
The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

# A Note on High-Dimensional Linear Regression With Interactions

Ning Hao & Hao Helen Zhang

**To cite this article:** Ning Hao & Hao Helen Zhang (2017) A Note on High-Dimensional Linear Regression With Interactions, The American Statistician, 71:4, 291-297, DOI: 10.1080/00031305.2016.1264311

**To link to this article:** https://doi.org/10.1080/00031305.2016.1264311

Accepted author version posted online: 15 Dec 2016.
Published online: 15 Dec 2016.

THE AMERICAN STATISTICIAN
2017, VOL. 71, NO. 4, 291–297: General
https://doi.org/10.1080/00031305.2016.1264311

Taylor & Francis
Taylor & Francis Group

GENERAL



# A Note on High-Dimensional Linear Regression With Interactions

Ning Hao and Hao Helen Zhang

Department of Mathematics, University of Arizona, Tucson, AZ

**ABSTRACT**

The problem of interaction selection in high-dimensional data analysis has recently received much attention. This note aims to address and clarify several fundamental issues in interaction selection for linear regression models, especially when the input dimension $p$ is much larger than the sample size $n$. We first discuss how to give a formal definition of "importance" for main and interaction effects. Then we focus on two-stage methods, which are computationally attractive for high-dimensional data analysis but thus far have been regarded as heuristic. We revisit the counterexample of Turlach and provide new insight to justify two-stage methods from the theoretical perspective. In the end, we suggest new strategies for interaction selection under the marginality principle and provide some simulation results.

**ARTICLE HISTORY**
Received December 2014
Revised October 2016

**KEYWORDS**
Heredity condition; Hierarchical structure; Interaction effects; Linear model; Marginality principle



## 1. Introduction

Given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, which is made up of independent and identically distributed copies of $(\mathbf{X}, Y)$, where $\mathbf{X} = (X_1, \ldots, X_p)^\top$ is the $p$-dimensional vector of covariates and $Y$ is the response variable, a standard linear regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{1}$$

where $\varepsilon$ is the random error term. In complex systems, the predictors often interact and their interaction effects may play a crucial role in model prediction and interpretation. Historically, models with two- or higher-order interaction terms have been considered for linear models and generalized linear models (Nelder 1977; McCullagh and Nelder 1989; Nelder 1994; McCullagh 2002), polynomial regression (Peixoto 1987, 1990), experimental designs (Hamada and Wu 1992; Chipman 1996; Chipman, Hamada, and Wu 1997), among others. A linear model with two-way interaction effects can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_{11} X_1^2 + \gamma_{12} X_1 X_2 + \cdots + \gamma_{pp} X_p^2 + \varepsilon, \tag{2}$$

where $\beta_0$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{pp})^\top$ are unknown regression parameters. In model (2), $X_1$,..., $X_p$ are *main* effects, $X_j^2$ $(1 \le j \le p)$ and $X_j X_k$ $(1 \le j < k \le p)$ are quadratic and two-way interaction effects, respectively. We refer to all of the degree-two terms as *interactions*. A special feature about model (2) is the intrinsic relationship among regressor terms, that is, $X_j X_k$ is the *child* of $X_j$ and $X_k$, and, $X_j$ and $X_k$ are *parents* of $X_j X_k$. This type of model structure is known as *hierarchy* or the *hierarchical* structure.

Historically, Nelder (1977) pointed out the importance of maintaining the hierarchical structure when identifying important effects in interaction models. He suggested using *the marginality principle*, which requires any interaction term be considered for selection only after its parents have entered the model. Nelder (1994) further provided a clear explanation for this principle as follows.

"When we fit sequences of quantitative terms such as $x_1, x_2, x_1x_2, x_1^2, x_2^2, \ldots$, we have to ask which sequences make sense. If we fit $x_1$ without an intercept, then the response must go through the origin, that is, zero must be a special point on the $x$-scale where $y$ is zero. Similarly, if $x_1^2$ is fitted without an $x_1$ term then the turning-point must occur at the origin (not impossible, but very unlikely). For if $x_1$ might just as well be $x_1 - a$ then $(x_1 - a)^2 = x_1^2 - 2ax_1 + a^2$ and the linear term reappears. Both terms must be fitted in the order $x_1$, then $x_1^2$, and we say that $x_1$ is $f$-marginal to $x_1^2$. With two continuous variables $x_1$ and $x_2$, new effects arise: if $x_1x_2$ is fitted without $x_1$ and $x_2$ then the response surface must be centered on a col (saddlepoint) for the process to make sense. In general, there is no reason to expect such a centering to occur, so $x_1$ and $x_2$ must be fitted before $x_1x_2$."

With the same spirit, Peixoto (1990) suggested that a well-formulated model should be invariant under simple coding transformations. The model $f(x_1, x_2) = \beta_0 + \gamma_{12}x_1x_2$ is not invariant, since one or more linear terms can be brought into the model after some coding transformation. For example, the transformation $\tilde{x}_1 = x_1 - 1$ will lead to $f(\tilde{x}_1, x_2) = \beta_0 + \gamma_{12}x_2 + \gamma_{12}\tilde{x}_1x_2$. Therefore, it is not sensible to consider the model $\{1, X_1X_2\}$ without $X_1$ or $X_2$.

In modern biological and medical research, gene–gene interactions, also called *epistatic effects*, and gene–environment interactions have been intensively studied in genome-wide association studies (GWAS; Evans et al. 2006; Manolio and Collins 2007; Kooperberg and LeBlanc 2008; Cordell 2009). To deal with such large and complex datasets, variable selection has been under rapid development over the past two decades. A comprehensive overview of modern variable selection theory and methods was given by Fan and Lv (2010) and the book by Bühlmann

**CONTACT** Hao Helen Zhang hzhang.work@gmail.com Department of Mathematics, University of Arizona, Tucson, AZ 85721.

and van de Geer (2011). Lately, research on interaction selection has been revived in the context of high-dimensional data analysis; see the recent works by Efron et al. (2004), Yuan, Joseph, and Zou (2009), Zhao, Rocha, and Yu (2009), Choi, Li, and Zhu (2010), Bien, Taylor, and Tibshirani (2013), and Hao and Zhang (2014). When the data dimension $p$ is comparable to or much larger than the sample size $n$, the problem of interaction selection for model (2) faces a number of challenges. Computationally, there are a total of $d = (p^2 + 3p)/2$ predictors, and therefore the number of candidate models $2^d$ can be enormously large and prohibitive for standard software. Second, to maintain the hierarchical structure in the final model, special optimization constraints are needed during the selection process, as suggested by several authors including Zhao, Rocha, and Yu (2009) and Yuan, Joseph, and Zou (2009). However, constrained programming demands high computational cost and is typically infeasible for large $p$. Furthermore, it is very challenging to study statistical inference and asymptotic properties of an interaction selection method, since interaction effects have much more complex covariance structures than main effects.

In this note, we first discuss some fundamental issues in interaction selection for model (2) in high-dimensional settings. When $p$ is extremely large, two-stage methods are possibly the only feasible choices in practice. However, the lack of a theoretical foundation for two-stage methods has been an issue for a long time in the literature. One goal of this note is to shed new light, at least in some situations, on the validity of two-stage methods. Furthermore, we discuss the marginality principle and suggest new strategies that are feasible for high-dimensional interaction selection. Throughout this note, we assume that $\mathbf{X} = (X_1, \ldots, X_p)^\top$ is a random vector following a continuous distribution $\mathcal{F}$, and the noise term $\varepsilon$ follows $\mathcal{N}(0, \sigma^2)$ independent of $\mathbf{X}$ with unknown variance $\sigma^2$.

## 2. Definition of "Importance"

We consider how to define *important effects* in a regression model. The answer is simple for a standard linear model containing main effects only, but not so straightforward for model (2) due to its hierarchical structure. In the following, we first discuss the invariance principle and then suggest a proper definition of *importance* for a model containing interaction terms.

### 2.1. Invariance Principle

In model (1), when $p$ is large, it is common to assume that the true model is sparse, that is, only a small number of variables are relevant to the response. Intuitively, the *relevance* or *importance* of a variable $X_j$ is reflected by the magnitude of its regression coefficient $\beta_j$. Formally, $X_j$ is *important* if and only if $\beta_j \neq 0$. Variable selection aims to identify the set of important variables, that is, the support of the coefficient vector $\boldsymbol{\beta}$, denoted by $\mathcal{S}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0,\ j = 1, \ldots, p\}$. For convenience, we define $\operatorname{sign}(\boldsymbol{\beta}) = (\operatorname{sign}(\beta_1), \ldots, \operatorname{sign}(\beta_p))^\top$.

In practice, it is common to center and rescale the predictors before variable selection is conducted. For example, before a shrinkage method like the LASSO (Tibshirani 1996) is applied, the predictors are usually standardized to have zero mean and unit variance so that they are on the same scale and their regression coefficients are directly comparable. Therefore, a proper definition of "importance" should satisfy the *invariance principle* with respect to the coding transformation of covariates (Peixoto 1990). To elaborate, consider the transformation $\tilde{X}_j = a_j(X_j - c_j)$ for $j = 1, \ldots, p$, where $a_j > 0$ and $c_j$ are arbitrary constants. Under this transformation, model (1) becomes

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1\tilde{X}_1 + \cdots + \tilde{\beta}_p\tilde{X}_p + \varepsilon = \left(\beta_0 + \sum_{j=1}^{p}\beta_j c_j\right) + a_1^{-1}\beta_1\tilde{X}_1 + \cdots + a_p^{-1}\beta_p\tilde{X}_p + \varepsilon.$$

It is clear that $\tilde{\beta}_j = a_j^{-1}\beta_j \neq 0$ if and only if $\beta_j \neq 0$. Furthermore, $\operatorname{sign}(\tilde{\boldsymbol{\beta}}) = \operatorname{sign}(\boldsymbol{\beta})$. Therefore, the definitions of $\mathcal{S}(\boldsymbol{\beta})$ and $\operatorname{sign}(\boldsymbol{\beta})$ both satisfy the invariance principle.

For high-dimensional variable selection, a number of model consistency criteria have been recently suggested to study asymptotic properties of an estimator $\hat{\boldsymbol{\beta}}$, including sure screening consistency (Fan and Lv 2010), model selection consistency (Fan and Li 2001; Zou 2006), and sign consistency (Zhao and Yu 2006). These three types of consistency amount to, with high probability, $\mathcal{S}(\hat{\boldsymbol{\beta}}) \supset \mathcal{S}(\boldsymbol{\beta})$, $\mathcal{S}(\hat{\boldsymbol{\beta}}) = \mathcal{S}(\boldsymbol{\beta})$, and $\operatorname{sign}(\hat{\boldsymbol{\beta}}) = \operatorname{sign}(\boldsymbol{\beta})$, respectively. Due to the invariance property of $\mathcal{S}(\boldsymbol{\beta})$ and $\operatorname{sign}(\boldsymbol{\beta})$, these consistency properties are also invariant under any coding transformation on predictors.

### 2.2. "Importance" of Interactions

We now define *important main effects* and *important interaction effects* for model (2). First, we point out that the traditional definition $\beta_j \neq 0$ or $\operatorname{sign}(\beta_j) \neq 0$ for "important main effects" is no longer proper for model (2), since it violates the invariance principle. This can be illustrated with Turlach's data-generating process (Turlach 2004),

$$Y = (X_1 - 0.5)^2 + X_2 + X_3 + X_4 + X_5 + \varepsilon. \tag{3}$$

Model (3) can be expressed in the following three different but equivalent equations,

$$\begin{aligned} Y &= X_1^2 - \mathbf{1}X_1 + \frac{1}{4} + X_2 + X_3 + X_4 + X_5 + \varepsilon, \\ Y &= \tilde{X}_1^2 + \mathbf{0}\tilde{X}_1 + X_2 + X_3 + X_4 + X_5 + \varepsilon, \\ &\quad \text{with} \quad \tilde{X}_1 = X_1 - 0.5, \\ Y &= \hat{X}_1^2 + \mathbf{1}\hat{X}_1 + \frac{1}{4} + X_2 + X_3 + X_4 + X_5 + \varepsilon, \\ &\quad \text{with} \quad \hat{X}_1 = X_1 - 1, \end{aligned}$$

where the last two expressions result from a simple coding transformation $X_1 - c$. In these three expressions, the coefficient of the first main effect is $-1$, 0, and 1, respectively. This would lead to three different interpretations about the effect of $X_1$: positive, null, or negative. Which one is correct? The answer depends on the coding system. The cause of inconsistent interpretations is that $X_1^2$ is a function of $X_1$. In general, as long as $\gamma_{jk} \neq 0$, there always exists some linear transformation that can change the signs of $\beta_j$ and $\beta_k$, making them independently positive, negative, or zero. Furthermore, under model (2), neither the support

$\mathcal{S}(\boldsymbol{\beta})$ nor $\text{sign}(\boldsymbol{\beta})$ is invariant under a coding transformation. It is problematic as the three expressions correspond to the same model. In general, the invariance principle can be violated when some deterministic relationship exists among predictors.

Next, we propose a proper definition for *important effects* in model (2) that satisfies the invariance principle.

*Definition 1.* For the data-generating process (2), we say that $X_j$ is important if and only if $\beta_j^2 + \sum_{k=1}^p \gamma_{jk}^2 > 0$, and $X_jX_k$ is important if $\gamma_{jk} \neq 0$. We define the set of important main effects by $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \{j : \beta_j^2 + \sum_{k=1}^p \gamma_{jk}^2 > 0,\ j = 1, \ldots, p\}$. The sign of main effects is defined as $\text{sign}(\boldsymbol{\beta})$ under any parameterization with $\text{E}(X_j) = 0,\ j = 1, \ldots, p$.

We show that Definition 1 is invariant under any coding transformation. Under the coding transformation $\tilde{X}_j = a_j(X_j - c_j)$ with $a_j > 0$, model (2) can be expressed as

$$Y = \left(\beta_0 + \sum_{j=1}^p \beta_j c_j + \sum_{1\le j\le k\le p} \gamma_{jk} c_j c_k\right) + \sum_{j=1}^p a_j^{-1}\left(\beta_j + \sum_{k=1}^p \gamma_{jk} c_k\right)\tilde{X}_j + \sum_{1\le j\le k\le p} \gamma_{jk} a_j^{-1} a_k^{-1} \tilde{X}_j \tilde{X}_k,$$

where $\gamma_{jk} = \gamma_{kj}$ for $j > k$. Under the new parameterization, we have

$$\begin{aligned}
\tilde{\beta}_0 &= \beta_0 + \sum_{j=1}^p \beta_j c_j + \sum_{1\le j\le k\le p} \gamma_{jk} c_j c_k,\\
\tilde{\beta}_j &= \sum_{j=1}^p a_j^{-1}\left(\beta_j + \sum_{k=1}^p \gamma_{jk} c_k\right),\\
\tilde{\gamma}_{jk} &= \gamma_{jk} a_j^{-1} a_k^{-1}.
\end{aligned}$$

It is easy to check the following facts:

(i) $\text{sign}(\tilde{\gamma}_{jk}) = \text{sign}(\gamma_{jk})$

(ii) $\beta_j^2 + \sum_{k=1}^p \gamma_{jk}^2 = 0, \Longleftrightarrow \beta_j = 0, \gamma_{jk} = 0, \forall j, k. \Longleftrightarrow \tilde{\beta}_j = 0, \tilde{\gamma}_{jk} = 0, \forall j, k. \quad \Longleftrightarrow \tilde{\beta}_j^2 + \sum_{k=1}^p \tilde{\gamma}_{jk}^2 = 0.$

Throughout this article, we focus on any parameterization obtained by a coding transformation from the original data. In the following proposition, we show that the sign of main effects is well-defined under Definition 1.

*Proposition 1.*

1. A main effect $X_j$ is important if and only if $\beta_j \neq 0$ or $\gamma_{jk} \neq 0$ for some $k$, under arbitrary parameterization. In particular, $\mathcal{S}(\boldsymbol{\beta}) \subset \mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma})$.
2. If $X_jX_k,\ j \neq k$ is important, so are its parent effects $X_j$ and $X_k$. If $X_j^2$ is important, so is $X_j$.
3. Assume $\text{E}(X_j) = 0$ for all $1 \le j \le p$. Under a scale transformation $\tilde{X}_j = a_jX_j$ with $a_j > 0$, we have $\text{sign}(\tilde{\beta}_j) = \text{sign}(\beta_j)$ for $j = 1, \ldots, p$.

"Important effects" in Definition 1 are valid and well-defined, as they eliminate possible inconsistent interpretations due to a coding transformation. More importantly, they provide us a rigorous framework to study theoretical properties of a variable selection procedure.

## 3. Myths About Two-Stage Methods

In the literature, there are two main types of interaction selection procedures: *one-stage* methods and *two-stage* methods. One-stage methods select main and interaction effects simultaneously, subject to a hierarchical constraint. Examples include several shrinkage methods such as Zhao, Rocha, and Yu (2009), Yuan, Joseph, and Zou (2009), Choi, Li, and Zhu (2010), and Bien, Taylor, and Tibshirani (2013). These methods use asymmetric penalty functions and inequality constraints to maintain the model hierarchy in the selection process. For the setting $p < n$, these estimators possess nice theoretical properties such as model selection consistency and oracle properties. However, the computational cost of one-stage methods can be very high or even infeasible for a large $p$ setting, as they require the solution of large scale and complex optimization problems. In contrast, two-stage methods are more attractive for high-dimensional problems, especially with $p \gg n$, due to their scalable computational algorithms (Wu et al. 2009, 2010).

Two-stage methods are widely used in practice, for example, in genomics data analysis. However, they are usually regarded as heuristic procedures because their theoretical foundation is not clearly understood. In addition, Turlach (2004) constructed a counterexample that cast further doubt on consistency of two-stage estimators. In the following, we will first revisit this counterexample to better understand the mechanism of two-stage methods and why they fail in this example. We then discuss situations where two-stage methods can actually be justified theoretically.

### 3.1. Turlach's Counterexample

Two-stage methods keep the model hierarchy in a natural selective manner by circumventing complex constraints on the models, and therefore they have computational advantages over one-stage methods. For example, Efron et al. (2004) suggested a two-stage least angle regression (LARS) for interaction selection. At stage one, only main effects are selected from model (1). Denote the set of selected main effects by $\widehat{\mathcal{M}} \subset \{1, \ldots, p\}$. At stage two, the two-stage LARS considers only interactions of those main effects in $\widehat{\mathcal{M}}$ and selects interaction terms based on the following reduced model

$$Y = \beta_0 + \sum_{j\in\widehat{\mathcal{M}}} \beta_j X_j + \sum_{j,k\in\widehat{\mathcal{M}};\, j\le k} \gamma_{jk} X_j X_k + \varepsilon. \qquad (4)$$

Since two-stage methods conduct variable selection under a misspecified model at stage one (by intentionally leaving out interaction effects), there has been doubt on their theoretical validity in the literature. Furthermore, Turlach (2004) constructed the following counterexample for the two-stage LARS by considering the data-generating process (3),

$$Y = (X_1 - 0.5)^2 + X_2 + X_3 + X_4 + X_5 + \varepsilon,$$

where $X_1, \ldots, X_{10}$ are independent and identically distributed from Unif$[0, 1]$, and they are independent of $\varepsilon$. Five variables, $X_1, \ldots, X_5$, are present in model (3). Because $\text{cov}(Y, X_1) = 0$, the two-stage LARS algorithm by Efron et al. (2004) does not select $X_1$ at stage one. Consequently, the procedure will fail to include the important quadratic term $X_1^2$ at stage two. In the

following two subsections, we explain why two-stage methods fail for this example and then discuss general conditions under which two-stage methods work.

### 3.2. New Insight from Turlach's Example

Based on Definition 1, the key to success for two-stage methods is to identify all of the important main effects at stage one, so that all of the important interactions are included for selection at stage two. Next we use Turlach's example to explain the working mechanism of two-stage methods.

Without loss of generality, we first center the predictors $\tilde{X}_j = X_j - \mathrm{E}(X_j) = X_j - 0.5$, $j = 1, \ldots, p$, and consider the model

$$Y = 2 + \tilde{X}_1^2 + \tilde{X}_2 + \tilde{X}_3 + \tilde{X}_4 + \tilde{X}_5 + \varepsilon. \tag{5}$$

In (5), the linear term $\tilde{X}_1$ disappears after centering. It turns out that, no variable selection methods based on (1) can identify $X_1$ unless by chance. To see this, let us consider the following least-square estimator based on the entire data population,

$$\begin{aligned}\boldsymbol{\beta}_{\mathrm{LS}} &= \underset{\beta_0,\ldots,\beta_5}{\operatorname{argmin}} \mathrm{E}(Y - \beta_0 - \sum_{j=1}^{5} \beta_j \tilde{X}_j)^2 \\ &= \underset{\beta_0,\ldots,\beta_5}{\operatorname{argmin}} \Big( \mathrm{E}(2 - \beta_0)^2 + \mathrm{E}(\tilde{X}_1^2 - \beta_1 \tilde{X}_1)^2 \\ &\quad + \sum_{j=2}^{5} \mathrm{E}(\tilde{X}_j - \beta_j \tilde{X}_j)^2 \Big) \\ &= (2, 0, 1, 1, 1, 1)^\top.\end{aligned}$$

The second coefficient in $\boldsymbol{\beta}_{LS}$ is zero, which implies that it is unlikely to select $X_1$ under model (1), even if the entire population were observed. However, if we simply change $(X_1 - 0.5)^2$ to $(X_1 - c)^2$ with $c \neq 0.5$ in (3), then two-stage methods would be able to identify $X_1$ successfully.

Motivated by Turlach's example, we can establish general conditions under which two-stage methods are valid. At stage one, two-stage methods essentially estimate the parameters

$$(\breve{\beta}_0, \breve{\boldsymbol{\beta}}) = \underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmin}} \mathrm{E} \left( Y - \beta_0 - \sum_{j=1}^{p} X_j \beta_j \right)^2, \tag{6}$$

instead of $\boldsymbol{\beta}$. Since model (1) is misspecified, $\breve{\boldsymbol{\beta}}$ is in general not the same as $\boldsymbol{\beta}$. Assume that $\breve{\boldsymbol{\beta}}$ is unique and sparse. A necessary condition for two-stage methods to work is that the set of important main effects $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is contained in $\mathcal{S}(\breve{\boldsymbol{\beta}})$, that is, $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \subset \mathcal{S}(\breve{\boldsymbol{\beta}})$. If a main effect is left out of $\mathcal{S}(\breve{\boldsymbol{\beta}})$, such as $X_1$ in Turlach's example, then it can be selected only by chance.

Is it possible to derive a sufficient condition to guarantee both $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\breve{\boldsymbol{\beta}})$ and $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathcal{S}(\boldsymbol{\beta})$? If so, then we have $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\breve{\boldsymbol{\beta}})$, which will validate two-stage methods. Hao and Zhang (2014) provided such a condition on the distribution of data to ensure $\breve{\boldsymbol{\beta}} = \boldsymbol{\beta}$. Their main result is reviewed next. Without loss of generality, assume $\mathrm{E}(Y) = 0$ and $\mathrm{E}(X_j) = 0$ for $j = 1, \ldots, p$ in model (2). Moreover, we center all interaction terms and denote them as $Z_{jk} = X_j X_k - \mathrm{E}(X_j X_k)$. Then model (2) is equivalent to

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_{11} Z_{11} + \gamma_{12} Z_{12} + \cdots + \gamma_{pp} Z_{pp} + \varepsilon. \tag{7}$$

Denote by $\Sigma$ the covariance matrix of vector $(X_1, \ldots, X_p, Z_{11}, \ldots, Z_{jk}, \ldots, Z_{pp})^\top$. First, it can be shown that, if the joint distribution of $(X_1, \ldots, X_p)^\top$, say, $\mathcal{F}$, is symmetric with respect to the origin $\mathbf{0}$, then the covariance matrix $\Sigma$ has a block structure as

$$\Sigma = \begin{pmatrix} \Sigma^{(1)} & 0 \\ 0 & \Sigma^{(2)} \end{pmatrix}, \tag{8}$$

where $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are the covariance matrices of $(X_1, \ldots, X_p)^\top$ and $(Z_{11}, \ldots, Z_{pp})^\top$, respectively. The block structure in (8) is because all the first and third moments of the joint distribution $\mathcal{F}$ are zero. The following proposition shows that the block structure of $\Sigma$ is a sufficient condition for $\breve{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

*Proposition 2.* If (8) holds, then $\breve{\boldsymbol{\beta}} = \boldsymbol{\beta}$. In particular, $\mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\breve{\boldsymbol{\beta}})$.

*Proof.* For (7), define $\omega = \gamma_{11} Z_{11} + \gamma_{12} Z_{12} + \cdots + \gamma_{pp} Z_{pp} + \varepsilon$. Based on (8), we have $\mathrm{cov}(\omega, X_j) = 0$ for any $1 \le j \le p$. Denote by $\boldsymbol{\beta}^*$ the true coefficient vector. Then,

$$\begin{aligned}\breve{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \mathrm{E} \left( Y - \sum_{j=1}^{p} X_j \beta_j \right)^2 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \mathrm{E} \left( \sum_{j=1}^{p} X_j \beta_j^* + \omega - \sum_{j=1}^{p} X_j \beta_j \right)^2 \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \mathrm{E} \left[ \left( \sum_{j=1}^{p} X_j \beta_j^* - \sum_{j=1}^{p} X_j \beta_j \right)^2 + \omega^2 \right] = \boldsymbol{\beta}^*,\end{aligned}$$

where the equal sign in the last line holds because all the variables are centered and $\mathrm{cov}(\omega, X_j) = 0$ for all $j = 1, \ldots, p$. □

*Remark 1.* Proposition 2 shows that two-stage methods can identify $\mathcal{S}(\boldsymbol{\beta})$ successfully at stage one under some conditions, even when the model is misspecified. One sufficient condition is the block structure of $\Sigma$. Furthermore, we point out that one condition on the data distribution to ensure the block structure is symmetry (with respect to the origin) of the joint distribution of $(X_1, \ldots, X_p)^\top$. The block structure condition may be restrictive from a practical perspective, but it sheds some light on future research directions.

Next, we consider the conditions that guarantee $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathcal{S}(\boldsymbol{\beta})$.

### 3.3. Strong Heredity Condition

Heredity conditions were first applied to experimental design by Hamada and Wu (1992), Chipman (1996), and Chipman, Hamada, and Wu (1997). Recently, these conditions have also been used to study interaction selection for linear regression models (Yuan, Joseph, and Zou 2009; Choi, Li, and Zhu 2010). The strong heredity condition for model (2) or (7) is expressed

as

$$\gamma_{jk} \neq 0 \quad \text{only if} \quad \beta_j \beta_k \neq 0, \quad \forall 1 \leq j, k \leq p, \tag{9}$$

and the weak heredity condition is expressed as

$$\gamma_{jk} \neq 0 \quad \text{only if} \quad \beta_j^2 + \beta_k^2 \neq 0, \quad \forall 1 \leq j, k \leq p. \tag{10}$$

Given any parameterization choice, the strong heredity condition (9) implies $\beta_j \neq 0$ for any important main effect $X_j$ (based on Definition 1), that is, $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathcal{S}(\boldsymbol{\beta})$. Altogether, Proposition 2, condition (8), and condition (9) guarantee that $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\breve{\boldsymbol{\beta}})$. Though heredity conditions (9) and (10) seem to be restrictive, they are actually natural in many practical settings. In the following, we provide some insight on heredity conditions to better understand them and appreciate their practical value.

First, the strong heredity condition is not that restrictive, as the set of models violating the strong heredity condition is generally "small." The following is a simple setting that illustrates this. Consider model (7) with $p = 2$ and three effects $X_1, X_2, X_1X_2$ (for simplicity, assume quadratic effects $X_1^2$ and $X_2^2$ are not involved). The parameter space of the coefficients $(\beta_1, \beta_2, \gamma_{12})^\top$ is $\mathbb{R}^3$. If the strong heredity condition is assumed, it only excludes the subset $\{\beta_1\beta_2 = 0, \gamma_{12}^2 > 0\}$ from $\mathbb{R}^3$. Since the excluded low-dimensional subset has a zero Lebesgue measure, the strong heredity condition almost surely covers the entire model space $\mathbb{R}^3$.

Second, whether a heredity condition holds depends on the model parameterization. This is a very important fact often overlooked in the literature. In linear regression, it is a common practice to center and rescale the data before variable selection. Since any coding transformation $X_j \to a_j(X_j - c_j)$ can lead to a new parameterization for the coefficient vector, it is meaningless to discuss heredity conditions without specifying the model parameterization. In Turlach's example with parameterization (5), condition (8) holds but condition (9) does not. It implies $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \supsetneq \mathcal{S}(\boldsymbol{\beta}) = \mathcal{S}(\breve{\boldsymbol{\beta}})$, which explains why two-stage methods may fail.

Third, the definitions of $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $\mathcal{S}(\breve{\boldsymbol{\beta}})$ are independent of model parameterization. In other words, whether all important main effects are in $\mathcal{S}(\breve{\boldsymbol{\beta}})$ does not depend on model parameterization. Nevertheless, a good model parameterization helps to conveniently connect these two sets via $\mathcal{S}(\boldsymbol{\beta})$.

In summary, as long as $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathcal{S}(\breve{\boldsymbol{\beta}})$ holds, two-stage methods can identify all important main effects at stage one under standard technical conditions. Recently, the screening and sign consistency results for two-stage methods have been established by Hao and Zhang (2014) and Hao, Feng, and Zhang (2014), respectively, in the context of forward selection and the LASSO.

## 4. Interaction Selection Under Marginality Principle

### 4.1. Marginality Principle

For interaction selection, both the marginality principle and the invariance principle, which were reviewed earlier, emphasize that a final model should maintain a hierarchical structure. For example, consider model (2) with $p = 2$. For simplicity, we tentatively ignore the quadratic terms $X_1^2$ and $X_2^2$. Based on these two principles, we should consider five candidate models: $\{1\}$, $\{1, X_1\}$, $\{1, X_2\}$, $\{1, X_1, X_2\}$, and the full model $\{1, X_1, X_2, X_1X_2\}$; no other sub-models are sensible. Note that the marginality principle does not exclude the case that the true data-generating process is indeed $Y = 1 + 2X_1X_2 + \varepsilon$ under a certain parameterization. In this case, we can fit the full model with a loss of 2 degrees of freedom, but it is risky to fit only $\{1, X_1X_2\}$. In short, the marginality principle provides a good guidance for interaction selection.

Next, it is worthwhile to point out the difference between the marginality principle and heredity conditions. The former provides a guide for variable selection in interaction models or other hierarchical models. The selected model must satisfy the hierarchical structure if the marginality principle is followed. On the other hand, heredity conditions are designed to exclude undesired data-generating processes, by putting some restrictions on the parameter space. They depend on the model parameterization.

For existing methods, there are two ways to ensure the hierarchical structure during variable selection. One-stage methods impose special penalties or inequality constrains on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to satisfy the strong heredity condition. For two-stage methods, the hierarchical structure is naturally preserved by their selection schemes.

### 4.2. New Strategies

Section 3 discusses the theoretical foundation for two-stage methods. In practice, two-stage methods still have their limitations and can be further improved. One problem of two-stage methods is that interaction effects are considered only after main effect selection. At stage one, the noise level can be quite high since interaction effects are treated as noise under a misspecified model, and therefore it is difficult to identify weak main effects. In the following, we propose two alternative strategies to overcome these drawbacks.

Many variable selection procedures produce a family of candidate models that is naturally nested or indexed by a tuning parameter. For example, for stepwise methods such as forward selection and LARS, a sequence of nested models is obtained; penalization approaches such as the LASSO produce a family of models indexed by a tuning parameter. These methods can be directly applied to a standard linear model (1) or an interaction model (2) by ignoring the hierarchical structure. The new strategy uses a family of dynamic candidate models $\{\mathcal{C}_t\}$ lying between models (1) and (2), which initiates at (1) and grows adaptively under the marginality principle. Now we sketch two ways of implementing this strategy. For a forward selection procedure, we denote by $\widehat{\mathcal{M}}_t$ the selected model after step $t$, and let $\mathcal{C}_t$ be the candidate set containing all main and interaction effects whose parents are both in $\widehat{\mathcal{M}}_t$. In particular, we set $\widehat{\mathcal{M}}_0 = \emptyset$ and $\mathcal{C}_0 = \{\text{all main effects}\}$. At step $t + 1$, a forward selection procedure selects one new variable from $\mathcal{C}_t$ and adds it to $\widehat{\mathcal{M}}_t$ to obtain $\widehat{\mathcal{M}}_{t+1}$. The idea was proposed and studied by Hao and Zhang (2014), resulting in a new method called iFORM. For a penalization procedure such as LASSO, we denote by $\lambda$ the tuning parameter. The coordinate descent algorithm can be used

to calculate the penalization estimator along a sequence $\lambda_{\max} = \lambda_0 > \lambda_1 > \cdots > \lambda_T > 0$. We let $\widehat{\mathcal{M}}_t$ be the selected model at step $t$ corresponding to $\lambda_t$ and define $\mathcal{C}_t$ based on $\widehat{\mathcal{M}}_t$ in the same way as above. At next step with parameter $\lambda_{t+1}$, we conduct the coordinate descent algorithm on the candidate model $\mathcal{C}_t$ to obtain $\widehat{\mathcal{M}}_{t+1}$. This idea was recently proposed by Hao, Feng, and Zhang (2014), and its advantageous performance over two-stage methods was shown by numerical experiments.

## 5. Numerical Analysis

We present three examples to illustrate performance of two-stage methods for interaction selection in high-dimensional linear regression settings. Example 1 considers continuous predictors only, and Examples 2 and 3 include categorical predictors for both main effects and interactions. For comparison, we consider three methods: two-stage forward selection (2FS), the new forward selection algorithm under marginality principle (iFORM) described in Section 4.2, and the oracle (Oracle) procedure (which is the gold standard but generally not available in practice). To select the tuning parameter, we use the standard BIC and its extension as proposed by Chen and Chen (2008). More examples can be found in Hao and Zhang (2014). In all the examples, we set $n = 200$ and $p = 1000$.

*Example 1.* Generate $\mathbf{X}$ from a multivariate Gaussian distribution with mean $\mathbf{0}$ and the autoregressive (AR) correlation structure $\text{cov}(X_j, X_k) = 0.5^{|j-k|}$ for $1 \le j, k \le p$. Generate the response $Y$ from model (2) with $\sigma = 2$, $\boldsymbol{\beta} = (2, 0, 2, 0, 2, 0, 2, 0, 2, \mathbf{0}_{991}^\top)^\top$, $\gamma_{13} = 1.5$, $\gamma_{17} = 1.7$, $\gamma_{57} = 1.9$, $\gamma_{79} = 2.1$, and the rest of the interaction effects being zero. Therefore, the important main effects are $\{X_1, X_3, X_5, X_7, X_9\}$ and the important interactions are $\{X_1X_3, X_1X_7, X_5X_7, X_7X_9\}$.

*Example 2.* Generate $\mathbf{X}$ in the same way as in Example 1. Then we dichotomize $X_7$ as 1 if $X_7 \le 0.2$, and 0 otherwise; dichotomize $X_9$ as 1 if $X_9 \le -0.2$, and 0 otherwise. Generate $Y$ from model (2) with $\sigma = 2$, the true $\boldsymbol{\beta} = (2, 0, 2, 0, 2, 0, 2, 0, 2, \mathbf{0}_{991}^\top)^\top$, $\gamma_{13} = \gamma_{17} = \gamma_{57} = \gamma_{79} = 4$, and the rest of the interaction effects being zero. The important main effects are $\{X_1, X_3, X_5, X_7, X_9\}$, and the important interaction effects are $\{X_1X_3, X_1X_7, X_5X_7, X_7X_9\}$.

*Example 3.* Generate $\mathbf{X}$ in the same way as in Example 1. Then we dichotomize the variables $\{X_j, 6 \le j \le p\}$ as 1 if $X_j \ge -0.3$ and 0 otherwise for each $j$. Generate $Y$ from model (2) with $\sigma = 2$, $\boldsymbol{\beta} = (2, 0, 2, 0, 2, 0, 2, 0, 2, \mathbf{0}_{991}^\top)^\top$, $\gamma_{13} = \gamma_{17} = \gamma_{57} = \gamma_{79} = 4$, and the rest of the interaction effects are zero. The important main effects are $\{X_1, X_3, X_5, X_7, X_9\}$, and important interaction effects are $\{X_1X_3, X_1X_7, X_5X_7, X_7X_9\}$. This dataset contains five continuous predictors and 995 categorical predictors.

For each setting, we run $M = 1000$ Monte Carlo simulations and report the average performance for each method, in terms of correctly selecting important linear and interaction effects, estimating nonzero regression coefficients, and making predictions. In particular, to evaluate linear effect selection, we report the probability of identifying important main effects (Cov), the percentage of correct zeros being identified (Cor0), the percentage of incorrect zeros being identified (Inc0), and the probability of selecting the exact set of important main effects (Ext). For interaction selection, we report the probability of identifying all important interaction effects (iCov), the percentage of correct zeros being selected (iCor0), the percentage of incorrect zeros (iInc0), and the probability of selecting the exact set of important interactions (iExt). We also report the average model size. To evaluate prediction and estimation results, we report the mean squared error (MSE) of estimated regression coefficients and the out-of-sample $R^2$ (Rsq) based on a test set of size $n$ from the same distribution as the training data. A larger Rsq suggests a better prediction.

**Table 1.** Numerical results for the simulated examples.

| | | Linear term selection | | | | Interaction selection | | | | Size and prediction | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cov | Cor0 | Inc0 | Ext | iCov | iCor0 | iInc0 | iExt | size | MSE | Rsq |
| Ex 1 | 2FS | 0.62 | 1.00 | 0.12 | 0.61 | 0.62 | 1.00 | 0.24 | 0.48 | 8.19 | 1.86 | 78.71 |
| | iFORM | 1.00 | 1.00 | 0.00 | 0.96 | 0.99 | 1.00 | 0.00 | 0.90 | 9.18 | 0.48 | 91.30 |
| | Oracle | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 9.00 | 0.47 | 91.32 |
| Ex 2 | 2FS | 0.27 | 1.00 | 0.22 | 0.26 | 0.27 | 1.00 | 0.33 | 0.22 | 7.46 | 5.33 | 79.60 |
| | iFORM | 0.96 | 1.00 | 0.01 | 0.89 | 0.94 | 1.00 | 0.03 | 0.58 | 9.48 | 1.33 | 93.90 |
| | Oracle | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 9.00 | 1.00 | 94.95 |
| Ex 3 | 2FS | 0.31 | 1.00 | 0.21 | 0.31 | 0.31 | 1.00 | 0.47 | 0.26 | 7.24 | 6.15 | 80.31 |
| | iFORM | 0.95 | 1.00 | 0.01 | 0.88 | 0.93 | 1.00 | 0.04 | 0.54 | 9.47 | 1.55 | 94.39 |
| | Oracle | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 9.00 | 1.05 | 95.72 |

Table 1 provides a summary of numerical results for the three examples. Overall, the two-stage FS (2FS) method works well for continuous predictors. In Example 1, the 2FS method can recover the exact set of important main effects with probability 61% and the exact set of important interactions with probability 48%, which is quite good for $p = 1000$, $d = 501, 500$, and $n = 200$. In the other two examples with categorical predictors, the 2FS method tends to miss some important main and interaction effects. For example, 2FS misses 22% of important variables and 33% of important interactions in Example 2. By contrast, iFORM performs markedly better than 2FS in all the examples. In Example 1, iFORM identifies the exact set of important main effects with probability 96% and the exact set of important interactions with probability 90%. In Examples 2 and 3, iFORM can identify important categorical predictors and their interactions with an accuracy higher than 90%. The true model size is 9 for all three examples. The final model sizes given by iFORM are 9.18, 9.48, 9.47, respectively, which are close to the true model size in all the three cases. In summary, the performance of iFORM is very close to that of the oracle procedure.

## 6. Discussion

This note aims to clarify some important issues in variable selection for linear models with interactions. The presented concepts and methods also apply to generalized linear models (GLM) and models with higher-order interaction terms or complex hierarchical structures. In practice, when choosing between main effect models, two-way interaction models, or higher-order interaction models, one also needs to consider the bias-variance tradeoff. In general, adding more interaction terms tends to reduce the modeling bias but increases the estimate variance.

In high-dimensional settings, many predictors tend to be highly correlated. Hao and Zhang (2014) observed promising performance of two-stage methods and iFORM under a variety of correlation structure settings. Quantitative predictors are commonly encountered in real world problems. Though the theoretical result of Proposition 2 is established for continuous variables, our numerical results shown in Examples 2 and 3 suggest that two-step methods still perform effectively when categorical predictors are involved in main effects and interactions. Very recently, Gosik et al. (2017) extended iFORM to identify significant eQTLs, which are categorical predictors with three distinct genotypes. Their results also suggested effective performance of iFORM in selecting quantitative predictors.

## Funding

The authors gratefully acknowledge the funding support of NSF DMS-1309507, NSF DMS-1418172, and NSFC-11571009.

## References

Bien, J., Taylor, J., and Tibshirani, R. (2013), "A Lasso for Hierarchical Interactions," *The Annals of Statistics*, 41, 1111–1141. [292,293]

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications (Springer Series in Statistics)*, New York: Springer. [292]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771. [296]

Chipman, H. (1996), "Bayesian Variable Selection with Related Predictors," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 24, 17–36. [291,294]

Chipman, H., Hamada, M., and Wu, C. F. J. (1997), "A Bayesian Variable-Selection Approach for Analyzing Designed Experiments with Complex Aliasing," *Technometrics*, 39, 372–381. [291,294]

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. [292,293,294]

Cordell, H. J. (2009), "Detecting Gene-Gene Interactions that Underline Human Diseases," *Nature Reviews Genetics*, 10, 392–404. [291]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–451. [292,293]

Evans, D. M., Marchini, J., Morris, A. P., and Cardon, L. R. (2006), "Two-Stage Two-Locus Models in Genome-Wide Association," *PLoS Genetics*, 2, e157. [291]

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. [292]

Fan, J., and Lv, J. (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101–148. [291,292]

Gosik, K., Kong, L., Chinchilli, V. M., and Wu, R. (2017), "iFORM/eQTL: An Ultrahigh-Dimensional Platform for Inferring the Global GeneticArchitecture of Gene Transcripts," *Briefings in Bioinformatics Advance Access*, 18, 250–259. [297]

Hamada, M., and Wu, C. F. J. (1992), "Analysis of Designed Experiments with Complex Aliasing," *Journal of Quality Technology*, 24, 130–137. [291,294]

Hao, N., Feng, Y., and Zhang, H. H. (2014), "Model Selection forHigh Dimensional Quadratic Regressions via Regularization," *Journal of the American Statistician*, to appear. DOI: 10.1080/01621459.2016.1264956 [295]

Hao, N., and Zhang, H. H. (2014), "Interaction Screening for Ultra-high Dimensional Data," *Journal of the American Statistical Association*, 109, 1285–1301. [292,294,295,296]

Kooperberg, C., and LeBlanc, M. (2008), "Increasing the Power of Identifying Gene × Gene Interactions in Genome-Wide Association Studies," *Genetic Epidemiology*, 32, 255–263. [291]

Manolio, T. A., and Collins, F. S. (2007), "Genes, Environment, Health, and Disease: Facing up to Complexity," *Human Heredity*, 63, 63–66. [291]

McCullagh, P. (2002), "What is a Statistical Model?," *The Annals of Statistics*, 30, 1225–1267. [291]

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models (Monographs onStatistics and Applied Probability)*, Cambridge, UK: Chapman and Hall. [291]

Nelder, J. A. (1977), "A Reformulation of Linear Models," *Journal of the Royal Statistical Society*, Series A, 140, 48–77. [291]

——— (1994), "The Statistics of Linear Models: Back to Basics," *Statistics and Computing*, 5, 221–234. [291]

Peixoto, J. L. (1987), "Hierarchical Variable Selection in Polynomial Regression Models," *The American Statistician*, 41, 311–313. [291]

Peixoto, J. L. (1990), "A Property of Well-Formulated Polynomial Regression Models," *The American Statistician*, 44, 26–30. [291,292]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [292]

Turlach, B. (2004), "Discussion of "Least Angle Regression"," *The Annals of Statistics*, 32, 481–490. [292,293]

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010), "Screen and Clean: A Tool for Identifying Interactions in Genome-Wide Association Studies," *Genetic Epidemiology*, 34, 275–285. [293]

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009), "Genome-Wide Association Analysis by Lasso Penalized Logistic Regression," *Bioinformatics*, 25, 714–721. [293]

Yuan, M., Joseph, V. R., and Zou, H. (2009), "Structured Variable Selection and Estimation," *Annals of Applied Statistics*, 3, 1738–1757. [292,293,294]

Zhao, P., Rocha, G., and Yu, B. (2009), "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection," *The Annals of Statistics*, 37, 3468–3497. [292,293]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [292]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [292]