbol{\beta}_j\right\|^2 + \sum_{jk\in\mathcal{K}'} \left\|\Phi_{jk}\boldsymbol{\beta}_{jk}\right\|^2} + \sum_{jk} \left(\lambda_1' 1_{\{j\notin\mathcal{K}'\}} + \lambda_1' 1_{\{k\notin\mathcal{K}'\}} + \lambda_2'\right) \left\|\Phi_{jk}\boldsymbol{\beta}_{jk}\right\|,$$

which in turn corresponds to a "modified criterion function". Note that we suppress the dependence of the penalty on λ_1' and λ_2' to simplify the notation. Observe that a VANISH estimator with a given support $\widehat{\mathcal{K}}$ satisfies the heredity constraint and minimizes the modified criterion function corresponding to $P_{\widehat{\mathcal{K}}}$. Write $\mathcal{K}$ for the collection of indexes in $\mathcal{K}_m$ and $\mathcal{K}_{in}$ and denote by $\mathsf{HS}(\mathcal{K})$ the collection of index sets $\mathcal{K}'$ that are subsets of $\mathcal{K}$ and satisfy the heredity constraint. To check that for a particular sample the VANISH estimator corresponding to the tuning parameters λ_1 and λ_2 recovers the correct support, we need to establish two results. First, we need to show that for each $\lambda_1' \geq \lambda_1, \lambda_2' \geq \lambda_2$ and each $\mathcal{K}' \in \mathsf{HS}(\mathcal{K})$ there exists a unique minimizer, $\widehat{\boldsymbol{\beta}}$, of the modified criterion function, for which $\widehat{\mathcal{K}} \subseteq \mathcal{K}$ and the dual feasibility conditions (discussed below) for the subgradient are satisfied as strict inequalities. Second, we need to show that for $\lambda_1' = \lambda_1$, $\lambda_2' = \lambda_2$ and each $\mathcal{K}' \subseteq \mathcal{K}$ we actually have $\widehat{\mathcal{K}} = \mathcal{K}$. The first result would imply that no noise terms can enter the model along the VANISH path up to the point specified by λ_1 and λ_2, while the second result would mean that at that point all the correct terms are in the model.

Note that a vector $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^D$ minimizes the modified criterion function if there exists a subgradient $\widehat{\mathbf{g}}$ that belongs to the subdifferential $\partial P_{\mathcal{K}'}(\widehat{\boldsymbol{\beta}})$ and

$$\Theta^T(\Theta\widehat{\boldsymbol{\beta}} - Y) + \widehat{\mathbf{g}} = 0, \tag{18}$$

where Θ is defined by analogy with $\Theta_{\mathcal{K}}$. Note that we suppress, for simplicity of the notation, the dependence of $\widehat{\mathbf{g}}$ and $\widehat{\boldsymbol{\beta}}$ on λ_1', λ_2' and $\mathcal{K}'$. The above display provides the stationary conditions for our criterion function. The minimizer is unique if the dual feasibility conditions for $\widehat{\mathbf{g}}_{\widehat{\mathcal{K}}^c}$ are satisfied as strict inequalities. We proceed by setting $\widehat{\boldsymbol{\beta}}_{\mathcal{K}^c} = 0$ and defining $\widehat{\boldsymbol{\beta}}_{\mathcal{K}}$ as the unique (due to invertibility of $\Theta_{\mathcal{K}}^T\Theta_{\mathcal{K}}$) minimizer of the modified criterion function restricted to the index set $\mathcal{K}$. We then derive $\widehat{\mathbf{g}}$ from the stationary conditions above. To complete the proof it is sufficient to establish that there exists a set of probability tending to one, on which the strict dual feasibility conditions for $\widehat{\mathbf{g}}_{\mathcal{K}^c}$,

$$\widehat{\mathbf{g}}_j^T(\Psi_j^T\Psi_j)^{-1}\widehat{\mathbf{g}}_j < (\lambda_1')^2, \quad j \notin \mathcal{K}_m \tag{19}$$

$$\widehat{\mathbf{g}}_{jk}^T(\Phi_{jk}^T\Phi_{jk})^{-1}\widehat{\mathbf{g}}_{jk} < \begin{cases} (2\lambda_1' + \lambda_2')^2, & jk \notin \mathcal{K}_{in},\ j \text{ and } k \notin \mathcal{K}_m \\ (\lambda_1' + \lambda_2')^2, & jk \notin \mathcal{K}_{in},\ \text{either } j \text{ or } k \in \mathcal{K}_m \\ (\lambda_2')^2, & jk \notin \mathcal{K}_{in},\ j \text{ and } k \in \mathcal{K}_m, \end{cases} \tag{20}$$

hold for all $\mathcal{K}' \in \mathsf{HS}(\mathcal{K})$ and all $\lambda_1' \geq \lambda_1,\ \lambda_2' \geq \lambda_2$, while inequality

$$\|\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K^*\|_\infty \leq b/2 \tag{21}$$

holds for all $\mathcal{K}' \in \mathsf{HS}(\mathcal{K})$ and $\lambda_1' = \lambda_1,\ \lambda_2' = \lambda_2$.

It follows from the stationary conditions for the modified criterion function restricted to $\mathcal{K}$ that $\widehat{\mathbf{g}}_j^T(\Psi_j^T\Psi_j)^{-1}\widehat{\mathbf{g}}_j \leq (\lambda_1')^2$ and $\widehat{\mathbf{g}}_{jk}^T(\Phi_{jk}^T\Phi_{jk})^{-1}\widehat{\mathbf{g}}_{jk} \leq (2\lambda_1' + \lambda_2')^2$, which in turn implies $\|\widehat{\mathbf{g}}_j\| \leq \lambda_1' C_{\mathcal{K}}^{1/2}$ for $j \in \mathcal{K}_m$ and $\|\widehat{\mathbf{g}}_{jk}\| \leq (2\lambda_1' + \lambda_2')C_{\mathcal{K}}^{1/2}$ for $jk \in \mathcal{K}_{in}$. Note that these bounds hold for all possible $\mathcal{K}' \in \mathsf{HS}(\mathcal{K})$. In the arguments that follow we will refer to $\widehat{\mathbf{g}}_{\mathcal{K}}$ only through these bounds, and hence all of these arguments will hold uniformly over $\mathcal{K}' \in \mathsf{HS}(\mathcal{K})$.

B Conditions (19) and (20)

The argument to establish these inequalities is very similar to the corresponding one in RLLW. The required conditions on the tuning parameters only impose lower bounds on their growth, and hence are satisfied for all λ_1' and λ_2'. The set of probability tending to one can be chosen as the set needed for λ_1 and λ_2, because it would work for the larger tuning parameters as well.

First we consider condition (19). The only major difference with the corresponding proof in RLLW is that we use inequality $\|\widehat{\mathbf{g}}_{jk}\| \leq (2\lambda_1' + \lambda_2')C_{\mathcal{K}}^{1/2}$ for $jk \in \mathcal{K}_{in}$ in addition to $\|\widehat{\mathbf{g}}_j\| \leq \lambda_1' C_{\mathcal{K}}^{1/2}$ for $j \in \mathcal{K}_m$, which results in the bound $\|\widehat{\mathbf{g}}_{\mathcal{K}}\|^2 \leq (s_m(\lambda_1')^2 + s_{in}[2\lambda_1' + \lambda_2']^2)C_{\mathcal{K}}$ as opposed to the bound $\|\widehat{\mathbf{g}}_{\mathcal{K}}\|^2 \leq (\lambda_1')^2 s C_{\mathcal{K}}$. Consequently, the right hand side in condition (14) is $(s_m + s_{in}[2 + \frac{\lambda_2}{\lambda_1}])^{-1/2}$ rather than simply $s^{-1/2}$ as in RLLW.

Next we consider the first inequality in condition (20). Now there are two major differences with RLLW. First, as in the above paragraph, we need to use the bound $\|\widehat{\mathbf{g}}_{\mathcal{K}}\|^2 \leq (s_m(\lambda_1')^2 + s_{in}[2\lambda_1' + \lambda_2']^2)C_{\mathcal{K}}$ instead of the bound $\|\widehat{\mathbf{g}}_{\mathcal{K}}\|^2 \leq (\lambda_1')^2 s C_{\mathcal{K}}$. And second, the inequality we are striving for here is $\|\Psi_j^T \Theta_{\mathcal{K}} \Sigma_{\mathcal{K}}^{-1}\| \leq (2\lambda_1' + \lambda_2') c_{\mathcal{K}}^{1/2} / \|\widehat{\mathbf{g}}_{\mathcal{K}}\|$ rather than $\|\Psi_j^T \Theta_{\mathcal{K}} \Sigma_{\mathcal{K}}^{-1}\| \leq \lambda_1' c_{\mathcal{K}}^{1/2} / \|\widehat{\mathbf{g}}_{\mathcal{K}}\|$. As a result, the right hand side in condition (15) is $(2\lambda_1 + \lambda_2)(s_m \lambda_1^2 + s_{in}[2\lambda_1 + \lambda_2]^2)^{-1/2}$ rather than $s^{-1/2}$ as in RLLW. The reasoning for the second and third inequalities in condition (20) is analogous.

C Condition (21)

Define $\mathbf{V} = \mathbf{Y} - \Theta_{\mathcal{K}} \boldsymbol{\beta}_{\mathcal{K}}^* - \boldsymbol{\epsilon}$, which gives the error due to finite truncation of the orthonormal basis. It follows directly from the stationary conditions that

$$\widehat{\boldsymbol{\beta}}_{\mathcal{K}} - \boldsymbol{\beta}_{\mathcal{K}}^* = \Sigma_{\mathcal{K}}^{-1}(\Theta_{\mathcal{K}}^T[\boldsymbol{\epsilon} + \mathbf{V}] - \widehat{\mathbf{g}}_{\mathcal{K}}). \tag{22}$$

The bounds we derived for $\widehat{\mathbf{g}}_{\mathcal{K}}$ at the end of Appendix A imply

$$\|\Sigma_{\mathcal{K}}^{-1} \widehat{\mathbf{g}}_{\mathcal{K}}\|_\infty \leq \lambda_{min}^{-1}(\Sigma_{\mathcal{K}}) \sqrt{s}(2\lambda_1 + \lambda_2)\sqrt{C_{\mathcal{K}}}. \tag{23}$$

We argue analogously to RLLW, using the assumptions on the rate of decay of the coefficients in $\boldsymbol{\beta}^*$, and derive the bound

$$\|\Sigma_{\mathcal{K}}^{-1} \Theta_{\mathcal{K}}^T \mathbf{V}\|_\infty \leq \|\Sigma_{\mathcal{K}}^{-1}\|_\infty \frac{sn}{d^{3/2}}. \tag{24}$$

Now consider the term $\Sigma_{\mathcal{K}}^{-1} \Theta_{\mathcal{K}} \boldsymbol{\epsilon}$, whose elements have a mean zero gaussian distribution. We again follow the argument in RLLW, using the Gaussian comparison results in Ledoux and Talagrand (1991), and derive

$$E\|\Sigma_{\mathcal{K}}^{-1} \Theta_{\mathcal{K}} \boldsymbol{\epsilon}\|_\infty \lesssim \sigma \sqrt{\frac{\log(sd)}{\lambda_{min}(\Sigma_{\mathcal{K}})}}, \tag{25}$$

where σ^2 is the variance of the error terms, and $\lambda_{min}(\Sigma_{\mathcal{K}})$ is the smallest eigenvalue of $\Sigma_{\mathcal{K}}$. Note that $\|\Sigma_{\mathcal{K}}^{-1}\|_\infty \leq \sqrt{s}d/\lambda_{min}(\Sigma_{\mathcal{K}})$ and $C_{\mathcal{K}}$ is bounded above by $\lambda_{max}(\Sigma_{\mathcal{K}})$, the largest eigenvalue of $\Sigma_{\mathcal{K}}$. Recall that $\lambda_{max}(\Sigma_{\mathcal{K}}) = O(n)$ and $\lambda_{min}^{-1}(\Sigma_{\mathcal{K}}) = O(1/n)$.

Consequently, if we use (22) together with bounds (23) through (25) and apply Markov's inequality, we can bound the probability of the event $\|\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}^*_K\|_\infty > b/2$ by a quantity that goes to zero under the assumption (16). This establishes condition (21) with probability tending to one.

D Proof of Theorem 3

Let ν_n denote the empirical process $h \mapsto \nu_n h = n^{1/2}(\mathbb{P}_n h - \mathbb{E}h)$, where the empirical measure $\mathbb{P}_n$ is defined with respect to the random vector $(Y, X_1, ..., X_p)$. Consider a functional class $\mathcal{G}$ that consists of all pairwise products of the form $yg_j(x_j)$, $yg_{jk}(x_j, x_k)$, $g_j(x_j)g_k(x_k)$, $g_j(x_j)g_{kl}(x_k, x_l)$, and $g_{jk}(x_j, x_k)g_{lm}(x_l, x_m)$, with all the g-functions coming from the Sobolev classes $S^2_C([0,1])$ and $S^2_C([0,1]^2)$. Let $\widehat{R}$ denote the empirical analog of the predictive risk. Note that inequality (17) in the definition of $\mathcal{M}_n$ guarantees $\|\boldsymbol{\beta}\|_1 \leq L_n$. This observation along with an argument analogous to the one in the proof of the persistence theorem in RLLW establish that

$$\sup_{m \in \mathcal{M}_n} |\widehat{R}(m) - R(m)| \leq n^{-1/2}(L_n + 1)^2 \sup_{\mathcal{G}} |\nu_n(\cdot)|.$$

It follows from, for example, Corollary 19.35 of van der Vaart (1998) that $\sup_{\mathcal{G}} |\nu_n(\cdot)|$ can be bounded above by a multiple of the bracketing integral for the functional class $\mathcal{G}$. Bracketing integrals for the Sobolev classes $S^2_C([0,1])$ and $S^2_C([0,1]^2)$ are finite by the classical results of Birman and Solomjak (1967) and hence the bracketing integral for the class of all possible pairwise products of the members of these two classes is finite as well. Because the true regression function is uniformly bounded, there exists a constant c such that $\max_{i \leq n} |Y_i| \leq c\sqrt{\log n}$ with probability tending to one. From here on we restrict our attention to the set where the above inequality is satisfied. On this set the bracketing integral for each of the classes $\{yg_j(x_j)\}$ and $\{yg_{jk}(x_j, x_k)\}$ is of order $(\log n)^{1/4}$. Consequently, the bracketing integral for the functional class $\mathcal{G}$ is of order $(\log p)^{1/2} + (\log n)^{1/4}$, and hence

$$\sup_{m \in \mathcal{M}_n} |\widehat{R}(m) - R(m)| \lesssim L_n^2 n^{-1/2} \Big[(\log p)^{1/2} + (\log n)^{1/4}\Big]. \tag{26}$$

For concreteness we will focus on the case $n = O(p)$; the case $p = o(n)$ can be handled analogously. The right hand side of the bound (26) simplifies to $L_n^2 \sqrt{(\log p)/n}$. Now we can use a comparison argument based on inequalities $R(m_n^*) \leq R(\widehat{m}_n)$ and $\widehat{R}(m_n^*) \geq \widehat{R}(\widehat{m}_n)$, exactly as in RLLW, to deduce $R(\widehat{m}_n) = R(m_n^*) + L_n^2 \sqrt{(\log p)/n}$ and complete the proof.

References

Birman, M. S. and Solomjak, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes w_p^α. *Math. USSR-Sbornik* **2(3)**, 295–317.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Annals of Statistics* **35**, 6, 2313–2351.

Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* **105**, 354–364.

Fan, J., Feng, Y., and Song, R. (2010). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Working Paper* .

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* **70**, 849–911.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 2, 109–135.

Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1**, 302–332.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Greenshtein, E. and Ritov, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Journal of Bernoulli* **10**, 971–988.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and Semiparametric Models.* Springer-Verlag Inc.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management* **5**, 81–102.

James, G. M. and Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika* **96**, 323–337.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces: Isoperimetry and Processes.* Springer-Verlag.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* **34**, 2272–2297.

Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics* **37**, 6B, 3779–3821.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis* **52**, 374–393.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D., Pollack, J., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics* **4**, 53–77.

Radchenko, P. and James, G. M. (2008). Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association* **103**, 1304–1315.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, B.* **71**, 1009–1030.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparcity recover using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory* **55**, 2183–2202.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* **2**, 224–244.

Yuan, M. (2007). Nonnegative garrote component selection in functional anova models. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings* 660–666.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37**, 6A, 3468–3497.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 301–320.