Stat Biosci (2017) 9:676–691
https://doi.org/10.1007/s12561-016-9184-6

Structured Detection of Interactions with the Directed Lasso

Hristina Pashova[1,2] · **Michael LeBlanc**[3] ·
Charles Kooperberg[3]

Received: 16 February 2016 / Accepted: 15 November 2016 / Published online: 29 November 2016

Abstract When considering low-dimensional gene–treatment or gene–environment interactions, we might suspect groups of genes to interact with treatment or environment in a similar way. For example, genes associated with related biological processes might interact with an environmental factor or a clinical treatment in its effect on a phenotype correspondingly. We use the idea of a structured interaction model together with penalized regression to limit the model complexity in a model in which we believe the interactions might behave in a similar way. We propose the directed lasso, a regression modeling strategy using a pairwise fused lasso penalty to encourage interaction model simplicity through fusion of effect size. We compare the performance of the directed lasso to the lasso and other methods in a simulation study and on data sampled from a breast cancer clinical trial.

Keywords Gene–environment interaction · Gene–treatment interaction · Interaction · Lasso · Fusion

✉ Hristina Pashova
hpashova@uw.edu

Michael LeBlanc
mleblanc@fredhutch.org

Charles Kooperberg
clk@fredhutch.org

[1] Department of Biostatistics, University of Washington, F-600 Health Sciences Building, Campus Mail Stop 357232, Seattle, WA 98195, USA

[2] Present Address: Axio Research, 2601 4th Ave Suite 200, Seattle, WA 98121, USA

[3] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N/M3-C102, Seattle, WA 98109, USA

1 Introduction

A common aspect of studying complex genetic associations, and specifically interactions, is that the power to detect them is usually limited. Being able to use the knowledge (or suspicion) about the form of the interactions can put a structure on the type of models that are considered and thereby significantly increase the power to identify such interactions. This idea has been used in other situations, for example, it was used in Tukey's one-degree-of-freedom interaction model for gene–gene and gene–environment interactions [7] and extended, for example, in [18].

The idea that multiple genes may interact with a treatment or environmental factor in a similar manner can be used to create a structure on the model that improves efficiency in identifying gene–environment interactions. For instance, there may be a linear combination of genetic or environmental variables that modify the risk in a similar fashion compared to the main effects. Penalized regression methods have been shown to be an effective tool to enforce such a structure (e.g., [8]).

Most methods for identification of interactions deal with both main effects and interactions in the same (symmetric) way. Instead, we propose to enforce a structure on the model by "fusing" the main effects with the interactions. The idea is to express the regression equation in the form of basis functions and to use a particular form of the basis functions which restrict the form of the interaction to be based on the form of the main effects. In particular, for a single treatment T and multiple genetic effects X, and a continuous outcome Y, we could choose to fit the model

$$Y = \beta_0 + \gamma T + f_1(X) + h f_1(X) \times T + \epsilon, \tag{1}$$

where $f_1(X)$ is modeled by a set of basis functions that depend on X and that could be (but does not need to be) as simple as a linear combination $\sum \beta_i X_i$. If one of the X_i is selected to be in the model for $f_1(X)$ then both the main effect and the interaction are included in the model for Y. A second parameter, h, identifies the strength and direction of the interactions compared to the main effects. Enforcing this structure on the interaction reduces the variance of the model and potentially simplifies its interpretation. We note here that instead of a treatment or environmental factor that interacts with multiple genes, we could have a treatment or a single gene interacting with multiple environmental factors. Therefore, in this paper, we will use T and X rather than G and E in our equations. We focus on a restricted set of predictors as confirmation studies of already preselected groups of gene expression variables.

The simplest model (1) involves a single h parameter, while the most flexible extension of the model could have a separate h for each interaction term and would in fact be equivalent to a full "saturated" model. Grouping the h's may increase the interpretability of the results, and in particular force different X's to interact with T on Y in a similar manner. To achieve grouping, we develop a version of the fused lasso [22] in problem (1) which is a generalization of the lasso (least absolute shrinkage and classification), initially proposed by Tibshirani [21].

Over the last 20 years, many adaptive regression methods have been developed that are specifically designed to identify interactions (e.g., [5,11,12,20]). These methods, however, typically do not make use of the type of information about the form of the interaction as is available in the situation above. During the last few years, regression penalization methods have been developed that are well suited to incorporate such types of information into the modeling [6,21,22]. These methods, however, have not been widely applied to the estimation of interactions.

1.1 Existing Techniques for Detecting Interactions

Our idea builds on the strong heredity interaction model (SHIM) [8] of Choi et al., which is a penalized regression method, that is specifically designed to identify interactions and enforce the strong heredity constraint. For instance, an interaction $X_i X_j$ can only be in the model if both X_i and X_j are in the model. Choi et al. develop an iterative procedure which uses the lasso at each step to fit a model of the form

$$g(X) = \beta_0 + \sum_i \beta_i X_i + \sum_i \sum_j \gamma_{ij} \beta_i \beta_j (X_i X_j).$$

SHIM does not distinguish between types of predictors so we identify them all as X. Note that if $\gamma_{ij} = \gamma$ for all i and j, this would be analogous to the Tukey one-degree-of-freedom model considered by Chatterjee et al. [7].

The SHIM method minimizes the penalized objective function

$$||Y - g(X)||^2 + \lambda_\beta \sum |\beta_j| + \lambda_\gamma \sum |\gamma_{ij}|,$$

with respect to (β, γ). The interaction terms are based on the main effects forcing the interactions to be zero when either main effect is zero. Choi et al. showed that the SHIM model has an asymptotic oracle property as n goes to infinity. As the sample size increases and the number of predictors remains fixed, under regularity conditions, the model performs as well as if the true model is known [8].

Under additional conditions (primarily that the number of predictors grows slow enough relative to the sample size), the same can be shown for the case when both the sample size and the number of predictors tend to infinity. Our approach builds on SHIM; we also force a strong heredity constraint. However, the difference between our proposal and SHIM is that we further enforce a structure on the γ_{ij} that relates the interaction to the main effects and leads to grouping of interaction effects.

An alternative penalized regression approach to identify interactions was proposed by Bien et al. [3]. They propose a lasso-like procedure that produces sparse estimates for the main effects and all two-way interactions, while satisfying heredity constraints. Instead of employing group lasso penalties, they add a set of convex constraints to the lasso model. A related idea is presented by Yuan et al., who propose non-negative garrote methods that can naturally incorporate hierarchical structural relationships between variables [23]. They incorporate the structural relationships as linear constraints on the corresponding penalties. This approach allows them to incorporate a variety of such structural relationships between predictors. Haris et al. develop the method FAMILY, a generalization to Bien et al., for a large set of predictors, enforcing strong heredity [14]. They employ the alternating direction method of multipliers algorithm to efficiently solve the problem. The method explores all possible interaction combinations and deals with large predictor space. Yet a different approach is taken by Lim and Hastie who first screen for candidate main effects and interactions and then do variable selection on the candidate set using a group lasso [16]. While these approaches all use penalized regression to identify interactions, none of them are focused on the more structured gene–treatment or gene–environment interaction problem that we try to solve.

Liu et al. propose the Bayesian mixture model [17] for binary disease status to simultaneously model gene–gene and gene–environment interactions following either strong or weak hierarchical interaction structure. They work with a limited number of main effects by first reducing the predictor space through other methods.

2 Directed Lasso: Lasso with Structured Interactions

We propose the directed lasso, a regression modeling strategy using a fused set of basis functions. Let T be a single treatment or environmental variable (which for convenience we will further refer to as "treatment") and let $X_1, \ldots, X_K$ be K genetic (or environmental) factors which we refer to as genetic below. The method is applicable to both classes of problems (treatment-gene and gene–environment interactions), the key statistical attribute is that the T variable is the univariate measure the multiple X_k are the higher dimensional features. We fuse each main effect X_k and the interaction term TX_k of this effect with a specific effect modifier into a single basis function. The least restrictive case is effectively a reparameterization of the full (saturated) interaction model

$$Y = \beta_0 + \gamma T + \sum_{k=1}^{K} \beta_k (1 + h_k T) X_k. \tag{2}$$

In terms of the traditional multiplicative interaction model, the parameter h_k estimates the strength and direction of the interaction TX_k in the model relative to the main effects X_k. (We assume that if T is not binary that it is normalized to facilitate interpretation of the h_k.) This model is in the same form as the SHIM model, except that not all pairwise interactions are considered. (We note that extensions to non-linear combinations of the X_k are immediate when we replace the X_k in (2) by a set of non-linear basis functions $B_k(X)$, e.g., splines.)

If we believe that some of the genetic terms interact with T in the same way, we can "fuse" the interaction basis functions TX_k for those X_k by letting the relevant h_k be equal to each other. Using such fused basis functions decreases the dimensionality of the model. In the extreme case, where we assume that all X_k interact with T in the same way, we obtain model (1) with $f_1(X) = \sum_k \beta_k X_k$. In this initial formulation, h is global for all the interactions estimated in the model. This is a rather restrictive model formulation.

If we knew a priori how the genetic factors X_k were grouped, we could fit these models using a slight variation of SHIM. In practice, however, we may assume that there is some grouping, but we may not know exactly which h_k are equal. We therefore are proposing a penalized regression method that encourages flexible grouping. In particular, to utilize the above model in our structured interactions scheme, we consider the pairwise fused lasso penalty for the difference between h_k's [19]. This particular regularization penalty will control variance by penalizing the size of the interaction and control the number of groups of interactions. We estimate the h_k's together with the estimation (and selection) of main effects. This way we are able to avoid having to pre-specify the number of groups or group membership in the model. Instead, the penalty we add controls the differences between h_k's and thus naturally encourages the formation of groups of interactions. This is reflected as the second to last term in the objective function below. The last term corresponds to individual predictor selection and shrinkage.

Set $\beta = (\beta_1, \ldots, \beta_K)$ and $h = (h_1, \ldots, h_K)$. Let $\phi = (\gamma, \beta, h)$ and $\hat{\phi}$ be the minimizer of

$$\hat{\phi} = \operatorname{argmin}_{\gamma,\beta,h} \left\| Y - \gamma T - \sum_{k=1}^{K} \beta_k X_k - \sum_{k=1}^{K} h_k \beta_k T X_k \right\|^2 + \lambda_h \left(\alpha \sum_{k=1}^{K} \sum_{j=1}^{K} |h_k - h_j| + \sum_{k=1}^{K} |h_k| \right) + \lambda_\beta \sum_{k=1}^{K} |\beta_k|. \tag{3}$$

Here α, λ_h, and λ_β are pre-specified constants.

2.1 Fitting Directed Lasso Models

To find $\hat{\phi}$ in (3), we can split the minimization problem in two parts and iterate between minimizing each until a solution is reached. In particular, our algorithm is

1. Initialize $\hat{\beta}_k^{(0)}$, $\hat{\gamma}^{(0)}$, and $\hat{h}_k^{(0)}$. Typically we initialize using (unpenalized) least squares estimates for all parameters.
2. Iterate between the following two steps:
   (a) Fix the $\hat{\beta}_k^{(i)}$'s and $\hat{\gamma}^{(i)}$ and estimate the $\hat{h}_k^{(i+1)}$'s by solving

$$\hat{h}^{(i+1)} = \operatorname{argmin}_h \left\| Y - \hat{\gamma}^{(i)} T - \sum_{k=1}^{K} \hat{\beta}_k^{(i)} X_k - \sum_{k=1}^{K} h_k (\hat{\beta}_k^{(i)} T X_k) \right\|^2 + \lambda_h \left(\alpha \sum_{k=1}^{K} \sum_{j=1}^{K} |h_k - h_j| + \sum_{k=1}^{K} |h_k| \right). \tag{4}$$

(b) Estimate the $\hat{\beta}_k^{(i+1)}$'s and the $\hat{\gamma}^{(i+1)}$ for fixed $\hat{h}_k^{(i+1)}$'s by solving

$$\left(\hat{\beta}^{(i+1)}, \hat{\gamma}^{(i+1)}\right) = \text{argmin}_{\gamma,\beta} \left\|Y - \gamma T - \sum_{k=1}^{K} \beta_k (X_k + \hat{h}_k^{(i+1)} T X_k)\right\|^2 + \lambda_\beta \sum_{k=1}^{K} |\beta_k|.$$

3. Stop when

$$\text{diff} = \frac{|M(\hat{\phi}^{(i)}) - M(\hat{\phi}^{(i+1)})|}{|M(\hat{\phi}^{(i)})|}$$

is less then a set small number, where

$$M(\phi) = \left\|Y - \hat{\gamma} T - \sum_{k=1}^{K} \hat{\beta}_k X_k - \sum_{k=1}^{K} \hat{h}_k \hat{\beta}_k T X_k\right\|^2 + \lambda_h \left(\alpha \sum_{k=1}^{K} \sum_{j=1}^{K} |\hat{h}_k - \hat{h}_j| + \sum_{k=1}^{K} |\hat{h}_k|\right) + \lambda_\beta \sum_{k=1}^{K} |\hat{\beta}_k|$$

is the fitted model for $\hat{\phi} = (\hat{\gamma}, \hat{\beta}_1, \ldots, \hat{\beta}_K, \hat{h}_1, \ldots, \hat{h}_K)$.

In step 2(a), the response is $Y - \hat{\gamma} T - \sum_{k=1}^{K} \hat{\beta}_k X_k$, and the predictors are $\hat{\beta}_k T X_k, k = 1, \ldots, K$. Because of the extra penalty on differences between the h parameters, this is not a standard lasso problem; we discuss a fitting algorithm in the next section. Step 2(b) is a standard lasso problem with predictors T, and $X_k + \hat{h}_k T X_k, k = 1, \ldots, K$.

We minimize the objective function with respect to either the set of γ and the β's or the h's and hence the objective function decreases at each step. The value of the objective function is then guaranteed to converge to a local minimum since it is bounded from below. However, similar to many penalized regression problems, convergence to the global optimum is not guaranteed (though in our computations presence of local minima never appeared to be a problem). In addition, we note that the rate of convergence is linear because of the alternating fashion of the minimization problem, so a large number of iterations may be needed, but this is not a practical limitation, as described in Sect. 2.3. The algorithm can be sped up by adding a step (c) where we find a parameter ρ to minimize our objective as a one-dimensional function of $(\hat{\beta}^{(i)}, \hat{\gamma}^{(i)}, \hat{h}^{(i)}) + \rho((\hat{\beta}^{(i+1)}, \hat{\gamma}^{(i+1)}, \hat{h}^{(i+1)}) - (\hat{\beta}^{(i)}, \hat{\gamma}^{(i)}, \hat{h}^{(i)}))$.

2.2 Alternating Direction Method of Multipliers (ADMM)

As mentioned before, the difficult part of solving (3) is the minimization in step 2(a) (Eq. 4). To solve Eq. (4), we use the Alternating Directions Method of Multipliers algorithm (ADMM) [4]. ADMM was developed in the 1970's and is closely related to dual decomposition [10] and the method of multipliers [15]. In recent years, ADMM has been used for some other complicated penalized regression problems (e.g., [9]). The computation of one step of ADMM is on the order of $O([Kn + K^3])$, where n is the sample size and K the number of basis functions.

The ADMM algorithm solves problems of the form

$$\text{minimize}_x (f(x) + g(z))$$

subject to $Ax + Bz = c$, where $x \in \mathbf{R}^n$ and $z \in \mathbf{R}^m$ with $A \in \mathbf{R}^{p\times n}$ and $B \in \mathbf{R}^{p\times m}$. It is assumed that both f and g are convex. The augmented Lagrangian is formed as

$$L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)||Ax + BZ - c||_2^2.$$

Then the algorithm consists of iterating between the following three steps: until convergence,

$$\begin{aligned} x^{k+1} &:= \text{argmin}_x L_\rho\left(x, z^k, y^k\right), \\ z^{k+1} &:= \text{argmin}_z L_\rho\left(x^{k+1}, z, y^k\right), \\ y^{k+1} &:= y^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right), \end{aligned}$$

with $\rho > 0$, a constant chosen a priori. Here x^k, z^k, and y^k are the solutions at the k-th iteration.

To apply ADMM to the pairwise fused lasso [22], we consider the problem

$$\text{minimize}_h \left(\frac{1}{2}||U - Lh||_2^2 + \lambda_h\left(\sum_k |h_k| + \alpha \sum_{1\le k<j\le K} |h_k - h_j|\right)\right), \tag{5}$$

where $U = Y - \gamma T - \sum_{k=1}^K \beta_k X_k$, and $L_k = \beta_k T X_k$. Above, we omit the hat ("$\widehat{\ }$") to simplify notation. Let F be a $p + \binom{p}{2} \times p$ matrix the first p row of which is an identity matrix and the latter part is a representation of all the pairwise differences between the elements of h. It is multiplied by a vector containing p 1's and $\binom{p}{2}$ α's. Therefore

$$F = [1 \ldots 1\, \alpha \ldots \alpha] \begin{bmatrix} 1 & 0 & 0 & 0 & 0\ldots 0 \\ 0 & 1 & 0 & 0 & 0\ldots 0 \\ \vdots & & & & \vdots \\ 0 & \ldots 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 0\ldots 0 \\ -1 & 0 & 1 & 0 & \ldots\ 0 \\ \vdots & & & & \vdots \\ 0 & \ldots 0 & -1 & 0 & 1 \\ 0 & \ldots 0 & 0 & -1 & 1 \end{bmatrix}.$$

Thus, we minimize

$$\text{minimize}_h \left(\frac{1}{2} ||U - Lh||_2^2 + \lambda_h ||Fh|| \right),$$

which in ADMM form looks like

$$\text{minimize}_{h,z} \left(\frac{1}{2} ||U - Lh||_2^2 + \lambda_h ||z|| \right),$$

subject to $Fh - z = 0$.

The three steps in the algorithm which we iterate between are then

$$\begin{aligned} h^{k+1} &:= \left(L^T L + \rho F^T F\right)^{-1} \left(L^T U + \rho F^T (z^k - u^k)\right), \\ z^{k+1} &:= S_{\lambda/\rho} \left(F h^{k+1} + u^k\right), \\ u^{k+1} &:= u^k + F h^{k+1} - z^{k+1}, \end{aligned}$$

where S is a smooth shrinkage function, such as the soft thresholding function $S_{\lambda/\rho}(a) = (a - \lambda/\rho)_+ - (-a - \lambda/\rho)_+$ is interpreted element-wise and u^k is a scaled version of the dual variable. In our simulations, we fixed $\rho = 1$ which gave us a good performance.

2.3 Tuning Parameter Selection

The directed lasso model has three tuning parameters $(\lambda_h, \lambda_\beta, \alpha)$ in Eq. (4). While convergence of the directed lasso problem is linear, as indicted above, the algorithm quite rapidly reaches a region “close” to the optimal solution, which in our experience, it is good enough for cross-validation of the tuning parameters. The tuning parameter α tunes the relative penalty for the h coefficients. If $\alpha = 1$ the h are grouped as much as possible, but are not shrunk to 0; if $\alpha = 0$ the h are not grouped but shrunk to 0. In some situations, it may make sense to set α a priori (see also our real data example); in other situations, it is reasonable to only consider a small number of possible values for α in a three-dimensional grid search. In our simulations, we treat α as a user-chosen constant, as it is primarily a scaling parameter of how much more penalization is applied to the differences between parameters than to the parameters themselves, and as such is mostly application dependent. We explore set values between 0.001 and 10. For each α, a two-dimensional grid search on $(\lambda_h, \lambda_\beta)$ is performed.

3 Breast Cancer Data

The data are generated from a phase III clinical trial for postmenopausal women with node-positive, ER-positive breast cancer and showed that chemo- therapy prior to tamoxifen added survival benefit to tamoxifen alone [1]. Optional tumor banking yielded specimens for gene expression determination by RT-PCR. Data are available on 367 individuals with tumor DNA. The outcome is disease-free survival. As part of follow-up studies, gene expression of a panel of 21 genes that are part a strong predictive factor of chemotherapy benefit compared to the tamoxifen DFS panel were measured. The genes of this panel are thought to be both prognostic and predictive of chemotherapy benefit [2]. Approximately half of the subjects were treated with each of the two treatment options. The goal of our analysis is to see whether some or all of the genes interact with the treatment in influencing the survival time of breast cancer patients. Because of the way the gene expression predictor panel is selected, it is quite conceivable that some genes interact with the treatment in a similar way influencing the outcome.

As our dataset has survival data, a formal analysis would be using either a Cox proportional hazards model, or some other (parametric) survival model. Rather than modifying the ADMM approach, we choose to use a martingale transformation of the survival outcomes that allow us to apply linear regression. For instance, in survival analysis, it is known that regressing martingale residuals from the null Cox model can be used to approximate the functional form of the regression function of the left out covariates (e.g., [13]). So we use the linear regression model to approximate the log hazard models. The martingale residual for the null model is $\delta_i - \hat{\Lambda}(T_i)$ where δ_i is the failure indicator; T_i is follow-up time; and $\hat{\Lambda}(\cdot)$ is the Nelson cumulative hazard estimator. We focus on events within the first 5 years so the maximum follow-up T_i was set at 5 years; 80 out of the 367 subjects are known to have died within five years.

Initial analysis of the dataset suggested that while there were a few suggestive interactions, none of those would be identified (with any flexible variable selection approach we considered). Given that we are analyzing survival data, with notorious low signal-to-noise ratios, and that we are searching for interactions, this is not surprising with a sample size of only 367. To "boost the signal," we decided to double the sample size using resampling. That is, we generate a dataset of size 734 by randomly resampling with replacement 734 observations from the original data. So in this way, this example should be viewed as a realistic but empirically justified simulation study. We note, an additional advantage of this approach is that we can resample multiple times, which allows us to assess the variability. Given we use a bootstrap simulated dataset, results are not expected to replicate prior published results on the 367 cases. To further avoid confusion with respect to the primary paper analysis, we have also chosen not to use the real gene label the individual gene components in this analysis.

With this augmented sample, we apply the directed lasso algorithm in search of groups of interactions. Table 1 gives estimated effects from the directed lasso. We chose the model parameters by fivefold cross-validation and the results are based on the augmented dataset of 734 observations. We also include standard errors for the coefficients, based on 25 bootstrap resamples of the augmented sample. For comparison, we present the lasso model selected by cross-validation and the full model with all main effects and all interactions. We note that the model is attempting to merge some of the interaction terms together, specifically several h's around -28 and 6 h's between -1 and -2, suggesting that the corresponding genes may interact with treatment in a similar manner.

Table 1 Model selection for the directed lasso and the augmented breast cancer data, for tuning parameters selected using cross-validation

Gene	Main effect		Interaction effect		h	Lasso	Full
		SE		SE			
Chemo	0.145	0.181					
p1	0.038	0.007	−0.062	0.010	−1.614	−0.080	−0.117
p2	−0.002	0.008	0.136	0.014	−63.340	0.145	0.198
p3	−0.002	0.005	−0.013	0.008	5.733	0.000	−0.010
p4	−0.002	0.006	0.085	0.012	−39.758	0.099	0.133
p5	0.047	0.006	−0.062	0.008	−1.319	−0.065	−0.089
b1	–	0.003	0.005	–	–	–	−0.003
g1	0.002	0.004	−0.011	0.006	−7.368	−0.010	−0.020
c1	−0.001	0.003	0.020	0.009	−13.447	0.012	0.028
e1	−0.048	0.004	0.053	0.004	−1.105	0.051	0.061
e2	−0.008	0.002	0.010	0.003	−1.306	0.009	0.018
e3	0.002	0.004	−0.041	0.007	−23.170	−0.029	−0.064
e4	0.002	0.002	−0.043	0.004	−28.087	−0.038	−0.046
i1	−0.009	0.003	0.054	0.005	−5.736	0.053	0.057
i2	0.044	0.007	−0.079	0.009	−1.798	−0.090	−0.102
h1	0.001	0.004	−0.039	0.006	−29.473	−0.037	−0.041
h2	0.058	0.005	−0.067	0.009	−1.152	−0.061	−0.082

Standard errors (SE) are based on 25 random bootstrap resamples of the dataset

4 Simulations

In this section, we present results from our simulation studies. For each of the set-ups, we simulate 100 observations from model

$$y = \rho T + \beta^T X + \gamma^T X T + \epsilon, \tag{6}$$

where X are standard normal uncorrelated continuous predictors, $T \sim \text{Bin}(0.6)$ and $\epsilon \sim N(0, 1)$ with $\rho = 1$. The model coefficients as presented in Table 2 and simulations are run 500 times. The models represent a range of potential interaction scenarios. For example, Model 1 has an interaction effect associated each non-zero main effect, while Models 3 and 4 have interactions effects associated with only a subset of the non-zero main effects. Model 5 is a null model with no interaction effects.

For Models 1, 2, and 3, we also examined sample sizes of 75 and 500. For the sample size of 500, we included a situation where all pairwise correlations between predictors, which were normally distributed, were 0.3. All simulations were performed 100 times. We added 15 unrelated predictors and 15 unrelated interactions to Model 1 to explore the performance of the directed lasso with a larger number of predictors. The results are recorded under Model 8.

Table 2 Simulation models: coefficients for interaction models

	β_1 $\vdots$ β_5	β_6 $\vdots$ β_{10}	β_{11} $\vdots$ β_{15}	γ_1 $\vdots$ γ_5	γ_6 $\vdots$ γ_{10}	γ_{11} $\vdots$ γ_{15}
Model 1	2	2	2	1	1	1
Model 2	2	2	2	1	0	0
Model 3	2	2	0	1	0	0
Model 4	2	2	0	0.25	0	0
Model 5	2	0	0	0	0	0
Model 6	2	1	1	1	0	0
Model 7	2	1	1	0.25	0	0

We report mean squared error (MSE) and the number of true positive (TP) and false positive (FP) interaction terms selected by the model based on averages of the simulations for each model set-up. The model is tuned on a training set, and optimal parameters are chosen based on performance on a validation set. A third sample, which is used to measure performance and is not otherwise used, had sample size equal to the training set.

We compare the performance of the directed lasso to the SHIM model, the lasso, and a full unpenalized model (e.g., the directed lasso model with $\lambda_\beta = \lambda_h = 0$). The lasso model was fit in two different ways. First we fit the lasso without any restrictions allowing all main effects and interactions to be included. This often results in fitted models that do not satisfy the heredity constraints; we refer to this approach as “lasso.” We also fit the lasso model, with the restriction that no penalty is applied to the main effects, forcing all of them in the model and automatically satisfying the heredity constraint; we refer to this approach as the “Restricted lasso.”

Table 3 presents the MSE from the seven simulated scenarios. When there is a big discrepancy in the size of main effects and interactions, the directed lasso outperforms SHIM, but the lasso models perform very similarly. When the interactions are half as big as the main effects, the directed lasso performs best. It also performs best when there are no interactions effects, though SHIM and the lasso have similarly good fits. Presumably the additional structured constraints of the directed lasso and SHIM are helpful in this setting, the simple lasso does well due to its good variable selection properties with only a small number of main effects present in the underlying model.

We note that in all but one of the examples with $n = 100$, the directed lasso has the smallest MSE, and in that one scenario, it is close to the lasso. The directed lasso performs much better than SHIM for Models 1, 3, and 6. In Model 3, SHIM has much lower false positive rates but ends up with higher MSE due to setting too often the set of all interactions to 0. The lasso and restricted lasso perform much worse than the directed lasso for Models 2 and 7, where they have much higher false positive rates. The restricted lasso (which includes heredity constraints) always does a little better than the full lasso and does much better when there are main effects that are zero. The full model is never competitive.

Table 3 Simulation results: MSE (SE)

	Dir. lasso	SHIM	Lasso	Res. lasso	Full
$n = 100$, cor $= 0$					
Model 1	**0.229**	0.509	0.502	0.506	0.511
	0.004	0.008	0.008	0.009	0.009
Model 2	**0.352**	0.366	0.435	0.508	0.527
	0.006	0.006	0.008	0.010	0.010
Model 3	0.380	0.475	**0.357**	0.458	0.527
	0.007	0.009	0.007	0.008	0.009
Model 4	**0.245**	0.296	0.315	0.368	0.513
	0.004	0.005	0.005	0.005	0.008
Model 5	**0.166**	0.219	0.197	0.308	0.517
	0.004	0.005	0.004	0.006	0.008
Model 6	**0.322**	0.475	0.428	0.514	0.524
	0.005	0.008	0.007	0.009	0.009
Model 7	**0.252**	0.306	0.405	0.481	0.525
	0.004	0.005	0.007	0.008	0.009
Model 8	**0.491**	0.713	0.932	1.840	2.590
	0.008	0.028	0.015	0.075	0.127
$n = 75$, cor $= 0$					
Model 1	**0.325**	0.793	0.815	0.826	0.901
	0.006	0.015	0.017	0.019	0.038
Model 2	**0.516**	0.528	0.680	0.771	0.888
	0.012	0.011	0.017	0.023	0.027
Model 3	0.562	0.667	**0.492**	0.655	0.856
	0.009	0.011	0.008	0.011	0.018
$n = 500$, cor $= 0$					
Model 1	**0.044**	0.071	0.071	0.071	0.071
	0.001	0.001	0.001	0.001	0.001
Model 2	**0.051**	0.060	0.065	0.069	0.072
	0.001	0.001	0.001	0.001	0.001
Model 3	**0.052**	0.059	0.055	0.065	0.071
	0.001	0.001	0.001	0.001	0.001
Model 8	**0.065**	0.086	0.102	0.131	0.138
	0.001	0.001	0.001	0.001	0.001

When the sample size is further restricted to $n = 75$, a similar performance is achieved. When the sample size is augmented to $n = 500$, the directed lasso performs best; however, the performance of all other models is improved as well. When some correlation is introduced between the predictors, the performance of the directed lasso is best in Models 1 and 2 and the regular lasso is best for Model 3. Model 8 performs similarly to Model 1. Thus, within the range that we explored, the relative perfor-

Table 3 continued

	Dir. lasso	SHIM	Lasso	Res. lasso	Full
$n = 500$, cor $= 0.3$					
Model 1	**0.042**	0.075	0.072	0.072	0.072
	0.001	0.001	0.001	0.001	0.001
Model 2	**0.055**	0.058	0.060	0.068	0.071
	0.001	0.001	0.001	0.001	0.001
Model 3	0.055	0.057	**0.049**	0.063	0.069
	0.001	0.001	0.001	0.001	0.001

Uncorrelated predictors “Full” is the full regression model that includes all predictors and interactions. “lasso” is the lasso model without any constraints. “Res. lasso” is the lasso model where the main effects are not penalized. “Dir. lasso” is the directed lasso. The boldfaced results are the best for a particular model

Table 4 MSE for the interactions only

	Dir. lasso	SHIM	Lasso	Res. lasso	Full
$n = 100$, cor $= 0$					
Model 1	**0.005**	0.061	0.066	0.066	0.067
Model 2	**0.032**	0.035	0.047	0.067	0.071
Model 3	0.039	0.058	**0.034**	0.061	0.070
Model 4	**0.014**	0.018	0.025	0.043	0.067
Model 5	0.003	**0.002**	0.009	0.028	0.067
Model 6	**0.025**	0.062	0.046	0.069	0.070
Model 7	**0.011**	0.023	0.038	0.062	0.070
Model 8	**0.006**	0.032	0.046	0.126	0.188
$n = 75$, cor $= 0$					
Model 1	**0.007**	0.095	0.109	0.112	0.121
Model 2	0.052	**0.051**	0.073	0.098	0.117
Model 3	0.069	0.082	**0.046**	0.089	0.119
$n = 500$, cor $= 0$					
Model 1	**0.002**	0.009	0.009	0.009	0.009
Model 2	**0.004**	0.006	0.006	0.008	0.009
Model 3	**0.005**	0.006	0.005	0.008	0.009
Model 8	**0.001**	0.004	0.005	0.009	0.009
$n = 100$, cor $= 0.3$					
Model 1	**0.002**	0.011	0.012	0.012	0.012
Model 2	**0.007**	0.008	0.008	0.012	0.013
Model 3	0.007	0.008	**0.005**	0.010	0.012

mance of the methods does not seem to depend much on sample size or correlation structure.

In Table 4, we show the mean squared error for the coefficients of the interactions. Since in all our scenarios we group interactions, it is not surprising that the directed

Table 5 Simulation results: average true positive and false positive coefficients for uncorrelated models

		Dir. lasso	SHIM	Lasso	Res. lasso	Full
$n = 100$, cor $= 0$						
Model 1	TP	1.00	1.00	1.00	1.00	1.00
Model 2	TP	1.00	1.00	1.00	1.00	1.00
	FP	0.25	0.55	0.71	0.98	1.00
Model 3	TP	1.00	1.00	1.00	1.00	1.00
	FP	0.98	0.54	0.59	0.99	1.00
Model 4	TP	0.84	0.48	0.83	0.49	1.00
	FP	0.55	0.16	0.55	0.98	1.00
Model 5	FP	0.36	0.04	0.38	0.73	1.00
Model 6	TP	1.00	1.00	1.00	1.00	1.00
	FP	0.27	0.68	0.71	0.99	1.00
Model 7	TP	0.86	0.30	0.86	0.85	1.00
	FP	0.65	0.15	0.67	0.97	1.00
Model 8	TP	1.00	1.00	1.00	0.99	1.00
	FP	0.84	0.01	0.54	0.95	1.00
$n = 74$, cor $= 0$						
Model 1	TP	1.00	1.00	1.00	1.00	1.00
Model 2	TP	0.99	1.00	1.00	1.00	1.00
	FP	0.27	0.54	0.71	0.88	1.00
Model 3	TP	1.00	0.99	1.00	0.99	1.00
	FP	0.64	0.41	0.56	0.90	1.00
$n = 500$, cor $= 0$						
Model 1	TP	1.00	1.00	1.00	1.00	1.00
Model 2	TP	1.00	1.00	1.00	1.00	1.00
	FP	0.09	0.51	0.72	0.90	1.00
Model 3	TP	1.00	1.00	1.00	1.00	1.00
	FP	0.24	0.51	0.58	0.87	0.99
Model 8	TP	1.00	1.00	1.00	1.00	1.00
	FP	0.83	0.00	0.49	0.91	0.99
$n = 500$, cor $= 0.3$						
Model 1	TP	1.00	1.00	1.00	1.00	1.00
Model 2	TP	1.00	1.00	1.00	1.00	1.00
	FP	0.20	0.53	0.59	0.90	0.99
Model 3	TP	1.00	1.00	1.00	1.00	1.00
	FP	0.38	0.50	0.44	0.83	0.99

"Dir. lasso" is the directed lasso

lasso, which groups interactions, is the best model, often by a substantial amount. The one exception is Model 5, in which there actually are no interactions at all.

To estimate the true positive (TP) (Table 5) coefficients each model selects, we average the number of true non-zero interaction coefficients that are estimated to be larger than 0.001 and average this over all simulations. This threshold was chosen to be small relative to the true underlying effects in the simulation study. Similarly, false positives (FP) are the average of the zero interaction coefficients which are estimated to be larger than 0.001 by the model, averaged over all simulation runs for each simulated scenario.

Interestingly, when the interaction coefficients are much smaller than the main effects, as is the case with Model 7, directed lasso outperforms the other methods in terms of RSS, but it has similar performance in terms of TP and FP. When the interaction terms are larger, as is the case in Model 6, then the directed lasso also has the best FP rate and best RSS.

As we expect, the Lasso model does not necessarily satisfy hereditary constrains in Models 3, 4, and 5, where some of the main effects are set exactly to 0. In only 28, 7, and 22% of the fitted Lasso models, for simulations of Model 3, 4, and 5, respectively, the models ended up satisfying heredity constraints. Within the fitted models that did not satisfy hereditary constrains, on average 1.7 interactions were fitted without main effects in Models 3 and 4 and 2.5 for Model 5.

5 Discussion

The directed lasso is a flexible interaction regression method, which utilizes model structure assumptions when appropriate to increase the power of identifying interactions. The directed lasso is designed for instances where we want to link the main effects and the interactions effects. We can impose constraints on how the interaction effects are associated with the main effects and control that relationship via one or more penalty parameters. In addition, we anticipate, there will advantages with respect to estimating interactions using this modeling strategy over unconstrained methods are when there are groups of interactions that modify the main effects in a similar fashion. We have shown that this is indeed the case in some simulated examples. In the context of biomedical studies, this is a plausible scenario when, for example, we have a treatment effect and we are investigating the interactions between the treatment and a group of genetic attributes. SNPs that are located on the same gene or genes that are associated with a similar process are likely to modify the treatment effect in a similar way. We found that the biggest gains for our modeling strategy are found when there is a group of factors with medium to large interaction effects.

Acknowledgements We thank Daniela Witten for introducing us to the ADMM algorithm. We thank Bill Barlow for access to the breast cancer dataset which was used to generate the simulation sample dataset analyzed in Sect. 3. This work was supported in part by National Institutes of Health Grants R01 HG006124, R01 CA90998, R01 HL114901, and P01 CA53996.

References

1. Albain K, Barlow W, O'Malley F et al (2005) Concurrent versus sequential chemohormonal therapy versus tamoxfen alone for postmenopausal, node-positive, ER and/or PgR-positive breast cancer: mature outcomes and new biologic correlates on phase III intergroup trial 0100 (S8814). Breast Cancer Res Treat 90:95

2. Albain K, Barlow W, Shak S, Hortobagyi G, Livingston R, Yeh I et al (2010) Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. Lancet Oncol 11(1):55–65
3. Bien J, Taylor J, Tibshirani R (2013) A lasso for hierarchical interactions. Ann Stat 41(3):1111–1141
4. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 3:1–122
5. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. CRC Press, Boca Raton
6. Bühlmann P (2006) Boosting for high-dimensional linear models. Ann Stat 34(2):559–583
7. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S (2006) Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 79(6):1002–1016
8. Choi N, William L, Zhu J (2010) Variable selection with the strong heredity constraint and its oracle property. J Am Stat Assoc 105(489):354–364
9. Danaher P, Wang P, Witten D (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc Ser B 76(2):373–397
10. Everett H (1963) Generalized lagrange multiplier method for solving problems of optimum allocation of resources. Oper Res 11(3):399–417
11. Friedman J (1991) Multivariate adaptive regression splines. Annu Stat 19(1):1–67
12. Friedman J, Stuetzle W (1981) Projection pursuit regression. J Am Stat Assoc 76(376):817–823
13. Grambsch P, Therneau T, Fleming T (1995) Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. Biometrics 51:1469–1482
14. Haris A, Witten D, Simon N (2016) Convex modeling of interactions with strong heredity. J Comput Graph Stat 25(4):981–1004. doi:10.1080/10618600.2015.1067217
15. Hestenes M (1969) Multiplier and gradient methods. J Optim Theory Appl 4:302–320
16. Lim M, Hastie T (2013) Learning interactions via hierarchical group-lasso regularization. arXiv:1308.2719
17. Liu C, Ma J, Amos C (2015) Bayesian variable selection for hierarchical gene-environment and gene-gene interactions. Hum Genet 134:23–36
18. Maity A, Carroll R, Mammen E, Chatterjee N (2009) Testing in semiparametric models with interaction, with applications to gene-environment interactions. J R Stat Soc Ser B 71(1):75–96
19. Petry S, Flexeder C, Tutz G (2011) Pairwise fused lasso. Technical report, University of Munich, Munich
20. Ruczinski I, Kooperberg C, LeBlanc M (2003) Logic regression. J Comput Graph Stat 12(3):475–511
21. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288
22. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B 67(1):91–108
23. Yuan M, Joseph R, Zou H (2009) Structured variable selection and estimation. Ann Appl Stat 3(4):1738–1757