Model Selection for High-Dimensional
Quadratic Regression via Regularization

Model Selection for High Dimensional Quadratic Regression via Regularization

Ning Hao, Yang Feng, and Hao Helen Zhang*

Abstract

Quadratic regression (QR) models naturally extend linear models by considering interaction effects between the covariates. To conduct model selection in QR, it is important to maintain the hierarchical model structure between main effects and interaction effects. Existing regularization methods generally achieve this goal by solving complex optimization problems, which usually demands high computational cost and hence are not feasible for high dimensional data. This paper focuses on scalable regularization methods for model selection in high dimensional QR. We first consider two-stage regularization methods and establish theoretical properties of the two-stage LASSO. Then, a new regularization method, called Regularization Algorithm under Marginality Principle (RAMP), is proposed to compute a hierarchy-preserving regularization solution path efficiently. Both methods are further extended to solve generalized QR models. Numerical results are also shown to demonstrate performance of the methods.

Keywords: Generalized quadratic regression, Interaction selection, LASSO, Marginality principle, Variable selection.

1 Introduction

Statistical models involving two-way or higher-order interactions have been studied in various contexts, such as linear models and generalized linear models (Nelder, 1977; McCullagh & Nelder, 1989), experimental design (Hamada & Wu, 1992; Chipman et al., 1997), and polynomial regression (Peixoto, 1987). In particular, a quadratic regression (QR) model formulated as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{1,1} X_1^2 + \beta_{1,2} X_1 X_2 + \cdots + \beta_{p,p} X_p^2 + \varepsilon \tag{1}$$

has been considered recently to analyze high dimensional data. In (1), $X_1,..., X_p$ are main effects, and order-2 terms $X_j X_k$ ($1 \le j \le k \le p$) include quadratic main effects ($j = k$) and two-way interaction effects ($j \ne k$). A key feature of model (1) is its hierarchical structure, as order-2 terms are derived from the main effects. To reflect their relationship, we call $X_j X_k$ the *child* of X_j and X_k, and X_j and X_k the *parents* of $X_j X_k$.

Standard techniques such as ordinary least squares can be applied to solve (1) for a small or moderate p. When p is large and variable selection becomes necessary, it is suggested that the selected model should keep the hierarchical structure. That is, interaction terms can be selected into the model only if their parents are in the model. This is referred to the marginality principle (Nelder, 1977). In general, a direct application of variable selection techniques to (1) can not automatically ensure the hierarchical structure in the final model. Recently, several regularization methods (Zhao et al., 2009; Yuan et al., 2009; Choi et al., 2010; Bien et al., 2013) have been proposed to conduct variable selection for (1) under the marginality principle by designing special forms of penalty functions. These methods are feasible when p is a few hundreds or less, and the resulting estimators have oracle properties when $p = o(n)$ (Choi et al., 2010). However, when p is much larger, these methods are not feasible since their implementation requires storing and manipulating the entire $O(p^2) \times n$ design matrix and solving complex constrained optimization problems. The memory and computational cost can be extremely high and prohibitive. Very recently, interaction screening for high-dimensional settings has drawn much attention, and a variety of interaction screening approaches have been proposed for regression and classification problems, including Hao & Zhang (2014a), Fan et al. (2015), and Kong et al. (2016). By contrast, the purpose of this

work is to develop scalable interaction selection approaches under a penalized framework for high dimensional data analysis.

In this paper, we study regularization methods on model selection and estimation for QR and generalized quadratic regression (GQR) models under the marginality principle. The main focus is the case $p \gg n$, which is a bottleneck for the existing regularization methods. We study theoretical properties of a two-stage regularization method based on the LASSO and propose a new efficient algorithm, RAMP, which produces a hierarchy-preserving solution path. In contrast to existing regularization methods, these procedures avoid storing $O(p^2) \times n$ design matrix and sidestep complex constraints and penalties, making them feasible to analyze data with many variables. In particular, our R package `RAMP` runs well on a desktop for data with $n = 400$ and $p = 10^4$ and it takes less than 30 seconds (with CPU 3.4 GHz Intel Core i7 and 32GB memory) to fit the QR model and get the whole solution path. The main contribution of this paper is threefold. First, we establish a variable selection consistency result of the two-stage LASSO procedure for QR and offer new insights on stage-wise selection methods. To our best knowledge, this is the first selection consistency result for high dimensional QR. Second, the proposed algorithms are computationally efficient and will make a valuable contribution to interaction selection tools in practice. Third, our methods are extended to interaction selection in GQR models, which are rarely studied in literature.

We define notations used in the paper. Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top$ be the $n \times p$ design matrix of main effects and $\mathbf{y} = (y_1, ..., y_n)^\top$ be the n-dimensional response vector. The linear term index set is $\mathcal{M} = \{1, 2, ..., p\}$, and the order-2 index set is $\mathcal{I} = \{(j, k) : 1 \leq j \leq k \leq p\}$. The regression coefficient vector $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_\mathcal{M}^\top, \boldsymbol{\beta}_\mathcal{I}^\top)^\top$, where $\boldsymbol{\beta}_\mathcal{M} = (\beta_1, ..., \beta_p)^\top$ and $\boldsymbol{\beta}_\mathcal{I} = (\beta_{1,1}, \beta_{1,2}, ..., \beta_{p,p})^\top$. For a subset $\mathcal{A} \subset \mathcal{M}$, use $\boldsymbol{\beta}_\mathcal{A}$ for the subvector of $\boldsymbol{\beta}_\mathcal{M}$ indexed in $\mathcal{A}$, and $\mathbf{X}_\mathcal{A}$ for the submatrix of $\mathbf{X}$ whose columns are indexed in $\mathcal{A}$. In particular, $\mathbf{X}_j$ is the jth column of $\mathbf{X}$. We treat the subscripts (j, k) and (k, j) as identical, i.e., $\beta_{j,k} = \beta_{k,j}$. Let c_1, c_2, ... and C_1, C_2, ... be positive constants which are independent of the sample size n. They are locally defined and their values may vary in different context. For a vector $\mathbf{v} = (v_1, ..., v_p)^\top$, $\|\mathbf{v}\| = \sqrt{\sum_{j=1}^p v_j^2}$ and $\|\mathbf{v}\|_1 = \sum_{j=1}^p |v_j|$. For a matrix $\mathbf{A}$, define $\|\mathbf{A}\|_\infty = \max_i \sum_j |A_{ij}|$ and $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{v}\|_2 = 1} \|\mathbf{A}\mathbf{v}\|_2$ as the standard operator norm, i.e., the square root of the largest eigenvalue of $\mathbf{A}^\top \mathbf{A}$.

The rest of the paper is organized as follows. Section 2 considers two-stage regularization methods for model selection in QR and studies theoretical properties of the two-stage LASSO. Section 3 proposes RAMP to compute the entire hierarchy-preserving solution path efficiently. Section 4 extends the proposed methods to generalized QR models. Section 5 presents numerical studies, followed by a discussion. Technical proofs are in the Appendix.

2 Two-stage Regularization Method

Variable selection and estimation via penalization is popular in high dimensional analysis. Examples include the LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), elastic net (Zou & Hastie, 2005), minimax concave penalty (MCP) (Zhang, 2010), among many others. Properties such as model selection consistency and oracle properties have been verified (Zhao & Yu, 2006; Wainwright, 2009; Fan & Lv, 2011). A general penalized estimator for linear models is defined as

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\mathcal{M}}) = \underset{(\beta_0, \boldsymbol{\beta}_{\mathcal{M}})}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y} - \mathbf{1}\beta_0 - \mathbf{X}\boldsymbol{\beta}_{\mathcal{M}}\|^2 + \sum_{j=1}^{p} J_\lambda(\beta_j), \tag{2}$$

where $\mathbf{y}$ is the response vector, $\mathbf{X}$ is the design matrix, $J_\lambda(\cdot)$ is a penalty function, and $\lambda \geq 0$ is a regularization parameter. The penalty $J(\cdot)$ and λ may depend on index j. For easy presentation, we use same penalty function and parameter for all j unless stated otherwise.

We consider the problem of variable selection for QR model (1). Define $\mathbf{X}^{\circ 2} = \mathbf{X} \circ \mathbf{X}$ as an $n \times \frac{p(p+1)}{2}$ matrix consisting of all pairwise column products. That is, for $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_p)$, $\mathbf{X}^{\circ 2} = \mathbf{X} \circ \mathbf{X} = (\mathbf{X}_1 \star \mathbf{X}_1, \mathbf{X}_1 \star \mathbf{X}_2, ..., \mathbf{X}_p \star \mathbf{X}_p)$, where $\star$ denotes the entry-wise product of two column vectors. For an index set $\mathcal{A} \subset \mathcal{M}$, define $\mathcal{A}^{\circ 2} = \mathcal{A} \circ \mathcal{A} = \{(j,k) : j \leq k; j, k \in \mathcal{A}\} \subset \mathcal{I}$, and $\mathcal{A} \circ \mathcal{M} = \{(j,k) : j \leq k; j \text{ or } k \in \mathcal{A}\} \subset \mathcal{I}$. We use $\mathbf{X}_{\mathcal{A}}^{\circ 2}$ as a short notation for $(\mathbf{X}_{\mathcal{A}})^{\circ 2}$, a matrix whose columns are indexed by $\mathcal{A}^{\circ 2}$.

Two-stage regularization methods for interaction selection have been considered in Efron et al. (2004); Wu et al. (2009), among others. However, their theoretical properties are not clearly understood. In the following, we first illustrate the general two-stage procedure for interaction selection.

Two-stage Regularization Method:

Stage 1: Solve (2). Denote the selected model by $\widehat{\mathcal{A}} = \{j : \hat{\beta}_j \neq 0, j = 1, \ldots, p\}$.

Stage 2: Solve

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{y} - \mathbf{1}\beta_0 - \mathbf{X}_{\widehat{\mathcal{A}}}\boldsymbol{\beta}_{\widehat{\mathcal{A}}} - \mathbf{X}_{\widehat{\mathcal{A}}}^{\circ 2}\boldsymbol{\beta}_{\widehat{\mathcal{A}}^{\circ 2}}\|^2 + \sum_{\alpha \in \widehat{\mathcal{A}}^{\circ 2}} J_\lambda(\beta_\alpha),$$

At Stage 1, only main effects are considered for selection, with all the order-2 terms being left out of the model. Denote the selected set by $\widehat{\mathcal{A}}$. At Stage 2, we expand $\widehat{\mathcal{A}}$ by including all the two-way interactions of those main effects within $\widehat{\mathcal{A}}$ and fit the new model. To keep the hierarchical structure, we do not penalize main effects at Stage 2, i.e., set $J_\lambda(\cdot) = 0$ for $j \in \widehat{\mathcal{A}}$. In order to keep the hierarchy, it is also possible to use other methods (Zhao et al., 2009; Yuan et al., 2009; Choi et al., 2010; Bien et al., 2013) at Stage 2.

One main advantage of this two-stage regularization procedure is its simple implementation. Existing R packages `lars` and `glmnet` can be directly used to carry out the procedure. Stage 1 serves as a dimension reduction step prior to Stage 2, so the two-stage method avoids estimating $O(p^2)$ parameters altogether, making the procedure feasible for very large p.

In spite of its computational advantages, theoretical properties of two-stage regularization methods are seldom studied in literature. A commonly raised concern is whether the important main effects can be consistently identified at Stage 1, when all order-2 terms are left out of the model on purpose. Next, we focus on the two-stage LASSO method and investigate its selection behavior at Stage 1. In particular, we establish the main-effect selection consistency result of the two-stage LASSO for QR under some regularity conditions.

The LASSO is a special case of (2) by using the ℓ_1 penalty

$$(\hat{\beta}_{0L}, \hat{\boldsymbol{\beta}}_L) = \operatorname*{argmin}_{(\beta_0, \boldsymbol{\beta}_{\mathcal{M}})} \frac{1}{2n} \|\mathbf{y} - \mathbf{1}\beta_0 - \mathbf{X}\boldsymbol{\beta}_{\mathcal{M}}\|^2 + \sum_{j=1}^{p} \lambda|\beta_j|.$$

In the following, we show that the LASSO solution $\hat{\boldsymbol{\beta}}_L$ is sign consistent at Stage 1, i.e., $\operatorname{sign}(\hat{\boldsymbol{\beta}}_L) = \operatorname{sign}(\boldsymbol{\beta}_{\mathcal{M}})$ with an overwhelming probability for a properly chosen tuning parameter. This result provides critical theoretical insight about the two-stage LASSO estimator.

Consider a sparse quadratic model with a Gaussian design. Assume that $\mathbf{x}_i$, $1 \leq i \leq n$, are independent and identically distributed (i.i.d.) from $\mathcal{N}(\mathbf{0}, \Sigma)$, and

$$y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_{\mathcal{M}} + (\mathbf{x}_i^\top)^{\circ 2} \boldsymbol{\beta}_{\mathcal{I}} + \varepsilon_i, \tag{3}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ is independent of $\{\mathbf{x}_i\}_{i=1}^n$. Without loss of generality, we further center y_i and $(\mathbf{x}_i^\top)^{\circ 2}$ and write

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_{\mathcal{M}} + \mathbf{u}_i^\top \boldsymbol{\beta}_{\mathcal{I}} + \varepsilon_i, \tag{4}$$

where y_i is the centered response and $\mathbf{u}_i^\top = (\mathbf{x}_i^\top)^{\circ 2} - \mathrm{E}(\mathbf{x}_i^\top)^{\circ 2}$ is a $p \times (p+1)/2$ row vector with all centered order-2 terms. Let $y_{\mathcal{M}i} = \mathbf{x}_i^\top \boldsymbol{\beta}_{\mathcal{M}}$ and $y_{\mathcal{I}i} = \mathbf{u}_i^\top \boldsymbol{\beta}_{\mathcal{I}}$. $\mathbf{y}_{\mathcal{M}} = (y_{\mathcal{M}1,}, ..., y_{\mathcal{M}n})^\top$, $\mathbf{y}_{\mathcal{I}} = (y_{\mathcal{I}1,}, ..., y_{\mathcal{I}n})^\top$, $\mathbf{U} = (\mathbf{u}_1, ..., \mathbf{u}_n)^\top$. Set $\tau^2 = \mathrm{Var}(y_{\mathcal{I}i})$. Define $\omega_i = \mathbf{u}_i^\top \boldsymbol{\beta}_{\mathcal{I}} + \varepsilon_i$ and $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)^\top$, which is treated as noise at Stage 1. Denote by $\Sigma_{\mathcal{AB}}$ the submatrix of Σ with row index $\mathcal{A}$ and column index $\mathcal{B}$. As illustrated in Hao & Zhang (2014b), the support and sign of the coefficient vector $\boldsymbol{\beta}_{\mathcal{M}}$ for a QR model depend on its parametrization because a coding transformation can change the support of $\boldsymbol{\beta}_{\mathcal{M}}$. Therefore, we follow Hao & Zhang (2014b) and define the index set of important main effects by $\mathcal{S} = \{j : \beta_j^2 + \sum_{k=1}^p \beta_{j,k}^2 > 0\}$. Let $s = |\mathcal{S}|$ and $\mathcal{T} = \{(k, \ell) : \beta_{k,\ell} \neq 0\}$. It follows this definition that $\mathcal{T} \subset \mathcal{S}^{\circ 2}$. Moreover, in order to make $\mathrm{sign}(\boldsymbol{\beta}_{\mathcal{M}})$ well-defined, we require that main effects are centered in (3). We refer to Hao & Zhang (2014b) for further explanations on the well-definedness of sign and support of the coefficient vector $\boldsymbol{\beta}_{\mathcal{M}}$ for a QR model.

Define $\Sigma_{\mathcal{S}^c | \mathcal{S}} = \Sigma_{\mathcal{S}^c \mathcal{S}^c} - \Sigma_{\mathcal{S}^c \mathcal{S}} (\Sigma_{\mathcal{S}\mathcal{S}})^{-1} \Sigma_{\mathcal{S}\mathcal{S}^c}$ where $\mathcal{S}^c = \mathcal{M} - \mathcal{S}$. Let $\Lambda_{\min}(\mathbf{A})$ be the smallest eigenvalue of $\mathbf{A}$ and $\rho_u(A) = \max_i A_{ii}$. Assume the following technical conditions:

(C1) (Irrepresentable Condition) $\|\Sigma_{\mathcal{S}^c \mathcal{S}} (\Sigma_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty \leq 1 - \gamma$, $\gamma \in (0, 1]$.

(C2) (Eigenvalue Condition) $\Lambda_{\min}(\Sigma_{\mathcal{S}\mathcal{S}}) \geq C_{\min} > 0$.

Theorem 1 *Consider the quadratic model with a random Gaussian design (4). Suppose that (C1)-(C2) hold. Consider the family of regularization parameters*

$$\lambda_n(\phi_p) = \sqrt{\frac{\phi_p \rho_u(\Sigma_{\mathcal{S}^c | \mathcal{S}})}{\gamma^2} \frac{4(\sigma^2 + \tau^2) \log(p)}{n}} \tag{5}$$

for some $\phi_p \geq 2$. If for some fixed $\delta > 0$, the sequence (n, p, s) and regularization sequence $\{\lambda_n\}$ satisfy

$$\frac{n}{2s \log(p - s)} > (1 + \delta)\frac{\rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}})}{C_{\min}\gamma^2}(1 + \frac{2(\sigma^2 + \tau^2)C_{\min}}{\lambda_n^2 s}), \tag{6}$$

then the following holds with probability greater than $1 - c_1 \exp(-c_2 \min\{s, \log(p - s), n^{\frac{1}{2}}\})$.

1. *The LASSO has a unique solution $\hat{\boldsymbol{\beta}}_L$ with support contained within $\mathcal{S}$.*

2. *Define the gap*

$$g(\lambda_n) = c_3\lambda_n \left\|\Sigma_{\mathcal{S}\mathcal{S}}^{-\frac{1}{2}}\right\|_\infty^2 + 20\sqrt{\frac{\sigma^2 s}{C_{\min}n}} + \frac{9\|\boldsymbol{\beta}_{\mathcal{I}}\|_2\sqrt{s}}{C_{\min}n^{\frac{1}{3}}}. \tag{7}$$

   Then if $\beta_{\min} = \min_{j \in \mathcal{S}} |\beta_j| > g(\lambda_n)$, then $\text{sign}(\hat{\boldsymbol{\beta}}_L) = \text{sign}(\boldsymbol{\beta}_{\mathcal{M}})$.

Furthermore, given (5), an alternative condition to (6) making the above results hold is

$$\frac{n}{2s \log(p - s)} > \frac{1 + \delta'}{1 - \frac{1}{\phi_p}}\frac{\rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}})}{C_{\min}\gamma^2} \tag{8}$$

for some $\delta' > 0$.

Remark 1. Conditions (C1)-(C2) are commonly used to show the model selection consistency of the LASSO estimator in the literature. Conditions (6) and (7) are key requirements on dimensionality and minimal signal strength $\beta_{\min}$, respectively. The normality assumption is used here to facilitate the proof and comparisons to existing results in linear regression. In the supplementary material, we establish Theorem 2, which extends the consistency result to non-Gaussian designs. Other extensions of theoretical results are discussed in Section 6.

Remark 2. The result in Theorem 1 generalizes Theorem 3 in Wainwright (2009) that is established in the context of linear regression. Theorem 1 implies that the two-stage LASSO can identify important main effects at Stage 1. The validity of the two-stage LASSO is then guaranteed, as the index set of important interactions $\mathcal{T} \subset \mathcal{S}^{\circ 2}$. That is, all important interaction effects can be included at Stage 2. Given Theorem 1, the interaction selection consistency result at Stage 2 can be obtained under some mild conditions on the matrix $\mathbf{X}_{\mathcal{S}}^{\circ 2}$, since the data dimensionality has been greatly reduced. One can also apply existing methods, e.g., Choi et al. (2010) at Stage 2, for which selection consistency has been established.

3 Regularization Path Algorithm under Marginality Principle (RAMP)

For linear regression models, regularization solution-path algorithms provide state-of-the-art computational tools to implement variable selection with high dimensional data. Popular algorithms include least angle regression (LARS) (Efron et al., 2004), its extensions (Park & Hastie, 2007; Wu, 2011; Zhou & Wu, 2014), and coordinate decent algorithm (CDA) (Friedman et al., 2007; Wu & Lange, 2008; Friedman et al., 2010; Yu & Feng, 2014). These computational tools can be used to implement two-stage methods for fitting QR. However, by the nature of two-stage approach, the whole solution-path highly depends on the selection result at Stage 1, which is obtained under considerably high noise level if interaction effects are strong. Therefore, it is desirable to develop a seamless path algorithm which can select main and interaction effects simultaneously subject to hierarchy. To achieve this, we propose a Regularization Algorithm under Marginality Principle (RAMP) via the coordinate descent to compute the solution path while preserving the model hierarchy along the path.

We first review the coordinate decent algorithm for the standard LASSO. Consider

$$\min \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta}_\mathcal{M})^2 + \lambda \|\boldsymbol{\beta}_\mathcal{M}\|_1.$$

There exists a penalty parameter $\lambda_{\max}$ such that the minimizer $\hat{\boldsymbol{\beta}}_L = \mathbf{0}$ if $\lambda \geq \lambda_{\max}$. As λ decreases from $\lambda_{\max}$ to 0, the LASSO solution $\hat{\boldsymbol{\beta}}_L = \hat{\boldsymbol{\beta}}_\lambda$ changes from $\mathbf{0}$ to the least squares estimator (if it exists). Usually, a sequence of values $\{\lambda_k\}_{k=1}^{K}$ between $\lambda_{\max}$ and $\zeta\lambda_{\max}$ is set, with $0 < \zeta < 1$, and a solution path $\hat{\boldsymbol{\beta}}_{\lambda_k}$ is calculated for each λ_k. For a fixed k, using $\hat{\boldsymbol{\beta}}_{\lambda_{k-1}}$ as the initial value, the CDA solves the optimization problem by cyclically minimizing each coordinate β_j until convergence. Define $\mathcal{M}_k = supp\{\hat{\boldsymbol{\beta}}_{\lambda_k}\}$, i.e., the active set for each λ_k.

In the following, we propose a coordinate descent algorithm to fit the quadratic model under regularization which obeys the marginality principle. Given a tuning parameter λ, the algorithm computes the ℓ_1 regression coefficients of main effects and interactions subject to the heredity condition. At step $k-1$, denote the current active main effect set as $\mathcal{M}_{k-1}$ and the interaction effect set as $\mathcal{I}_{k-1}$. Define $\mathcal{H}_{k-1}$ as the parent set of $\mathcal{I}_{k-1}$, i.e., it contains the

8

main effect which has at least one interaction effect (child) in $\mathcal{I}_{k-1}$. Set $\mathcal{H}_{k-1}^c = \mathcal{M} - \mathcal{H}_{k-1}$.

Regularization Algorithm under Marginality Principle (RAMP):

Initialization: Set $\lambda_{\max} = n^{-1} \max |\mathbf{X}^\top \mathbf{y}|$ and $\lambda_{\min} = \zeta \lambda_{\max}$ with some small $\zeta > 0$. Generate an exponentially decaying sequence $\lambda_{\max} = \lambda_1 > \lambda_2 > \cdots > \lambda_K = \lambda_{\min}$. Initialize the main effect set $\mathcal{M}_0 = \emptyset$ and the interaction effect set $\mathcal{I}_0 = \emptyset$.

Path-building: Repeat the following steps for $k = 1, \cdots, K$. Given $\mathcal{M}_{k-1}, \mathcal{I}_{k-1}, \mathcal{H}_{k-1}$, add the possible interactions among main effects in $\mathcal{M}_{k-1}$ to the current model. Then with respect to $(\beta_0, \boldsymbol{\beta}_\mathcal{M}^\top, \boldsymbol{\beta}_{\mathcal{M}_{k-1}^{\circ 2}}^\top)^\top$, we minimize

$$\frac{1}{2n} \sum_{i=1}^n \left(y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta}_\mathcal{M} - (\mathbf{x}_i^\top)_{\mathcal{M}_{k-1}}^{\circ 2} \boldsymbol{\beta}_{\mathcal{M}_{k-1}^{\circ 2}} \right)^2 + \lambda_k \| \boldsymbol{\beta}_{\mathcal{H}_{k-1}^c} \|_1 + \lambda_k \| \boldsymbol{\beta}_{\mathcal{M}_{k-1}^{\circ 2}} \|_1, \qquad (9)$$

where the penalty is imposed on the candidate interaction effects and $\mathcal{H}_{k-1}^c$, which contains the main effects not enforced by the strong heredity constraint. Record $\mathcal{M}_k$, $\mathcal{I}_k$ and $\mathcal{H}_k$ according to the solution. Add the corresponding main effects from $\mathcal{I}_k$ into $\mathcal{M}_k$ to enforce the heredity constraint, and calculate the OLS based on the current model.

Different from two-stage approaches, RAMP allows at each step the interaction effects $\mathcal{M}_{k-1}^{\circ 2}$ to enter the model for selection. Following the same strategy, we propose a weak hierarchy version of RAMP, denoted by RAMP-w, as a flexible relaxation. The main difference is that we use the set $\mathcal{M}_{k-1} \circ \mathcal{M}$ instead of $\mathcal{M}_{k-1}^{\circ 2}$ in (9) and solve the optimization problem with respect to $(\beta_0, \boldsymbol{\beta}_\mathcal{M}^\top, \boldsymbol{\beta}_{\mathcal{M}_{k-1} \circ \mathcal{M}}^\top)^\top$. In this way, an interaction term can enter the model for selection immediately after one of its parents has been selected at a previous step. Therefore, RAMP-w is particularly useful in the scenario when only one parent of an important effect is strong. Both RAMP and RAMP-w are implemented in our R package `RAMP`, which is available on the CRAN website for researchers to use. Moreover, other penalty options such as SCAD and MCP also included in the `RAMP` package.

Figure 1 illustrates two hierarchy-preserving solution paths obtained by the RAMP under strong and weak heredity constraints, respectively. In this toy example, $n = 500, p = 100$, and $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $Y = X_1 + 3X_6 + 4X_1X_3 + 5X_1X_6 + \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$. Without

the marginality principle, X_1X_6 would be the most significant predictor as it has the highest marginal correlation with Y. On the other hand, RAMP with the strong heredity selects X_1 and X_6 before picking up X_1X_6 on the solution path. Note that RAMP does not select X_1X_3 until at a very late stage on the solution path due to the strong heredity assumption. Under the weak heredity, RAMP-w is able to select in sequence X_6, X_1X_6, X_1, and X_1X_3. The reason is that after X_6 is selected, X_1X_6 is immediately added into the candidate interaction set and then successfully selected, even before X_1 is selected. Similarly, X_1X_3 is picked up by the algorithm after one of its parents X_1 has been selected.
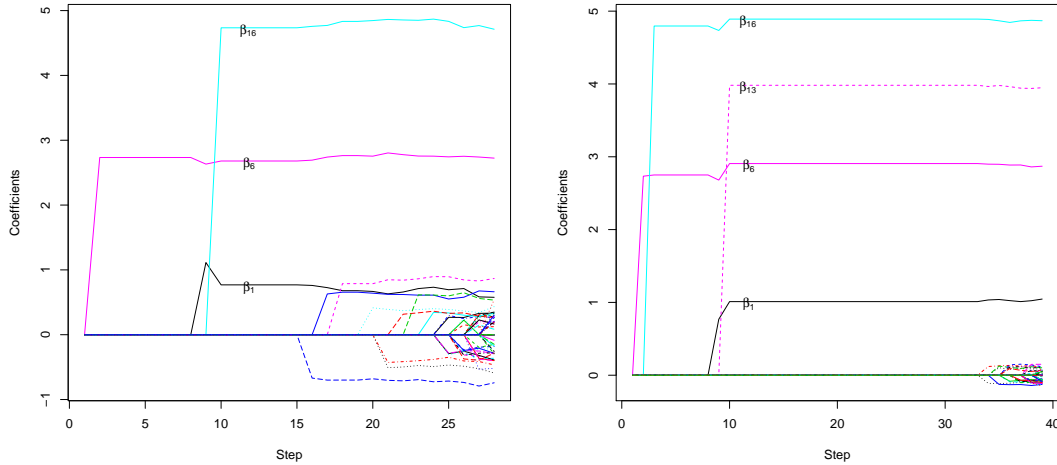
Figure 1: Two hierarchy-preserving solution paths for a toy example produced by the RAMP and RAMP-w, respectively. Left Panel: strong hierarchy. Right Panel: weak hierarchy.

4 Extension to Generalized QR Models

4.1 Generalized Quadratic Regression

A standard generalized linear model (GLM) assumes that the conditional distribution of $\mathbf{y}$ given $\mathbf{X}$ belongs to the canonical exponential family with density

$$f_n(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^n f_0(y_i; \theta_i) = \prod_{i=1}^n \left\{ c(y_i) \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} \right] \right\},$$

10

where $\phi > 0$ is a dispersion parameter, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ are the regression coefficients, and

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\top = \mathbf{X}\boldsymbol{\beta}. \tag{10}$$

The function $b(\boldsymbol{\theta})$ is twice continuously differentiable with a positive second-order derivative. In sparse high dimensional modeling, $\boldsymbol{\beta}$ is a long vector with a small number of nonzero entries. In the context of QR, the design matrix is $(\mathbf{X}, \mathbf{X}^{\circ 2})$. A natural generalization of GLM is to modify (10) as

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\top = \mathbf{X}\boldsymbol{\beta}_{\mathcal{M}} + \mathbf{X}^{\circ 2}\boldsymbol{\beta}_{\mathcal{I}}. \tag{11}$$

In the literature, there are very few computational tools available to fit high dimensional GQR models. Next, we illustrate how the aforementioned algorithms can be used for GQR.

4.2 Two-stage Regularization Methods

For high dimensional data, the penalized likelihood method is commonly used to fit GLM. Given the systematic component (10), the penalized likelihood estimator is defined as

$$\underset{\boldsymbol{\beta}}{\mathrm{argmin}}\, Q_n(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\mathrm{argmin}}\, -\ell_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} J_\lambda(|\beta_j|),$$

where $\ell_n(\boldsymbol{\beta}) = \log f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) = \frac{1}{n}\left(\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^\top \mathbf{b}(\mathbf{X}\boldsymbol{\beta})\right)$ is the log-likelihood up to a scalar, $J_\lambda(\cdot)$ is a penalty function and $\lambda \geq 0$ is the regularization parameter.

For GQR with systematic component (11), we propose the two-stage approach as follows. At Stage 1, only main effects are selected by the penalization method with order-2 terms being left out. Denote the selected main-effect set by $\widehat{\mathcal{A}}$. At Stage 2, we expand $\widehat{\mathcal{A}}$ by adding all the two-way interactions (children) of those main effects (parents) within $\widehat{\mathcal{A}}$ and solve

$$\underset{\boldsymbol{\beta}}{\mathrm{argmin}}\, Q_n(\boldsymbol{\beta}, \widehat{\mathcal{A}}) = \underset{\boldsymbol{\beta}}{\mathrm{argmin}}\, -\ell_n(\boldsymbol{\beta}, \widehat{\mathcal{A}}) + \sum_{\alpha \in \widehat{\mathcal{A}}^{\circ 2}} J_\lambda(\beta_\alpha),$$

where

$$\ell_n(\boldsymbol{\beta}, \widehat{\mathcal{A}}) = \frac{1}{n}\left[\mathbf{y}^T(\mathbf{X}_{\widehat{\mathcal{A}}}\boldsymbol{\beta}_{\widehat{\mathcal{A}}} + \mathbf{X}_{\widehat{\mathcal{A}}}^{\circ 2}\boldsymbol{\beta}_{\widehat{\mathcal{A}}^{\circ 2}}) - \mathbf{1}^\top \mathbf{b}(\mathbf{X}_{\widehat{\mathcal{A}}}\boldsymbol{\beta}_{\widehat{\mathcal{A}}} + \mathbf{X}_{\widehat{\mathcal{A}}}^{\circ 2}\boldsymbol{\beta}_{\widehat{\mathcal{A}}^{\circ 2}})\right].$$

At Stage 2, we intentionally do not impose penalty on main effects in $\widehat{\mathcal{A}}$, so that all the selected main effects at Stage 1 will stay in the final model. This will assure the hierarchical structure of main effects and interactions in the final model.

4.3 New Path Algorithm for Generalized QR

The RAMP proposed in Section 3 can be easily extended to fit the GQR. The major difference is to replace the penalized least squares by the penalized likelihood function at each step. The CDA algorithm is used to minimize the penalized likelihood function iteratively.

RAMP Algorithm for GQR:

Initialization: Set $\lambda_{\max} = n^{-1} \max |\mathbf{X}^\top \mathbf{y}|$ and $\lambda_{\min} = \zeta \lambda_{\max}$ with $1 > \zeta > 0$. Generate an exponentially decaying sequence $\lambda_{\max} = \lambda_1 > \lambda_2 > \cdots > \lambda_K = \lambda_{\min}$. Initialize the main effect set $\mathcal{M}_0 = \emptyset$ and the interaction effect set $\mathcal{I}_0 = \emptyset$.

Path-building: Repeat the following steps for $k = 1, \cdots, K$. Given $\mathcal{M}_{k-1}, \mathcal{I}_{k-1}, \mathcal{H}_{k-1}$, add the possible interactions among main effects in $\mathcal{M}_{k-1}$ to the current model. Then with respect to $(\beta_0, \boldsymbol{\beta}^\top, \boldsymbol{\beta}_{\mathcal{M}_{k-1}^{\circ 2}}^\top)^\top$, we maximize

$$\ell_n(\boldsymbol{\beta}, \mathcal{M}_{k-1}) - \lambda_k \|\boldsymbol{\beta}_{\mathcal{H}_{k-1}^c}\|_1 - \lambda_k \|\boldsymbol{\beta}_{\mathcal{M}_{k-1}^{\circ 2}}\|_1,$$

where

$$\begin{aligned}
\ell_n(\boldsymbol{\beta}, \mathcal{M}_{k-1}) = \frac{1}{n} \Big[& \mathbf{y}^T (\mathbf{X}_{\mathcal{M}_{k-1}} \boldsymbol{\beta}_{\mathcal{M}_{k-1}} + \mathbf{X}_{\mathcal{M}_{k-1}}^{\circ 2} \boldsymbol{\beta}_{\mathcal{M}_{k-1}^{\circ 2}}) \\
& - \mathbf{1}^\top \mathbf{b} (\mathbf{X}_{\mathcal{M}_{k-1}} \boldsymbol{\beta}_{\mathcal{M}_{k-1}} + \mathbf{X}_{\mathcal{M}_{k-1}}^{\circ 2} \boldsymbol{\beta}_{\mathcal{M}_{k-1}^{\circ 2}}) \Big].
\end{aligned}$$

Calculate $\mathcal{M}_k, \mathcal{I}_k$ and $\mathcal{H}_k$ according to the solution. Add the main effects from $\mathcal{I}_k$ into $\mathcal{M}_k$ to enforce the heredity constraint, and calculate the MLE based on the current model.

5 Numerical Studies

5.1 Simulation Examples

We consider data generating processes with varying signal-to-noise ratios, different covariate structures, error distributions, and heredity structures. In particular, Example 1 is a QR model under a $p \gg n$ settings with strong heredity considered in Hao & Zhang (2014a). Example 2 is a high-dimensional logistic regression model with interaction effects. Examples 3 and 4 consider QR models with the weak and strong heredity structures respectively, where we consider a relatively small p to make the comparison possible with the hierarchical lasso (Bien et al., 2013). Example 5 considers a QR model with a heavy tail error distribution to demonstrate the robustness of our methods.

For comparison, we consider RAMP and two two-stage methods, i.e., two-stage LASSO (2-LASSO) and two-stage SCAD (2-SCAD). We also include existing methods iFORT and iFORM (Hao & Zhang, 2014a), the hierarchical lasso (Bien et al., 2013), and the benchmark method ORACLE for which the true sparse model is known.

When computing the solution paths of two-stage methods and RAMP, we choose the tuning parameter by EBIC with $\gamma = 1$ (Chen & Chen, 2008). We also implemented other parameter tuning criteria including AIC, BIC, and GIC (Fan & Tang, 2013), and observed that the EBIC tends to work the best among most of the simulation settings that we considered. For easy presentation, we report only the results for EBIC.

Let $\mathcal{S} = \{j : \beta_j \neq 0\}$ and $\mathcal{T} = \{(j, k) : \beta_{j,k} \neq 0\}$ with cardinality $s = |\mathcal{S}|$ and $t = |\mathcal{T}|$. For each example, we run $M = 100$ Monte-Carlo simulations for each method and make a comparison. For the m-th simulation, denote the estimated subsets as $\widehat{\mathcal{S}}^{(m)}$ and $\widehat{\mathcal{T}}^{(m)}$, the estimated coefficient vector as $\widehat{\boldsymbol{\beta}}^{(m)}$, the main effects and interaction effects as $\widehat{\beta}_j^{(m)}$ and $\widehat{\beta}_{j,k}^{(m)}$. We evaluate variable selection and model estimation performance based on the following criteria.

- Main effects coverage percentage (main.cov): $M^{-1} \sum_{m=1}^{M} I(\mathcal{S} \subset \widehat{\mathcal{S}}^{(m)})$.

- Interaction effects coverage percentage (inter.cov): $M^{-1} \sum_{m=1}^{M} I(\mathcal{T} \subset \widehat{\mathcal{T}}^{(m)})$.

- Main effects exact selection percentage (main.exact): $M^{-1} \sum_{m=1}^{M} I(\mathcal{S} = \widehat{\mathcal{S}}^{(m)})$.

- Interaction effects exact selection percentage (inter.exact): $M^{-1} \sum_{m=1}^{M} I(\mathcal{T} = \widehat{\mathcal{T}}^{(m)})$.

- Model size (size): $M^{-1} \sum_{m=1}^{M} (|\widehat{\mathcal{S}}^{(m)}| + |\widehat{\mathcal{T}}^{(m)}|)$.

- Root mean squared error (RMSE): $\{M^{-1} \sum_{m=1}^{M} [\sum_{j=0}^{p} (\hat{\beta}_j^{(m)} - \beta_j)^2 + \sum_{j=1}^{p} \sum_{k=j}^{p} (\hat{\beta}_{j,k}^{(m)} - \beta_{j,k})^2]\}^{1/2}$.

Example 1 Set $(n, p, s, t) = (400, 5000, 10, 10)$. Generate the covariates $\{\mathbf{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\Sigma_{jk} = 0.5^{|j-k|}$ and generate the response $\mathbf{y}$ by model (1). $\mathcal{S} = \{1, 2, \cdots, 10\}$ with the true regression coefficients $\boldsymbol{\beta}_{\mathcal{S}} = (3, 3, 3, 3, 3, 2, 2, 2, 2, 2)^{\top}$. The set of important interaction effects is $\mathcal{T} = \{(1, 2), (1, 3), (2, 3), (2, 5), (3, 4), (6, 8), (6, 10), (7, 8), (7, 9), (9, 10)\}$ with the corresponding coefficients $(2, 2, 2, 2, 2, 1, 1, 1, 1, 1)$.

To have different signal-to-noise ratio situations, we consider $\sigma \in \{2, 3, 4\}$. The results are summarized in Table 1. With regard to model selection, the proposed RAMP has a high coverage percentage in selecting both main effects and interaction effects. The 2-LASSO tends to miss some important main effects while picking up some noise variables, ending up with the largest model size on average. On the other hand, the 2-SCAD has a high exact selection percentage with a low coverage percentage. Compared to RAMP, the iFORM tends to have a lower coverage on interaction effects. The iFORT is the worst in terms both variable selection and model estimation. With regard to parameter estimation, RAMP has the smallest root mean square error (RMSE) when $\sigma = 3$ and 4.

Example 2 We consider a logistic regression model with

$$\log \frac{P(Y = 1|X)}{P(Y = 0|X)} = \beta_1 X_1 + 3X_6 + 3X_{10} + 3X_1 X_6 + 3X_6 X_{10},$$

where $(n, p, s, t) = (400, 2000, 3, 2)$ and $\mathbf{X} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. For different signal-to-noise ratios, we vary the coefficient $\beta_1 \in \{1, 2, 3\}$.

The results are summarized in Table 2, which lead to the following observations. When the signal is strong ($\beta_1 = 2, 3$), RAMP, 2-LASSO and 2-SCAD perform similarly in selecting

Table 1: Selection and estimation results for Example 1.

	σ	main effects		interaction effects			
		coverage	exact	coverage	exact	size	RMSE
RAMP	2	1.00	0.96	1.00	0.35	20.98	0.87
	3	0.99	0.91	0.83	0.17	21.25	1.29
	4	0.92	0.77	0.47	0.11	20.83	1.96
2-LASSO	2	0.78	0.60	0.78	0.01	24.77	1.56
	3	0.75	0.56	0.75	0.01	24.64	1.85
	4	0.72	0.51	0.69	0.01	24.40	2.20
2-SCAD	2	0.70	0.58	0.70	0.53	19.92	1.81
	3	0.69	0.55	0.62	0.26	20.31	2.06
	4	0.65	0.52	0.43	0.14	20.56	2.42
iFORT	2	0.00	0.00	0.00	0.00	14.54	6.64
	3	0.00	0.00	0.00	0.00	13.74	7.02
	4	0.00	0.00	0.00	0.00	12.72	7.52
iFORM	2	1.00	0.98	0.98	0.40	20.71	0.59
	3	1.00	0.97	0.34	0.17	19.94	1.40
	4	0.97	0.97	0.02	0.01	18.71	2.16
ORACLE	2	1.00	1.00	1.00	1.00	20.00	0.55
	3	1.00	1.00	1.00	1.00	20.00	0.83
	4	1.00	1.00	1.00	1.00	20.00	1.11

main effects; while RAMP and 2-SCAD is much better in selecting interactions than 2-LASSO. When the signal is weak ($\beta_1 = 1$), 2-LASSO and 2-SCAD fail to identify the correct main effects most of time, which in turn leads to low coverage of important interaction effects. On the other hand, RAMP performs reasonably well in terms of selecting both main effects and interaction effects. With regard to RMSE, RAMP outperforms 2-LASSO and 2-SCAD in all scenarios. Note that the iFORT and iFORM are omitted in this example, as they do not handle binary responses.

In the next two examples, we compare RAMP and hierNet algorithms for both strong

Table 2: Selection and estimation results for Example 2.

	β_1	main effects		interaction effects			
		coverage	exact	coverage	exact	size	RMSE
RAMP	1	0.92	0.78	0.92	0.91	4.98	1.80
	2	1.00	0.93	1.00	1.00	5.08	1.16
	3	1.00	0.92	0.99	0.99	5.13	1.36
2-LASSO	1	0.45	0.41	0.45	0.14	4.05	3.97
	2	1.00	0.93	1.00	0.29	6.58	1.41
	3	1.00	0.80	1.00	0.42	6.31	1.66
2-SCAD	1	0.49	0.43	0.49	0.49	3.58	3.76
	2	1.00	0.81	1.00	0.94	5.28	1.03
	3	1.00	0.74	1.00	0.86	5.52	1.22
ORACLE	1	1.00	1.00	1.00	1.00	5.00	0.84
	2	1.00	1.00	1.00	1.00	5.00	0.78
	3	1.00	1.00	1.00	1.00	5.00	0.83

and weak hierarchy scenarios.

Example 3 Set $(n, p, s, t) = (400, 100, 10, 10)$. Generate the covariates $\{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{jk} = 0.5^{|j-k|}$ and generate the response $\mathbf{y}$ by model (1). $\mathcal{S} = \{1, 2, \cdots, 10\}$ with the true regression coefficients $\boldsymbol{\beta}_{\mathcal{S}} = (3, 3, 3, 3, 3, 2, 2, 2, 2, 2)^{\top}$. The set of important interaction effects is $\mathcal{T} = \{(1, 2), (1, 13), (2, 3), (2, 15), (3, 4), (6, 10), (6, 18), (7, 9), (7, 18), (10, 19)\}$ with the corresponding coefficients $(2, 2, 2, 2, 2, 1, 1, 1, 1, 1)$.

In this example, the strong heredity does not hold while the weak heredity is satisfied. Note that we take p to be relatively small due to the heavy computational cost of hierNet (Bien et al., 2013). Here, we compare RAMP and RAMP-w (RAMP with the weak heredity constraint) with hierNet-s and hierNet-w, and the results are summarized in Table 3. As expected, when applying RAMP with strong heredity (RAMP), it always misses some important interaction effects. However, the RAMP with weak heredity (RAMP-w) successfully

Table 3: Selection and estimation results as well as average computing time (in seconds) per replicate for Example 3.

| | σ | main effects | | interaction effects | | size | RMSE | Time |
		coverage	exact	coverage	exact			
	2	1.00	0.71	0.00	0.00	19.45	3.54	37.49
RAMP	3	1.00	0.83	0.00	0.00	16.86	3.71	34.74
	4	0.98	0.89	0.00	0.00	15.28	3.87	34.88
	2	1.00	1.00	0.99	0.25	21.33	0.79	47.02
RAMP-w	3	1.00	0.99	0.63	0.12	21.16	1.31	46.51
	4	1.00	0.98	0.16	0.00	20.07	1.98	46.10
	2	1.00	0.00	1.00	0.00	133.45	5.69	3143.30
hierNet-s	3	1.00	0.00	0.96	0.00	119.62	5.33	3232.62
	4	1.00	0.00	0.74	0.00	95.06	5.01	3507.85
	2	1.00	0.00	1.00	0.00	126.83	6.60	295.88
hierNet-w	3	1.00	0.01	0.98	0.00	96.59	6.17	346.83
	4	1.00	0.04	0.75	0.00	65.31	5.73	444.99

recovers the important interaction effects with a high proportion, especially when the error variance is small. Comparing with the hierNet, the RAMP-w in general selects a much smaller model with a smaller RMSE. In particular, the computation time of hierNet is much longer than RAMP for both the strong and weak versions.

Example 4 Set $(n, p, s, t) = (400, 200, 10, 10)$. The rest setup is same as Example 1.

In this example, we consider the case where the strong heredity holds and compare RAMP and RAMP-w with hierNet-s and hierNet-w. From Table 4, it is clear that RAMP outperforms RAMP-w in terms of both the coverage percentage and the exact selection percentage for interaction effect. This is not surprising as the RAMP-w searches for additional interaction effects compared with RAMP. In addition, the RMSE of RAMP is the smallest among the four methods throughout all noise levels. Both hierNet-s and hierNet-w have very good

Table 4: Selection and estimation results as well as average computing time (in seconds) per replicate for Example 4.

	σ	main effects		interaction effects				
		coverage	exact	coverage	exact	size	RMSE	Time
RAMP	2	1.00	1.00	1.00	0.35	20.97	0.86	34.58
	3	1.00	0.98	0.93	0.23	21.31	1.18	32.95
	4	0.97	0.92	0.64	0.10	21.35	1.72	32.28
RAMP-w	2	1.00	1.00	0.99	0.25	21.25	0.87	56.01
	3	1.00	1.00	0.78	0.18	21.14	1.25	54.58
	4	1.00	1.00	0.30	0.06	20.02	1.92	53.71
hierNet-s	2	1.00	0.00	1.00	0.00	120.99	5.53	15847.28
	3	1.00	0.00	0.99	0.00	115.69	5.15	16552.18
	4	1.00	0.00	0.92	0.00	90.55	4.79	16864.49
hierNet-w	2	1.00	0.01	1.00	0.00	97.62	5.79	1467.46
	3	1.00	0.02	0.98	0.00	61.04	5.41	1798.27
	4	1.00	0.01	0.90	0.00	53.31	5.24	2156.99

coverage percentage but with almost zero exact selection percentage for both main effects and interaction effects. As a result, they select a large number of noise variables in the final model. Note that the computation time for hierNet-s is over 4 hours for a single replicate. As a result, we omit the comparison with hierNet for the other higher dimensional examples.

Example 5 We use the same setting as in Example 1 except for the error distribution, which is changed to a t distribution with degrees of freedom 3.

This example is designed to examine the robustness of proposed methods under heavy tail error distributions. For brevity, we report only the performance of the vanilla RAMP with strong heredity enforced. It is clear from Table 5 that under the heavy tail error distribution, RAMP has a similar performance as in Example 1.

Table 5: Selection and estimation results for Example 5.

		main effects		interaction effects			
	σ	coverage	exact	coverage	exact	size	RMSE
	2	1.00	0.94	0.98	0.29	21.59	1.02
RAMP	3	0.97	0.92	0.84	0.18	21.37	1.52
	4	0.90	0.76	0.49	0.08	21.00	2.33

5.2 Real Data Example: Supermarket Data

We consider the supermarket dataset analyzed in Wang (2009) and Hao & Zhang (2014a). The data set contains the daily sale information of a major supermarket located in northern China, with $n = 464$ and $p = 6,398$. The total number of interaction effects is about 2.0×10^7. The response Y is the number of customers on a particular day with the predictor X measuring sale volumes of a selection of products. The supermarket manager would like to find out which products are most informative in predicting the response, which would be useful to design promotions around those products.

Here, we randomly split the data into a training set ($n_1 = 400$) and a test set ($n_2 = 64$) to evaluate the prediction performance of different methods. We also compare the performance of RAMP with the regular LASSO without taking interaction effects into account. Because of the issue of tuning parameter selection, we report the results using different tuning methods including AIC, BIC, EBIC (Chen & Chen, 2008), and GIC (Fan & Tang, 2013) for both RAMP and the LASSO.

For each random split, we calculate the number of selected variables, the number of selected interaction effects, and the out-of-sample R^2 on the test set. The average performance over 100 random splits is presented in Table 6. When we use BIC, EBIC and GIC, RAMP selects a model with higher out-of-sample R^2 values than the LASSO. When using more stringent tuning parameter criteria like the EBIC and GIC, it is observed that the RAMP performs significantly better than the LASSO. For example, when GIC is used, RAMP selects 30 variables on average with around 3 of them being interaction effects, and has an average out-of-sample R^2 value of 90.08, which is much higher than the corresponding LASSO

Table 6: Mean selection and prediction results on the supermarket data set over 100 random splits. The standard errors are in parentheses.

	RAMP			LASSO		
	size	size.inter	R^2	size	size.inter	R^2
AIC	229.12(1.68)	94.53(1.06)	90.48(0.23)	264.28(0.91)	0.00(0.00)	92.04(0.18)
BIC	101.17(3.25)	34.36(1.65)	91.18(0.20)	63.47(0.77)	0.00(0.00)	90.76(0.20)
EBIC	29.27(1.01)	3.07(0.29)	89.67(0.31)	15.62(0.46)	0.00(0.00)	72.09(0.53)
GIC	30.71(0.92)	3.20(0.30)	90.08(0.28)	19.19(0.74)	0.00(0.00)	75.05(0.58)

results. It is clear that by using RAMP with the inclusion of possible interaction effects, we can obtain a more interpretable model with a reasonably good prediction performance. Moreover, from Table 8 in Hao & Zhang (2014a), the out-of-sample R^2 values with the associated standard error for iFORT and iFORM are 88.91 (0.17) and 88.66 (0.18), respectively, both of which are outperformed by RAMP with any tuning parameter selection method.

6 Discussion

We study regularization methods for interaction selection subject to the marginality principle for QR and GQR models. One main advantage of these algorithms is their computational efficiency and feasibility for high and ultra-high dimensional data. In particular, a key feature of RAMP is that it can select main and interaction effects simultaneously while still keeping the hierarchy structure. The strategy of RAMP can be used to extend other algorithms, e.g., LARS, to build the entire solution path when fitting the regularized QR models. All algorithms considered in this paper utilize the hierarchy structures. Such structures are natural for quadratic models (Nelder, 1977; Hao & Zhang, 2014b). Nevertheless, in certain applications, some main effects may not be strong enough to be selected first without incorporating the interaction effects. Other approaches (Zhao et al., 2009; Yuan et al., 2009; Choi et al., 2010; Bien et al., 2013) can be applied in this scenario, as these methods keep the hierarchy in different ways. However, a drawback is that most of these algorithms are

relatively slow when p is large. Recently, there have been studies on interaction selection which do not rely on the strong or weak hierarchy. Based on the idea of sure independence screening (Fan & Lv, 2008; Fan et al., 2011; Cheng et al., 2014), Jiang & Liu (2014) proposed Sliced Inverse Regression for Interaction Detection (SIRI) for screening interaction variables; Fan et al. (2016) introduced a new approach called interaction pursuit for interaction identification using screening and variable selection. It would be interesting to incorporate these screening based methods into our path algorithm to handle general scenarios.

We demonstrate theoretical properties of the two-stage LASSO method for QR. As a referee pointed out, selection consistency results on the LASSO often rely on the irrepresentable condition, which is not realistic in applications. In order to extend current results, it is desirable to investigate a broad range of penalty functions for GQR, e.g., under frameworks similar to Fan & Lv (2011) and Fan & Lv (2013).

An R package `RAMP` has been developed and is available from the CRAN website.

7   Appendix

The main results are shown in Appendix A, and a related lemma is put in Appendix B.

7.1   Appendix A

Proof of Theorem 1. We will apply the primal-dual witness (PDW) method and use $(W1)$, $(W2)$, etc. to denote the formula (1), (2),... in Wainwright (2009). Recall in our paper, the n-vector $\boldsymbol{\omega}$ is the imaginary noise at Stage 1, which is the sum of the Gaussian noise $\boldsymbol{\varepsilon}$ and the interaction effects $(\mathbf{u}_1^\top \boldsymbol{\beta}_{\mathcal{I}}, ..., \mathbf{u}_n^\top \boldsymbol{\beta}_{\mathcal{I}})^\top$, and hence it is not independent of the design matrix $\mathbf{X}$.

Part I: Verifying strict dual feasibility.

The goal is to show that, with overwhelming probability, under condition (6), inequality $|Z_j| < 1$ holds for each $j \in \mathcal{S}^c$, where Z_j is defined in (W10). For every $j \in \mathcal{S}^c$, conditional

on $\mathbf{X}_{\mathcal{S}}$, (W37) gives a decomposition $Z_j = A_j + B_j$ where

$$
\begin{aligned}
A_j &= \mathbf{E}_j^\top \left\{ \mathbf{X}_{\mathcal{S}} (\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1} \check{\mathbf{z}}_{\mathcal{S}} + \Pi_{\mathbf{X}_{\mathcal{S}}^\perp} \left(\frac{\boldsymbol{\omega}}{\lambda_n n} \right) \right\} \\
B_j &= \Sigma_{j\mathcal{S}} (\Sigma_{\mathcal{S}\mathcal{S}})^{-1} \check{\mathbf{z}}_{\mathcal{S}},
\end{aligned}
$$

where $\mathbf{E}_j^\top = \mathbf{X}_j^\top - \Sigma_{j\mathcal{S}} (\Sigma_{\mathcal{S}\mathcal{S}})^{-1} \mathbf{X}_{\mathcal{S}}^\top \in \mathbb{R}^n$ with $E_{ij} \sim \mathcal{N}(0, [\Sigma_{\mathcal{S}^c|\mathcal{S}}]_{jj})$.

Condition (C1) implies

$$
\max_{j \in \mathcal{S}^c} |B_j| \leq 1 - \gamma.
$$

Conditioned on $\mathbf{X}_{\mathcal{S}}$ and $\boldsymbol{\omega}$, A_j is Gaussian with mean zero and variance $\mathrm{Var}(A_j) \leq \rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}}) M_n$ where

$$
M_n = \frac{1}{n} \check{\mathbf{z}}_{\mathcal{S}}^\top \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \check{\mathbf{z}}_{\mathcal{S}} + \left\| \Pi_{\mathbf{X}_{\mathcal{S}}^\perp} \left(\frac{\boldsymbol{\omega}}{\lambda_n n} \right) \right\|_2^2 .
$$

The following lemma, proved in appendix B, generalizes Lemma 4 in Wainwright (2009).

Lemma 1 *For any $\epsilon \in (0, \frac{1}{2})$, define the event $\overline{\mathcal{T}}(\epsilon) = \{ M_n > \overline{M}_n(\epsilon) \}$, where*

$$
\overline{M}_n(\epsilon) = \left(1 + \max \left\{ \epsilon, \frac{8}{C_{\min}} \sqrt{\frac{s}{n}} \right\} \right) \left(\frac{s}{C_{\min} n} + \frac{2(\sigma^2 + \tau^2)}{\lambda_n^2 n} \right).
$$

Then $\mathbf{P}(\overline{\mathcal{T}}(\epsilon)) \leq C_1 \exp(-C_2 \min\{\sqrt{n}\epsilon^2, s\})$ for some $C_1, C_2 > 0$.

By Lemma 1,

$$
\begin{aligned}
\mathbf{P} \left(\max_{j \in \mathcal{S}^c} |Z_j| \geq 1 \right) &\leq \mathbf{P} \left(\max_{j \in \mathcal{S}^c} |A_j| \geq \gamma \right) \\
&\leq \mathbf{P} \left(\max_{j \in \mathcal{S}^c} |A_j| \geq \gamma \mid \overline{\mathcal{T}}^c(\epsilon) \right) + C_1 \exp(-C_2 \min\{\sqrt{n}\epsilon^2, s\}). \quad (12)
\end{aligned}
$$

Note that the goal is to show the probability in (12) is exponentially decayed. Conditional on $\overline{\mathcal{T}}^c(\epsilon)$, $\mathrm{Var}(A_j) \leq \rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}}) \overline{M}_n(\epsilon)$, so

$$
\mathbf{P} \left(\max_{j \in \mathcal{S}^c} |A_j| \geq \gamma \mid \overline{\mathcal{T}}^c(\epsilon) \right) \leq 2(p - s) \exp \left(-\frac{\gamma^2}{2\rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}}) \overline{M}_n(\epsilon)} \right).
$$

The assumptions of Theorem 1 imply $\frac{s}{n} = o(1)$ and $\frac{1}{\lambda_n^2 n} = o(1)$, so $\overline{M}_n(\epsilon) = o(1)$. Therefore, it suffices to show that the decaying rate of the exponential term dominates $p - s$. It is easy to check that (6) can guarantee that $\max_{j \in \mathcal{S}^c} |Z_j| < 1$ holds with probability at least $1 - c_1 \exp(-c_2 \min\{s, \log(p - s), n^{\frac{1}{2}}\})$.

Now we show the sufficiency of the alternative condition (8). In particular, we show (5) and (8) imply (6), which is equivalent to

$$\frac{n}{1 + \delta} > 2s \log(p - s) \frac{\rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}})}{C_{\min}\gamma^2}(1 + \frac{2(\sigma^2 + \tau^2)C_{\min}}{\lambda_n^2 s}).$$

Plugging in (5), we have

$$
\begin{aligned}
\frac{n}{1 + \delta} &> 2s \log(p - s) \frac{\rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}})}{C_{\min}\gamma^2} + 2s \log(p - s) \frac{\rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}})}{C_{\min}\gamma^2} \frac{2(\sigma^2 + \tau^2)C_{\min}}{\lambda_n^2 s} \\
&= 2s \log(p - s) \frac{\rho_u(\Sigma_{\mathcal{S}^c|\mathcal{S}})}{C_{\min}\gamma^2} + \frac{n}{\phi_p} \frac{\log(p - s)}{\log p}.
\end{aligned}
\tag{13}
$$

Following the same argument after (W40) in Wainwright (2009), (13) is implied by (8) for $\phi_p \geq 2$.

Part II: Sign consistency.

In order to show sign consistency, we need to show that (W13) holds. That is

$$\text{sign}(\beta_j + \Delta_j) = \text{sign}(\beta_j), \quad \text{for all } j \in \mathcal{S}, \tag{14}$$

where

$$\Delta_j = \mathbf{e}_j^\top \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n}\right)^{-1} \left[\frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\omega} - \lambda_n \text{sign}(\boldsymbol{\beta}_{\mathcal{S}})\right].$$

From definition, we have

$$\max_{j \in \mathcal{S}} |\Delta_j| \leq F_1 + F_2 \leq F_1 + (F_{2,1} + F_{2,2}),$$

where

$$F_1 = \lambda_n \left\| \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}}) \right\|_\infty$$

$$F_2 = \left\| \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \frac{1}{n} \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\omega} \right\|_\infty$$

$$F_{2,1} = \left\| \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \frac{1}{n} \mathbf{X}_{\mathcal{S}}^\top \boldsymbol{\varepsilon} \right\|_\infty$$

$$F_{2,2} = \left\| \left(\frac{\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}{n} \right)^{-1} \frac{1}{n} \mathbf{X}_{\mathcal{S}}^\top \mathbf{y}_{\mathcal{I}} \right\|_\infty$$

(W41) and a correction version of (W42) give upper bounds of tail probability of F_1 and $F_{2,1}$, respectively. That is

$$\mathbf{P}\left(F_1 > c_3 \lambda_n \left\| \Sigma_{\mathcal{S}\mathcal{S}}^{-\frac{1}{2}} \right\|_\infty^2 \right) \leq 4 \exp(-c_2 \min\{s, \log(p-s)\}), \tag{15}$$

$$\mathbf{P}\left(F_{2,1} \geq 20 \sqrt{\frac{\sigma^2 s}{C_{\min} n}} \right) \leq 4 \exp(-c_1 s). \tag{16}$$

Now we work on the addition term $F_{2,2}$. By (W60),

$$\mathbf{P}\left(\left\| \left(\frac{1}{n} \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} \right)^{-1} \right\|_2 \geq \frac{9}{C_{\min}} \right) \leq 2 \exp(-n/2).$$

$$\left\| \frac{1}{n} \mathbf{X}_{\mathcal{S}}^\top \mathbf{y}_{\mathcal{I}} \right\|_2 \leq \|\boldsymbol{\beta}_{\mathcal{I}}\|_2 \max_{j \in \mathcal{S}; (k,\ell) \in \mathcal{T}} \left\{ \left| \frac{1}{n} \mathbf{X}_j^\top (\mathbf{X}_k \star \mathbf{X}_\ell) \right| \right\}.$$

$\frac{1}{n} \mathbf{X}_j^\top (\mathbf{X}_k \star \mathbf{X}_\ell)$ is a sample third moment, so by Lemma B.5 in Hao & Zhang (2014a),

$$\mathbf{P}\left(\left| \frac{1}{n} \mathbf{X}_j^\top (\mathbf{X}_k \star \mathbf{X}_\ell) \right| > \epsilon \right) \leq c_4 \exp(-c_5 n^{\frac{2}{3}} \epsilon^2).$$

Therefore, we have

$$\mathbf{P}\left(\left\| \frac{1}{n} \mathbf{X}_{\mathcal{S}}^\top \mathbf{y}_{\mathcal{I}} \right\|_2 \geq \|\boldsymbol{\beta}_{\mathcal{I}}\|_2 \epsilon \right) \leq s^3 c_4 \exp(-c_5 n^{\frac{2}{3}} \epsilon^2).$$

Overall,

$$\mathbf{P}\left(F_{2,2} \geq \frac{9}{C_{\min}}\|\boldsymbol{\beta}_{\mathcal{I}}\|_2\epsilon\right) \leq s^3 c_6 \exp(-c_7 n^{\frac{2}{3}}\epsilon^2).$$

Setting $\epsilon = \frac{s^{\frac{1}{2}}}{n^{\frac{1}{3}}}$, we have

$$\mathbf{P}\left(F_{2,2} \geq \frac{9\|\boldsymbol{\beta}_{\mathcal{I}}\|_2\sqrt{s}}{C_{\min}n^{\frac{1}{3}}}\right) \leq c_8 \exp(-c_9 s). \tag{17}$$

Combining (15), (16) and (17), we have that with probability greater than $1 - c_1'\exp(-c_2' \min\{s, \log(p-s)\})$,

$$\max_{j\in\mathcal{S}}|\Delta_j| \leq c_3\lambda_n\left\|\Sigma_{\mathcal{S}\mathcal{S}}^{-\frac{1}{2}}\right\|_\infty^2 + 20\sqrt{\frac{\sigma^2 s}{C_{\min}n}} + \frac{9\|\boldsymbol{\beta}_{\mathcal{I}}\|_2\sqrt{s}}{C_{\min}n^{\frac{1}{3}}} = g(\lambda_n).$$

Therefore (14) holds when $\beta_{\min} > g(\lambda_n)$. $\square$

7.2   Appendix B

Proof of Lemma 1. The first summand of M_n can be controlled exactly the same way as in Wainwright (2009), i.e.,

$$\frac{1}{n}\check{\mathbf{z}}_{\mathcal{S}}^\top\left(\frac{\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}}}{n}\right)^{-1}\check{\mathbf{z}}_{\mathcal{S}} \leq \left(1 + \frac{8}{C_{\min}}\sqrt{\frac{s}{n}}\right)\frac{s}{nC_{\min}}$$

with probability at least $1 - 2\exp(-s/2)$.

Turning to the second summand, we observe that $\Pi_{\mathbf{X}_{\mathcal{S}}^\perp}$ is an orthogonal projection matrix and $\boldsymbol{\omega} = \boldsymbol{\varepsilon} + \mathbf{y}_{\mathcal{I}}$, so

$$\left\|\Pi_{\mathbf{X}_{\mathcal{S}}^\perp}\left(\frac{\boldsymbol{\omega}}{\lambda_n n}\right)\right\|_2^2 \leq \frac{\|\boldsymbol{\omega}\|_2^2}{\lambda_n^2 n^2} \leq \frac{2}{\lambda_n^2 n}\frac{\|\boldsymbol{\varepsilon}\|_2^2 + \|\mathbf{y}_{\mathcal{I}}\|_2^2}{n}.$$

Note that $\|\boldsymbol{\varepsilon}\|_2^2/\sigma^2 \sim \chi_n^2$, by (W54a),

$$\mathbf{P}\left(\frac{\|\boldsymbol{\varepsilon}\|_2^2}{n} \leq (1+\epsilon)\sigma^2\right) \leq \exp\left(-\frac{3n\epsilon^2}{16}\right). \tag{18}$$

Moreover,

$$\|\mathbf{y}_{\mathcal{I}}\|_2^2 - n\tau^2 = \sum_{i=1}^n(\mathbf{u}_i^\top\boldsymbol{\beta}_{\mathcal{I}})^2 - \tau^2,$$

25

is a sum of mean zero independent random variables. Define $\mathbf{B} = (B_{jk})$ is the coefficient matrix with $B_{jk} = \beta_{j,k}/2$, $(j \neq k)$ and $B_{jj} = \beta_{j,j}$.

For each i, we can write

$$\mathbf{u}_i^\top \boldsymbol{\beta}_{\mathcal{I}} = \mathbf{x}_i^\top \mathbf{B} \mathbf{x}_i - \mathrm{E}(\mathbf{x}_i^\top \mathbf{B} \mathbf{x}_i) = \mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i - \mathrm{tr}(\mathbf{A}),$$

where $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{A} = (\Sigma)^{\frac{1}{2}} \mathbf{B} (\Sigma)^{\frac{1}{2}}$.

The moment generating function $M(t)$ of the quadratic form $\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i$ is

$$M(t) = \mathrm{E} e^{t \mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i} = \det(\mathbf{I} - 2t\mathbf{A})^{-\frac{1}{2}} = \prod_{j=1}^{s} (1 - 2t\lambda_j)^{-\frac{1}{2}}, \tag{19}$$

where $\{\lambda_j\}_{j=1}^{s}$ are eigenvalues of $\mathbf{A}$ with ascending order. From (19), we have

$$\mathrm{E}(\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i) = \mathrm{tr}(\mathbf{A}), \quad \mathrm{Var}(\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i) = 2\mathrm{tr}(\mathbf{A}^2) = \tau^2,$$

and

$$\mathrm{Var}\left((\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i - \mathrm{tr}(\mathbf{A}))^2 \right) = 48\mathrm{tr}(\mathbf{A}^4) + 8\mathrm{tr}^2(\mathbf{A}^2).$$

Define $W_i = \frac{(\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i - \mathrm{tr}(\mathbf{A}))^2}{\tau^2}$, then $\mathrm{E}(W_i) = 1$, $\mathrm{Var}(W_i) = 12\frac{\mathrm{tr}(\mathbf{A}^4)}{\mathrm{tr}^2(\mathbf{A}^2)} + 2 \leq 14$. Moreover,

$$
\begin{aligned}
\mathrm{E} e^{t|W_i|^{\frac{1}{2}}} &= \mathrm{E} e^{t \frac{|\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i - \mathrm{tr}(\mathbf{A})|}{\tau}} \\
&\leq \mathrm{E} e^{t \frac{\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i - \mathrm{tr}(\mathbf{A})}{\tau}} + E e^{-t \frac{\mathbf{e}_i^\top \mathbf{A} \mathbf{e}_i - \mathrm{tr}(\mathbf{A})}{\tau}} \\
&= e^{-t \frac{\mathrm{tr}(\mathbf{A})}{\tau}} M(\frac{t}{\tau}) + e^{t \frac{\mathrm{tr}(\mathbf{A})}{\tau}} M(\frac{-t}{\tau}) \\
&= \left(\prod_{j=1}^{s} \frac{e^{-\sqrt{2}ta_j}}{1 - \sqrt{2}ta_j} \right)^{\frac{1}{2}} + \left(\prod_{j=1}^{s} \frac{e^{\sqrt{2}ta_j}}{1 + \sqrt{2}ta_j} \right)^{\frac{1}{2}}
\end{aligned}
$$

where $a_j = \lambda_j / \sqrt{\sum_{j=1}^{s} \lambda_j^2}$, so $\sum_{j=1}^{s} a_j^2 = 1$. It is easy to see $\frac{e^{-x}}{1-x} \leq 1 + x^2$ for $x \in [-\frac{1}{2}, \frac{1}{2}]$. For $0 \leq t \leq \frac{\sqrt{2}}{4}$, $|\sqrt{2}ta_j| \leq \frac{1}{2}$, so both summand in the last formula can be controlled by

$$\left(\prod_{j=1}^{s} (1 + 2t^2 a_j^2) \right)^{\frac{1}{2}} \leq \left(\prod_{j=1}^{s} (1 + a_j^2/4) \right)^{\frac{1}{2}} \leq \left(\prod_{j=1}^{s} e^{a_j^2/4} \right)^{\frac{1}{2}} = e^{\frac{1}{2} \sum_{j=1}^{s} \frac{a_j^2}{4}} = e^{\frac{1}{8}}.$$

Therefore, $\mathrm{E} e^{t|W_i|^{\frac{1}{2}}} \leq 2e^{\frac{1}{8}}$ for $0 \leq t \leq \frac{\sqrt{2}}{4}$. And $\mathrm{E} e^{t|W_i - 1|^{\frac{1}{2}}} \leq \mathrm{E} e^{t(|W_i|+1)^{\frac{1}{2}}} \leq \mathrm{E} e^{t+t|W_i|^{\frac{1}{2}}} \leq 2e^{\frac{\sqrt{2}}{4} + \frac{1}{8}}$.

By Lemma B.4 in Hao & Zhang (2014a),

$$\mathbf{P}\left(\left|\sum_{i=1}^{n}(W_i - 1)\right| > n\epsilon\right) \leq c_1 \exp(-c_2 n^{\frac{1}{2}}\epsilon^2),$$

for some positive constants c_1, c_2. That is

$$\mathbf{P}\left(\left|\|\mathbf{y}_\mathcal{I}\|_2^2 - n\tau^2\right| \geq \tau^2 n\epsilon\right) \leq c_1 \exp(-c_2 n^{\frac{1}{2}}\epsilon^2),$$

which implies

$$\mathbf{P}\left(\frac{\|\mathbf{y}_\mathcal{I}\|_2^2}{n} \leq (1+\epsilon)\tau^2\right) \leq c_1 \exp\left(-c_2 n^{\frac{1}{2}}\epsilon^2\right). \tag{20}$$

(18) and (20) imply

$$\mathbf{P}\left(\left\|\Pi_{\mathbf{X}_{\bar{\mathcal{S}}}^{\perp}}\left(\frac{\boldsymbol{\omega}}{\lambda_n n}\right)\right\|_2^2 \geq (1+\epsilon)\frac{2(\sigma^2 + \tau^2)}{\lambda_n^2 n}\right) \leq c_3 \exp\left(-c_4 n^{\frac{1}{2}}\epsilon^2\right).$$

And the conclusion of Lemma 1 follows. $\square$

References

BIEN, J., TAYLOR, J. & TIBSHIRANI, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics* **41**, 1111–1141.

CHEN, J. & CHEN, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.

CHENG, M.-Y., HONDA, T., LI, J., PENG, H. et al. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *The Annals of Statistics* **42**, 1819–1849.

CHIPMAN, H., HAMADA, M. & WU, C. F. J. (1997). A bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics* **39**, pp. 372–381.

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

EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics* **32**, pp. 407–451.

FAN, J., FENG, Y. & SONG, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association* **106**, 544–557.

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

FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

FAN, J. & LV, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on* **57**, 5467 –5484.

FAN, Y., KONG, Y., LI, D. & LV, J. (2016). Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933* .

FAN, Y., KONG, Y., LI, D. & ZHENG, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics* **43**, 1243–1272.

FAN, Y. & LV, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association* **108**, 1044–1061.

FAN, Y. & TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 531–552.

FRIEDMAN, J., HASTIE, T., HÖFLING, H. & TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1**, 302–332.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1.

HAMADA, M. & WU, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology* **24**, 130–137.

HAO, N. & ZHANG, H. H. (2014a). Interaction screening for ultra-high dimensional data. *Journal of the American Statistical Association* **109**, 1285–1301.

HAO, N. & ZHANG, H. H. (2014b). A note on high dimensional linear regression with interactions. *arXiv preprint arXiv:1412.7138* .

JIANG, B. & LIU, J. S. (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics* **42**, 1751–1786.

KONG, Y., LI, D., FAN, Y. & LV, J. (2016). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics* **to appear**.

MCCULLAGH, P. & NELDER, J. (1989). *Generalized Linear Models.* Monographs on Statistics and Applied Probability. Chapman and Hall.

NELDER, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)* **140**, pp. 48–77.

PARK, M. Y. & HASTIE, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 659–677.

PEIXOTO, J. L. (1987). Hierarchical variable selection in polynomial regression models. *The American Statistician* **41**, pp. 311–313.

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

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

WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512–1524.

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. & LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.

WU, T. T. & LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics* **2**, 224–244.

WU, Y. (2011). An ordinary differential equation-based solution path algorithm. *Journal of nonparametric statistics* **23**, 185–199.

YU, Y. & FENG, Y. (2014). Apple: Approximate path for penalized likelihood estimators. *Statistics and Computing* **24**, 803–819.

YUAN, M., JOSEPH, V. R. & ZOU, H. (2009). Structured variable selection and estimation. *Annals of Applied Statistics* **3**, 1738.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942.

ZHAO, P., ROCHA, G. & YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* , 3468–3497.

ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.

ZHOU, H. & WU, Y. (2014). A generic path algorithm for regularized statistical estimation. *Journal of the American Statistical Association, to appear* .

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301320.